ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ
ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Τεχνικές και μηχανισμοί συσταδοποίησης χρηστών και κειμένων για την προσωποποιημένη πρόσβαση περιεχομένου στον παγκόσμιο ιστό

Τσόγκας Βασίλειος
Μηχανικός Η/Υ κ' Πληροφορικής, M.Sc.
Α.Μ. 558

Πάτρα, Δεκέμβριος 2014

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Τεχνικές και μηχανισμοί συσταδοποίησης χρηστών και κειμένων για την προσωποποιημένη πρόσβαση περιεχομένου στον παγκόσμιο ιστό

Τσόγκας Βασίλειος
Μηχανικός Η/Υ κ' Πληροφορικής, M.Sc.
Α.Μ. 558

Επιβλέπων Καθηγητής:
Χρήστος Μπούρας, Καθηγητής

Τριμελής Επιτροπή:
Ευστράτιος Γαλλόπουλος, Καθηγητής
Χρήστος Μακρής, Επίκουρος Καθηγητής
Χρήστος Μπούρας, Καθηγητής

Επταμελής Επιτροπή:
Νικόλαος Αβούρης, Καθηγητής
Ευστράτιος Γαλλόπουλος, Καθηγητής
Ιωάννης Γαροφαλάκης, Καθηγητής
Χρήστος Μακρής, Επίκουρος Καθηγητής
Βασίλειος Μεγαλοοικονόμου, Καθηγητής
Χρήστος Μπούρας, Καθηγητής
Αθανάσιος Τσακαλίδης, Καθηγητής

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

"αφιερωμένη στο γιο μου, το νόημα και το φως της ζωής μου"

Για την συγγραφή της διδακτορικής διατριβής χρησιμοποιήθηκε λογισμικό XeLaTeX

ΠΕΡΙΕΧΟΜΕΝΑ

1 Εισαγωγή **33**
1.1 Γενικά . . . 35
1.2 Υπάρχουσα κατάσταση . . . 35
1.3 Περιγραφή της εργασίας . . . 35
1.4 Δομή της εργασίας . . . 38

2 Καθορισμός του προβλήματος **40**
2.1 Γενικά . . . 42
2.1.1 Άρθρα νέων . . . 43
2.1.2 Web, News και Meta portals . . . 43
2.1.2.1 Web portals . . . 43
2.1.2.2 News portals . . . 44
2.1.2.3 Meta portals . . . 44
2.2 Συστήματα προτάσεων . . . 45
2.3 Προεπεξεργασία δεδομένων . . . 45
2.3.1 Χρήση εξωτερικής βάσης γνώσης . . . 46
2.3.1.1 WordNet . . . 46
2.3.1.1.1 Υπερώνυμα/Υπώνυμα . . . 47
2.3.1.1.2 Μερόνυμα/Ολόνυμα . . . 47
2.3.2 n-grams . . . 48
2.4 Συσταδοποίηση κειμένων . . . 48
2.4.1 Τυπικός ορισμός συσταδοποίησης . . . 50
2.4.2 Πλήθος συστάδων . . . 51
2.5 Συσταδοποίηση χρηστών . . . 51
2.6 Προσωποποίηση στο χρήστη . . . 52
2.6.1 Συμμετοχή του χρήστη στις διαδικασίες του συστήματος . . . 52
2.7 Το Πρόβλημα του νέου χρήστη . . . 52

3 Ερευνητικά Θέματα **55**
3.1 Φυσική Επεξεργασία Γλώσσας . . . 57
3.1.1 Σύνηθες NLP εργασίες . . . 58
3.2 Ανάκτηση Πληροφορίας . . . 59
3.2.1 Μοντελοποίηση ανάκτησης πληροφορίας . . . 60

3.2.1.1 Μοντέλα ανάκτησης πληροφορίας . . . 60
3.2.1.2 Διάσταση μαθηματικής βάσης μοντέλων ανάκτησης πληροφορίας . 61
3.2.1.3 Διάσταση ιδιοτήτων του μοντέλου . . . 62
3.2.1.4 Vector Space Model . . . 62
3.2.2 Αξιολόγηση αποτελεσμάτων ανάκτησης πληροφορίας . . . 63
3.2.2.1 Ανάκληση και ακρίβεια . . . 63
3.2.2.2 Fall-out . . . 64
3.2.2.3 F-measure . . . 64
3.2.2.4 Μέση τιμή ακρίβειας . . . 65
3.2.2.5 R-Ακρίβεια . . . 65
3.3 Φιλτράρισμα Πληροφορίας . . . 65
3.3.1 Εξόρυξη από τον παγκόσμιο ιστό . . . 66
3.3.2 Συνεργατικό φιλτράρισμα - Collaborative Filtering . . . 67
3.3.2.1 Ροή πληροφορίας CF . . . 68
3.3.2.2 Απαιτήσεις CF . . . 69
3.3.2.3 Κατηγορίες CF . . . 70
3.3.3 Φιλτράρισμα βάσει περιεχομένου . . . 70
3.4 Συστήματα προτάσεων . . . 71
3.5 Προεπεξεργασία κειμένου . . . 72
3.5.1 Εξαγωγή λέξεων κλειδιών . . . 73
3.5.2 Εξαγωγή n-grams . . . 73
3.6 Ταξινόμηση κειμένων . . . 75
3.7 Συσταδοποίηση κειμένων . . . 75
3.7.1 Αλγόριθμοι συσταδοποίησης . . . 76
3.7.1.1 Ιεραρχικοί αλγόριθμοι . . . 76
3.7.1.1.1 Τυπικές ιεραρχικές μέθοδοι συσταδοποίησης . . . 77
3.7.1.1.2 Πολυπλοκότητα . . . 78
3.7.1.2 Μερισματικοί αλγόριθμοι . . . 78
3.7.1.3 Οικογένεια k-means . . . 79
3.7.1.3.1 Expectation Maximization . . . 80
3.7.1.3.2 Spherical k-means . . . 82
3.7.1.3.3 Πολυπλοκότητα k-means . . . 83
3.7.1.3.4 Προβλήματα k-means . . . 83
3.7.1.4 Άλλες προσεγγίσεις συσταδοποίησης . . . 84
3.7.1.4.1 Ασαφής συσταδοποίηση . . . 84
3.7.1.4.2 Παραγωγικοί Αλγόριθμοι . . . 85
3.7.1.4.3 Gaussian Μοντέλα . . . 85
3.7.1.4.4 Μείωση διαστατικότητας . . . 86
3.7.1.4.5 Συσταδοποίηση δέντρου επιθεμάτων . . . 86
3.7.1.4.6 DBSCAN . . . 87
3.7.1.5 Μετρικές απόστασης (ομοιότητας) . . . 88
3.7.1.5.1 Ευκλείδεια απόσταση . . . 88
3.7.1.5.2 City-block / απόσταση Manhattan . . . 88
3.7.1.5.3 Απόσταση Pearson . . . 88
3.7.1.5.4 Ομοιότητα συνημιτόνου . . . 89
3.7.1.5.5 Απόσταση Spearman-rank . . . 89
3.7.1.5.6 Απόσταση Kendall's . . . 90
3.7.1.6 Μετρικές αξιολόγησης συσταδοποίησης . . . 90

3.7.1.6.1 Δείκτης συσταδοποίησης (Clustering Index) . . . 91
3.7.1.6.2 Μέσο απόλυτο σφάλμα . . . 91
3.7.2 Αξιοποίηση Εξωτερικών Βάσεων Γνώσης . . . 91
3.7.2.1 WordNet . . . 91
3.7.2.1.1 Χρήση του WordNet στην συσταδοποίηση . . . 92
3.7.3 Πλήθος συστάδων . . . 92
3.7.4 Ονοματοδοσία συστάδων . . . 94
3.8 Προσωποποίηση στον Χρήστη . . . 95
3.9 Το Πρόβλημα του νέου Χρήστη . . . 96
3.9.1 Ερωτήσεις προς, και βαθμολογήσεις από τον χρήστη . . . 97

4 Αρχιτεκτονική **101**
4.1 Στόχοι του συστήματος . . . 103
4.2 Γενική αρχιτεκτονική . . . 103
4.3 Ροή Πληροφορίας . . . 104
4.3.1 Προεπεξεργασία κειμένου . . . 107
4.3.2 Συσταδοποίηση . . . 109
4.3.2.1 Συσταδοποίηση W-kmeans . . . 109
4.3.2.2 Συσταδοποίηση άρθρων νέων . . . 110
4.3.2.3 Μοντελοποίηση και συσταδοποίηση χρηστών . . . 111
4.3.2.4 Υπολογισμός πλήθους συστάδων . . . 112
4.3.3 Πρόβλημα νέου χρήστη . . . 113
4.3.4 Προσωποποίηση στο χρήστη . . . 115

5 Ανάλυση και Αλγοριθμική Προσέγγιση **117**
5.1 Υποσύστημα προεπεξεργασίας κειμένου . . . 119
5.1.1 Αξιοποίηση n-grams . . . 119
5.1.2 Ζύγιση άρθρων . . . 120
5.1.2.1 Ζύγιση keywords για την συσταδοποίηση . . . 122
5.2 Υποσύστημα συσταδοποίησης . . . 123
5.2.1 Αλγόριθμος W-kmeans . . . 123
5.2.2 Συσταδοποίηση άρθρων νέων . . . 124
5.2.2.1 Εξαγωγή και ζύγιση υπερωνύμων . . . 124
5.2.2.2 Αλγόριθμος ενίσχυσης άρθρων νέων με υπερώνυμα . . . 128
5.2.3 Ονοματοδοσία συστάδων . . . 129
5.3 Προσωποποίηση στο χρήστη . . . 130
5.3.1 Εύρεση συνεδρίων χρηστών . . . 131
5.3.2 Συσταδοποίηση Χρηστών με χρήση του W-kmeans . . . 133
5.3.3 Προφίλ χρηστών και προσωποποίηση με χρήση συσταδοποίησης . . . 135
5.4 Πρόβλημα νέου χρήστη . . . 140

6 Τεχνολογίες υλοποίησης και προδιαγραφές του συστήματος **146**
6.1 Γλώσσα υλοποίησης βασικών υποσυστημάτων . . . 148
6.2 Προεπεξεργασία . . . 148
6.2.1 Εξαγωγή n-grams . . . 148
6.2.2 Υπερώνυμα του WordNet . . . 148
6.3 Συσταδοποίηση . . . 149
6.3.1 Υλοποιήσεις αλγορίθμων συσταδοποίησης . . . 149

6.3.1.1 CLUTO . . . 149
6.3.1.2 SenseClusters . . . 150
6.3.1.3 Συσταδοποίηση στη MATLAB . . . 150
6.3.1.3.1 Text to Matrix Generator . . . 151
6.3.1.4 C Clustering Library . . . 151
6.4 Βάση δεδομένων . . . 152
6.4.1 MySQL . . . 152
6.4.2 Βάση δεδομένων του συστήματος . . . 153
6.4.2.1 Νέοι πίνακες . . . 154
6.4.2.1.1 Πίνακες συσταδοποίησης άρθρων νέων . . . 154
6.4.2.1.1.1 clustering_passes . . . 154
6.4.2.1.1.2 clusters . . . 157
6.4.2.1.1.3 article2cluster . . . 157
6.4.2.1.1.4 cluster_similarities . . . 157
6.4.2.1.2 Πίνακες συσταδοποίησης χρηστών . . . 158
6.4.2.1.2.1 clustering_passes_sesions . . . 158
6.4.2.1.2.2 session_clusters . . . 158
6.4.2.1.2.3 session2cluster . . . 159
6.4.2.1.2.4 cluster_similarities_sessions . . . 159
6.4.2.1.2.5 user_sessions . . . 159
6.4.2.1.2.6 user_sessions_articles . . . 160
6.4.2.1.3 Πίνακες n-grams . . . 160
6.4.2.1.3.1 extraction_ng . . . 160
6.4.2.1.3.2 extraction_ng2ar . . . 160
6.5 Διασύνδεση μηχανισμών . . . 160
6.6 Προδιαγραφές . . . 161
6.6.1 Συλλογή άρθρων και εξαγωγή χρήσιμου κειμένου . . . 161
6.6.2 Προεπεξεργασία κειμένου . . . 161
6.6.3 Κατηγοριοποίηση εξαγωγή περίληψης και συσταδοποίησης . . . 162
6.6.4 Προσωποποίηση . . . 163
6.7 Απαιτήσεις του συστήματος . . . 163
6.7.1 Λογισμικό και βιβλιοθήκες . . . 163
6.7.2 Υλικό . . . 163

7 Αξιολόγηση Αλγορίθμων και Υποσυστημάτων 165
7.1 Υποσύστημα Προεπεξεργασίας κειμένου . . . 167
7.1.1 Αξιοποίηση n-grams . . . 167
7.1.1.1 Σύνολο δεδομένων . . . 167
7.1.1.2 Αποτελέσματα και ανάλυση . . . 167
7.2 Συσταδοποίηση . . . 170
7.2.1 Συσταδοποίηση άρθρων νέων . . . 170
7.2.1.1 Αξιολόγηση βασικών αλγορίθμων βιβλιογραφίας . . . 170
7.2.1.1.1 Σύνολο δεδομένων . . . 170
7.2.1.1.2 Αποτελέσματα και ανάλυση . . . 171
7.2.1.2 Αξιολόγηση W-kmeans . . . 181
7.2.1.2.1 Σύνολο δεδομένων . . . 181
7.2.1.2.2 Αποτελέσματα και ανάλυση . . . 181
7.2.1.3 Αξιολόγηση ονοματοδοσίας συστάδων . . . 185

7.2.1.3.1 Σύνολο δεδομένων . . . 185
7.2.1.3.2 Αποτελέσματα και ανάλυση . . . 185
7.2.2 Συσταδοποίηση χρηστών . . . 186
7.2.2.1 Σύνολο δεδομένων . . . 186
7.2.2.2 Αποτελέσματα και ανάλυση . . . 186
7.3 Πρόβλημα νέου χρήστη . . . 191
7.3.1 Σύνολο δεδομένων . . . 191
7.3.2 Αποτελέσματα και ανάλυση . . . 192
7.4 Προσωποποίηση στο χρήστη / παραγωγή προτάσεων . . . 195
7.4.1 Σύνολο δεδομένων . . . 195
7.4.2 Αποτελέσματα και ανάλυση . . . 196

8 Συμπεράσματα **200**
8.1 Το πρόβλημα... . . . 202
8.2 ...και η αντιμετώπισή του . . . 202
8.3 Αξιοποίηση n-grams . . . 203
8.4 Συσταδοποίηση . . . 204
8.4.1 Αξιολόγηση αλγορίθμων βιβλιογραφίας . . . 204
8.4.2 W-kmeans για συσταδοποίηση άρθρων νέων . . . 204
8.4.3 Συσταδοποίηση χρηστών συστήματος . . . 205
8.4.4 Πρόβλημα νέου χρήστη . . . 205
8.5 Προσωποποίηση στο χρήστη και σύστημα προτάσεων . . . 205

9 Μελλοντική εργασία **209**
9.1 Γενικές περιοχές μελλοντικής έρευνας . . . 211
9.2 Προεπεξεργασία . . . 211
9.3 Συσταδοποίηση . . . 212
9.4 Προσωποποίηση και παραγωγή προτάσεων . . . 212
9.5 Παρουσίαση πληροφορίας . . . 213
9.6 Πρόβλημα νέου χρήστη . . . 213

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1 Stemmed keywords με τις συχνότητες εμφάνισής τους όπως εξάγονται από ένα τυχαίο άρθρο 120
2 Τα πιο συχνά εμφανιζόμενα n-grams όπως εξάγονται από το ίδιο άρθρο 120
3 Βάρος ορισμένων υπερωνύμων του σχήματος 14 128

4 Σύνθεση υλικού για ανάπτυξη του συστήματος 164
5 Σύνθεση υλικού του εξυπηρετητή του συστήματος προτάσεων άρθρων νέων 164

6 Σημειογραφία ιεραρχικής συσταδοποίησης 172
7 Επίδραση της εξαγωγής ουσιαστικών και stemming στις μεθοδολογίες συσταδοποίησης 180
8 Αξιολόγηση των μεθοδολογιών συσταδοποίησης σε σχέση με την συσταδοποίηση των ίδιων των χρηστών 181
9 Σύγκριση του W-kmeans με CLUTO και SenseCluster σε σχέση με CI και χρόνο εκτέλεσης. 183
10 Αποτελέσματα ακρίβειας της ονοματοδοσίας συστάδων του W-kmeans ανά κατηγορία 186
11 Σύγκριση μεθοδολογιών CF 191
12 Αλλάζοντας την μεθοδολογία παραγωγής προτάσεων με βάση το χρόνο 197

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

1 Δένδρο υπερωνύμων του όρου dog 47
2 Κατηγοριοποίηση και συσταδοποίηση 50

3 Ακρίβεια - Ανάκληση. Με C είναι τα σχετικά άρθρα που ανακτήθηκαν. 64
4 Τυπικό δενδρόγραμμα ιεραρχικής συσταδοποίησης 77
5 Ο αλγόριθμος EM σε τέσσερις επαναλήψεις του 81
6 Ευαισθησία του k-means στις αρχικές συνθήκες 84
7 Τυπικές συστάδες του αλγορίθμου DBSCAN 87
8 Εκτιμώμενη αύξηση διακύμανσης με παράλληλη αύξηση του πλήθους των συστάδων 93

9 Αρχιτεκτονική του συστήματος προτάσεων άρθρων νέων 104
10 Προεπεξεργασία κειμένου που οδηγεί στην εξαγωγή keywords και n-grams 107
11 Συσταδοποίηση άρθρων νέων και χρηστών 110
12 Συσταδοποίηση άρθρων νέων - τυπικοί αλγόριθμοι και W-kmeans 111
13 Ροή πληροφορίας κατά την εγγραφή νέου χρήστη 114

14 Αθροιστικό δέντρο υπερωνύμων για τρεις λέξεις: ‘pie’, ‘apple’ και ‘orange’ 125
15 Γραφική αναπαράσταση της sigmoid συνάρτησης 42 που χρησιμοποιείται από τον αλγόριθμο W-kmeans 127

16 Διάγραμμα E-R της ΒΔ χωρίς τους νέους πίνακες 155
17 Διάγραμμα E-R των νέων πινάκων της ΒΔ 156

18 Η επίδραση της αξιοποίησης των n-grams στην διαδικασία συσταδοποίησης για διάφορες τιμές του n 168
19 Αποτελέσματα απόδοσης των αλγορίθμων W-kmeans και k-means για διάφορες τιμές ζυγίσματος των εξαγόμενων n-grams 169
20 Αποτελέσματα συσταδοποίησης με χρήση της Ευκλείδειας απόστασης 173
21 Αποτελέσματα συσταδοποίησης με χρήση της απόστασης συνημιτόνου 174
22 Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Pearson 175
23 Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Spearman 176
24 Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Kendals’ τ 177
25 Αποτελέσματα συσταδοποίησης με χρήση της απόστασης City-block 178
26 Χρόνοι εκτέλεσης διαμερισματικών αλγορίθμων σε σχέση με τα πλήθη συστάδων . . 180

27 Σύγκριση W-kmeans και k-means για διάφορες κατηγορίες και πλήθη άρθρων . . . 182
28 Σύγκριση W-kmeans και k-means για συσταδοποίηση άρθρων νέων και για διάφορα πλήθη συστάδων . . . 184
29 Σύγκριση W-kmeans και k-means για συσταδοποίηση συνεδριών χρηστών και διάφορα πλήθη συστάδων . . . 187
30 Τιμές MAE των προτάσεων του συστήματος με και χωρίς την χρήση του W-kmeans 188
31 Σύγκριση της απόδοσης του συστήματος προτάσεων με χρήστη της πληροφορίας συσταδοποίησης χρηστών και μη . . . 189
32 F-measure τιμές των προτάσεων του συστήματος με και χωρίς την χρήση του W-kmeans . . . 190
33 Αξιολόγηση των επιλογών του συστήματος για πρόταση προς το χρήστη ώστε να συγκεντρωθούν οι απαραίτητες βαθμολογήσεις άρθρων νέων . . . 192
34 Σύγκριση μεθοδολογιών πρότασης άρθρων σε σχέση με την τεχνική μας που βασίζεται στη συσταδοποίηση . . . 194
35 Σύγκριση μεθοδολογιών πρότασης άρθρων σε σχέση με την τεχνική μας που βασίζεται στη συσταδοποίηση . . . 195
36 Τιμές MAE των προτάσεων με χρήση των διαφόρων ευρετικών . . . 197
37 Μέσες τιμές F-measure προτάσεων προς τον χρήστη με χρήση των διαφόρων ευρετικών . . . 198

ΕΠΙΤΕΛΙΚΗ ΣΥΝΟΨΗ

Με την πραγματικότητα των υπέρογκων και ολοένα αυξανόμενων πηγών κειμένου στο διαδίκτυο, καθίστανται αναγκαία η ύπαρξη μηχανισμών οι οποίοι βοηθούν τους χρήστες ώστε να λάβουν γρήγορες απαντήσεις στα ερωτήματά τους. Η δημιουργία περιεχομένου, προσωποποιημένου στις ανάγκες των χρηστών, κρίνεται απαραίτητη σύμφωνα με τις επιταγές της συνδυαστικής έκρηξης της πληροφορίας που είναι ορατή σε κάθε "γωνία" του διαδικτύου. Ζητούνται άμεσες και αποτελεσματικές λύσεις ώστε να "τιθασευτεί" αυτό το χάος πληροφορίας που υπάρχει στον παγκόσμιο ιστό, λύσεις που είναι εφικτές μόνο μέσα από ανάλυση των προβλημάτων και εφαρμογή σύγχρονων μαθηματικών και υπολογιστικών μεθόδων για την αντιμετώπισή τους.

Η παρούσα διδακτορική διατριβή αποσκοπεί στο σχεδιασμό, στην ανάπτυξη και τελικά στην αξιολόγηση μηχανισμών και καινοτόμων αλγορίθμων από τις περιοχές της ανάκτησης πληροφορίας, της επεξεργασίας φυσικής γλώσσας καθώς και της μηχανικής εκμάθησης, οι οποίοι θα παρέχουν ένα υψηλό επίπεδο φιλτραρίσματος της πληροφορίας του διαδικτύου στον τελικό χρήστη. Πιο συγκεκριμένα, στα διάφορα στάδια επεξεργασίας της πληροφορίας αναπτύσσονται τεχνικές και μηχανισμοί που συλλέγουν, δεικτοδοτούν, φιλτράρουν και επιστρέφουν κατάλληλα στους χρήστες κειμενικό περιεχόμενο που πηγάζει από τον παγκόσμιο ιστό. Τεχνικές και μηχανισμοί που σκοπό έχουν την παροχή υπηρεσιών πληροφόρησης πέρα από τα καθιερωμένα πρότυπα της υφιστάμενης κατάστασης του διαδικτύου.

Πυρήνας της διδακτορικής διατριβής είναι η ανάπτυξη ενός μηχανισμού συσταδοποίησης (clustering) τόσο κειμένων, όσο και των χρηστών του διαδικτύου. Στο πλαίσιο αυτό μελετήθηκαν κλασικοί αλγόριθμοι συσταδοποίησης οι οποίοι και αξιολογήθηκαν για την περίπτωση των άρθρων νέων προκειμένου να εκτιμηθεί αν και πόσο αποτελεσματικός είναι ο εκάστοτε αλγόριθμος.

Σε δεύτερη φάση υλοποιήθηκε αλγόριθμος συσταδοποίησης άρθρων νέων που αξιοποιεί μια εξωτερική βάση γνώσης, το WordNet, και είναι προσαρμοσμένος στις απαιτήσεις των άρθρων νέων που πηγάζουν από το διαδίκτυο.

Ένας ακόμη βασικός στόχος της παρούσας εργασίας είναι η μοντελοποίηση των κινήσεων που ακολουθούν κοινοί χρήστες καθώς και η αυτοματοποιημένη αξιολόγηση των συμπεριφορών, με ορατό θετικό αποτέλεσμα την πρόβλεψη των προτιμήσεων που θα εκφράσουν στο μέλλον οι χρή-

στες. Η μοντελοποίηση των χρηστών έχει άμεση εφαρμογή στις δυνατότητες προσωποποίησης της πληροφορίας με την πρόβλεψη των προτιμήσεων των χρηστών. Ως εκ' τούτου, υλοποιήθηκε αλγόριθμος προσωποποίησης ο οποίος λαμβάνει υπ' όψιν του πληθώρα παραμέτρων που αποκαλύπτουν έμμεσα τις προτιμήσεις των χρηστών.

Οι παραπάνω μηχανισμοί αφού αξιολογήθηκαν ξεχωριστά, στη συνέχεια ενσωματώθηκαν στην πλατφόρμα αποδελτίωσης άρθρων νέων[1] που είχε υλοποιηθεί στα πλαίσια της μεταπτυχιακής διπλωματικής εργασίας, μετασχηματίζοντάς την έτσι σε ένα σύστημα προτάσεων άρθρων νέων (news articles recommendation system).

Οι τεχνικές που προτείνονται σε αυτή τη διδακτορική διατριβή επεκτείνουν και διαφοροποιούν εργασίες άλλων ερευνητών, προσθέτοντας νέες μεθόδους αντιμετώπισης του προβλήματος προτάσεων άρθρων νέων. Η εργασία που πραγματοποιήθηκε στα πλαίσια της παρούσας διδακτορικής διατριβής αναφέρεται συνοπτικά παρακάτω.

- *Μελέτη αλγορίθμων συσταδοποίησης και αξιολόγησή τους για την περίπτωση των άρθρων νέων από το διαδίκτυο*
  Αυτό το κομμάτι της διδακτορικής διατριβής αφορά στην μελέτη αλγορίθμων συσταδοποίησης κειμένων και αξιολόγηση της εφαρμογής αυτών στην περίπτωση των άρθρων νέων (news articles) που πηγάζουν από το διαδίκτυο. Στόχος αυτής της μελέτης ήταν η εφαρμογή διαφόρων τεχνικών συσταδοποίησης και η σύγκριση των αποτελεσμάτων όσον αφορά στο μεγάλο πλήθος και ποικιλομορφία που παρουσιάζουν τα άρθρα νέων του διαδικτύου. Συγκεκριμένα, μελετήθηκαν ιεραρχικοί (hierarchical) αλγόριθμοι με διάφορες μετρικές απόστασης μεταξύ των σχηματιζόμενων συστάδων: pairwise single, maximum, average, centroid linkage καθώς επίσης και πολλοί διαμερισματικοί (partitional) αλγόριθμοι: k-means, k-medoids, k-means++. Παράλληλα, για κάθε έναν από τους παραπάνω αλγορίθμους συσταδοποίησης χρησιμοποιήθηκαν και διάφορες μετρικές ομοιότητας: Euclidian, City-block, Pearson correlation coefficient, Cosine similarity, Spearman-rank, Kendall's tau. Για την αξιολόγηση των παραπάνω αλγόριθμων – μετρικών χρησιμοποιήθηκαν άρθρα νέων τα οποία συλλέχθηκαν από διάφορα online ειδησεογραφικά πρακτορεία (news portals). Επίσης, για την σύγκριση της ποιότητας των παραγόμενων συστάδων χρησιμοποιήθηκε η μετρική του Clustering Index και του F-measure. Τέλος, έγινε αξιολόγηση από πραγματικούς χρήστες ως προς την ποιότητα των παραγόμενων συστάδων.

- *Σχεδιασμός και υλοποίηση υβριδικού αλγορίθμου συσταδοποίησης άρθρων νέων (W-kmeans)*
  Έχοντας τα αποτελέσματα από την προαναφερθείσα έρευνα υπόψη, στα πλαίσια της διδακτορικής διατριβής, προχωρήσαμε στον σχεδιασμό και υλοποίηση νέου αλγορίθμου για την συσταδοποίηση άρθρων νέων. Το αποτέλεσμα αυτής της έρευνας ήταν ο αλγόριθμος W-kmeans ο οποίος αποτελεί μία προέκταση του κλασικού k-means αλγορίθμου ενώ παράλληλα ενισχύεται από την εξωτερική γνώση που μπορεί να προσφέρει το WordNet, ένας από τους πιο ευρέως διαδεδομένους θησαυρούς λέξεων για την Αγγλική γλώσσα. Το WordNet,

[1] http://perssonal.cti.gr

οργανώνει διάφορες γλωσσολογικές σχέσεις σε ιεραρχίες οι οποίες μπορούν να αναπαρασταθούν σε δενδροειδής δομές. Κάνοντας χρήση αυτών των δομών, αναζητούμε στο WordNet για τα υπερώνυμα (hypernyms) των σημαντικότερων λέξεων που απαρτίζουν ένα άρθρο νέου και έτσι επεκτείνουμε το συνολικό νοηματικό περιεχόμενό του. Επί της ουσίας με αυτή τη διαδικασία εισάγουμε "νέα γνώση" στην υπάρχουσα λίστα λέξεων κάτι που κάνει την διαδικασία συσταδοποίησης λιγότερο ασαφή και περισσότερο αποτελεσματική. Αθροίζοντας τις δενδροειδής δομές των υπερώνυμων των σημαντικότερων όρων ενός κειμένου, αυτό που παρατηρήσαμε είναι ότι όσο πιο πολύ πλησιάζουμε στην ρίζα του δέντρου (οντότητα - entity), τόσο πιο συχνά εμφανίζεται το υπερώνυμο αλλά και τόσο πιο γενικού νοήματος γίνεται αυτό. Επομένως τυπικά υπάρχουν δύο παράμετροι που πρέπει να ληφθούν υπ' όψιν στην διαδικασία της επιλογής των υπερωνύμων που θα ενισχύσουν το κείμενο: η συχνότητα εμφάνισης και το βάθος. Η ζύγιση των παραπάνω παραμέτρων έγινε βάσει μίας σιγμοειδούς (sigmoid) συνάρτησης της οποίας η παράμετρος που εκφράζει το πόσο απότομη είναι περιλαμβάνει τόσο το βάθος όσο και την συχνότητα του υπερωνύμου.

Μια ακόμη σημαντική χρήση της εφαρμογής του WordNet η οποία μελετήθηκε ήταν η εξαγωγή ετικετών (labeling) εκ' των παραγόμενων συστάδων. Η διαδικασία του labeling λειτουργεί ατομικά σε κάθε συστάδα άρθρων λαμβάνοντας υπόψιν αρχικά το 10% των σημαντικότερων λέξεων-κλειδιών των άρθρων της συστάδας. Στη συνέχεια, και για κάθε μία από τις λέξεις-κλειδιά, παράγονται τα δέντρα υπερωνύμων τους τα οποία και συνδυάζονται σε ένα συνολικό δέντρο. Οι κόμβοι που προκύπτουν ζυγίζονται και ταξινομούνται βάσει του βάρους τους, με τα 5 πρώτα υπερώνυμα να επιστρέφονται ως αντιπροσωπευτικά της συστάδας. Αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία ετικετών που καλύπτουν νοηματικά την συστάδα και που μάλιστα πολλές φορές δεν είναι μέρος των λέξεων-κλειδιών των άρθρων που απαρτίζουν τη συστάδα.

Συνδυάζοντας τις παραπάνω τεχνικές, καταλήξαμε στο αλγόριθμο W-kmeans, ο οποίος αξιολογήθηκε σε σχέση με παρόμοιους partitional αλγορίθμους χρησιμοποιώντας την μετρική του Clustering Index. Τα αποτελέσματα της διαδικασίας αξιολόγησης έδειξαν σημαντική βελτίωση της απόδοσης σε σχέση με τον κλασικό k-means αλγόριθμο. Παράλληλα, οι παραγόμενες ετικέτες έχουν υψηλή ποιότητα και θα μπορούσαν να αποτελέσουν ένα σημαντικό εργαλείο για online υπηρεσίες δεικτοδότησης άρθρων νέων και όχι μόνο.

- *Επέκταση και χρήση του αλγορίθμου W-kmeans για την περίπτωση των χρηστών*
  Στο τμήμα αυτό της διδακτορικής διατριβής έγινε επέκταση/προσαρμογή του αλγορίθμου W-kmeans στην περίπτωση συσταδοποίησης χρηστών που παρακολουθούν άρθρα νέων του διαδικτύου. Πιο συγκεκριμένα, μελετήθηκε και υλοποιήθηκε η επέκταση της εφαρμογής του αλγορίθμου για τις καταγεγραμμένες συνεδρίες των χρηστών που είναι εγγεγραμμένοι στην online υπηρεσία δεικτοδότησης. Παράλληλα, έγινε αξιολόγηση των συνεπειών που έχει η προσέγγιση αυτή στην μηχανή προτάσεων του συστήματός μας, μετρώντας την συνολική επίδοση που έχει αυτή όσον αφορά στην ακρίβεια και ανάκληση (precision/recall) των πα-

ραγόμενων αποτελεσμάτων.

Ο αλγόριθμος W-kmeans για την περίπτωση εφαρμογής του σε χρήστες, προχωράει ως εξής: αρχικά εξάγονται οι συνεδρίες (sessions) από άρθρα τα οποία ο χρήστης επέλεξε να δει σε συγκεκριμένου μεγέθους χρονικά παράθυρα. Στη συνέχεια, για κάθε συνεδρία αθροίζουμε τα άρθρα που απαρτίζουν την συνεδρία και στη συνέχεια εμπλουτίζουμε τις λέξεις-κλειδιά με σχετικά υπερώνυμα που εξάγονται από το WordNet με τον τρόπο που περιγράφεται στην συνέχεια. Αρχικά για κάθε μία από τις λέξεις-κλειδιά παράγουμε τις δενδροειδής δομές από υπερώνυμα που οδηγούν στο υπερώνυμο - ρίζα (οντότητα - entity) και στη συνέχεια αθροίζουμε όλες τις δενδροειδής δομές σε μία. Πρακτικά, υπάρχουν δύο παράμετροι οι οποίες πρέπει να ληφθούν υπ' όψιν όσον αφορά στη σημαντικότητα του κάθε υπερώνυμου: το βάθος του στο δέντρο και η συχνότητα εμφάνισής του. Ζυγίζοντας τις παραπάνω παραμέτρους με μία σιγμοειδή (sigmoid) συνάρτηση και στη συνέχεια ταξινομώντας βάσει του βάρους, καταλήγουμε σε μία λίστα από υπερώνημα τα οποία εκφράζουν το προφίλ του χρήστη βάσει τις επιλογές που έχει κάνει. Η λίστα αυτή χρησιμοποιείται έπειτα κατά το στάδιο προτάσεων στο χρήστη για την παρουσίαση αποτελεσμάτων τα οποία με μεγάλη πιθανότητα τον ενδιαφέρουν.

Για την πειραματική αξιολόγηση της εφαρμογής του αλγορίθμου W-kmeans στα προφίλ των χρηστών, χρησιμοποιήθηκε μεγάλο πλήθος από άρθρα νέων προερχόμενα διάφορα διαδικτυακά ειδησεογραφικά πρακτορεία καθώς και αρκετούς εγγεγραμμένοι χρήστες του συστήματος. Επίσης ως κριτήριο αξιολόγησης των σχηματιζόμενων συστάδων χρησιμοποιήθηκε το Clustering Index καθώς και το F-measure. Τα αποτελέσματα έδειξαν μία σημαντική βελτίωση σε σχέση με τον κλασικό k-means αλγόριθμο. Παράλληλα, οι προσφερόμενες προτάσεις άρθρων στους χρήστες ήταν σημαντικά βελτιωμένες σε σχέση με πριν όπου δεν εφαρμόζονταν η συσταδοποίηση χρηστών.

- *Προσωποποίηση των προτεινόμενων άρθρων νέων βάσει της πληροφορίας συσταδοποίησης*
  Με βάση τα παραπάνω αποτελέσματα σε σχέση με την συσταδοποίηση άρθρων νέων, καθώς και των χρηστών αυτών, στο τμήμα αυτό της διδακτορικής διατριβής αναπτύχθηκε τεχνική προσωποποίησης των προτεινόμενων προς τους χρήστες άρθρων νέων, η οποία αξιοποιεί την πληροφορία των συστάδων χρηστών του συστήματος. Ο αλγόριθμος προσωποποίησης που αναπτύχθηκε, μπορεί να χαρακτηριστεί ως υβριδικός καθώς βασίζεται τόσο στο ίδιο το περιεχόμενο των άρθρων (content-based) όσο και στο συνεργατικό φιλτράρισμα (collaborative filtering) αξιοποιώντας την συσταδοποίηση και τις επιλογές των χρηστών του συστήματος. Παράλληλα, έχει τη δυνατότητα της προσαρμογής στα μεταβαλλόμενα ενδιαφέροντα του χρήστη με σχετικά μικρές αλλά διαρκείς μεταβολές στα προφίλ των χρηστών. Ο αλγόριθμος ενσωματώνει αρκετά ευρετικά, όπως τα επιλεγμένα προς ανάγνωση άρθρα νέων από τον χρήστη, τον χρόνο που ξοδεύει διαβάζοντάς τα, την κατηγορία των άρθρων, καθώς και την γνώση της συστάδας που ανήκει ο χρήστης.

  Η εφαρμογή της προαναφερθείσας τεχνικής προσωποποίησης με χρήση συσταδοποίησης, οδήγησε σε βελτιωμένα αποτελέσματα όσον αφορά τόσο στην ικανότητα του συστήματος να

συγκλίνει γρηγορότερα στις πραγματικές προτιμήσεις των χρηστών, όσο και στην ποιότητα των προτάσεων για άρθρα νέων που προσφέρει προς τους χρήστες.

- *Το πρόβλημα του "νέου χρήστη" και αντιμετώπισή του*
  Ένα σύστημα συστάσεων (recommendation system), μπορεί να βρεθεί σε μία κατάσταση κατά την οποία δεν έχει αρκετή πληροφορία στην οποία να βασίσει τις αποφάσεις/προτάσεις του. Αυτού του είδους η κατάσταση είναι γνωστή στην βιβλιογραφία ως cold start problem και διακρίνεται σε τρεις περιπτώσεις: α) πρόβλημα νέου στοιχείου (new item problem) όπου ένα νέο στοιχείο (στην περίπτωσή μας ένα άρθρο νέου) προστίθεται στο σύστημα χωρίς να υπάρχουν ακόμη αξιολογήσεις για αυτό, β) πρόβλημα νέου χρήστη (new user problem) όπου ένας νέος χρήστης εγγράφεται στο σύστημα χωρίς να είναι γνωστό κάτι για τις προτιμήσεις του, γ) πρόβλημα νέου συστήματος όπου αποτελεί συνδυασμό των παραπάνω περιπτώσεων. Στο τμήμα αυτό της διδακτορικής διατριβής αναπτύχθηκε μια προσωποποιημένη μεθοδολογία για την αντιμετώπιση του "προβλήματος νέου χρήστη" (new user problem).

  Η τεχνική που υλοποιήθηκε, είναι αρχικά παρόμοια με την "στοιχείο προς στοιχείο" στρατηγική (item by item strategy). Στη συνέχεια, δεδομένης μία τουλάχιστον επιλογής για αξιολόγηση άρθρου από τον χρήστη, αξιοποιείται η πληροφορία της συσταδοποίησης άρθρων, και πιο συγκεκριμένα τα αποτελέσματα του W-kmeans αλγορίθμου που υπάρχουν στη βάση δεδομένων για την μετέπειτα επιλογή προτάσεων. Έπειτα, και εφόσον δεν έχουν ήδη επιλεγεί αρκετά άρθρα για αξιολόγηση, χρησιμοποιούμε τα αποτελέσματα του W-kmeans αλγορίθμου όσον αφορά στην συσταδοποίηση χρηστών του συστήματος για τις προτάσεις που ακολουθούν. Η διαδικασία συνεχίζεται έως ότου ο συνολικός αριθμός αξιολογήσεων από τον χρήστη φτάσει σε κάποιο όριο στο οποίο μπορούμε να θεωρήσουμε ότι η διαδικασία εκτίμησης των προτιμήσεων του χρήστη έχει ολοκληρωθεί.

  Η πειραματική αξιολόγηση της προαναφερθείσας τεχνικής έδειξε ότι με τη χρήση κατά μέσο όρο 5 άρθρων από κάθε σχετική συστάδα άρθρου ή χρήστη, παίρνουμε τα καλύτερα αποτελέσματα και την ταχύτερη σύγκλιση στο προφίλ του χρήστη. Χρησιμοποιώντας αυτό το συμπέρασμα, υπολογίσαμε ότι η τεχνική μας χρειάζεται κατά μέσο όρο 37.5 άρθρα προς παρουσίαση στη χρήστη προκειμένου να πάρει 20 επιτυχείς αξιολογήσεις – ένα αποτέλεσμα σημαντικά καλύτερο από τις τυπικές υπάρχουσες μεθόδους της βιβλιογραφίας σχετικά με την αντιμετώπιση του προβλήματος νέου χρήστη.

- *Αξιοποίηση word n-grams για βελτίωση της συσταδοποίησης άρθρων νέων*
  Ένα n-gram ορίζεται ως η ακολουθία κειμένου μεγέθους 'n' που αποτελείται από συνεχόμενα γράμματα ή λέξεις. Για την περίπτωση των word n-grams, ενδιαφερόμαστε μόνο για σειρές το πολύ n συνεχόμενων λέξεων στις ακολουθίες κειμένων. Για παράδειγμα ένα 4-gram είναι το εξής: economic situation in Greece.

  Στο τμήμα αυτό της διδακτορικής διατριβής αναπτύχθηκε τεχνική ενίσχυσης του αλγορίθμου συσταδοποίησης άρθρων νέων από το διαδίκτυο (W-kmeans) με χρήση n-grams λέξεων (word n-grams) κατά την διαδικασία της εξαγωγής λέξεων κλειδιών (keyword extraction). Για την

ενίσχυση του αλγορίθμου W-kmeans, χρησιμοποιήθηκε μία προσέγγιση ζυγίσματος η οποία αξιοποιεί τόσο την συχνότητα εμφάνισης των keywords (bag of words representation) όσο και αυτή των n-grams. Πιο συγκεκριμένα, ο αλγόριθμος αναθέτει βάρη στα n-grams του κειμένου (όπου $2 < n < 6$) παρόμοια με τα tf-idf (term frequency – inverse document frequency) βάρη των keywords, κατά τη διαδικασία της εξαγωγής λέξεων-κλειδιών (keyword extraction), και έπειτα συνδυάζει τα συνολικά βάρη για να αξιολογήσει ποια keywords και n-grams είναι πιο σημαντικά ώστε να λαμβάνονται υπόψη κατά την συσταδοποίηση.

Η εκτίμηση της σημαντικότητας των keywords και n-grams στη διαδικασία της συσταδοποίησης αποτέλεσε αντικείμενο της πειραματικής διαδικασίας, από την οποία προέκυψε ότι η ζύγιση keywords / n-grams σε λόγο 7/3 έδινε τα καλύτερα αποτελέσματα για την συσταδοποίηση (συστάδες καλύτερα διαχωρισμένες και με μεγαλύτερη συνοχή). Παράλληλα βρέθηκε ότι για $n = 3$, δηλαδή όταν λαμβάνονται υπόψη τόσο τα 2-grams όσο και τα 3-grams για την διαδικασία ζυγίσματος, έχουμε καλύτερα αποτελέσματα για την συσταδοποίηση άρθρων από το διαδίκτυο (κάτι που επιβεβαίωσε την υπάρχουσα σχετική βιβλιογραφία).

EXECUTIVE SUMMARY

With the reality of the ever increasing information sources from the internet, both in sizes and indexed content, it becomes necessary to have methodologies that will assist the users in order to get the information they need, exactly the moment they need it. The delivery of content, personalized to the user needs is deemed as a necessity nowadays due to the combinatoric explosion of information visible to every corner of the world wide web. Solutions effective and swift are desperately needed in order to deal with this information overload. These solutions are achievable only via the analysis of the refereed problems, as well as the application of modern mathematics and computational methodologies.

This Ph.d. dissertation aims to the design, development and finally to the evaluation of mechanisms, as well as, novel algorithms from the areas of information retrieval, natural language processing and machine learning. These mechanisms shall provide a high level of filtering capabilities regarding information originating from internet sources and targeted to end users. More precisely, through the various stages of information processing, various techniques are proposed and developed. Techniques that will gather, index, filter and return textual content well suited to the user tastes. These techniques and mechanisms aim to go above and beyond the usual information delivery norms of today, dealing via novel means with several issues that are discussed.

The kernel of this Ph.d. dissertation is the development of a clustering mechanism that will operate both on news articles, as well as, users of the web. Within this context several classical clustering algorithms were studied and evaluated for the case of news articles, allowing as to estimate the level of efficiency of each one within this domain of interest. This left as with a clear choice as to which algorithm should be extended for our work.

As a second phase, we formulated a clustering algorithm that operates on news articles and user profiles making use of the external knowledge base of WordNet. This algorithm is adapted to the requirements of diversity and quick churn of news articles originating from the web.

Another central goal of this Ph.d. dissertation is the modeling of the browsing behavior of system users within the context of our recommendation system, as well as, the automatic

evaluation of these behaviors with the obvious desired outcome or predicting the future preferences of users. The user modeling process has direct application upon the personalization capabilities that we can over on information as far as user preferences predictions are concerned. As a result, a personalization algorithm we formulated which takes into consideration a plethora or parameters that indirectly reveal the user preferences.

The above mechanisms, after being evaluated separately, were later incorporated as modules within the online news indexing service[2] that was implemented as part of my M.Sc. thesis, transforming it into a complete news articles recommendation system. The techniques that are proposed in this Ph.d. dissertation extend and diversify over works from other researchers, adding new methodologies in order to deal with the problem of recommending news articles. The work covered as part of the Ph.d. dissertation is shortly outlined below.

- *Study of existing news clustering algorithms and evaluation for the case of news articles originating from the web*
  This part of the Ph.d. dissertation has to do with the study of clustering algorithms which operate upon texts and the evaluation of this application for the case of news articles. The goal of this study was the application of various clustering methodologies and then the comparison of their performance as far as the great numbers and diversity that news articles exhibit, are concerned. In particular, hierarchical clustering algorithms were studied: pairwise single, maximum, average, centroid linkage. In addition, several partitional clustering algorithms were also studied: k-means, k-medoids, k-means++. For each of the above clustering algorithms various distance measures for calculating the distance among the formulated clusters were used: Euclidian, City-block, Pearson correlation coefficient, Cosine similarity, Spearman-rank, Kendall's tau. For the evaluation of the above combination of clustering algorithms and distance measures, news articles collected from numerous news portals were useds. Furthermore, for comparing the quality of the generated clusters the Clustering Index and F-measure metrics were utilized. Finally, the quality of the generated clusters was evaluated by real system users, giving some useful feedback about the performance of the wining clustering methodology.

- *Design and implementation of a hybrid news articles clustering algorithm (W-kmeans)*
  Having the results of the aforementioned research in mind, within the scope of this Ph.d. dissertation, we moved the design and implementation of a new news articles clustering algorithm. The outcome of this research was the W-kmeans algorithm which is an extension of the classical k-means clustering algorithm, assisted by the external knowledge that WordNet, one of the most widely used English language thesauri, can offer. WordNet, by organizing the various linguistic relationships into hierarchies can be represented into tree-like structures. Using these structures, we seek into WordNet for the hypernyms of the words which constitute a news article, enhancing thus its overall

[2] http://perssonal.cti.gr

context meaning. In essence, via this process, we are introducing 'new knowledge' into the existing keywords lists, something that makes the clustering process less fuzzy and more effective. By aggregating the hypernym structures of the text's keywords, what we observed was that the more we got closer to the root of this tree (called 'entity' within WordNet), the more frequently the hypernym would appear but also the more generic its meaning would become. As a results there are typically two parameters that should be taken into consideration with regards to the process of hypernym selection/weighting that shall enhance the text: the frequency of appearance and its depth. The weighting scheme of these parameters was done using a sigmoid function of which the parameter that defines how steep it is includes the both the weight and the frequency of the hypernym.

Another important use for the application of WordNet that was studied, is the labeling generation process regarding the produced clusters. The labeling process operates within each individual news articles cluster initially taking into consideration the top 10% of the most important keywords of the articles belonging to the particular cluster. Next, of each of those keywords the WordNet hypernym tree is generated and those trees are aggregated together into a global tree. The nodes that are produced by this process are then weighted and sorted according to their weight, and the top 5 hypernyms are returned as representatives of the cluster. The outcome of the above process is cluster labels which cover the sense of each clyster and which might not even be part of the keywords that make us the cluster.

Combining the above techniques into a single process, we named the algorithm as W-kmeans. W-kmeans was then evaluated against similar partitinal algorithms use the Clustering Index metric. The results of the evaluation process showed significant improvement compared with the classical k-means algorithm. Furthermore, the generated labels are of high quality and can constitute an important tool for inline services which index news articles (amongst other things).

- *Expansion and use of the W-kmeans clustering algorithm for the case of system users*
  Within this part of the Ph.d. dissertation, the adaptation of the W-kmeans algorithm for the case of user clustering was performed (as far as users browsing news articles are concerned). In particular, an expansion of the clustering algorithm was investigated and implemented that would take into account the system users as registered into our recommendation system. In addition, we evaluated the consequences of this approach into the recommendation engine of the system, evaluating thus the overall performance improvement that this has with regards to precision/recall metrics on the produced results. The W-kmeans algorithm for the case of users proceeds as follows: initial the user sessions are extracted using news articles for which the user has expressed interest into reading within specific time windows. Following, for each user session, we sum up the articles that make it up and then we enrich the extracted keywords using WordNet hypernyms

in the way that is described next. Initially for each of the keywowrds we generate the tree-like structures of hypernymes that lead to the hypernym-root and we then aggregate all these structures into a combined one. There are practically two parameters that need to be taken into consideration as far as the importance of each hypernym is concerned: its depth in the tree and its frequency of appearance. By weighting the above parameters into a sigmoid function and then by sorting them by weight, we end up with a list of hypernymes that express the user profile based on the choices that he did. This like can be used later in multiple ways (like in the personalization/recommendation phase, or for dealing with the new user problem).

For the experimental evaluation of the application of the W-kmeans algorithm to the user profiles, we used a good number of news articles originating from online news portals, as well as data from registered system users. Again as an evaluative criterion we used the Clustering Index and the F-measure. The results showed a significant improvement compared to the classical k-keans algorithm. In addition, the article recommendations towards the users were significantly improved compared to the case when user clustering was not employed.

- *Personalization of the proposed articles based on clustering information*
  Using the above results regarding news articles and user clustering, in this part of the Ph.d. dissertation we developed a personalization technique that lead to the actual recommendations made by the system. This technique makes use of several heuristics that had been investigated before, but is now enhanced to also incorporate clustering into the weighting scheme.

  The personalization algorithm that was developed can be characterized as hybrid since it's based both onto the context of the articles themselves, as well as the collaborative filtering, using continuously the clustering information along with the previous user choices. Moreover, it has the capability of adapting to the always evolving user interests with relatively small but continuous profile updates. The algorithm incorporates a multitude of heuristics like the previously viewed articles, the times spent by the user reading them, the articles categorization along with the previously mentioned clustering information. The application of the aforementioned personalization technique resulted in improved results with regards to both the ability of the system to quickly converge to the real user interests, and to the quality of the news articles suggestions offered to the end users.

- *Addressing the new user problem*
  A recommendation system can be found in a situation where it does not have enough information on which to rely its decisions/recommendations. This kind of state is commonly known as the cold start problem and is made up of three individual cases: a) the new item problem, where a new item (in our case a news article) is added to the system without any ratings or choices yet available for it, b) the new user problem, where a new user would register into the system without any kind of information regarding his preferences

made available, making any future recommendation completely a luck experiment, c) the new system problem which is a combination of a) and b).

In this part of the Ph.d. dissertation we developed a personalized methodology for dealing with the new user problem. The technique that was implemented is initially similar to the item by item personalized strategy. However, given at least one successful user rating, the information regarding news clustering, and in particular the W-kmeans clustering results stored in the database, are taken advantage of for the follow-up suggestions for rating. Next, and as long as not enough news articles have been selected by the user for rating, we use the results of the W-kmeans algorithm with regards to user clustering for selecting the upcoming queries for rating. The process continues until the total number of user ratings reaches a particular limit upon which we can assume that the estimation of user interests has completed.

The experimental procedure of the aforementioned methodology revealed that by using, on average, 5 articles from each of the relative cluster, either the articles one, or the users one, we get the best results and the fastest convergence to the actual user profile. Making use of this conclusion, we calculated that the proposed technique needs, on average, 37.5 articles to be presented to the user in order to gather 20 successful evaluations - a result far better than the typical methods proposed in the literature regarding the problem.

- *Making use of word n-grams in order to improve the news clustering results*
  An n-gram is the textual sequence of size n which consists of continuous letter or words. For the case of word n-grams, we are interested in sequences of at most n continuous words into the texts. For example, a 4-gram would be the following: economic situation in Greece.

  Into this part of the Ph.d. dissertation, a technique for improving the process of news article clustering was developed that makes use of word n-grams during the keyword extraction phase. For improving associating n-grams with W-kmeans, we used a weighting scheme which takes advantage of the information of both the article keywords (bag of words representation), as well a similar n-grams representation. More specifically, the algorithm assigns weights to the text n-grams (where $2 < \mathrm{n} < 6$) similar to the tf-idf (term frequency – inverse document frequency) keyword weights during the keyword extraction phase, and then combines the aggregate weights in order to evaluate which n-grams and keywords are important and how so as to be taken under consideration for the clustering process that follows.

  The assessment of the importance of the keywords and n-grams within the clustering process constituted an area of experimentation from which we found that the weighting of keywords/n-grams in a ratio of 7/3, would give the best clustering results (clusters well connected within and well separated from outside). In addition, we found that for $n = 3$, meaning that when we kept both 2-grams and 3-grams during the weighting process, we

would have the best results as far as news clustering is concerned (a result confirming existing bibliography).

ΔΗΜΟΣΙΕΥΣΕΙΣ ΚΑΙ ΑΝΑΦΟΡΕΣ

Δημοσιεύσεις σχετικές με την διδακτορική διατριβή

Δημοσιεύσεις σε διεθνή περιοδικά

1. Improving News Articles Recommendations via User Clustering. *International Journal of Machine Learning and Cybernetics* (to appear) C. Bouras, V. Tsogkas, 2015

   Abstract

   Παρότι συχνά μόνο η συσταδοποίηση αντικειμένων συχνά προτείνεται από τεχνικές Web mining για συστήματα προτάσεων άρθρων νέων, μία από τις ποικίλες διεργασίες την προσωποποίησης προτάσεων είναι η συσταδοποίηση των ίδιων των χρηστών. Με την συνδυαστική έκρηξη των online άρθρων νέων, η πρόβλεψη των συνηθειών πλοήγησης των χρηστών με χρήση συνεργατικού φιλτραρίσματος (CF) έχει κερδίσει αρκετά έδαφος στην περιοχή της προσωποποίησης του ιστού. Παρόλα αυτά, οι κοινές CF τεχνικές υποφέρουν από χαμηλή ακρίβεια και απόδοση. Η παρούσα έρευνα προτείνει μία νέα προσωποποιημένη προσέγγιση για παραγωγή προτάσεων, η οποία ενσωματώνει την συσταδοποίηση τόσο σε επίπεδο περιεχομένου όσο και χρηστών. Βασίζεται στο αλγόριθμο W-kmeans καθώς και άλλες IR τεχνικές, όπως η κατηγοριοποίηση και περίληψη κειμένου, προκειμένου να προσφέρει στους χρήστες άρθρα που ταιριάζουν στα προφίλ τους. Το σύστημα προτάσεων που αναπτύχθηκε μπορεί γρήγορα να προσαρμόζεται στα χρονικά μεταβαλλόμενα ενδιαφέροντα των χρηστών. Επιπλέον, τα πειραματικά αποτελέσματα έδειξαν ότι η αξιοποίηση συσταδοποίησης αντικειμένων και χρηστών επιφέρει σημαντικά οφέλη στο σύστημα προτάσεων.

2. Assisting cluster coherency via N-grams and clustering as a tool to deal with the new user problem. *International Journal of Machine Learning and Cybernetics*: 1-14, Springer Verlang, C. Bouras, V. Tsogkas, 2014

   Abstract

Οι τεχνικές ''συνεργατικού φιλτραρίσματος'' (collaborative filtering techniques) πάσχουν από το λεγόμενο πρόβλημα ''νέου χρήστη''. Αυτή η κατάσταση συμβαίνει όταν ένας νέος χρήστης προστίθεται σε ένα σύστημα προτάσεων (recommendation system) και δεν υπάρχει αρκετή πληροφορία την οποία μπορεί να χρησιμοποιήσει το σύστημα για να στηρίξει τις προτάσεις του. Το σύστημα χρειάζεται επομένως κάποια δεδομένα σχετικά με τον νέο χρήστη προκειμένου να μπορεί να κάνει τις προσωποποιημένες προτάσεις. Σε αυτή τη δημοσίευση επιχειρούμε να αντιμετωπίσουμε το πρόβλημα νέου χρήστη χρησιμοποιώντας μία προσωποποιημένη στρατηγική σχετικά με τις προτάσεις που γίνονται στο χρήστη προκειμένου να βαθμολογηθούν αυτές κατά της διαδικασία αρχικής εκμάθησης. Η προσέγγισή μας κάνει χρήση υπερωνήμων τα οποία εξάγονται από το WordNet και προσεγγίζει γρήγορα στα πραγματικά ενδιαφέροντα του χρήστη βασιζόμενη παράλληλα σε λίγες βαθμολογήσεις από την πλευρά του χρήστη. Παράλληλα ερευνούμε την βελτίωση που μπορεί να έχει στα αποτελέσματα της συσταδοποίησης άρθρων νέων από το διαδίκτυο η αξιοποίηση n-grams λέξεων κατά την διαδικασία εξαγωγής λέξεων-κλειδιών. Η τεχνική αυτή συγκρίνεται με την τυπική ''bag of words'' αναπαράσταση που χρησιμοποιούσε προηγούμενα ο αλγόριθμος W-kmeans. Η πειραματική διαδικασία δείχνει ότι μέσω του κατάλληλου ζυγίσματος της βαρύτητας των keywords, των n-grams καθώς και της τιμής n, μία σημαντική βελτίωση μπορεί να επιτευχθεί σχετικά με τα αποτελέσματα της συσταδοποίησης.

3. A clustering technique for news articles using WordNet. *Knowledge-Based Systems Journal, Elsevier Science*, Vol. 36, C. Bouras, V. Tsogkas, 2012, 115 - 128

Abstract

Η συσταδοποίηση κειμενικής πληροφορίας αποτελεί μία ισχυρή τεχνική αντιμετώπισης του προβλήματος διαχείρισης της παραγόμενης ποσότητας άρθρων νέων που κατακλύζουν το διαδίκτυο. Μέσω αυτής, μπορούμε να οργανώσουμε δεδομένα σε μικρότερους και πιο διαχειρίσημους ''πυρήνες'' πληροφορίας. Πληθώρα προσεγγίσεων έχουν προταθεί στη βιβλιογραφία με τυπικά προβλήματα να παραμένουν η συνωνυμία, η ασάφεια καθώς και η έλλειψη συγκεκριμένων αντιπροσωπευτικών περιγράφων των συστάδων (labels). Στην παρούσα έρευνα, ερευνούμε την εφαρμογή ενός φάσματος αλγορίθμων συσταδοποίησης, καθώς και μετρικών σύγκρισης, στον τομέα των άρθρων νέων που προέρχονται από το διαδίκτυο. Παράλληλα προτείνουμε μία τροποποίηση/βελτιστοποίηση του αλγορίθμου k-means κάνοντας χρήση την εξωτερική γνώση από υπερώνημα (hypernyms) του WordNet με διττό τρόπο: εμπλουτίζοντας τις λέξεις κλειδιά (bag of words) οι οποίες χρησιμοποιούνται προηγούμενα από την διαδικασία συσταδοποίησης, και επίσης, αξιοποιώντας αυτή την πληροφορία προκειμένου να υποβοηθηθεί η παραγωγή αντιπροσωπευτικών τίτλων για κάθε συστάδα. Παράλληλα, εξετάζουμε την επίδραση που έχει η προεπεξεργασία κειμένου στη διαδικασία συσταδοποίησης. Χρησιμοποιώντας ένα σώμα (corpus) άρθρων νέων που πηγάζουν από μείζονα ηλεκτρονικά ειδησεογραφικά πρακτορεία, η σύγκριση των υπαρχόντων αλγορίθμων συσταδοποίησης έδειξε ότι η k-means δίνει καλύτερα συνολικά αποτελέσματα σε σχέση με την αποδοτικότητά του.

Αυτό ενισχύεται όταν ο αλγόριθμος συνοδεύεται από προκαταρκτικά βήματα για καθαρισμό δεδομένων και κανονικοποίηση, παρά την θεωρητικά απλοϊκή του φύση. Εκτός αυτού, ο προτεινόμενος W-kmeans αλγόριθμος συσταδοποίησης βελτιώνει σημαντικά τον τυπικό k-means παράγοντας επίσης χρήσιμες και ποιοτικές ετικέτες (cluster tags) βάσει της διαδικασίας που περιγράφεται στην συγκεκριμένη δημοσίευση.

Δημοσιεύσεις σε διεθνή συνέδρια

1. Evaluating the Unification of Multiple Information Retrieval Techniques into a News Indexing Service. *3rd Intenational Conference on Data Management Technologies and Applications*, Vienna, Austria, C. Bouras, V. Tsogkas, Aug. 29 - 31 2014

   Abstract

   Όσο οι online πηγές ειδησεογραφικών νέων αυξάνονται, τόσο αυξάνεται και ο όγκος της σχετικής πληροφορίας. Πολλαπλές προσεγγίσεις έχουν προταθεί για την οργάνωση αυτού του όγκου πληροφορίας. Στην παρούσα δημοσίευση, ερευνούμε την ενοποίηση πολλαπλών τεχνικών ανάκτησης πληροφορίας, όπως προεπεξεργασία κειμένου, επέκταση n-grams, περίληψη κειμένου, καθώς και συσταδοποίηση στοιχείων/χρηστών, σε έναν μηχανισμό σχεδιασμένο να ενοποιεί και να δεικτοδοτεί άρθρα νέων που πηγάζουν από το διαδίκτυο. Στόχος μας είναι να επιτρέψουμε στους χρήστες να μπορούν απρόσκοπτα και γρήγορα να πάρουν την ειδησεογραφική ενημέρωση η οποία τους ταιριάζει. Δείχνουμε πως, η χρήση καθεμίας από τις προτεινόμενες τεχνικές, βελτιώνει την ακρίβεια του συστήματος σε σχέση με τα προτεινόμενα άρθρα για τους εγγεγραμμένους χρήστες. Τέλος εξετάζουμε πως αυτές οι τεχνικές συνολικά μπορούν να αποτελέσουν μία ενοποιημένη λύση για ένα σύστημα προτάσεων (recommendation system).

2. Enhancing news articles clustering using word n – grams. *2nd Intenational Conference on Data Management Technologies and Applications*, Reykjavvk, Iceland, C. Bouras, V. Tsogkas, July 29 - 31 2013, 53 – 60

   Abstract

   Σε αυτή την εργασία, ερευνούμε την πιθανή βελτίωση των αποτελεσμάτων της συσταδοποίησης κειμένων, και εν' προκειμένω, άρθρων νέων που προέρχονται από το διαδίκτυο, μέσω της χρήσης n-grams λέξεων κατά την διαδικασία της εξαγωγής λέξεων κλειδιών. Παρουσιάζουμε και αξιολογούμε μία προσέγγιση ζυγίσματος η οποία συνδυάζει την συσταδοποίηση άρθρων νέων με χρήση n-grams τα οποία εξάγονται offline και χρησιμοποιούνται παράλληλα με τις λέξεις κλειδιά του εκάστοτε κειμένου. Η συγκεκριμένη τεχνική συγκρίνεται με την απλοϊκή bag-of-words αναπαράσταση (όπου αξιοποιούνται μόνο οι λέξεις κλειδιά) την οποία χρησιμοποιούσε προηγούμενα ο αλγόριθμος συσταδοποίησης W-kmeans. Η πειραματική διαδικασία έδειξε ότι μέσω της ρύθμισης των παραμέτρων ζυγίσματος μεταξύ λέξεων κλειδιών και n-grams, καθώς και του n, μπορεί να δώσει σημαντικές βελτιώσεις όσον αφορά την επίδοση

του αλγορίθμου συσταδοποίησης.

3. Clustering to Deal with the New User Problem. *15th IEEE International Conference on Computational Science and Engineering*, Paphos, Cyrus, C. Bouras, V. Tsogkas, 5 - 7 December 2012, pp. 58 – 65

   Abstract

   Οι τεχνικές συνεργατικού φιλτραρίσματος (collaborative filtering) επιχειρούν να ανακουφίσουν τον χρήστη από την υπερ-τροφοδότηση πληροφορίας με το να εντοπίζουν ποια στοιχεία ένας χρήστης θα έβρισκε ενδιαφέροντα. Εστιάζουν στον εντοπισμό χρηστών με παρόμοια ενδιαφέροντα και χρησιμοποιούν τις προηγούμενες επιλογές τους προκειμένου να προτείνουν στοιχεία. Συχνά όμως, οι τεχνικές αυτές πάσχουν από το αναφερόμενο πρόβλημα ''νέου χρήστη'' το οποίο λαμβάνει χώρα όταν ένας χρήστης προστίθεται στο σύστημα χωρίς εκείνο να έχει αρκετές πληροφορίας ώστε να κάνει προτάσεις. Το σύστημα επομένως θα πρέπει να αποκτήσει ορισμένα δεδομένα σχετικά με τον χρήστη προκειμένου να αρχίζει να προσφέρει προτάσεις. Σε αυτή την δημοσίευση, παρουσιάζουμε έναν καινοτόμο αλγόριθμο ο οποίος συνδυάζει προηγούμενα αποκτημένη γνώση από την συσταδοποίηση τόσο άρθρων νέων όσο και χρηστών συστήματος προκειμένου να συμπεράνει όσο πιο γρήγορα γίνεται τις προτιμήσεις του χρήστη. Επιχειρούμε να αντιμετωπίσουμε το πρόβλημα ''νέου χρήστη'' προσφέροντας μία προσωποποιημένη στρατηγική παρουσίασης άρθρων νέων στον χρήστη προκειμένου να τα βαθμολογήσει. Η προσέγγισή μας επίσης κάνει χρήση υπερωνύμων τα οποία εξάγονται από το WordNet και φαίνεται να προσεγγίζει γρήγορα στα πραγματικά ενδιαφέροντα του χρήστη με τις λιγότερο δυνατόν απαιτούμενες βαθμολογήσεις άρθρων νέων.

4. User Personalization via W – kmeans. KES2012 - *The 16th International Conference on Knowledge Based & Intelligent Information & Engineering Systems*, San Sebastian, Spain, C. Bouras, V. Tsogkas, 10-12 September 2012, 555 – 564

   Abstract

   Με την ραγδαία ''έκρηξη'' τον online άρθρων νέων, η πρόβλεψη των προτιμήσεων του χρήστη με την χρήση τεχνικών συνεργατικού φιλτραρίσματος έχει εγείρει αρκετό ενδιαφέρον σε σχέση με την προσωποποιημένη πρόσβαση. Παρόλα αυτά, οι συνηθισμένες τεχνικές συνεργατικού φιλτραρίσματος πάσχουν από χαμηλή ακρίβεια και απόδοση. Η δημοσίευση αυτή εστιάζει σε μία νέα προσωποποιημένη προσέγγιση προτάσεων που ενσωματώνει την συσταδοποίση άρθρων νέων και χρηστών, μέσω του αλγορίθμου W-kmeans, μαζί με άλλες τεχνικές ανάκτησης πληροφορίας, όπως κατηγοριοποίηση και περίληψη κειμένου. Το προτεινόμενο σύστημα μπορεί εύκολα να προσαρμόζεται σε διαφοροποιημένες προτιμήσεις χρηστών.

5. Clustering user preferences using W – kmeans. *The 7th International Conference on Signal Image Technology & Internet Based Systems (SITIS 11)*, Dijion - France, C. Bouras, V. Tsogkas, November 28 - December 1 2011, pp. 75 – 82

Abstract

Παρότι συχνά μόνο η συσταδοποίηση κειμένων χρησιμοποιείται ως τεχνική εξόρυξης πληροφορίας από το Web σε συστήματα προτάσεων (recommenders), ένα από τα τμήματα της προσωποποίησης προτάσεων είναι επίσης η συσταδοποίηση των χρηστών. Σε αυτή τη δημοσίευση προτείνουμε μια μεθοδολογία συσταδοποίησης των μοτίβων των χρηστών του Web. Πιο συγκεκριμένα, προσαρμόζουμε τον W-kmeans αλγόριθμο, ο οποίος προηγούμενα χρησιμοποιήθηκε για την περίπτωση της συσταδοποίησης κειμένων, στην περίπτωση της συσταδοποίησης προφίλ χρηστών αναλύοντας τα προηγούμενα μοτίβα τους. Παράλληλα ερευνούμε την επίδραση που έχει αυτή η βελτίωσε όσον αφορά στον μηχανισμό προτάσεων του συστήματος και αξιολογούμε την απόδοσή του σε σχέση με την ακρίβεια – ανάκληση των παραγόμενων προτάσεων προς τους χρήστες.

6. W - kmeans: Clustering News Articles using WordNet. *Advanced Knowledge - based Systems, Invited Session of the 14th International Conference on Knowledge – based and Intelligent Information & Engineering Systems*, Cardiff Wales, UK, C. Bouras, V. Tsogkas, September 8 - 10 2010, pp. 379 – 388

Abstract

Το Web είναι ''γεμάτο'' από άρθρα νέων, μία συντριπτική πηγή πληροφορίας τόσο λόγο της πληθώρας της όσο και της ποικιλομορφίας της. Αντιθέτως, η ανάθεση άρθρων νέων σε παρόμοιες κατηγορίες αποτελεί με μια ισχυρή τεχνική ανάκτησης πληροφορίας και διαχείρισης δεδομένων για αναζήτηση θεματικών κατηγοριών σε κείμενα. Σε αυτή τη δημοσίευση ερευνούμε την εφαρμογή ενός εύρους αλγορίθμων συσταδοποίησης, καθώς με μετρικών ομοιότητας, για την περίπτωση άρθρων νέων τα οποία πηγάζουν από το διαδίκτυο, ενώ παράλληλα συγκρίνουμε την αποδοτικότητά τους για την χρήση μας. Παράλληλα ερευνούμε την επίδραση που έχει η προεπεξεργασία κειμένου στην αργότερα συσταδοποίησή του. Τα πειραματικά αποτελέσματα έδειξαν ότι ο αλγόριθμος k-means, παρά την απλοϊκότητα του, συνδυαζόμενος από ορισμένα βήματα προεπεξεργασίας για τον καθαρισμό, κανονικοποίηση και ενίσχυση των λέξεων κλειδιών του κειμένου, μπορεί να δώσει σημαντικά βελτιωμένα αποτελέσματα όσον αφορά στην ποιότητά τους.

7. Assigning Web News to Clusters. *The Fifth International Conference on Internet and Web Applications and Services, (ICIW 2010)*, Barcelona, Spain, C. Bouras, V. Tsogkas, May 9 - 15 2010

Abstract

Η συσταδοποίηση κειμένου (document clustering) αποτελεί μια ισχυρή τεχνική η οποία έχει χρησιμοποιηθεί ευρέως για την οργάνωση δεδομένων σε μικρότερους και πιο διαχειρίσημους ''πυρήνες'' πληροφορίας. Πολλαπλές προσεγγίσεις έχουν προταθεί στην βιβλιογραφία

οι οποίες υποφέρουν από προβλήματα όπως η συνωνυμία, η αμφισημία καθώς και η έλλειψη μιας περιγραφής των παραγόμενων συστάδων. Σε αυτή τη δημοσίευση προτείνουμε την βελτίωση του τυπικού αλγορίθμου k-means χρησιμοποιώντας την εξωτερική γνώση από υπερώνυμα του WordNet με διττό τρόπο: ενισχύοντας την λίστα από λέξεις (bag of words) που χρησιμοποιούνται πριν από τη διαδικασία συσταδοποίησης και υποβοηθώντας την παραγωγή περιγραφών που ακολουθεί. Η πειραματική μας διαδικασία έδειξε μία σημαντική βελτίωση σε σχέση με τον κλασικό k-means αλγόριθμο για ένα σύνολο άρθρων νέων τα οποία ανακτήθηκαν από πολλαπλά online ειδησεογραφικά πρακτορεία. Παράλληλα η διαδικασία παραγωγής περιγραφών των συστάδων είναι αρκετά αποτελεσματική.

Λοιπές δημοσιεύσεις

Κεφάλαια σε βιβλία

1. Squeak Etoys: Interactive and Collaborative Learning Environment. *Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends, IGI Global,* Chapter 37, C. Bouras, V. Poulopoulos, V. Tsogkas, 2010, pp. 417 - 427

Διεθνή περιοδικά

1. Adaptation of RSS feeds based on the user profile and on the end device. *Journal of Network and Computer Applications, Elsevier Science*, Vol. 33, C. Bouras, V. Poulopoulos, V. Tsogkas, 2010, pp. 410 – 421

2. Noun Retrieval Effect on Text Summarization and Delivery of Personalized News Articles to the User's Desktop. *Data and Knowledge Engineering, Elsevier Science*, Special Issue Advanced Knowledge, Vol. 69, C. Bouras, V. Tsogkas, 2010, pp. 664 – 677

3. Networking and Security Issues for Remote Gaming: The Approach of G@L International Journal on Advances in Security, *IARIA*, Vol. 2, No. 2, 3, C. Bouras, V. Poulopoulos, V. Tsogkas, 2009, pp. 171 - 181

4. PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries. *Data and Knowledge Engineering Journal, Elsevier Science*, 2008, Vol. 64, Issue 1 , C. Bouras, V. Poulopoulos, V. Tsogkas, 2008, pp. 330 - 345

Διεθνή συνέδρια

1. Caching News Channels on the User's Desktop. *IADIS International Conference Applied Computing*, Rome, Italy, C. Bouras, G. Tsichritzis, V. Tsogkas, November 19 - 21 2009, pp. 35 – 42

2. Personalization Mechanism for Delivering News Articles on the User's Desktop. *The Fourth International Conference on Internet and Web Applications and Services – ICIW 2009*, Venice, Italy, C. Bouras, V. Tsogkas, 24 - 28 May 2009, pp. 157 – 162

3. Networking Aspects for the Security of Game Input. *5th IEEE International Workshop on Networking Issues in Multimedia Entertaiment - NIME09*, Las Vegas, USA, C. Bouras, V. Poulopoulos, V. Tsogkas, 13 January 2009

4. Evaluating PeRSSonal: A Medium for Personalized Dynamically Created News Feeds. *IADIS International Conference WWW/Internet Freiburg*, Germany, C. Bouras, V. Poulopoulos, V. Tsogkas, 13 - 15 October 2008

5. Improving text summarization using noun retrieval techniques. *Advanced Knowledge – based Systems, Invited Session of the 12nd International Conference on Knowledge – based and Intelligent Information & Engineering Systems(KES 2008)*, Zagreb, Croatia, C. Bouras, V. Tsogkas, 3 - 5 September 2008, pp. 593 – 600

6. Creating dynamic personalized RSS summaries. *8th Industrial Conference on Data Mining – ICDM 2008*, , Leipzig, Germany, C. Bouras, V. Poulopoulos, V. Tsogkas, 16 - 18 July 2008, pp. 1 – 15

7. Networking Aspects for Gaming Systems. *Third International Conference on Internet and Web Applications (ICIW 2008)*, Athens, Greece, C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas, 8 - 13 June 2008, pp. 650 – 655

8. Efficient Summarization Based On Categorized Keywords. *The 2007 International Conference on Data Mining (DMIN07)*, Las Vegas, Nevada, USA, C. Bouras, V. Poulopoulos, V. Tsogkas, 25 - 28 June 2007

9. Personalizing text summarization based on sentence weighting. *IADIS European First International Conference Data Mining (ECDM 2007)*, Lisbon, Portugal, C. Bouras, V. Poulopoulos, V. Tsogkas, 3 - 8 July 2007, pp. 3 – 10

10. Input here - Execute there through networks: the case of gaming. *The 15th Workshop on Local and Metropolitan Area Networks (LANMAN 2007)*, Princeton, NJ, USA, C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas, 10 - 13 June 2007

11. The importance of the difference in text types to keyword extraction: Evaluating a mechanism. *7th International Conference on Internet Computing 2006 (ICOMP 2006)*, Las Vegas, Nevada, USA, C. Bouras, C. Dimitriou, V. Poulopoulos, V. Tsogkas, 26 - 29 June 2006, pp. 43 - 49

Αναφορές από άλλους ερευνητές

- PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries. Data and Knowledge Engineering Journal, Elsevier Science, 2008, Vol. 64, Issue 1 , C. Bouras, V. Poulopoulos, V. Tsogkas, 2008, pp. 330 - 345

    1. Web News Portal Content Personalization using Information Extraction Techniques and Weighted Voronoi Diagrams. Ševa, J., 2014.
    2. Hybridization of EM and SVM clusters for efficient text categorization, Murugan, S. A., & Suresh, P. 2014.
    3. A Knowledge Document Structured Summarization Model. International Journal of Electronic Business 11.1, Yang, Shih-Ting, and Yu-Ting Gong., 2013, pp.60-72
    4. Combining summaries using unsupervised rank aggregation. Computational Linguistics and Intelligent Text Processing. Palshikar, Girish Keshav, Shailesh Deshpande, and G. Athiappan Springer Berlin Heidelberg, 2012, pp. 378-389

- Improving text summarization using noun retrieval techniques. Advanced Knowledge – based Systems, Invited Session of the 12nd International Conference on Knowledge – based and Intelligent Information & Engineering Systems(KES 2008), Zagreb, Croatia, C. Bouras, V. Tsogkas, 3 - 5 September 2008, pp. 593 – 600

    1. Latent semantic sentence clustering for multi-document summarization. Geiß, Johanna. University of Cambridge, Computer Laboratory, Technical Report UCAM-CL-TR-802 (2011).
    2. Topic-Dependent-Class-Based-Gram Language Model. Audio, Speech, and Language Processing. Naptali, Welly, Masatoshi Tsuchiya, and Seiichi Nakagawa., IEEE Transactions on 20.5 (2012): 1513-1525.
    3. An alternative approach for statistical single-label document classification of newspaper articles. Mamakis, Georgios, Athanasios G. Malamos, and J. Andrew Ware. Journal of Information Science (2011).
    4. A review of retrospective news event detection. Semantic Technology and Information Retrieval (STAIR), Ramadan, Qusai Hussein, and Masnizah Mohd. 2011 International Conference on. IEEE, 2011.
    5. i-JEN: visual interactive Malaysia crime news retrieval system. Visual Informatics: Sustaining Research and Innovations. Ali, Nazlena Mohamad, et al. Springer Berlin Heidelberg, 2011. 284-294.
    6. A Framework for Progressive Trusting Services. International Journal On Advances in Intelligent Systems 3.3 and 4. Dini, Oana, Pascal Lorenz, and Hervé Guyennet. (2011): 326-346.

7. Document Classification in Summarization. Journal of Information and Computing Science 7.1. Mamakis, Georgios, et al. (2012): 025-036.
8. Online Service Similarities and Reputation-based Selection. The Second International Conferences on Advanced Service Computing. 2010. Dini, Oana, et al. SERVICE COMPUTATION 2010

- Personalization Mechanism for Delivering News Articles on the User's Desktop. The Fourth International Conference on Internet and Web ASpplications and Services – ICIW 2009, Venice, Italy, C. Bouras, V. Tsogkas, 24 - 28 May 2009, pp. 157 – 162

1. Content-based news recommendation. E-commerce and web technologies. Kompan, Michal, and Mária Bieliková. Springer Berlin Heidelberg, 2010. 61-72.
2. Effective hierarchical vector-based news representation for personalized recommendation. Computer Science and Information Systems 9.1. Bieliková, Mária, Michal Kompan, and Dušan Zeleník (2012): 303-322.
3. Semantic metadata in the news production process: achievements and challenges. Proceeding of the 16th International Academic MindTrek Conference. Pellegrini, Tassilo. ACM, 2012.
4. Integrating linked data into the content value chain: a review of news-related standards, methodologies and licensing requirements. Proceedings of the 8th International Conference on Semantic Systems. Pellegrini, Tassilo. ACM, 2012.
5. The Economics of Big Data: A Value Perspective on State of the Art and Future Trends. Big Data Computing. Pellegrini, Tassilo. New York: Chapman and Hall/CRC (2013): 343-371.
6. Classifying News Headlines for Providing User Centered E-Newspaper Using SVM. Deshmukh, R. R., and Mr DK Kirange
7. Vector-based tree news recommendation. Bielikova, Mária, Michal Kompan, and Dušan Zelenik.

- A clustering technique for news articles using WordNet. Knowledge-Based Systems Journal, Elsevier Science, Vol. 36, C. Bouras, V. Tsogkas, 2012, 115 – 128

1. Subset K-Means Approach for Handling Imbalanced-Distributed Data., Kumar, Ch N. Santhosh, et al. Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Springer International Publishing, 2015.
2. Undersampled K-means approach for handling imbalanced distributed data. Progress in Artificial Intelligence. Kumar, N. Santhosh, et al., 2014: 1-10.
3. Ninaus, G., Reinfrank, F., Stettinger, M., & Felfernig, A. Content-Based Recommendation Techniques for Requirements Engineering., 2014

4. An updated literature review on the problem of Class Imbalanced Learning in Clustering. Kumar, Ch N. Santhosh, et al.
5. Clustering based on Cuckoo Optimization Algorithm. Intelligent Systems (ICIS). Ameryan, Mahya, Mohammad Reza Akbarzadeh Totonchi, and Seyyed Javad Seyyed Mahdavi. Iranian Conference on. IEEE, 2014.
6. Locality mutual clustering for document retrieval. Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication. Nguyen, Khu Phi, and Hong Tuyet Tu. ACM, 2014.
7. Performance Evaluation of Semantic Approaches for Automatic Clustering of Similar Web Services. Computing and Communication Technologies (WCCCT), Vadivelou, G., and E. Ilavarasan. 2014 World Congress on. IEEE, 2014.
8. Clustering-based topical Web crawling using CFu-tree guided by link-context. Frontiers of Computer Science: 1-15. Liu, Lu, and Tao Peng
9. Imbalanced K-Means: An algorithm to cluster imbalanced-distributed data. Kumar, Ch N. Santhosh, et al

- Assigning Web News to Clusters. The Fifth International Conference on Internet and Web Applications and Services, (ICIW 2010), Barcelona, Spain, C. Bouras, V. Tsogkas, May 9 - 15 2010

1. A survey of techniques for event detection in Twitter. Computational Intelligence (2013). Atefeh, Farzindar, and Wael Khreich
2. A review of retrospective news event detection. Semantic Technology and Information Retrieval (STAIR) 2011 International Conference on. IEEE, Ramadan, Qusai Hussein, and Masnizah Mohd., 2011.
3. OPTIMAL INITIAL CENTROID IN K-MEANS FOR CRIME TOPIC. Mohd, Masnizah. (2010).
4. i-JEN: visual interactive Malaysia crime news retrieval system. Visual Informatics: Sustaining Research and Innovations. Ali, Nazlena Mohamad, et al. Springer Berlin Heidelberg, 2011. 284-294.
5. Feedback-driven clustering for automated linking of web pages. 8th International Conference for Internet Technology and Secured Transactions (ICITST), Oest, Adam, and Manjeet Rege. IEEE, 2013.
6. Information Integration in News Articles from Various Sources. Holub, Michal
7. An Intelligent Document Clustering Approach to Detect Crime Patterns. Procedia Technology 11. Bsoul, Qusay, Juhana Salim, and Lailatul Qadri Zakaria. (2013): 1181-1187.
8. Article Recommendations for News Feed. Shen, Minghan

- Networking Aspects for Gaming Systems. Third International Conference on Internet and Web Applications (ICIW 2008), Athens, Greece, C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas, 8 - 13 June 2008, pp. 650 – 655

    1. Games@ large distributed gaming system. Proc. of Networked & Electronic Media Summit (NEM2009). Laikari, Arto, et al. Saint-Malo, France (2009).
    2. Gaming platform for running games on low-end devices. User Centric Media. Laikari, Arto, et al. Springer Berlin Heidelberg, 2010. 259-262.
    3. Graph of Game Worlds: New Perspectives on Video Game Architectures. Zhu, M. E. N. G., et al. Manuscript submitted for publication (2012).
    4. Entertainment Services-Distributed 3D Gaming System. Laikari, Arto, Editor: Pentti Vähä Graphic design: Tuija Soininen (2009): 68.
    5. Game Streaming Prototypen mit Hilfe von Serverseitigem Rendering. Moser, Mario. Entwurf eines. na, 2010.
    6. Software Architectures and the Creative Processes in Game Development. Wang, Alf Inge, and Njål Nordmark, 2014

- Clustering user preferences using W – kmeans. The 7th International Conference on Signal Image Technology & Internet Based Systems (SITIS 11), Dijion - France, C. Bouras, V. Tsogkas, November 28 - December 1 2011, pp. 75 – 82

    1. Semantic preserving text tepresentation and its applications in text clustering. Howard, Michael. (2012).

- Noun Retrieval Effect on Text Summarization and Delivery of Personalized News Articles to the User's Desktop. Data and Knowledge Engineering, Elsevier Science, Special Issue Advanced Knowledge, Vol. 69, C. Bouras, V. Tsogkas, 2010, pp. 664 – 677

    1. SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering. Oliva, Jesús, et al. 70.4 (2011): 390-405.
    2. Analysis and study on text representation to improve the accuracy of the normalized compression distance. Granados, Ana. AI Communications 25.4 (2012): 381-384.
    3. Is the contextual information relevant in text clustering by compression?. Granados, Ana, David Camacho, and Francisco Borja Rodríguez. Expert Systems with Applications 39.10 (2012): 8537-8546.
    4. COMPENDIUM: A text summarization system for generating abstracts of research papers. Natural Language Processing and Information Systems. Lloret, Elena, María Teresa Romá-Ferri, and Manuel Palomar. Springer Berlin Heidelberg, 2011. 3-14.
    5. Analysis and study on text representation to improve the accuracy of the normalized compression distance. Granados Fontecha, Ana (2012).

6. Web Service to Execute A Datamining Task. Velkumar, R., A. Muthukumaravel, and N. Sathya

- W - kmeans: Clustering News Articles using WordNet. Advanced Knowledge - based Systems, Invited Session of the 14th International Conference on Knowledge – based and Intelligent Information & Engineering Systems, Cardiff Wales, UK, C. Bouras, V. Tsogkas, September 8 - 10 2010, pp. 379 – 388

    1. Keen-Means: A Web Page Clustering Tool Based on an Self-Adjustable K-Means Algorithm. Tseng, Chun Hsiung, et al. Ubi-Media Computing and Workshops (UMEDIA), 2014 7th International Conference on. IEEE, 2014.
    2. Semantic Framework to Text Clustering with Neighbors. ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II. Lalitha, Y. Sri, and A. Govardhan. Springer International Publishing, 2014.
    3. Beyond cluster labeling: Semantic interpretation of clusters' contents using a graph representation. Knowledge-Based Systems 56. Role, François, and Mohamed Nadif. (2014): 141-155.
    4. Clustering system based on text mining using the K-means algorithm: news headlines clustering. Lama, Prabin (2013).

- Adaptation of RSS feeds based on the user profile and on the end device. Journal of Network and Computer Applications, Elsevier Science, Vol. 33, C. Bouras, V. Poulopoulos, V. Tsogkas, 2010, pp. 410 – 421

    1. Development and performance evaluation of a new RSS tool for a Web-based system: RSS_PROYECT. Journal of Network and Computer Applications 36.1. De La Torre-DíEz, Isabel, et al (2013): 255-261.
    2. Automatic multi-label categorization of news feeds. Darabi, Majid, Hossein Adeli, and Nasseh Tabrizi

- Creating dynamic personalized RSS summaries. 8th Industrial Conference on Data Mining – ICDM 2008, , Leipzig, Germany, C. Bouras, V. Poulopoulos, V. Tsogkas, 16 - 18 July 2008, pp. 1 – 15

    1. RSS feeds behavior analysis, structure and vocabulary. Travers, Nicolas, et al. International Journal of web information systems 10.3 (2014): 5-5.
    2. Characterizing web syndication behavior and content. Web Information System Engineering–WISE 2011. Hmedeh, Zeinab, et al. Springer Berlin Heidelberg, 2011. 29-42.
    3. Everything you would like to know about RSS feeds and you are afraid to ask. BDA'11, Base de Données Avancées. Hmedeh, Zeinab, et al. (2011): 1-20.

ΠΡΟΛΟΓΟΣ

Όταν το καλοκαίρι του 2002 μάθαινα, με απερίγραπτη χαρά, ότι γινόμουν δεκτός στο ΤΜΗΥΠ του Πανεπιστημίου Πατρών, ποτέ δεν θα περίμενα ότι 12 χρόνια αργότερα, θα ολοκλήρωνα ένα κείμενο σαν το παρόν. Μία διδακτορική διατριβή, η οποία αποτελεί το επιστέγασμα προσπαθειών, χρόνου αλλά και προσωπικής χαράς και ικανοποίησης από κάθε άποψη.

Η θεματολογία με την οποία ασχολήθηκα από την προπτυχιακή διπλωματική μου εργασία, στην μεταπτυχιακή μου εργασία και τώρα στην διδακτορική μου διατριβή, ήταν μία βήμα προς βήμα προσέγγιση, ένα υπέροχο ταξίδι στους συγκεκριμένους τομείς της επιστήμης των υπολογιστών που θεωρώ ότι μου προσέφερε σημαντικά εφόδια σαν μηχανικό, επιστήμονα, αλλά και πάνω απ' όλα σαν άνθρωπο. Θεωρώ τον εαυτό μου εξαιρετικά τυχερό που είχα την ευκαιρία να κάνω αυτό το ταξίδι σε αυτή τη σχολή και με αυτόν τον τρόπο.

Βρισκόμενος λοιπόν πριν από το 'τέλος του δρόμου' θα ήθελα να ευχαριστήσω ορισμένους ανθρώπους που πραγματικά με βοήθησαν όλα αυτά τα χρόνια, είτε σε ακαδημαϊκό, είτε σε προσωπικό επίπεδο, και χωρίς τους οποίους ίσως να μην βρισκόμουν σε αυτή την ευχάριστη για μένα θέση. Θα ήθελα λοιπόν να ευχαριστήσω τον καθηγητή μου Χρήστο Μπούρα για την στήριξη και υπομονή που έδειξε προς το πρόσωπό μου όλα αυτά τα χρόνια. Ο ιδιαίτερος τρόπος με τον οποίο αντιμετώπιζε ότι είχε να κάνει με την ακαδημαϊκή μου σταδιοδρομία, μου έδινε ώθηση και όραμα για να συνεχίζω την προσπάθεια. Επιπλέον, ευχαριστώ θερμά τον καθηγητή κ. Ευστράτιο Γαλλόπουλο και τον επίκουρο καθηγητή κ. Χρήστο Μακρή για την συμμετοχή και στήριξή τους ως μέλη της τριμελούς επιτροπής, τόσο στην μεταπτυχιακή μου εργασία, όσο και στην παρούσα. Επίσης, ευχαριστώ τους καθηγητές κ. Νικόλαο Αβούρη, Αθανάσιο Τσακαλίδη, Ιωάννη Γαροφαλάκη και Βασίλειο Μεγαλοοικονόμου για την συμμετοχή τους στην επταμελή επιτροπή αξιολόγησης της παρούσας διδακτορικής διατριβής.

Ευχαριστώ από τα βάθη της καρδιάς μου τους γονείς μου, Θρασύβουλο και Θεοδώρα, που με την αγάπη και τις αξίες που με μεγάλωσαν, με έκαναν έναν χρήσιμο, σκεπτόμενο και πάνω απ' όλα Άνθρωπο. Την αδερφή μου, Αλεξάνδρα, για την καταπληκτική παιδική ηλικία που μοιραστήκαμε και συχνά αναπολώ και για το χαμόγελό της.

Η εργασία αυτή είναι αφιερωμένη στους δύο ανθρώπους που μοιράζονται την πρώτη θέση στην

καρδιά μου. Στην γυναίκα μου, Αντιγόνη, που είναι πάντα δίπλα μου, συνοδοιπόρος, στα καλά και στα άσχημα, με υπομονή και αγάπη για να με στηρίζει. Και φυσικά στο γιο μου, το νόημα και το φως της ζωή μου, τον λόγο για τον οποίο αισθάνομαι πραγματικά υπερήφανος σε αυτό τον κόσμο.

Κλείνοντας θα ήθελα να εκφράζω την ελπίδα μου προς τον αναγνώστη ότι η ανάγνωση της διδακτορικής διατριβής θα είναι τόσο ευχάριστη, ενδιαφέρουσα και δημιουργική όσο ήταν η συγγραφή της.

Τσόγκας Βασίλης, Πάτρα, Δεκέμβριος 2014

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

> Above all things, reverence yourself.
>
> Pythagoras, Greek Mathematician, 497 BC

Το παρόν κεφάλαιο παρουσιάζει γενικά στοιχεία για την διδακτορική διατριβή που πραγματοποιήθηκε, δίνει ορισμένες εισαγωγικές πληροφορίες αγγίζοντας τις ερευνητικές περιοχές με τις οποίες καταπιάνεται και παραθέτει τη δομή της εργασίας.

1.1 Γενικά

Ζούμε σε μια κοινωνία αλλαγής και προόδου. Σε μια κοινωνία που χαρακτηρίζεται από τον τεράστιο όγκο της πληροφορίας που διακινείται μέσα στις τάξεις της. Κυρίως όμως διανύουμε την εποχή της κατάργησης των συνόρων και της αδιάλειπτης επικοινωνίας μεταξύ των ανθρώπων. Το διαδίκτυο αποτελεί τον τροχό γι' αυτές τις αλλαγές, η ποσότητα όμως των δεδομένων που υπάρχουν και διακινούνται μέσω αυτού είναι τόσο τεράστια, ώστε να αποσπά τους πολίτες της κοινωνίας αυτής στην προσπάθειά τους να βρουν χρήσιμη πληροφορία και επομένως να μετατρέπεται σε τροχοπέδη της αλλαγής.

1.2 Υπάρχουσα κατάσταση

Τα άρθρα νέων πλημμυρίζουν το διαδίκτυο τόσο με το ακραία μεγάλο πλήθος τους, τόσο και από την ολοένα και αυξανόμενη συχνότητα εμφάνιση των πηγών τους. Είναι πρακτικά αδύνατο για έναν χρήστη του διαδικτύου σήμερα να μπορέσει να παρακολουθήσει χωρίς βοήθεια (π.χ. φιλτράρισμα ή μέσω προτάσεων) ένα γεγονός ή μια σειρά γεγονότων που τον ενδιαφέρουν. Παράλληλα, η αμεροληψία στην ενημέρωση είναι ένα μείζον θέμα το οποίο δύσκολα αντιμετωπίζεται δίχως σφαιρική ενημέρωση επί των εν' λόγω γεγονότων από πολλαπλές πηγές.

Από την άλλη πλευρά, η συσταδοποίηση άρθρων νέων παρέχει ένα ισχυρό εργαλείο από το πεδίο της ανάκτησης πληροφορίας για τον εντοπισμό θεμάτων (συστάδων) πληροφορίας σε κείμενα. Η συσταδοποίηση μπορεί συνεπώς να αποτυπώσει την υποκείμενη ιεραρχία περιεχομένου μεγάλου πλήθους αντικειμένων, παρέχοντας έτσι στα συστήματα ανάκτησης πληροφορίας (π.χ. συστήματα προτάσεων) την δυνατότητα διευκόλυνσης των χρηστών, βοηθώντας έτσι στην αντιμετώπιση της προαναφερθείσας κατάστασης.

1.3 Περιγραφή της εργασίας

Η παρούσα διδακτορική διατριβή προσβλέπει στο σχεδιασμό, στην ανάπτυξη και τελικά στην αξιολόγηση μηχανισμών και καινοτόμων αλγορίθμων από τις περιοχές της ανάκτησης πληροφορίας, της επεξεργασίας φυσικής γλώσσας καθώς και της μηχανικής εκμάθησης που θα παρέχουν ένα υψηλό επίπεδο φιλτραρίσματος των άρθρων νέων του διαδικτύου προς τον τελικό χρήστη. Πιο συγκεκριμένα, στα διάφορα στάδια επεξεργασίας της πληροφορίας αναπτύσσονται τεχνικές και μηχανισμοί που συλλέγουν, δεικτοδοτούν, φιλτράρουν και επιστρέφουν κατάλληλα στους χρήστες κειμενικό περιεχόμενο που πηγάζει από τον παγκόσμιο ιστό.

Πυρήνας της διδακτορικής διατριβής είναι η ανάπτυξη ενός μηχανισμού συσταδοποίησης (clustering) τόσο κειμένων, όσο και των χρηστών του διαδικτύου. Στο πλαίσιο αυτό μελετήθηκαν κλασικοί αλγόριθμοι συσταδοποίησης οι οποίοι και αξιολογήθηκαν για την περίπτωση των άρθρων, κειμένου προκειμένου να εκτιμηθεί αν και πόσο αποτελεσματικός είναι ο εκάστοτε αλγόριθμος. Σε δεύτερη

φάση υλοποιήθηκε αλγόριθμος συσταδοποίησης άρθρων νέων που αξιοποιεί μια εξωτερική βάση γνώσης, το WordNet, και είναι προσαρμοσμένος στις απαιτήσεις των άρθρων νέων που πηγάζουν από το διαδίκτυο. Ένας ακόμη βασικός στόχος της παρούσας εργασίας είναι η μοντελοποίηση των κινήσεων που ακολουθούν κοινοί χρήστες καθώς και η αυτοματοποιημένη αξιολόγηση των συμπεριφορών, με ορατό θετικό αποτέλεσμα την πρόβλεψη των προτιμήσεων που θα εκφράσουν στο μέλλον οι χρήστες. Η μοντελοποίηση των χρηστών έχει άμεση εφαρμογή στις δυνατότητες προσωποποίησης της πληροφορίας με την πρόβλεψη των προτιμήσεων των χρηστών. Ως εκ' τούτου, υλοποιήθηκε αλγόριθμος προσωποποίησης ο οποίος λαμβάνει υπ' όψιν του πληθώρα παραμέτρων που αποκαλύπτουν έμμεσα τις προτιμήσεις των χρηστών.

Σκοπός της διδακτορικής διατριβής είναι η επέκταση και η βελτίωση, προς συγκεκριμένες κατευθύνσεις, του μηχανισμού που δημιουργήθηκε στα πλαίσια της μεταπτυχιακής διπλωματικής εργασίας που εκπόνησα με τίτλο *"Προσωποποιημένη Προβολή Περιεχομένου του διαδικτύου σε Desktop Εφαρμογή με Τεχνικές ανάκτησης δεδομένων, προεπεξεργασίας κειμένου, αυτόματης κατηγοριοποίησης και εξαγωγής περίληψης"* [235]. Στα πλαίσια της παραπάνω μεταπτυχιακής εργασίας, δημιουργήθηκε ένας ολοκληρωμένος μηχανισμός ο οποίος μπορεί αυτόματα να κάνει ανάλυση σε κείμενα του διαδικτύου προκειμένου να εξάγει λέξεις-κλειδιά. Μέσα από αυτή την ανάλυση προκύπτουν οι σημαντικότερες προτάσεις του κειμένου που το χαρακτηρίζουν, και οι οποίες μπορούν, αν συνενωθούν, να αποτελέσουν μια σύντομη περίληψη του κειμένου. Ο μηχανισμός αξιοποιεί γνώσεις για την κατηγορία του κειμένου καθώς και για τις προτιμήσεις που παρουσιάζουν οι χρήστες προκειμένου να βελτιώσει και να φιλτράρει τα αποτελέσματα που παρουσιάζονται. Το σύστημα που κατασκευάστηκε έχει τα εξής βασικά υποσυστήματα: μηχανισμός ανάκτησης δεδομένων και εξαγωγής χρήσιμου κειμένου από τον παγκόσμιο ιστό, μηχανισμός εξαγωγής λέξεων-κλειδιών από το πηγαίο κείμενο, μηχανισμός κατηγοριοποίησης κειμένου, ο οποίος μπορεί να συμμετάσχει στη διαδικασία εξαγωγής περίληψης και να ενδυναμώσει τα αποτελέσματά της, μηχανισμοί προσωποποίησης περιεχομένου στο χρήστη και φυσικά, μηχανισμός εξαγωγής περίληψης. Οι παραπάνω μηχανισμοί είναι ενσωματωμένοι στο σύστημα αποδελτίωσης PeRSSonal [171], το οποίο χρησιμοποιείται για την ανάκτηση, προεπεξεργασία, κατηγοριοποίηση, προσωποποίηση και περίληψη άρθρων από ειδησεογραφικούς τόπους του διαδικτύου.

Για τη δημιουργία του μηχανισμού PeRSSonal συμμετείχαν οι Βασίλης Πουλόπουλος (συντονισμός εργασίας, κατασκευή ολοκληρωμένου διαδικτυακού περιβάλλοντος PeRSSonal, δημιουργία αλγορίθμων συγκέντρωσης κειμένων, εξαγωγής εικόνων, κατηγοριοποίησης, εξαγωγής περιλήψεων, προσωποποίησης, προσαρμογής στο χρήστη) [36] [35] [8] [37] [14], Γεώργιος Αδάμ (advaRSS, cuter, m-cuter + υποστήριξη συνολικά του συστήματος) [5] [6] [7] [4], Κωνσταντίνος Ασημάκης (greek stemmer and tagger) [5], Γεώργιος Τσιχριτζής (garbage article location) [38] και Βασίλης Τσόγκας (οι δημοσιεύσεις δίνονται στην επιτελική σύνοψη), ενώ για κομμάτια τα οποία δεν μπήκαν ποτέ στο μηχανισμό για ερευνητικούς λόγους έχουν εργασθεί οι Αντωνέλλης Ιωάννης και Σιλιντζήρης Παναγιώτης, εργασίες των οποίων έχουν δημοσιευθεί.

Η παρούσα διδακτορική διατριβή επομένως "χτίζει" πάνω και επεκτείνει τα αποτελέσματα της μεταπτυχιακής εργασίας και ως εκ' τούτου μοιράζεται ένα αρκετά μεγάλο κομμάτι των μηχανισμών

και αλγορίθμων. Κατά συνέπεια, ερευνητικά θέματα ή αλγοριθμικά κομμάτια που παραμένουν αμετάβλητα δεν αναλύονται διεξοδικά στην παρούσα διδακτορική διατριβή. Αντ' αυτού, αναφέρονται συνοπτικά ή προτείνεται στον αναγνώστη να ανατρέξει στα σχετικά εδάφια της μεταπτυχιακής εργασίας όπου αυτό κρίνεται αναγκαίο.

Ο σκοπός λοιπόν της παρούσας εργασίας είναι διττός. Πρώτον, η ενίσχυση ορισμένων από των υπαρχόντων διαδικασιών του μηχανισμού που δημιουργήθηκε πρότερα με αποτελεσματικότερες μεθόδους, ευρετικά και αλγορίθμους. Δεύτερο, η ανάπτυξη και αξιοποίηση αλγορίθμου συσταδοποίησης άρθρων νέων και χρηστών του συστήματος καθώς και η μελέτη της βέλτιστης αλληλεπίδρασης των υποσυστημάτων με την νέα παράμετρο της συσταδοποίησης πληροφορίας. Φυσικά τα παραπάνω αξιολογούνται τόσο αυτοτελώς όσο και σε συνδυασμό μεταξύ τους προκειμένου να αποδειχθεί η χρησιμότητά τους συγκεκριμένα για το σύστημά μας αλλά και γενικά για τα συστήματα προτάσεων άρθρων νέων.

Πιο συγκεκριμένα λοιπόν, στο στάδιο προεπεξεργασίας κειμένου, οι αλγόριθμοι αναγνώρισης και εξαγωγής χρήσιμου κειμένου έχουν ενισχυθεί και βελτιστοποιηθεί ώστε να εκτελούνται ταχύτερα και να επιστρέφουν με υψηλότερη ακρίβεια το περιεχόμενο που ανταποκρίνεται στο ωφέλιμο κείμενο μιας ιστοσελίδας. Συνοπτικά, η βελτίωση αφορά στη ανάκτηση και αξιοποίηση n-grams λέξεων καθώς και στην χρήση της εξωτερικής βάσης γνώσης WordNet. Η εφαρμογή των νέων τεχνικών προεπεξεργασίας κειμένου έχει ως αποτέλεσμα την καλύτερη νοηματική απεικόνιση των άρθρων νέων στον διανυσματικό χώρο των λέξεων κλειδιών και n-grams που ανακτούνται, κάτι που όπως αποδεικνύεται και πειραματικά, έχει αξιόλογα οφέλη για τις διαδικασίες που ακολουθούν. Ιδιαίτερα δε για την προσωποποιημένη επιλογή άρθρων νέων στα μέτρα του εκάστοτε χρήστη, η οποία και είναι ο βασικός στόχος ενός συστήματος προτάσεων.

Στη συνέχεια ακολουθεί το ολοκαίνουργιο υποσύστημα συσταδοποίησης δεδομένων που λειτουργεί τόσο σε άρθρα νέων όσο και χρήστες του συστήματος. Για το υποσύστημα αυτό, αφού μελετήθηκε και αξιολογήθηκε μια πληθώρα αλγορίθμων συσταδοποίησης, τόσο ιεραρχικών (hierarchical) όσο και διαιρετικών (partitional), ερευνήθηκε και υλοποιήθηκε μία νέα παραλλαγή του πασίγνωστου αλγορίθμου συσταδοποίησης, k-means. Ο αλγόριθμος αυτός, τον οποίο και ονομάσαμε W-kmeans (WordNet-enabled k-means), αξιοποιεί την εξωτερική βάση γνώσης WordNet προκειμένου να ενισχύσει την υπάρχουσα κειμενική πληροφορία με παρόμοια/παραπλήσια, αξιοποιώντας την σχέση υπερωνύμων/υποωνύμων που ανακτάται από το WordNet. Στοχεύει επομένως στην εύρεση υποκείμενων σχέσεων μεταξύ άρθρων ή χρηστών που συχνά δεν καταγράφονται μόνο με την χρήση των λέξεων κλειδιών που αποτελούν μέρος αυτών.

Η προσωποποιημένη παρουσίαση των αποτελεσμάτων στη μεριά του χρήστη επίσης ενισχύεται μέσω των τεχνικών συσταδοποίησης. Ο αλγόριθμος προσωποποίησης λαμβάνει υπ' όψιν του πολλές παραμέτρους, μεταξύ των οποίων το ιστορικό περιήγησης, οι χρόνοι που μένει ο χρήστης σε κάποιο άρθρο, οι επιλογές του και φυσικά τα αποτελέσματα της συσταδοποίησης, με σκοπό να παράγει το προφίλ του. Ο αλγόριθμος προσωποποίησης που προτείνεται ουσιαστικά "μαθαίνει" από τις επιλογές του χρήστη και προσαρμόζεται στις πραγματικές προτιμήσεις του με το πέρασμα του χρόνου. Έτσι το σύστημα μπορεί να ανταποκρίνεται στις διαρκώς μεταβαλλόμενες προτιμήσεις των χρηστών,

στοιχείο εξαιρετικά επωφελές για ένα σύστημα προτάσεων.

Μία ακόμη άμεση αξιοποίηση του νέου αλγορίθμου W-kmeans που αποτέλεσε επίσης τμήμα της διδακτορικής διατριβής ήταν η αντιμετώπιση του προβλήματος νέου χρήστη. Το πρόβλημα αυτό αποτελεί μια κατάσταση με την οποία έρχονται συχνά αντιμέτωπα τα συστήματα προτάσεων και που επηρεάζει αρνητικά την απόδοσή τους. Η αξιοποίηση της πληροφορίας συσταδοποίησης ως προς αυτή την κατεύθυνση μας βοήθησε μέσω συγκεκριμένων αλγοριθμικών βημάτων να αντιμετωπίσουμε πρακτικά και με λίγα βήματα το εν' λόγω πρόβλημα, αξιοποιώντας μία λογική ανατροφοδότηση σχετικά με τις επιλογές αξιολόγησης που πραγματοποιούν οι χρήστες.

Συνολικά, μέσα από την εργασία προέκυψαν αποτελέσματα που έχουν να κάνουν με σύγκριση αλγορίθμων σε όλα τα παραπάνω στάδια του μηχανισμού αλλά και ανταπόκριση του μηχανισμού στις ανάγκες του χρήστη. Τα αποτελέσματα αυτά, τα οποία και παρουσιάζονται, είναι ιδιαίτερα ενθαρρυντικά και μας παρακινούν για περαιτέρω έρευνα στα θέματα με τα οποία καταπιαστήκαμε, καθώς και στα γενικότερα ερευνητικά πεδία που αυτά αναφέρονται.

1.4 Δομή της εργασίας

Η υπόλοιπη εργασία δομείται ως εξής: στο κεφάλαιο 2 γίνεται μία γενικότερη καταγραφή των προβλημάτων στα οποία απευθύνεται η διδακτορική διατριβή. Στο κεφάλαιο 3 παρουσιάζονται οι τρέχουσες εξελίξεις στα ερευνητικά πεδία που μας αφορούν (State of the Art) καθώς και οι σχετικές εργασίες πάνω στις οποίες βασίζεται η διδακτορική διατριβή. Στο κεφάλαιο 4 γίνεται μια γενικότερη περιγραφή της αρχιτεκτονικής και των χαρακτηριστικών που προτείνεται για ένα σύστημα προτάσεων άρθρων νέων - το σύστημα δηλαδή που προϋπήρχε και η παρούσα διατριβή αναβαθμίζει. Ακολουθεί η παρουσίαση των αλγορίθμων που αναπτύχθηκαν για καθένα από τα υποσυστήματα (κεφάλαιο 5). Στο κεφάλαιο 6 παρουσιάζονται οι τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος καθώς και οι προδιαγραφές του. Στο κεφάλαιο 7 γίνεται μια αναλυτική παρουσίαση των δεδομένων και των πειραματικών αποτελεσμάτων που αφορούν στην αξιολόγηση του συστήματος. Στο κεφάλαιο 8 δίνονται τα συμπεράσματα που προέκυψαν από την εργασία και τέλος στο κεφάλαιο 9 παρουσιάζονται κάποιες προτάσεις για μελλοντική επέκταση του μηχανισμού, καθώς και η γενικότερη μελλοντική εργασία που θα μπορούσε να γίνει σε καθένα από τα υποσυστήματα με τα οποία καταπιαστήκαμε.

ΚΕΦΑΛΑΙΟ 2

ΚΑΘΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

> The only true wisdom is in
> knowing you know nothing.
>
> *Socrates, Greek Philosopher, 469 BC*

Στο παρόν κεφάλαιο γίνεται μία συνοπτική παρουσίαση των θεμάτων με τα οποία καταπιάνεται η διδακτορική διατριβή. Αναφέρουμε τα προβλήματα που αφορούν στην καθημερινή χρήση του διαδικτύου και εξηγούμε πως και γιατί προσπαθούμε να τα επιλύσουμε. Πιο συγκεκριμένα, παρουσιάζονται ορισμένες προβληματικές καταστάσεις οι οποίες είναι συχνές στο διαδίκτυο και αφορούν: α) στο τρόπο που μπορεί να γίνει αποτελεσματικότερο το φιλτράρισμα πληροφορίας σε άρθρα νέων (news articles), β) στην βελτιστοποίηση διαδικασιών που τυπικά χρησιμοποιεί ένα σύστημα προτάσεων και γ) σε πιο πρακτικά ζητήματα που αντιμετωπίζουν αυτά - όπως για παράδειγμα η εκτίμηση του πλήθους των συστάδων σε ένα πλήθος κειμένων ή η αντιμετώπιση του προβλήματος νέου χρήστη.

2.1 Γενικά

Το διαδίκτυο είναι πλέον παντού: σε κάθε συσκευή, σε κάθε μεριά του σπιτιού στην κοινωνία ολόκληρη. Εξάλλου, το διαδίκτυο των πραγμάτων (Internet of Things (IoT)), στο οποίο η συνδεσιμότητα συσκευών από παντού με στο διαδίκτυο έχει ωριμάσει αρκετά ώστε να αποτελεί πλέον μια καθημερινότητα.

Με νούμερα, η χρήση του διαδικτύου την δεκαετία 2004-2014 έχει αυξηθεί κατά το ασύλληπτο ποσοστό του 220% [101] και το δεικτοδοτημένο μέγεθός του από τις μηχανές αναζήτησης *Google* [83] και *Bing* [31], το 2014 τουλάχιστον, ξεπερνά τις 50 δισεκατομμύρια σελίδες [221]. Και αυτό αποτελεί μόνο το περιεχόμενο που είναι προσβάσιμο, ή αλλιώς, δεικτοδοτείται, από τις μηχανές αναζήτησης - μη υπολογίζοντας επομένως το περιεχόμενο του *Deep Web*.

Παράλληλα, η συνδυαστική έκρηξη που λαμβάνει χώρα όσον αφορά στις τεχνολογίες που χρησιμοποιούνται στο διαδίκτυο και κατ' επέκταση στις νέες υπηρεσίες, τα νέα κοινωνικά δίκτυα που ολοένα και αυξάνονται σε πλήθος καθώς και η διείσδυση της ευρυζωνικότητας σε ολοένα και μεγαλύτερα ποσοστά του πληθυσμού, φυσικά κάνει την δημιουργία νέου περιεχομένου πιο απλή και γρηγορότερη από ποτέ. Χαρακτηριστικό παράδειγμα εδώ αποτελεί το *YouTube* [225], στο οποίο κάθε λεπτό που περνάει "ανεβαίνουν" βίντεο αθροιστικής διάρκειας 100 ωρών!

Όλα αυτά τα στοιχεία μας οδηγούν στο συμπέρασμα ότι η διαδικασία αναζήτησης και η επιτυχής εύρεση πληροφορίας που μας ενδιαφέρει στο διαδίκτυο είναι αν μη τι άλλο μια υπόθεση δύσκολη. Θα μπορούσε εύκολα να ειπωθεί ότι όπως κάθε κοινωνία, έτσι και το διαδίκτυο, έχει τα δικά του προβλήματα. Πηγή αυτών των προβλημάτων μπορεί να θεωρηθεί η "άναρχη δόμησή του", η έλλειψη σαφούς νομοθεσίας αλλά και η αίσθηση ελευθερίας που αφήνει τους "κατοίκους" του να ενεργούν ουσιαστικά κατά βούληση, βρίσκοντας στο διαδίκτυο μία επανάσταση που θέλουν στην πραγματική τους ζωή, έναν τρόπο έκφρασης ιδεών, έναν τρόπο έκφρασης της γνώσης και της μάθησης.

Τη σήμερον ημέρα, η ελευθερία της έκφρασης και του λόγου παγκοσμίως διασφαλίζεται από τον τρόπο με τον οποίο διακινείται το περιεχόμενο στο διαδίκτυο. Η διάχυση γνώσης και εμπειρίας θα μπορούσαν επίσης να χαρακτηριστούν σαν θετικά επακόλουθα από την ύπαρξη μεγάλου όγκου πληροφορίας στον παγκόσμιο ιστό. Θα πρέπει όμως κανείς να αναλογιστεί κατά πόσο όλος αυτός ο όγκος πληροφορίας και όλες οι πηγές ενημέρωσης του διαδικτύου είναι έγκυρες. Δεν υπάρχει απολύτως κανένας μηχανισμός που να μπορεί να διασφαλίσει σε κάθε επισκέπτη του διαδικτύου πως οι σελίδες που παρακολουθεί και το περιεχόμενο που συλλέγει είναι αξιόπιστο και ποιοτικό. Πλέον, ακόμα και ο μέσος χρήστης, γνωρίζει μηχανισμούς μέσα από τους οποίους μπορεί να βρει στοιχεία για οποιοδήποτε θέμα. Κανείς όμως δε μπορεί να του εγγυηθεί επιτυχία και ταχύτητα στη διαδικασία ανεύρεσης αλλά πάνω απ' όλα, ποιότητα στα αποτελέσματα της εκάστοτε αναζήτησής του. Απαιτούνται καινοτόμες τεχνικές, νέες ιδέες και νέες προσεγγίσεις για να αντιμετωπιστεί το πρόβλημα. Οι χρήστες δεν θέλουν απλά πληροφορία, θέλουν να μπορούν να εντοπίζουν εύκολα και γρήγορα ποιοτική πληροφορία, πληροφορία που τους ενδιαφέρει και ταιριάζει με το ύφος τους.

Ακόμα περισσότερο, επιθυμούν αυτή η πληροφορία να τους προσφέρετε μέσα από αυτόματους μηχανισμούς που έχουν τη δυνατότητα να φιλτράρουν το "χάος" του διαδικτύου.

Η έλλειψη ποιότητας στης τάξεις του διαδικτύου έχει κεντρίσει το ενδιαφέρον της επιστημονικής κοινότητας εδώ και αρκετά χρόνια. Πολλά πεδία της επιστήμης της πληροφορικής, και όχι μόνο, βρίσκονται στο επίκεντρο του ενδιαφέροντος: data mining, text analysis, text categorization, semantic web και πολλά ακόμα, τα οποία αν και ήταν γνωστά ακόμα και πριν την εξάπλωση του διαδικτύου, επανεξετάζονται καθώς φαίνεται να είναι αυτά που δίνουν λύσεις στα μειονεκτήματά του.

2.1.1 Άρθρα νέων

Στην παρούσα διδακτορική διατριβή δε θα αναλωθούμε στην καταγραφή των πολλών, αν μη τι άλλο, προβλημάτων του διαδικτύου αλλά θα επικεντρωθούμε σε ένα κομμάτι των προβλημάτων που προκύπτουν από την αέναη, καθημερινή και καταιγιστική παραγωγή πληροφορίας σε αυτό. Ακόμα περισσότερο, θα εστιάσουμε την προσοχή μας στις πληροφορίες που δημιουργούνται σε καθημερινή βάση από την πληθώρα των ενημερωτικών δικτυακών πυλών που κατακλύζουν στην κυριολεξία το διαδίκτυο. Ο λόγος για τα γνωστά άρθρα νέων ή αλλιώς news articles, τα οποία αποτελούν κειμενική πληροφορία ενημέρωσης που πηγάζει από *news portals* του διαδικτύου.

Ένα άρθρο νέου καταγράφει πρόσφατη ή τρέχουσα πληροφορία σχετικά με ένα γεγονός το οποίο παρουσιάζει γενικό (ή μη) ενδιαφέρον ή συσχετίζεται με συγκεκριμένη θεματολογία (π.χ. πολιτική ή αθλητική). Μπορεί να περιλαμβάνει ή να μην περιλαμβάνει αυτόπτες μάρτυρες οι οποίοι "είδαν" το γεγονός. Επίσης, μπορεί να περιλαμβάνει φωτογραφικό υλικό, στατιστικά στοιχεία, γραφικές αναπαραστάσεις, συνεντεύξεις, δημοσκοπήσεις, αντιπαραθέσεις σε κάποιο θέμα, κ.λπ. Επικεφαλίδες συχνά χρησιμοποιούνται για να τραβήξουν το ενδιαφέρον των αναγνωστών σε ένα συγκεκριμένο μέρος του άρθρου ή και σε όλο. Ο συγγραφέας ενός άρθρου νέου μπορεί να παραθέτει γεγονότα και αναλυτικές πληροφορίες που απαντούν σε ερωτήσεις όπως: ποιος, τι, πότε, που, γιατί και πως.

Αν και ο παραπάνω ορισμός μοιάζει να ταιριάζει σε άρθρα νέων που δημοσιεύονται στον έντυπο τύπο, η ηλεκτρονική τους εκδοχή δεν διαφέρει σε τίποτα.

2.1.2 Web, News και Meta portals

Στην παρούσα ενότητα αναφέρουμε ορισμένες πληροφορίες για τις πύλες πληροφόρησης στο διαδίκτυο, γνωστές και ως portals.

2.1.2.1 Web portals

Ένα web portal είναι συχνά ένας ειδικά σχεδιασμένος ιστότοπος ο οποίος συνδυάζει και αθροίζει πληροφορία από διάφορες πηγές με έναν ενιαίο τρόπο. Συνήθως κάθε πηγή πληροφορίας έχει μία συγκεκριμένη θέση στον ιστότοπο για την απεικόνιση πληροφορίας (συχνά αναφέρεται ως portlet). Ο χρήστης μπορεί να ρυθμίζει τις πληροφορίες που θα φαίνονται σε αυτό. Ο ενιαίος τρόπος με τον οποίο η πληροφορία απεικονίζεται σε ένα web portal εξαρτάται συχνά τόσο από τον χρήστη στον οποίο απευθύνεται, όσο και από την ποικιλομορφία του περιεχομένου.

Ένα web portal μπορεί να έχει μία διεπαφή αναζήτησης, (search API) η οποία επιτρέπει στους χρήστες να αναζητούν περιεχόμενο μέσα στο ίδιο το portal. Άλλες υπηρεσίες που μπορεί να παρέχει ένα web portal είναι η δυνατότητα ανταλλαγής μηνυμάτων (e-mail ή IM), απεικόνιση πληροφορίας πραγματικού χρόνου (π.χ. τιμές μετοχών), πληροφορίες από Βάση Δεδομένων (ΒΔ) ή ακόμα και περιεχόμενο ψυχαγωγίας (π.χ. βιβλία ή ταινίες).

Μερικά παραδείγματα από web portals (κάποια από τα οποία πλέον δεν υπάρχουν) είναι τα εξής: AOL [15], Excite [66], Netvibes [155], iGoogle [99], MSN [148], Naver [153], Lycos [134], Indiatimes [100], Rediff [180], Yahoo! [223], κ. α.

2.1.2.2 News portals

Μια ειδική υποκατηγορία από web portals αποτελούν τα news portals , τα οποία και επικεντρώνονται στην δεικτοδότηση άρθρων νέων από διάφορες πηγές. Πρόκειται επομένως για Δικτυακούς τόπους που σαν στόχο έχουν την ενημέρωση των χρηστών του διαδικτύου για τα επίκαιρα κυρίως νέα σε παγκόσμιο επίπεδο. Μερικά και πολύ σημαντικά από αυτά είναι το CNN[52], το BBC[25], το Reuters[182], το FoxNews[70], καθώς και οι υπηρεσίες που προσφέρονται από τους πολυπληθείς και από τους πλέον αναγνωρίσιμους δικτυακούς τόπους Google[83] και Yahoo[223].

Οι Δικτυακοί αυτοί τόποι εστιάζονται στο να ενημερώνουν τους χρήστες τους για ότι συμβαίνει καθημερινά στον πλανήτη. Τα νέα/άρθρα παρουσιάζονται με δομημένο τρόπο στις συγκεκριμένες σελίδες, ωστόσο το πλήθος τους είναι τέτοιο ώστε να είναι σχεδόν αδύνατο από κάποιον χρήστη να μπορέσει εντός του εικοσιτετραώρου να παρακολουθήσει όλες τις ειδήσεις που δημοσιεύονται στις πολλές διαφορετικές κατηγορίες. Ακόμα και η εστίαση σε μία συγκεκριμένη κατηγορία απαιτεί τη συνεχή και διαρκή παρακολούθηση κάθε δικτυακού τόπου προκειμένου να υπάρχει πλήρης ενημέρωση. Επίσης, πολλά από αυτά τα νέα παρουσιάζονται από την οπτική γωνία του αρθρογράφου καθώς σπάνια - πλέον - δημοσιεύονται ακέραια ακόμα και τα δελτία τύπου, με αποτέλεσμα να χάνεται συχνά το κριτήριο της αντικειμενικότητας μίας είδησης. Απόρροια όλων των παραπάνω είναι το εξής: οι χρήστες του διαδικτύου δυσκολεύονται στον εντοπισμό μίας είδησης που τους ενδιαφέρει με αποτέλεσμα να αναλώνουν το χρόνο τους στην αναζήτηση της είδησης, του νέου, του άρθρου, παρά στην ανάγνωση του ίδιου του άρθρου. Σημαντικό είναι επίσης ότι η ενημέρωση που έχουν, κάθε άλλο παρά σφαιρική είναι, μιας και τελικά προτιμούν έναν και μόνο ιστότοπο για την ενημέρωσή τους.

2.1.2.3 Meta portals

Όπως αναφέρθηκε και νωρίτερα, η παρακολούθηση άρθρων νέων από μία σφαιρική και αντικειμενική άποψη απαιτεί την ενημέρωση από πολλαπλές πηγές. Ως εκ' τούτου, στα πλαίσια της μεταπτυχιακής μου εργασίας, δημιουργήθηκε η υπηρεσία *PeRSSonal* [171] η οποία παρέχει ακριβώς αυτό: εντοπίζοντας άρθρα νέων από πηγές τις οποίες ορίζει είτε ο χρήστης, είτε το ίδιο το σύστημα, παρέχει την συνδυασμένη πληροφορία στον χρήστη, εύκολα και γρήγορα. Καθότι ένα τέτοιο σύστημα αποτελεί κάτι περισσότερο από ένα απλό news portal (βάση του ορισμού στην παράγραφο 2.1.2.2), αθροίζοντας ουσιαστικά άρθρα νέων από news portals, θα μπορούσαμε να

το χαρακτηρίσουμε ως ένα meta portal. Παρόμοια συστήματα, γνωστά και ως συστήματα αποδελτίωσης άρθρων νέων του παγκόσμιου ιστού είναι τα εξής: Google News [84], NewsMe [157], NewsJunkies [156], personews [170], κ. α.

2.2 Συστήματα προτάσεων

Τα συστήματα προτάσεων (recommendation systems) αποτελούν μία υποκατηγορία των συστημάτων φιλτραρίσματος πληροφορίας τα οποία αποσκοπούν στην πρόβλεψη βαθμολογιών ή γενικά προτιμήσεων που πρόκειται να έχει ο χρήστης προς ένα αντικείμενο (π.χ. άρθρο νέου) [184]. Τα συστήματα προτάσεων έχουν γίνει εξαιρετικά συνηθισμένα στις μέρες μας, μίας και βρίσκουν εφαρμογές σε μια πληθώρα προβλημάτων. Τα πιο συνηθισμένα είναι πιθανά εκείνα που προτείνουν ταινίες, μουσική, νέα, βιβλία, ερευνητικά άρθρα, ερωτήματα προς μηχανές αναζήτησης και προϊόντων στη γενική περίπτωση.

Τα συστήματα προτάσεων τυπικά παράγουν μία λίστα από προτάσεις με βάση έναν από τους παρακάτω δύο τρόπους [102]:

- Συνεργατικό φιλτράρισμα (collaborative filtering)
- Φιλτράρισμα βασισμένο στο περιεχόμενο (content-based filtering)

Οι CF προσεγγίσεις χτίζουν ένα μοντέλο με βάση την προηγούμενη συμπεριφορά ενός χρήστη (π.χ. τα αντικείμενα που αγόρασε ή επέλεξε ή βαθμολόγησε), καθώς και παρόμοιες αποφάσεις οι οποίες έγιναν από άλλους χρήστες. Στη συνέχεια χρησιμοποιούν αυτό το μοντέλο προκειμένου να προβλέψουν αντικείμενα (ή βαθμολογήσεις αντικειμένων) για τα οποία ο χρήστης μπορεί να ενδιαφέρεται [142]. Αντίθετα οι προσεγγίσεις που κάνουν φιλτράρισμα βασισμένο στο περιεχόμενο κάνουν χρήση διακριτών χαρακτηριστικών των αντικειμένων προκειμένου να προτείνουν επιπρόσθετα αντικείμενα με παρόμοιες ιδιότητες. Ο συνδυασμός και των παραπάνω δύο τεχνικών (υβριδική προσέγγιση) είναι επίσης συχνός στης μέρες μας και είναι εξάλλου και η λογική επιλογή την οποία ακολουθήσαμε και για το σύστημα που υλοποιήθηκε.

2.3 Προεπεξεργασία δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί τον συνδυασμό των τεχνικών εκείνων που χρησιμοποιούνται από ένα σύστημα που βασίζεται σε κειμενικά ή άλλου είδους πρωτογενή δεδομένα, προκειμένου να καταλήξει σε πληροφορία αξιοποιήσιμη από τα υποσυστήματα ανάκτησης πληροφορίας που συνήθως ακολουθούν. Με βάση τον παραπάνω γενικό ορισμό, για την περίπτωση ενός συστήματος που βασίζεται σε χρήση λέξεων κλειδιών (Keywords (KWs)) η προεπεξεργασία δεδομένων αφορά σε μία σειρά τεχνικών στις οποίες υπόκεινται το χρήσιμο κείμενο:

- αφαίρεση των σημείων στίξης καθώς και των αριθμών που τυχόν περιέχει
- αφαίρεση λέξεων οι οποίες δεν περικλείουν κάποιο νόημα, για παράδειγμα άρθρα

- εύρεση της ρίζας μίας λέξης (Stemming)
- εύρεση των μερών του λόγου των λέξεων του κειμένου (Part of Speech (POS) tagging)
- πιθανή αξιοποίηση μιας ή περισσοτέρων εξωτερικών βάσεων γνώσης
- εντοπισμός και καταγραφή n-grams

Σαν αποτέλεσμα, η προεπεξεργασία δεδομένων έχει λοιπόν την δομικής πληροφορίας από το κείμενο, ικανή για την νοηματική αναπαράστασή του. Τυπικά, πρόκειται για τις λέξεις-κλειδιά που υπάρχουν στο κείμενο, συνοδευόμενες από τη συχνότητα με την οποία παρουσιάζονται μέσα σε αυτό, αλλά και το σημείο του κειμένου στο οποίο εντοπίζονται. Για την περαιτέρω ενίσχυση των διαδικασιών ανάκτησης πληροφορίας που ακολουθούν, στις τεχνικές προεπεξεργασίας κειμένου θα εντάξουμε και την ανάκτηση των ουσιαστικών του κειμένου μέσω τεχνικών POS tagging, μιας και είναι γενικά αποδεκτό ότι τα ουσιαστικά του κειμένου φέρουν το μεγαλύτερο ποσοστό της χρήσιμης πληροφορίας αυτού.

Για τους μηχανισμούς εξαγωγής κειμένου, η απόρριψη οποιασδήποτε πληροφορίας δεν σχετίζεται με το κείμενο, και γενικά η προεπεξεργασία πληροφορίας, αποτελεί μία μεγάλη πρόκληση. Παρά το γεγονός ότι επιφανειακά βασίζεται σε συγκεκριμένα και σταθερά βήματα, θα πρέπει να γίνει εκτενής ανάλυση του είδους της πληροφορίας που είναι επιθυμητή προκειμένου το βήμα της προεπεξεργασίας να καταλήξει σε σημαντικά αποτελέσματα και πιο συγκεκριμένα στην εξαγωγή των σωστών λέξεων κλειδιών. Πολλά ευρετικά έχουν ερευνηθεί στη βιβλιογραφία σχετικά με το συγκεκριμένο θέμα. Η εύρεση των καταλλήλων για την περίπτωση των άρθρων νέων καθώς και η σωστή αξιοποίησή τους αποτελεί σημαντικό τμήμα της διδακτορικής διατριβής.

2.3.1 Χρήση εξωτερικής βάσης γνώσης

Πέρα από την ίδια την γνώση που μπορούν οι μηχανισμοί να αντλήσουν από τα ίδια τα κείμενα, μία ενδιαφέρουσα προσέγγιση αποτελεί η εξόρυξη πληροφορίας από εξωτερικές πηγές. Η γνώση που εξάγεται με αυτόν τον τρόπο προστίθεται στην υπάρχουσα για την παραγωγή ενός αποτελεσματικότερου μοντέλου ανάκτησης πληροφορίας στον εκάστοτε τομέα.

2.3.1.1 WordNet

Το WordNet αποτελεί μία από τις πιο ευρέως διαδεδομένες και μεγαλύτερες λεξιλογικές βάσεις δεδομένων της Αγγλικής γλώσσας. Επιχειρεί με άλλα λόγια να μοντελοποιήσει την λεξιλογική γνώση των ανθρώπων που μιλούν την αγγλική (ως μητρική γλώσσα). Παρότι το WordNet είναι προσβάσιμο από τον καθένα μέσω των πολλαπλών διεπαφών του (web-based, εφαρμογή ή κλήση βιβλιοθηκών), η βασική του χρησιμότητα είναι στην αυτοματοποιημένη ανάλυση κειμένου και σε εφαρμογές τεχνητής νοημοσύνης Artificial Intelligence (AI). Περιέχοντας πάνω από 150.000 όρους, το WordNet παρέχει σύντομους ορισμούς και παραδείγματα χρήσης. Επίσης ομαδοποιεί ουσιαστικά, ρήματα, επίθετα και επιρρήματα σε ομάδες συνωνύμων τα οποία και ονομάζει synsets. Το WordNet μπορεί επομένως να ερμηνευθεί ως ένας συνδυασμός λεξικού και θησαυρού της Αγγλικής.

Τα synsets οργανώνονται σε:

- έννοιες - περιέχοντας έτσι τα συνώνυμα κάθε λέξης
- υπερώνυμα/υπώνυμα
- μερόνυμα/ολόνυμα δίνοντας έτσι μία ιεραρχικές δενδρικές δομές για κάθε όρο που υπάρχει στο WordNet.

2.3.1.1.1 Υπερώνυμα/Υπώνυμα

Η σχέση υπερώνυμου/υπωνύμου (hypernym/hyponym) αποτελεί μία βασική συσχέτιση μεταξύ των όρων του WordNet η οποία και θα μας απασχολήσει αρκετά στη συνέχεια. Πιο συγκεκριμένα, και για την περίπτωση των ουσιαστικών ισχύει ο ορισμός 2.3.1.

Ορισμός 2.3.1. Έστω δύο όροι του WordNet: X και Y, τότε:
Ο Y είναι ένα υπερώνυμο του X αν κάθε X είναι ένα είδος από το Y, π.χ. το φρούτο ένα υπερώνυμο του μήλου.
Ο Y είναι ένα υπώνυμο του X αν για κάθε Y είναι ένα είδος από το X, π.χ. το μήλο ένα υπώνημο του φρούτου.

Ένα γράφημα υπερωνύμων αποτελεί την δενδρική απεικόνιση της συσχέτισης υπερωνύμου/υπωνύμου που αναφέρθηκε. Για παράδειγμα, το δένδρο υπερωνύμων του όρου dog, φαίνεται στο σχήμα 1.

dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore
=> placental, placental mammal, eutherian, eutherian mammal
=> mammal
=> vertebrate, craniate
=> chordate
=> animal, animate being, beast, brute, creature, fauna
=> ...

Σχήμα 1: Δένδρο υπερωνύμων του όρου dog

2.3.1.1.2 Μερόνυμα/Ολόνυμα

Για την σχέση μερονύμου/ολονύμου του WordNet για την περίπτωση των ουσιαστικών ισχύει ο ορισμός 2.3.2

Ορισμός 2.3.2. Έστω δύο όροι του WordNet: X και Y, τότε:
Ο Y είναι ένα μερόνυμο του X αν το Y είναι ένα μέρος του X, π.χ. το παράθυρο είναι ένα μερόνυμο του κτηρίου.

Ο Υ είναι ένα ολόνυμο του Χ αν το είναι ένα είδος από το , π.χ. το κτήριο είναι ένα ολόνυμο του παραθύρου.

2.3.2 n-grams

Ένα n-gram είναι μία συνεχόμενη ακολουθία από n αντικείμενα σε μία δεδομένη ακολουθία από γραπτό κείμενο ή προφορικό λόγο. Τα αντικείμενα μπορεί να είναι φωνήματα, συλλαβές, γράμματα, λέξεις ή σύνολα λέξεων αναλόγως την εφαρμογή. Ένα n-gram μεγέθους 1, συχνά αναφέρεται και ως "unigram', μεγέθους 2 ως "bigram" η "digram" , μεγέθους 3 ως "trigram" .

Ένα μοντέλο n-gram είναι ένα είδους πιθανοτικό μοντέλο γλώσσας το οποίο υπολογίζει την πιθανότητα του επομένου αντικειμένου σε μία τέτοια ακολουθία της μορφής (n-1) μοντέλου Markov. Τα μοντέλα n-gram χρησιμοποιούνται στις μέρες μας ευρύτατα στην πιθανοτική θεωρία, στη θεωρία επικοινωνίας, στην υπολογιστική γλωσσολογία (π.χ. στατιστική φυσική επεξεργασία γλώσσας), στην υπολογιστική βιολογία (π.χ. ανάλυση βιολογικών σειρών), καθώς και στην συμπίεση πληροφορίας. Τα βασικά θετικά στοιχεία των n-gram μοντέλων (και των αλγορίθμων που τα χρησιμοποιούν) είναι η σχετική απλότητά τους, καθώς και η ικανότητα κλιμακοσιμότητας που έχουν, επιτρέποντας έτσι σε μικρά πειράματα να κλιμακώνονται αρκετά αποδοτικά.

Η αξιοποίηση της πληροφορίας των n-grams των κειμένων, και πιο συγκεκριμένα, ο τρόπος ζύγισής τους, αποτελεί ένα σημαντικό τμήμα της διδακτορικής διατριβής όπως θα παρουσιαστεί στα επόμενα κεφάλαια.

2.4 Συσταδοποίηση κειμένων

Η κειμενική πληροφορία είναι η πιο συνηθισμένη μορφή πληροφορίας που διακινείται στο διαδίκτυο και τα κοινωνικά δίκτυα. Τα κείμενα τυπικά αναπαρίστανται στο vector space μοντέλο όπου η ακριβής σειρά των όρων απαλείφεται και τα δεδομένα αντιμετωπίζονται ως λίστα από λέξεις (Bag of Words (BOW)). Τα άρθρα νέων έχουν μία σειρά από ιδιότητες οι οποίες πρέπει να ληφθούν υπόψιν κατά την αξιοποίησή των δεδομένων τους:

- είναι πολύ μεγάλης διαστατικότητας και αραιά. Αυτό συνάγει με το γεγονός ότι μία γλώσσα αποτελείται τυπικά από εξαιρετικά πολλούς όρους (λέξεις), ενώ κάθε κείμενο περιλαμβάνει ένα σχετικά απειροελάχιστο ποσοστό αυτών των όρων. Επομένως, τα περισσότερα από τα χαρακτηριστικά της αναπαράστασης είναι μηδενικά.

- οι τιμές των χαρακτηριστικών αντιστοιχούν σε συχνότητες λέξεων και είναι επομένως τυπικά μη-μηδενικές. Αυτό είναι κάτι σημαντικό για τις τεχνικές εκείνες που αξιοποιούν ακριβώς αυτό το χαρακτηριστικό.

Ένας από τους σύνηθες τρόπους οργάνωσης μεγάλου όγκου δεδομένων, όπως στην περίπτωση μας τα άρθρα νέων ύστερα από την ανάκτησή τους από το διαδίκτυο, είναι η χρήση τεχνικών συσταδοποίησης. Η συσταδοποίηση αντικειμένων αναφέρεται στην διαδικασία διαχωρισμού των αντικειμένων μιας συλλογής σε πολλαπλές υπο-συλλογές, βασιζόμενοι στην ομοιότητα των αντικειμένων

μεταξύ τους. Γενικά η συσταδοποίηση έχει αποδειχθεί ως μία εξαιρετικά χρήσιμη Information Retrieval (IR) τεχνική αφού εντοπίζει ενδιαφέροντες πυρήνες πληροφορίας και κατανομών στα υποκείμενα δεδομένα. Βοηθά στην κατασκευή ουσιαστικών διαμερισμάτων σε μεγάλους όγκους δεδομένων με χρήση πολλαπλών μεθοδολογιών και ευρετικών που έχουν αναπτυχθεί ανά τα χρόνια. Τυπικές χρήσης της συσταδοποίησης είναι οι:

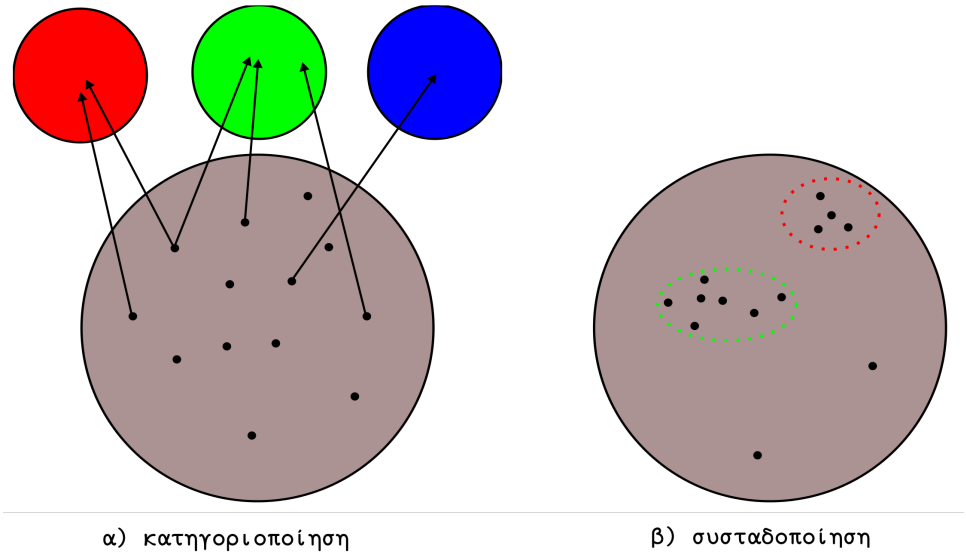
- για την δόμηση αποτελεσμάτων που προκύπτουν από ερωτήματα χρηστών
- για τον σχηματισμό της βάσης για περαιτέρω επεξεργασία των οργανωμένων ομάδων με χρήση άλλων τεχνικών IR, όπως η προσωποποίηση
- μέσα στο εύρος συστημάτων προτάσεων επηρεάζοντας άμεσα την απόδοσή τους όσον αφορά στις προτάσεις που κάνουν αυτά στους τελικούς χρήστες

Σε έναν πιο γενικό ορισμό μία αποτελεσματικής τεχνικής συσταδοποίησης, θα λέγαμε ότι είναι εκείνη που οργανώνει μία συλλογή από κείμενα σε ομάδες, τέτοιες ώστε τα κείμενα μέσα στην εκάστοτε ομάδα να είναι τόσο παρόμοια μεταξύ τους, όσο και διαφορετικά από εκείνα των άλλων ομάδων [107]. Η συσταδοποίηση μπορεί να παράγει είτε διαχωρισμένες, είτε αλληλεπικαλυπτόμενες συστάδες. Στην δεύτερη περίπτωση, είναι δυνατό για ένα κείμενο να εμφανίζεται σε πολλαπλές συστάδες.

Η συσταδοποίηση κειμένων (ή εγγράφων) αποτελεί ουσιαστικά ένα υποσύνολο από ένα ευρύτερο πεδίο συσταδοποίησης δεδομένων το οποίο μοιράζεται ιδέες από τα πεδία της ανάκτησης πληροφορίας (IR), φυσικής επεξεργασίας γλώσσας (Natural Language Processing (NLP)) και μηχανικής μάθησης (Machine Learning (ML)) μεταξύ άλλων. Συχνά αναφέροντας την έννοια "συσταδοποίηση" αναφερόμαστε απλά στην συσταδοποίηση κειμένων. Η διαδικασία της συσταδοποίησης στοχεύει στην εύρεση φυσικών ομαδοποιήσεων και επομένως παρουσιάζει μια γενική εικόνα των κλάσεων (νοηματικές θεματολογίες) σε μια συλλογή από κείμενα. Στο πεδίο της τεχνητής νοημοσύνης (AI) αναφέρεται ως μη-εποπτευόμενη μηχανική μάθηση (unsupervised machine learning).

Η συσταδοποίηση δεν πρέπει να συγχέεται με την κατηγοροποίηση κειμένων όπου το πλήθος των κλάσεων (και οι ιδιότητές τους) είναι γνωστά εκ' των προτέρων, και επομένως, τα κείμενα αντιστοιχίζονται σε αυτές τις κλάσεις. Αντιθέτως, σε ένα πρόβλημα συσταδοποίησης, ούτε το πλήθος των κλάσεων (συστάδες), ούτε οι ιδιότητές τους είναι γνωστές από πριν. Η διαφοροποίηση αυτή απεικονίζεται στο σχήμα 2, όπου στην περίπτωση α) οι τρεις κλάσεις στις οποίες αντιστοιχίζονται τα κείμενα είναι γνωστές από πριν. Αντίθετα στην περίπτωση β) ένας άγνωστος αριθμός συστάδων συνεπάγεται από τα ίδια τα κείμενα βάσει κάποιο κριτηρίου ομοιότητας (στην περίπτωση αυτή το κριτήριο είναι η απόσταση). Η κατηγοριοποίηση επομένως αποτελεί ένα παράδειγμα εποπτευόμενης μηχανικής μάθησης.

Παρόλα αυτά, υπάρχουν πολλές προκλήσεις στις οποίες οι τεχνικές συσταδοποίησης πρέπει να αντεπεξέλθουν. Μεταξύ αυτών και η αποδοτικότητα: οι παραγόμενες συστάδες θα πρέπει να είναι καλά συνδεδεμένες νοηματικά, παρά την ποικιλομορφία του περιεχομένου καθώς και το μέγεθος των αρχικών κειμένων. Για παράδειγμα, είναι συχνό φαινόμενο κάποια άρθρα νέων να ανήκουν στην



Σχήμα 2: Κατηγοριοποίηση και συσταδοποίηση

ίδια νοηματική συστάδα, παρότι δεν μοιράζονται κοινές λέξεις. Το αντίστροφο είναι επίσης πιθανό: άρθρα νέων που μοιράζονται κοινές λέξεις, είναι όμως άσχετα μεταξύ τους.

Η ασάφεια και η συνωνυμία είναι επομένως δύο από τα βασικά προβλήματα που οι τεχνικές συσταδοποίησης κειμένων αποτυγχάνουν συχνά να αντιμετωπίσουν αποτελεσματικά. Επίσης, το να έχουμε συστήματα IR απλά να παράγουν συστάδες κειμένων δεν είναι αρκετό από μόνο του. Και ο λόγος γι' αυτό είναι ότι είναι κυριολεκτικά αδύνατο για τους ανθρώπους να αντιληφθούν την πληροφορία απλά και μόνο κοιτάζοντας μέσα σε εκατοντάδες ή χιλιάδες κείμενα. Αντιθέτως, αναθέτοντας νοηματικές ετικέτες - επικεφαλίδες στις συστάδες έχει περισσότερο νόημα καθώς επιτρέπει στους χρήστες εύκολα και γρήγορα να αναγνωρίσουν σε τι αναφέρεται η κάθε συστάδα καθώς και να μπορέσουν εν' συνεχεία να αναλύσουν τα αποτελέσματα της συσταδοποίησης.

Στην παρούσα διδακτορική διατριβή, περιγράφουμε μία πληθώρα τεχνικών, αλγορίθμων και μηχανισμών συσταδοποίησης και αξιολογούμε την εφαρμογή τους στην περίπτωση των άρθρων νέων που πηγάζουν από το διαδίκτυο. Ο στόχος μας δεν είναι να παρουσιάσουμε διεξοδικά οτιδήποτε έχει ερευνηθεί σε αυτόν τον τομέα, αλλά να συγκρίνουμε τα αποτελέσματα των παραπάνω πειραμάτων συσταδοποίησης ώστε να εκτιμήσουμε ποια τεχνική ταιριάζει καλύτερα στην μεγάλη ποικιλομορφία και ποσότητα των άρθρων νέων του διαδικτύου.

2.4.1 Τυπικός ορισμός συσταδοποίησης

Ο τυπικός ορισμός του προβλήματος συσταδοποίσης έχει ως εξής:

Ορισμός 2.4.1. Δεδομένου ενός συνόλου κειμένων D, επιθυμούμε την ανάθεση καθενός από τα κείμενα $d \in D$ σε συστάδες παρόμοιων κειμένων ανακαλύπτοντας έτσι τις φυσικές τους κατηγορίες. Βασιζόμενοι στο vector-space μοντέλο, μπορούμε να αναπαραστήσουμε κάθε κείμενο $d \in D$ ως έναν πίνακα συχνοτήτων από τα χαρακτηριστικά που περιέχει: $\vec{d} = (f_1, ..., f_n)$.

Συνήθως τα χαρακτηριστικά των κειμένων είναι οι όροι από τους οποίους αποτελείται, π.χ. λέξεις κλειδιά, n-grams, κ.λπ. Μπορούμε να εκφράσουμε το σύνολο των κειμένων D σαν έναν $m \times n$ πίνακα, όπου m το πλήθος των κειμένων στο D και n το πλήθος των χαρακτηριστικών. Το στοιχείο (i, j) περιέχει το πλήθος εμφάνισης του χαρακτηριστικού j στο κείμενο i.

2.4.2 Πλήθος συστάδων

Ο προσδιορισμός του πλήθους των συστάδων σε ένα σύνολο δεδομένων, μία ποσότητα η οποία συχνά αναφέρεται ως k, όπως στην περίπτωση του k-means αλγορίθμου, είναι ένα σύνηθες πρόβλημα στην συσταδοποίηση δεδομένων, τόσο μάλιστα που αποτελεί και ξεχωριστό πεδίο έρευνας ανεξάρτητα από τους αλγορίθμους συσταδοποίησης. Για μία συγκεκριμένη κατηγορία αλγόριθμων συσταδοποίησης (οικογένεια k-means/Expectation Maximization (EM) αλγόριθμος), ο εκ' τον προτέρων καθορισμός του πλήθους των συστάδων αποτελεί βασική προϋπόθεση. Άλλοι αλγόριθμοι όπως οι Density-based spatial clustering of applications with noise (DBSCAN) και Ordering points to identify the clustering structure (OPTICS) δεν απαιτούν τον καθορισμό μίας τέτοιας παραμέτρου, ενώ η ιεραρχική συσταδοποίηση αποφεύγει το πρόβλημα εξολοκλήρου.

Η σωστή επιλογή του k είναι συχνά διφορούμενη, με ερμηνείες οι οποίες εξαρτώνται από το σχήμα και την κλίμακα της κατανομής των σημείων στο σύνολο δεδομένων, καθώς και την επιθυμητή λύση από τον χρήστη. Παράλληλα, η αύξηση του k χωρίς κάποιον έλεγχο, πάντα θα μειώνει το μέγεθος του σφάλματος στην τελική συσταδοποίηση, έως την ακραία περίπτωση του μηδενικού σφάλματος, όπου κάθε σημείο θεωρείται και ως μία συστάδα ($k = n$).

Διαισθητικά επομένως, η βέλτιστη επιλογή του k θα ισορροπεί ανάμεσα στην μέγιστη συμπίεση των δεδομένων με όσο το δυνατόν μαζικότερες συστάδες, και την μέγιστη ακρίβεια με όσο το δυνατόν περισσότερες συστάδες. Εάν μία προφανής τιμή για το k δεν είναι γνωστή εκ' των προτέρων από τις ιδιότητες των ίδιων των δεδομένων, θα πρέπει κάπως να επιλεχθεί - και προς αυτή την κατεύθυνση αρκετές μέθοδοι, οι οποίες και παρουσιάζονται στο επόμενο κεφάλαιο, έχουν ερευνηθεί στη βιβλιογραφία.

2.5 Συσταδοποίηση χρηστών

Ότι αναφέρθηκε στην ενότητα 2.4 για την συσταδοποίηση αντικειμένων (άρθρων νέων) ισχύει και για την περίπτωση συσταδοποίησης χρηστών με την βασική διαφορά ότι η συσταδοποίηση ενεργεί πάνω στις προτιμήσεις, ή αλλιώς προφίλ, των χρηστών. Έτσι, κάποιο τμήμα της συλλογής ονομάζεται συστάδα χρήστη και περιλαμβάνει χρήστες που έχουν εκφράσει παρόμοια ενδιαφέροντα σε ότι έχει να κάνει με τις προτιμήσεις τους σε άρθρα νέων ενώ πλοηγούνται σε μία συλλογή.

Η συσταδοποίηση χρηστών αποτελεί ένα κομβικό τμήμα της διδακτορικής διατριβής, μιας και αποτελεί ουσιαστικά τον μοχλό με τον οποίο η απόδοση του συστήματος προτάσεων αυξάνεται σημαντικά.

Ο τρόπος που αντιμετωπίζουμε τις συστάδες χρηστών έχει ως εξής: ξεκινώντας από τις καταγεγραμμένες συνεδρίες χρηστών και θέτοντας σαφή χρονικά όρια πλοήγησης, αναλύουμε τα επι-

λεγμένα άρθρα τα οποία και συσταδοποιούμε με χρήση του αλγορίθμου W-kmeans. Κατά συνέπεια, το πρόβλημα της συσταδοποίησης χρηστών ανάγεται στο αντίστοιχο της συσταδοποίησης άρθρων νέων μέσα σε συγκεκριμένα πλαίσια και επιλογές που θα αναλυθούν στις επόμενες ενότητες.

2.6 Προσωποποίηση στο χρήστη

Η προσωποποίηση στο χρήστη είναι η διαδικασία κατά την οποία τα αποτελέσματα που εμφανίζονται τελικά στο χρήστη προσαρμόζονται προκειμένου να ανταποκρίνονται στις ανάγκες του. Πιο συγκεκριμένα, τα στάδια της προσωποποίησης αφορούν τον εντοπισμό άρθρων τα οποία ενδιαφέρουν το χρήστη και παρουσίασή τους με τέτοιον τρόπο ώστε να ταιριάζουν στις ανάγκες του χρήστη. Το πρόβλημα που τίθεται είναι ένας "έξυπνος" αλγόριθμος ο οποίος θα μπορεί να αξιοποιεί όλες τις πληροφορίες που μπορούν να συγκεντρωθούν από την περιήγηση του χρήστη στο δικτυακό τόπο και αξιοποίηση αυτών των πληροφοριών προκειμένου να εμφανιστούν όσο το δυνατόν καλύτερα και πιο ποιοτικά αποτελέσματα.

2.6.1 Συμμετοχή του χρήστη στις διαδικασίες του συστήματος

Ο χρήστης είναι αυτός που δέχεται την τελική πληροφορία και αυτός που ουσιαστικά διαμορφώνει την πληροφορία για τον εαυτό του. Αυτό σημαίνει πως ο χρήστης θα πρέπει να είναι αναπόσπαστο κομμάτι του συστήματος. Θα πρέπει να είναι σε θέση να διαμορφώσει διαδικασίες του πυρήνα του συστήματος με βάση τις πληροφορίες που δίνει άμεσα ή έμμεσα στο σύστημα ως ανάδραση.

Στα περισσότερα συστήματα τα οποία αντιμετωπίστηκαν κατά τη διάρκεια της μελέτης για τη συγκεκριμένη εργασία, παρατηρήθηκε πως ο χρήστης συμμετέχει μόνο στα επιτελικά στάδια των συστημάτων ενώ έχουν ήδη εκτελεστεί τα βασικά βήματα του πυρήνα των μηχανισμών. Η συμμετοχή του χρήστη στις διαδικασίες πυρήνα ενός large scale συστήματος είναι επίπονη διαδικασία η οποία απαιτεί αλγορίθμους που θα μπορούν να εκτελούνται αποδοτικά σε πραγματικό χρόνο προκειμένου ο χρήστης να διαμορφώνει όχι μόνον τα τελικά αποτελέσματα που εμφανίζονται σε αυτόν αλλά και συγκεκριμένες διαδικασίες ολόκληρου του συστήματος.

2.7 Το Πρόβλημα του νέου χρήστη

Ένα κοινό πρόβλημα από το οποίο όλα τα συστήματα συνεργατικού φιλτραρίσματος συχνά πάσχουν είναι αυτό της κρύας εκκίνησης (cold start problem). Το πρόβλημα αυτό έχει τρεις εκφάνσεις:

- το πρόβλημα νέου αντικειμένου, όπου ένα νέο αντικείμενο πρωτο-εισάγεται στο σύστημα και δεδομένου ότι δεν έχει αξιολογηθεί από κανέναν, το σύστημα δεν μπορεί να το προτείνει (και επομένως περνάει στην αφάνεια)
- το πρόβλημα νέου χρήστη, όπου ένας νέος χρήστης χρησιμοποιεί το σύστημα για πρώτη φορά και ως εκ' τούτου δεν υπάρχουν προτάσεις από το σύστημα προς αυτόν. Το πρόβλημα

παραμένει τουλάχιστον έως ότου το σύστημα αποκτήσει κάποια γνώση για τις προτιμήσεις του χρήστη

- το πρόβλημα του νέου συστήματος το οποίο αποτελεί συνδυασμό των δύο παραπάνω περιπτώσεων

Στην διδακτορική διατριβή ασχοληθήκαμε με το πρόβλημα του νέου χρήστη, για την επίλυση του οποίου προτείνουμε μια συγκεκριμένη αλγοριθμική προσέγγιση.

ΚΕΦΑΛΑΙΟ 3

ΕΡΕΥΝΗΤΙΚΑ ΘΕΜΑΤΑ

> Beware of false knowledge; it is more dangerous than ignorance.
>
> *George Bernard Shaw, Irish Dramatist, 1856*

Στο παρόν κεφάλαιο περιγράφεται η τρέχουσα κατάσταση σε σχέση με τα θέματα που καταπιάνεται η διδακτορική διατριβή. Παρουσιάζεται επομένως το state of the art με βάση τις τελευταίες εξελίξεις στους τομείς αυτούς, εργασίες παραπλήσιες καθώς και αλγοριθμικές προσεγγίσεις.

3.1 Φυσική Επεξεργασία Γλώσσας

Η φυσική επεξεργασία γλώσσας (NLP) είναι ένα πεδίο της επιστήμης υπολογιστών, της τεχνητής νοημοσύνης, καθώς και της γλωσσολογίας, το οποίο ασχολείται με τις διεπαφές μεταξύ γλωσσών υπολογιστών και φυσικών (ανθρωπίνων) γλωσσών. Ως εκ' τούτου, το NLP σχετίζεται με την περιοχή της αλληλεπίδρασης ανθρώπου-υπολογιστή. Στις πολλές προκλήσεις που πρέπει να αντιμετωπίσει το NLP περιλαμβάνονται: η κατανόηση φυσικής γλώσσας η οποία επιτρέπει στους υπολογιστές να εξάγουν νόημα από την ανθρώπινη γλώσσα, καθώς και άλλες που εμπεριέχουν παραγωγή φυσικής γλώσσας.

Οι σύγχρονοι NLP αλγόριθμοι βασίζονται στη μηχανική εκμάθηση, και ειδικότερα στην στατιστική μηχανική εκμάθηση [137]. Προηγούμενες υλοποιήσεις της επεξεργασίας γλωσσών αφορούσαν στην άμεση καταγραφή και χρήση συγκεκριμένων κανόνων. Μέσω της χρήσης μηχανικής εκμάθησης όμως, γίνεται χρήση γενικών αλγορίθμων εκπαίδευσης οι οποίοι συχνά βασίζονται σε στατιστικά συμπεράσματα ώστε να μάθουν αυτόματα τους κανόνες μέσω της ανάλυσης μεγάλου πλήθους από βάσεις γνώσης (corpus) και τυπικά πραγματικά παραδείγματα χρήσης. Οι βάσεις γνώσης αυτές αποτελούνται από ένα σύνολο κειμένων τα οποία έχουν προ-σημειωθεί ώστε να εμπεριέχουν τις σωστές τιμές με τις οποίες πρέπει να γίνει εκμάθηση.

Πολλές διαφορετικές κατηγορίες αλγορίθμων μηχανικής εκμάθησης έχουν εφαρμοστεί σε NLS εργασίες. Αυτοί οι αλγόριθμοι δέχονται ως είσοδο ένα μεγάλο σύνολο χαρακτηριστικών τα οποία παράγονται από τα δεδομένα εισόδου. Ορισμένοι από τους αρχικά χρησιμοποιούμενους αλγορίθμους, όπως τα δένδρα απόφασης, παρήγαγαν συστήματα κανόνων εάν-τότε (if-then rules). Όλο και συχνότερα όμως η έρευνα επικεντρώθηκε σε στατιστικά μοντέλα, τα οποία παίρνουν πιθανοτικές αποφάσεις βασισμένα στην εφαρμογή πραγματικών βαρών σε καθένα από τα χαρακτηριστικά εισόδου. Αυτά τα μοντέλα έχουν το πλεονέκτημα ότι μπορούν να εκφράσουν την σχετική βεβαιότητα από πολλές πιθανές απαντήσεις σε σχέση με μόνο μία, παράγοντας έτσι πιο αποδοτικά αποτελέσματα - ειδικά όταν ένα τέτοιο μοντέλο συμπεριλαμβάνεται ως ένα στοιχείο σε ένα μεγαλύτερο σύστημα.

Τα συστήματα που βασίζονται σε αλγορίθμους μηχανικής εκμάθησης έχουν πολλαπλά πλεονεκτήματα σε σχέση με τους χειροκίνητα παραγόμενους κανόνες:

- Οι διαδικασίες εκμάθησης που χρησιμοποιούνται κατά τη διαδικασία της μηχανικής εκμάθησης εστιάζουν αυτόματα στις πιο συνηθισμένες περιπτώσεις, ενώ οι χειροκίνητοι κανόνες συχνά είναι μη κατανοητό που πρέπει να εστιάσουν

- Οι αυτόματες διαδικασίες εκμάθησης μπορούν να κάνουν χρήση αλγορίθμων στατιστικής συμπερασματολογίας για να παράξουν μοντέλα τα οποία είναι ισχυρά σε μη συνηθισμένη είσοδο (π.χ. που περιέχουν λέξεις ή δομές που δεν έχουν συναντηθεί παλαιότερα). Γενικά, ο χειρισμός τέτοιας εισόδου με αποτελεσματικό τρόπο με χρήση χειροκίνητων κανόνων είναι εξαιρετικά δύσκολος, επιρρεπής σε λάθη και χρονοβόρος.

- Τα συστήματα που βασίζονται σε αυτόματη εκμάθηση των κανόνων μπορούν να γίνουν πιο ακριβή απλά παρέχοντας περισσότερα δεδομένα. Αντίθετα, τα συστήματα που βασίζονται σε χειροκίνητους κανόνες μπορούνε να γίνουν πιο ακριβή μόνο αυξάνοντας την πολυπλοκότητα των κανόνων, το οποίο είναι αρκετά δυσκολότερο.

3.1.1 Σύνηθες NLP εργασίες

Παρακάτω είναι μία λίστα από μερικές από τις πιο μελετημένες στη βιβλιογραφία εργασίες (tasks) NLP. Να σημειώσουμε ότι ορισμένες από αυτές έχουν άμεσες πραγματικές εφαρμογές, ενώ άλλες πιο συχνά εξυπηρετούν ως υπο-εργασίες οι οποίες χρησιμοποιούνται για την επίλυση μεγαλύτερων εργασιών.

- Αυτόματη εξαγωγή περίληψης (Automatic summarization)
- Ανάλυση συναναφορών (Coreference resolution)
- Ανάλυση λόγου ομιλίας (Discourse analysis)
- Μηχανική μετάφραση (Machine translation)
- Μορφολογική τμηματοποίηση (Morphological segmentation)
- Αναγνώριση κανονικών ονομάτων (Named entity recognition (NER))
- Παραγωγή φυσικής γλώσσας (Natural language generation)
- Κατανόηση φυσικής γλώσσας (Natural language understanding)
- Οπτική αναγνώριση χαρακτήρων (Optical character recognition (OCR))
- Εύρεση μερών του λόγου (POS tagging)
- Διαπέρασμα προτάσεων (Parsing)
- Απάντηση ερωτήσεων (Question answering)
- Εξαγωγή συσχετίσεων (Relationship extraction)
- Διαχωρισμός προτάσεων (Sentence breaking - boundary disambiguation)
- Συναισθηματική ανάλυση (Sentiment analysis)
- Αναγνώριση λόγου (Speech recognition)
- Τμηματοποίηση λόγου (Speech segmentation)
- Τμηματοποίηση και αναγνώριση θεμάτων (Topic segmentation and recognition)
- Τμηματοποίηση λέξεων (Word segmentation)

- Αποσαφήνιση νοήματος λέξεων (Word sense disambiguation)
- Ανάκτηση πληροφορίας (IR)
- Εξαγωγή πληροφορίας (Information Extraction (IE))
- Οντολογική και λεξικογραφική ανάλυση (ontological and lexical analysis)
- Επεξεργασία λόγου (Speech processing)
- Εξαγωγή ρίζας λέξεων (Stemming)
- Απλούστευση κειμένου (Text simplification)
- Κείμενο σε λόγο (Text-to-speech)
- Ορθογραφικός έλεγχος κειμένου (Text-proofing)
- Αναζήτηση φυσικής γλώσσας (Natural language search)
- Επέκταση ερωτημάτων (Query expansion)

Στα πλαίσια της μεταπτυχιακής μου εργασίας [235] ασχολήθηκα με τα ακόλουθα NLP tasks: *αυτόματη εξαγωγή περίληψης, εξαγωγή και ανάκτηση πληροφορίας, εύρεση μερών του λόγου, διαχωρισμός προτάσεων* και *εξαγωγή ρίζας λέξεων*. Στο πλαίσιο του συστήματος προτάσεων που αναπτύχθηκε στην διδακτορική διατριβή, τα NLP tasks που μας αφορούν είναι επιπλέον: η επέκταση ερωτημάτων, καθώς και η οντολογική και λεξικογραφική ανάλυση.

3.2 Ανάκτηση Πληροφορίας

Η Ανάκτηση Πληροφορίας (ΑΠ)(IR) είναι η διαδικασία αποτελεσματικής εύρεσης πηγών πληροφόρησης σχετικών με μία ανάγκη από μία δεδομένη συλλογή (π.χ. κειμένων). Οι αναζητήσεις προκειμένου να επιτευχθεί η ΑΠ μπορεί να βασίζονται σε μετα-πληροφορία ή σε δεικτοδότηση του πλήρους κειμένου. Τυπικά, η διαδικασία ανάκτησης πληροφορίας ξεκινά όταν ένας χρήστης εισάγει ένα ερώτημα στο σύστημα. Τα ερωτήματα είναι σύνολα από πληροφοριακές ανάγκες, όπως αυτές παρουσιάζονται από τους χρήστες, όπως για παράδειγμα, συμβολοσειρές σε μηχανές αναζήτησης. Στην ΑΠ ένα ερώτημα γενικά δεν χαρακτηρίζει μοναδικά ένα και μόνο αντικείμενο στην συλλογή. Αντιθέτως, πολλά αντικείμενα που ταιριάζουν με το ερώτημα, ίσως με διαφορετικούς βαθμούς ομοιότητας, επιστρέφονται από το σύστημα. Ένα αντικείμενο είναι απλά μία οντότητα η οποία αναπαρίσταται από κάποια πληροφορία στη βάση δεδομένων. Τα ερωτήματα των χρηστών επομένως 'ταιριάζονται' με αυτή την πληροφορία. Τα περισσότερα συστήματα ΑΠ υπολογίζουν μία αριθμητική μετρική, ή αλλιώς σκορ, το οποίο αντιπροσωπεύει πόσο καλά κάθε αντικείμενο ταιριάζει με το ερώτημα, και στη συνέχει ταξινομεί τα αντικείμενα με βάση αυτή το το σκορ. Τα αντικείμενα με το μεγαλύτερο σκορ έπειτα επιστρέφονται στον χρήστη.

3.2.1 Μοντελοποίηση ανάκτησης πληροφορίας

Ακολουθεί ένας τυπικός ορισμό (3.2.1) ενός μοντέλου ανάκτησης πληροφορίας.

Ορισμός 3.2.1. Ένα μοντέλο ανάκτησης πληροφορίας [21] είναι η τετράδα $[D, Q, F, R(q_i, d_j)]$ όπου:

1. D είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής
2. Q είναι ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες του χρήστη. Αυτές οι αναπαραστάσεις καλούνται ερωτήματα
3. F είναι ένα υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους
4. $R(q_i, d_j)$ είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in Q$ και μια αναπαράσταση κειμένου $d_j \in D$. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα q_i.

Αξιοποιώντας λοιπόν τον παραπάνω ορισμό ενός μοντέλου ΑΠ, θα λέγαμε ότι ξεκινούμε από έναν τρόπο αναπαράστασης των κειμένων και των πληροφοριακών αναγκών του χρήστη. Στη συνέχεια (βήμα 3) ορίζουμε ένα υπόβαθρο πάνω στο οποίο αναπαρίσταται τα κείμενα και τα ερωτήματα. Είναι σημαντικό το υπόβαθρο να οριστεί με τρόπο τέτοιο ώστε να υποστηρίζει σύγκριση μεταξύ των αντικειμένων/ερωτημάτων ώστε να καταλήγουμε σε μία δεδομένη κατάταξη των αποτελεσμάτων του εκάστοτε ερωτήματος. Κάθε μοντέλο διαχειρίζεται το υπόβαθρο διαφορετικά. Ο τρόπος που γίνεται αυτό σε ότι έχει να κάνει με τα πιο διαδεδομένα μοντέλα, περιγράφεται στην επόμενη ενότητα.

3.2.1.1 Μοντέλα ανάκτησης πληροφορίας

Τα κλασσικά μοντέλα ΑΠ, πάνω στα οποία βασίζονται και πολλές παραλλαγές τους, είναι τα:

- *Boolean*
- *Vector Space*
- Πιθανοτικό

Το Boolean μοντέλο ΑΠ βασίζεται στη δυαδική (boolean) λογική καθώς και στην θεωρία συνόλων, δεδομένου ότι τόσο τα αντικείμενα προς αναζήτηση, όσο και τα ερωτήματα του χρήστη αντιμετωπίζονται ως σύνολα από όρους. Η ανάκτηση βασίζεται στο αν τα αντικείμενα περιέχουν τους όρους αναζήτησης.

Το μοντέλο Vector Space είναι ένα αλγεβρικό μοντέλο αναπαράστασης των αντικειμένων ως πίνακες χαρακτηριστικών, και άρα, όρους δεικτοδότησης. Έχοντας την αναπαράσταση των αντικειμένων στον $n-$διάστατο χώρο (όπου n τα συνολικά χαρακτηριστικά όλων των κειμένων), μπορούμε να υπολογίσουμε αποστάσεις και ομοιότητες μεταξύ των αντικειμένων.

Τέλος το πιθανοτικό μοντέλο, το οποίο βασίζεται στη θεωρία πιθανοτήτων, αντιστοιχίζει πιθανότητες σε κάθε ένα από τα αντικείμενα δεδομένου του ερωτήματος.

Πέρα από τα τρία παραπάνω κλασσικά μοντέλα, στην βιβλιογραφία έχουν προταθεί αρκετά νέα ή ακόμα και παραλλαγές αυτών. Για την καλύτερη αναπαράσταση και απεικόνιση, τα μοντέλα ΑΠ συχνά κατηγοριοποιούνται σε δύο διαστάσεις: σε σχέση με την μαθηματική τους βάση και σε σχέση με τις ιδιότητες του μοντέλου.

3.2.1.2 Διάσταση μαθηματικής βάσης μοντέλων ανάκτησης πληροφορίας

Σε σχέση με την μαθηματική τους βάση, τα μοντέλα ΑΠ ταξινομούνται στις εξής κατηγορίες:

- Τα συνολοθεωρητικά μοντέλα (Set-theoretic models), που αναπαριστούν τα κείμενα ως σύνολα λέξεων ή φράσεων. Οι ομοιότητες συχνά αντλούνται από συνολοθεωρητικές πράξεις πάνω σε αυτά τα σύνολα. Τέτοια μοντέλα είναι τα:
  - Τυπικό δυαδικό μοντέλο (Standard Boolean model) [124]
  - Εκτεταμένο δυαδικό μοντέλο (Extended Boolean model) [190]
  - Ασαφής ανάκτηση (Fuzzy retrieval) [227]
- Τα αλγεβρικά μοντέλα, τα οποία αναπαριστούν τα κείμενα και τα ερωτήματα συχνά ως διανύσματα, πίνακες ή πλειάδες. Η ομοιότητα μεταξύ ενός διανύσματος ερωτήματος και διανύσματος κειμένου αναπαρίσταται ως μια τιμή. Αλγεβρικά μοντέλα είναι τα:
  - Μοντέλο διανυσματικού χώρου (Vector Space Model (VSM)) [191]
  - Γενικευμένο Μοντέλο διανυσματικού χώρου (Generalized VSM) [219]
  - (Ενισχυμένο) θεματικό μοντέλο διανυσματικού χώρου (Enhanced Topic-based VSM) [26]
  - Εκτεταμένο δυαδικό μοντέλο (Extended Boolean model) [190]
  - Latent Semantic Indexing (LSI) που συχνά αναφέρεται και ως Latent Semantic Analysis (LSA) [62]
- Τα πιθανοτικά μοντέλα, τα οποία αντιμετωπίζουν τη διαδικασία της ΑΠ ως μία πιθανοτική συμπερασματολογία. Οι ομοιότητες υπολογίζονται ως πιθανότητα του κειμένου να είναι σχετικό για ένα δεδομένο ερώτημα. Πιθανοτικά θεωρήματα, όπως του Bayes, αποτελούν συχνά τη βάση για αυτά τα μοντέλα. Πιθανοτικά μοντέλα είναι τα:
  - Δυαδικό μοντέλο ανεξαρτησίας (Binary Independence Model) [226]
  - Πιθανοτικά μοντέλα που βασίζονται στην okapi (BM25) συνάρτηση συσχέτισης [186]
  - Αβέβαιης συμπερασματολογίας (Uncertain inference models) [213]
  - Μοντέλα γλώσσας (Language models) [174]
  - Μοντέλα απόκλισης από την τυχαιότητα (Divergence-from-randomness model) [88]

– Μοντέλα λανθάνουσας κατανομής Dirichlet (Latent Dirichlet allocation) [32]

- Τα μοντέλα ανάκτησης που βασίζονται σε χαρακτηριστικά, αντιμετωπίζουν τα κείμενα ως διανύσματα τιμών συναρτήσεων χαρακτηριστικών (ή απλά ως χαρακτηριστικά) και αναζητούν τον βέλτιστο τρόπο για να συνδυάσουν αυτά τα χαρακτηριστικά σε ένα μόνο σκορ συσχέτισης [130]. Οι συναρτήσεις χαρακτηριστικών είναι άσχετες με το κείμενο ή το ερώτημα και επομένως μπορούν εύκολα να ενσωματώσουν σχεδόν καθένα από τα υπόλοιπα μοντέλα ΑΠ απλά ως ένα νέο χαρακτηριστικό.

3.2.1.3 Διάσταση ιδιοτήτων του μοντέλου

Τα μοντέλα δίχως αλληλεξάρτηση όρων, αντιμετωπίζουν τους όρους/λέξεις ως μη εξαρτημένες μεταξύ τους. Αυτό το γεγονός συνήθως αναπαρίσταται στα μοντέλα VSM μέσω της υπόθεσης ορθογωνιότητας των διανυσμάτων όρων ή στα πιθανοτικά μοντέλα μέσω της υπόθεσης ανεξαρτησίας των μεταβλητών όρων. Τα μοντέλα με έμφυτη την ανεξαρτησία των όρων επιτρέπουν μία αναπαράσταση των ανεξαρτησιών μεταξύ των όρων. Παρόλα αυτά, ο βαθμός ανεξαρτησίας μεταξύ δύο όρων ορίζεται από το ίδιο το μοντέλο. Συνήθως συνεπάγεται άμεσα ή έμμεσα (π.χ. με την μείωση των διαστάσεων) από την συν-εμφάνιση αυτών των όρων στο σύνολο των κειμένων. Τα μοντέλα αυτής της κατηγορίας, ναι μεν επιτρέπουν την αναπαράσταση των αλληλεξαρτήσεων μεταξύ των όρων, δεν κάνουν κάποια υπόθεση όμως σε σχέση με το πως ορίζεται η αλληλεξάρτηση μεταξύ δύο όρων. Αντίθετα βασίζονται σε εξωτερική πηγή για αυτή την πληροφορία (για παράδειγμα ανθρώπινη αλληλεπίδραση ή εξελιγμένους αλγορίθμους)

3.2.1.4 Vector Space Model

Το VSM αναπτύχθηκε στην αρχική του μορφή για αυτόματη δεικτοδότηση δεδομένων [191]. Σύμφωνα με το VSM, μία συλλογή από n κείμενα με m μοναδικούς όρους αναπαρίσταται ως ένας πίνακας όρων-κειμένων $n \times m$ όπου δηλαδή κάθε κείμενο είναι ένα διάνυσμα από m συντεταγμένες. Παρότι το μοντέλο αυτό καθ' αυτό είναι είναι καλά εδραιωμένο, αποτελεί την βάση για πολλά μοντέλα και σχετική έρευνα στο χώρο. Επίσης αποτελεί την βάση για την ανάλυσή μας στην διδακτορική διατριβή και επομένως αξίζει να εμβαθύνουμε λίγο περισσότερο σε αυτό.

Πολλά σχήματα ζυγίσματος όρων έχουν χρησιμοποιηθεί στο VSM, συμπεριλαμβανομένου του δυαδικού ζυγίσματος συχνότητας όρου και της απλής εκδοχής ζυγίσματος βάση της συχνότητας (δηλαδή πόσες φορές εμφανίζονται οι λέξεις στο κείμενο). Στο πιο διαδεδομένο σχήμα, τα διανύσματα αναπαράστασης του κειμένου, αποτελούνται από βάρη που αντιστοιχούν στις συχνότητες των όρων του, πολλαπλασιαζόμενα με το αντίστροφο της συχνότητας τους στην όλη συλλογή κειμένων ($td \times idf$). Η υπόθεση πίσω από αυτό είναι ότι οι λέξεις οι οποίες εμφανίζονται συχνά σε ένα κείμενο, αλλά σπάνια στην συνολική συλλογή κειμένων έχουν υψηλή δυνατότητα αναπαράστασης της πληροφορίας. Σε όλα αυτά τα σχήματα βέβαια είναι σύνηθες να γίνεται μία κανονικοποίηση των διανυσμάτων των κειμένων σε μοναδιαία κλίμακα.

Οι περισσότεροι αλγόριθμοι συσταδοποίησης χρησιμοποιούν κάποιας μορφής VSM αναπαρά-

στασης παρότι πρέπει να αναφερθεί ότι δεν καταγράφεται με αυτό το μοντέλο οποιαδήποτε πληροφορία σε σχέση με την σειρά εμφάνισης των λέξεων, γι' αυτό και το VSM αναφέρεται συχνά και ως αναπαράσταση λίστας λέξεων (BOW representation), η μοντέλο λεξικού.

Δύο σημαντικές ιδιότητες του μοντέλου θα πρέπει να τονιστούν. Πρώτον, σε μία συλλογή από ετερογενή θέματα (κάτι εξαιρετικά σύνηθες για την περίπτωση της συσταδοποίησης), ο αριθμός των μοναδικών όρων μπορεί, και συχνά είναι, εξαιρετικά μεγάλος. Αυτό έχει ως αποτέλεσμα τα διανύσματα των κειμένων να είναι πολλών διαστάσεων. Για την αντιμετώπιση αυτού του προβλήματος ένα πλήθος τεχνικών προεπεξεργασίας έχουν ερευνηθεί στην βιβλιογραφία. Δεύτερον, ο πίνακας που παράγεται από μία τυπική βάση κειμένων είναι σε γενικές γραμμές πολύ αραιός με το VSM, διότι η βάση κειμένων περιέχει πολύ περισσότερους όρους σε σχέση με το καθένα ξεχωριστό κείμενο που την απαρτίζει.

3.2.2 Αξιολόγηση αποτελεσμάτων ανάκτησης πληροφορίας

Ένα από τα βασικά στοιχεία αξιολόγησης του IR είναι η μέτρηση του κατά πόσο τα ανακτημένα κείμενα είναι σχετικά με το ερώτημα που κάνουμε. Έτσι λοιπόν, ένα βασικό στοιχείο στο οποίο εστιάζουμε είναι η εύρεση μετρικών που θα μπορούν να αναπαραστήσουν αριθμητικά τη σχετικότητα των αποτελεσμάτων ενός συστήματος IR. Πολλές μετρικές έχουν αναπτυχθεί ανά καιρούς και στην παρούσα ενότητα θα καταγράψουμε συνοπτικά τις σημαντικότερες και πιο συνηθισμένες από αυτές.

3.2.2.1 Ανάκληση και ακρίβεια

Ίσως οι πιο γνωστές μετρικές αξιολόγησης των αποτελεσμάτων ενός συστήματος ανάκτησης πληροφορίας να είναι η ανάκληση και η ακρίβεια. Η ακρίβεια μας δίνει το ποσοστό (%) των σχετικών κειμένων εν συγκρίσει με αυτά που ανακτήθηκαν, ενώ η ανάκληση μας δίνει το ποσοστό (%) των κειμένων που ανακτήθηκαν εν συγκρίσει με μία συλλογή που γνωρίζουμε ότι περιέχει όλα τα σχετικά.

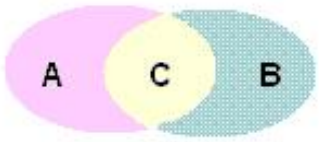
Φορμαλιστικά, οι σχέσεις που ισχύουν για τις δύο αυτές μετρικές είναι οι παρακάτω:

$$R = \frac{|\{A\} \cap \{B\}|}{A} \tag{1}$$

$$P = \frac{|\{A\} \cap \{B\}|}{B} \tag{2}$$

όπου R η ανάκληση, P η ακρίβεια, A τα σχετικά κείμενα που βρέθηκαν και B όλα τα άρθρα που ανακτήθηκαν. Οι παραπάνω συσχετίσεις είναι εμφανείς στο σχήμα 3.

Θα λέγαμε επομένως ότι η ανάκληση μας δίνει ένα μέτρο για το πόσο καλά μια αναζήτηση εντοπίζει αυτό που θέλουμε, ενώ η ακρίβεια μετράει το πόσο καλά απορρίπτουμε αυτό που δεν θέλουμε. Αυτές οι μετρικές, παρότι πολύ χρήσιμες για την αξιολόγηση, είναι δύσχρηστες από τη φύση τους. Πρώτα απ' όλα η έννοια της ακρίβειας είναι συνήθως αποκλειστικά υποκειμενικό κριτήριο και όχι μια αντικειμενική θετική ή αρνητική απάντηση. Δεύτερον, για κάθε βάση πληροφορίας που είναι αρκετά μεγάλη για να κατασκευαστεί μια μηχανή αναζήτησης πάνω της, θα είναι δύσκολο να



Σχήμα 3: Ακρίβεια - Ανάκληση. Με C είναι τα σχετικά άρθρα που ανακτήθηκαν.

υπολογιστούν πραγματικές τιμές ανάκλησης λόγω του μεγέθους της βάσης (για να υπολογιστεί επακριβώς η ανάκληση θα πρέπει να γνωρίζουμε ακριβώς πόσα matches έγιναν, και αν γνωρίζαμε κάτι τέτοιο, ποιος ο λόγος να έχουμε μια μηχανή αναζήτησης;). Τρίτον, η ακρίβεια και η ανάκληση δεν είναι στον πραγματικό κόσμο απλά αριθμοί· είναι δύο έννοιες που σχετίζονται στενά. Για παράδειγμα ενώ ψάχνουμε στις σελίδες απάντησης μιας μηχανής αναζήτησης για ένα ερώτημα που δώσαμε, περιμένουμε καθώς περνάμε τις σελίδες η ανάκληση να βελτιώνεται ενώ παράλληλα η ακρίβεια να χειροτερεύει.

3.2.2.2 Fall-out

Η μετρική Fall-out ορίζεται ως η αναλογία των μη σχετικών κειμένων τα οποία ανακτούνται, σε σχέση με όλα τα μη σχετικά κείμενα τα οποία υπάρχουν:

$$\text{fall-out} = \frac{|\{\text{μη σχετικά κείμενα}\} \cap \{\text{ανακτημένα κείμενα}\}|}{|\text{μη σχετικά κείμενα}|} \tag{3}$$

Φυσικά μπορούμε να παρατηρήσουμε ότι εύκολα μπορούμε να παράγουμε μηδενικές τιμές για την fall-out μετρική: απλά δεν επιστρέφουμε αποτελέσματα.

3.2.2.3 F-measure

Η μετρική F-measure, ή αλλιώς F-score, αποτελεί τον αρμονικό μέσο της ακρίβειας και ανάκλησης ή γενικά ένας ζυγισμένος συνδυασμός των δύο αυτών μετρικών:

$$\text{F} = 2 \times \frac{\text{ακρίβεια} \times \text{ανάκληση}}{\text{ακρίβεια} + \text{ανάκληση}} \tag{4}$$

Έστω λοιπόν ένα σύνολο από κείμενα C που ένα recommendation system προτείνει στον χρήστη, καθώς και ένα σύνολο από κείμενα $\bar{C}$ τα οποία επισκέπτεται ο χρήστης μετά τις προτάσεις του συστήματος. Επίσης έστω ότι με $r(c', c)$ είναι το πλήθος των κειμένων τα οποία ανήκουν και στα δύο παραπάνω σύνολα. Τότε:

$$F(c', c) = \frac{2r(c', c)p(c', c)}{r(c', c) + p(c', c)} \tag{5}$$

όπου:

$$r(c', c) = \frac{\text{doc}(c', c)}{\text{doc}(c')} \tag{6}$$

και:

$$p(c', c) = \frac{\text{doc}(c', c)}{\text{doc}(c)} \tag{7}$$

Η μετρική (4) ονομάζεται και F1-measure. Στη γενική περίπτωσή λοιπόν:

$$\text{F}_\beta = \frac{(1 + \beta^2) \times (\text{ακρίβεια} \times \text{ανάκληση})}{(\beta^2 \times \text{ακρίβεια}) + \text{ανάκληση}} \tag{8}$$

Τέλος, η φυσική σημασία αυτής της μετρικής αφορά στην αποτελεσματικότητα του συστήματος που αξιολογούμε σε σχέση με κάποια εφαρμογή που θεωρεί β φορές πιο σημαντική την ανάκληση σε σχέση με την ακρίβεια.

3.2.2.4 Μέση τιμή ακρίβειας

Η ακρίβεια και η ανάκληση είναι μετρικές μίας και μόνο τιμής, βασισμένες στην πλήρη λίστα από κείμενα που επιστρέφεται από το σύστημα. Για συστήματα που επιστρέφουν μία ταξινομημένη σειρά από κείμενα, είναι επιθυμητό να λαμβάνεται υπόψιν επίσης και η σειρά με την οποία τα επιστρεφόμενα αντικείμενα παρουσιάζονται. Υπολογίζοντας την ακρίβεια και την ανάκληση σε κάθε θέση της σειράς κατάταξης των κειμένων, μπορούμε να σχεδιάσουμε την καμπύλη ακρίβειας-ανάκλησης, ζωγραφίζοντας την ακρίβεια $p(r)$ σαν συνάρτηση της ανάκλησης r. Η μέση τιμής της ακρίβειας είναι επομένως:

$$\text{AveragePr} = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{9}$$

όπου k είναι η σειρά στην ταξινόμηση των ανακτημένων κειμένων, n είναι το πλήθος των ανακτημένων κειμένων, $P(k)$ είναι η ακρίβεια στο σημείο αποκοπής k στη λίστα και $\Delta r(k)$ είναι η αλλαγή στην ανάκληση από τα σημεία $k-1$ έως k [232].

3.2.2.5 R-Ακρίβεια

Η μετρική αυτή [20] καταγράφει την ακρίβεια στην R-ιοστή θέση στην κατάταξη των αποτελεσμάτων για ένα ερώτημα που έχει R σχετικά κείμενα. Η R-ακρίβεια είναι υψηλά συσχετιζόμενη με την μέση ακρίβεια. Επίσης, η ακρίβεια είναι ίση με την ανάκληση στην R-ιοστή θέση.

3.3 Φιλτράρισμα Πληροφορίας

Ένα σύστημα IR δύσκολα μπορεί να πετύχει πολύ υψηλές τιμές τόσο ακρίβειας όσο και ανάκλησης. Οι τιμές αυτές μάλιστα δεν έχουν καμία σύγκριση με ένα σύστημα DataBase Management System (DBMS) που τα ποσοστά αυτά βρίσκονται στο 100%. Ωστόσο θα μπορούσε κανείς να πει πως και τα δύο συστήματα πραγματοποιούν την ίδια διαδικασία, δηλαδή ανάκτηση πληροφορίας. Αυτό βέβαια έχει να κάνει με τον τρόπο με τον οποίο δομείται ένα σύστημα DBMS και ο οποίος είναι τέτοιος ώστε να εξυπηρετεί απόλυτα τις ανάγκες ενός χρήστη.

Αυτή η δυσκολία που αντιμετωπίζουν τα συστήματα IR (μικρές τιμές ανάκλησης και ακρίβειας) γεννούν ένα άλλο επιστημονικό πεδίο το οποίο υπάρχει παράλληλα με το IR και είναι το Information Filtering (IF). Σε ένα κλασσικό άρθρο οι Belkin και Croft παρουσίασαν δύο διαφορετικούς ορισμούς για τα δύο παραπάνω θέματα οι οποίοι έχουν κοινές τεχνικές αλλά διαφέρουν σε τρία βασικά στοιχεία [27]. Πρώτον, στο IR όταν ο χρήστης κάνει ένα ερώτημα περιμένει άμεση απόκριση. Στο IF ο χρήστης μπορεί να περιμένει, εν γνώσει του, για μεγάλο χρονικό διάστημα μέχρι να του παρουσιαστεί μία απάντηση. Επιπρόσθετα, το IF χειρίζεται και θέματα που από τη φύση τους είναι δυναμικά και εντάσσει στο μηχανισμό του στοιχεία εκμάθησης σύμφωνα με τα κείμενα που προσθέτει στη συλλογή του. Τελευταίο και βασικότερο, είναι πως το IR αναζητά παραπλήσια κείμενα από μία μεγάλη συλλογή κειμένων σε αντίθεση με το IF, το οποίο προσπαθεί να αφαιρέσει από μία συλλογή τα εισερχόμενα κείμενα που δεν είναι σχετικά, κρατώντας έτσι μόνο ότι θεωρεί σχετικό με τον εκάστοτε χρήστη. Παρ' όλες τις διαφορές που έχουν τα δύο αυτά πεδία δεν πρέπει να αμελούμε πως έχουν παραπλήσιο σκοπό: να εξασφαλίσουν ότι τα κείμενα που θα παρουσιαστούν στο χρήστη είναι σχετικά με το ερώτημά του.

Τα διαγράμματα ακρίβειας/ανάκλησης είναι χρήσιμα εφόσον μελετούμε την απόδοση ανάκτησης διαφορετικών αλγορίθμων σε ένα σύνολο από πρότυπες πληροφοριακές ανάγκες. Ωστόσο υπάρχουν περιπτώσεις στις οποίες θα θέλαμε να συγκρίνουμε την απόδοση αλγορίθμων ανάκτησης για ατομικές πληροφοριακές ανάγκες. Οι λόγοι για να το κάνουμε αυτό είναι δύο:

1. η χρήση μέσων τιμών που προκύπτουν από την εκτέλεση διαφόρων ερωτημάτων μπορεί να αποκρύπτει σημαντικές ανωμαλίες στον αλγόριθμο ανάκτησης,

2. όταν συγκρίνουμε δύο αλγορίθμους, μπορεί να θέλουμε να μελετήσουμε κατά πόσο ο ένας είναι καλύτερος του άλλου για κάθε μία από τις πληροφοριακές ανάγκες που έχουμε και όχι συνολικά.

Σε τέτοιες περιπτώσεις υπολογίζουμε μία μόνο τιμή ακρίβειας για κάθε ερώτημα, η οποία θα μπορούσε να θεωρηθεί σαν σύνοψη του συνολικού διαγράμματος ακρίβειας/ανάκλησης. Συνήθως αυτή η τιμή είναι η ακρίβεια σε κάποιο συγκεκριμένο επίπεδο ανάκλησης. Φυσικά αυτές είναι λίγες από τις πολλές προσεγγίσεις αξιολόγησης που μπορούν να γίνουν.

3.3.1 Εξόρυξη από τον παγκόσμιο ιστό

Η εξόρυξη από τον παγκόσμιο ιστό (Web mining) εστιάζει στην εύρεση φυσικών οντοτήτων και συσχετισμό αυτών από πηγές του διαδικτύου ή χρήστες αυτού. Θα μπορούσαμε να χωρίσουμε χοντρικά το Web mining σε τρεις βασικές κατηγορίες [54]. Αρχικά, στο Web content mining, όπου η πληροφορία εξάγεται από το περιεχόμενο των σελίδων και των υπερσυνδέσμων (hyperlinks) αυτών, όχι επομένως από τους χρήστες καθ' αυτούς. Δεύτερων, στο Web Structure Mining, όπου η δομική πληροφορία σχετικά με τα hyperlinks και η οργάνωση των σελίδων παίζει κυρίαρχο ρόλο. Και τρίτων, στο Web Usage Mining, το οποίο εστιάζει στην εξαγωγή χρήσιμων προτύπων χρήσης από την συμπεριφορά των χρηστών.

Η συσταδοποίηση των χρηστών του διαδικτύου αποτελεί ένα ξεχωριστό ερευνητικό πεδίο στην υποκατηγορία του Web Usage Mining το οποίο αποσκοπεί στην περιγραφή γενικών τάσεων στην συμπεριφορά των χρηστών μέσα σε ένα δεδομένο χρονικό πλαίσιο. Όπως εξηγείται στο [168], το Web mining είναι ουσιαστικά η εξαγωγή ενδιαφερόντων και πιθανά χρησίμων προτύπων και έμμεσης πληροφορίας από αντικείμενα ή συμπεριφορές σχετικές με τον παγκόσμιο ιστό.

Το πεδίο έχει επίσης μελετηθεί και στο πλαίσιο της προσωποποίησης του ιστού από πολλούς ερευνητές, π.χ. [63], [71]. Στο [147] λαμβάνονται υπόψιν βασικά δύο τύποι από πρότυπα χρήσης και γίνεται συσταδοποίηση πάνω σε αυτά προκειμένου να κατασκευαστούν γενικά προφίλ πλοήγησης των χρηστών, χωρίς μάλιστα να έχει κάποια επίπτωση η σειρά πρόσβασης. Στο [71] παρουσιάζεται μία μέθοδος η οποία κάνει χρήση επαγωγής με βάση τα χαρακτηριστικά των χρηστών, όπου οι συνεδρίες των χρηστών αναπαρίστανται ως πίνακες στον n-διάστατο Ευκλείδειο χώρο των όρων. Η οπτικοποίηση των επιλογών του χρήστη έχει επίσης μελετηθεί στο [41] για πρότυπα πλοήγησης. Στο [90] εισάγεται μία μεθοδολογία στοίχισης ακολουθίας (Sequence Alignment) η οποία συσταδοποιεί τους χρήστες με βάσει τα πρότυπα πλοήγησής τους. Αυτή η μέθοδος βασίζεται στην σειρά με την οποία τα γεγονότα πλοήγησης λαμβάνουν χώρα από τους χρήστες.

Το Web usage mining ουσιαστικά οδηγεί στο συνεργατικό φιλτράρισμα όταν κάνει χρήση των γνωστών προτιμήσεων από ένα σύνολο χρηστών προκειμένου να κάνει προτάσεις ή προβλέψεις σχετικά με άγνωστες προτιμήσεις χρηστών.

3.3.2 Συνεργατικό φιλτράρισμα - Collaborative Filtering

Το συνεργατικό φιλτράρισμα (collaborative filtering) έχει δύο έννοιες [184], μία στενή και μία πιο ευρύτερη [204]. Γενικά, το συνεργατικό φιλτράρισμα είναι η διαδικασία φιλτραρίσματος της πληροφορίας με χρήση τεχνικών που εμπεριέχουν συνεργασία μεταξύ πηγών, αντιλήψεων, κ.λπ. Η προσαρμογή των CF συστημάτων στις προτιμήσεις του χρήστη, μειώνει την προσπάθεια αναζήτησης από την πλευρά του. Οι εφαρμογές του συνεργατικού φιλτραρίσματος τυπικά εμπεριέχουν πολύ μεγάλα σύνολα δεδομένων. Μέθοδοι CF έχουν εφαρμοστεί σε πολλά διαφορετικά είδη δεδομένων, συμπεριλαμβανομένων των: αίσθηση και παρακολούθηση δεδομένων, οικονομικά δεδομένα, ηλεκτρονικό εμπόριο, κ. α. Στην νεότερη, πιο στενή έννοια, το CF είναι μία μέθοδος για αυτόματες προβλέψεις (φιλτράρισμα) σε σχέση με τα ενδιαφέροντα του χρήση, με χρήση συλλογή των ενδιαφερόντων ή των προτιμήσεών πολλών άλλων χρηστών (συνεργασία). Αξίζει να σημειωθεί ότι οι προβλέψεις που κάνει ένα CF είναι στοχευμένες για τον συγκεκριμένο χρήστη, όμως η αρχική πληροφορία πηγάζει από πολλούς άλλους. Αυτό διαφέρει από την απλούστερη προσέγγιση η οποία δίνει ένα μέσο (όχι συγκεκριμένο) σκορ για κάθε είδος ενδιαφέροντος, που βασίζεται για παράδειγμα στο πλήθος των ψήφων.

Η έννοια του συνεργατικού φιλτραρίσματος εισήχθηκε από τους ερευνητές ενός εκ' των πρώτων συστημάτων προτάσεων, του Tapestry [80], προκειμένου να περιγράψουν αυτή την τεχνική προσωποποιημένων προτάσεων που βασίζεται στην ομοιότητα των ενδιαφερόντων των χρηστών. Το συνεργατικό φιλτράρισμα στοχεύει επομένως στο να περιγράψει γενικά τις διάφορες τεχνικές προσωποποιημένων προτάσεων. Από τότε, έχει ευρέως υιοθετηθεί και εξελιχθεί σε τέτοιο βαθμό

ώστε τα συστήματα προτάσεων να προτείνουν ιδιαίτερα ενδιαφέροντα αποτελέσματα στους χρήστες, ενώ παράλληλα να φιλτράρουν αποτελεσματικά τον όγκο δεδομένων που διαχειρίζονται. Η βασική υπόθεση ενός CF συστήματος [121] είναι ότι:

Υπόθεση 1. *αν οι χρήστες X και Y βαθμολογούν n αντικείμενα παρόμοια, ή γενικά έχουν παρόμοιες συνήθειες (π.χ. αγοραστικές, ακουστικές, κ.λπ.), τότε θα βαθμολογήσουν η θα ενεργήσουν σε άλλα αντικείμενα παρόμοια*

Οι CF αλγόριθμοι συχνά απαιτούν:

1. την ενεργή συμμετοχή των χρηστών στη διαδικασία - συχνά με απαντήσεις σε σχέση ή ενέργειες που φανερώνουν τις προτιμήσεις τους
2. έναν εύκολο τρόπο αναπαράστασης των ενδιαφερόντων των χρηστών στο σύστημα
3. αλγόριθμους οι οποίοι είναι ικανοί να ταιριάξουν ανθρώπους με παρόμοια ενδιαφέροντα.

3.3.2.1 Ροή πληροφορίας CF

Τυπικά η ροή πληροφορίας σε ένα σύστημα CF έχει ως εξής:

- ένας χρήστης εκφράζει τα ενδιαφέροντά του βαθμολογώντας/αξιολογώντας αντικείμενα (π.χ. βιβλία, ταινίες, άρθρα νέων) του συστήματος. Αυτές οι βαθμολογίες μπορούν να ειδωθούν ως μια "στο περίπου" αναπαράσταση των ενδιαφερόντων του χρήστη στο συγκεκριμένο τομέα ενδιαφέροντος.
- το σύστημα ταιριάζει τα ενδιαφέροντα του χρήστη με εκείνα άλλων χρηστών και βρίσκει εκείνους με "παρόμοια" ενδιαφέροντα
- έχοντας τους παρόμοιους χρήστες, το σύστημα προτείνει αντικείμενα τα οποία οι παρόμοιοι χρήστες έχουν βαθμολογήσει υψηλά αλλά δεν έχουν ακόμα βαθμολογηθεί από τον χρήστη (υποθέτοντας ότι η απουσία βαθμολόγησης συχνά φανερώνει μη γνώση για το συγκεκριμένο αντικείμενο)

Ένα βασικό πρόβλημα του συνεργατικού φιλτραρίσματος είναι το πως να συνδυαστούν και να ζυγιστούν οι προτιμήσεις των παρόμοιων χρηστών. Μερικές φορές, οι χρήστες μπορούν να βαθμολογήσουν άμεσα τα προτεινόμενα αντικείμενα. Ως αποτέλεσμα, με το πέρασμα του χρόνου, το σύστημα κερδίζει μία ολοένα και αυξανόμενη αναπαράσταση των προτιμήσεων του χρήστη. Ένα ακόμη πρόβλημα του CF είναι ότι τα σκορ ομοιότητας τυπικά δεν λαμβάνουν υπόψιν τους τα μεταβαλλόμενα ενδιαφέροντα χρήστη. Επίσης δεν μπορούν να υπολογίζουν την αξιοπιστία των επιλογών των χρηστών, κάτι που μπορεί εύκολα να οδηγήσει σε άσχημα αποτελέσματα προτάσεων, ακόμη και για τους καλύτερους αλγορίθμους. Στην διδακτορική διατριβή προσπαθούμε να αντιμετωπίσουμε το παραπάνω πρόβλημα κάνοντας μικρές αλλά συνεχείς αλλαγές στα προφίλ χρηστών βάσει των εκάστοτε επιλογών τους.

Ένα ακόμη πρόβλημα που επίσης έχουν τα συστήματα συνεργατικού φιλτραρίσματος είναι ότι δεν δουλεύουν πάντα καλά λόγω του φαινομένου της αραιότητας των διαθέσιμων δεδομένων (data scarcity). Κάθε χρήστης του συστήματος έχει δει ένα μικρό μέρος μόλις των δεδομένων και επομένως ακριβείς προβλέψεις δεν μπορούν εύκολα να γίνουν, τουλάχιστον έως ότου η κάλυψη των χρηστών στα δεδομένα έχει αυξηθεί σε κάποιο βαθμό. Ένας τρόπος αντιμετώπισης αυτής της κατάστασης είναι η ομαδοποίηση των χρηστών σε ομάδες παρομοίων ενδιαφερόντων. Έτσι, αξιοποιώντας την πιθανή συμμετρία στις επιλογές των χρηστών που βρίσκονται στις ίδιες συστάδες, θα μπορούσαμε να ομαδοποιήσουμε άρθρα νέων βασιζόμενοι στο ποιος τα βλέπει - χρησιμοποιώντας έτσι ομάδες άρθρων αντί για μεμονωμένους χρήστες. Η αντίστροφη προσέγγιση είναι επίσης πιθανή: έστω μία ομάδα χρηστών οι οποίοι έχουν προηγουμένως εκφράσει το ενδιαφέρον τους για ένα συγκεκριμένο θέμα. Ένα πρόσφατο άρθρο με ομοιότητα προς κάποια από τα άρθρα που προηγουμένως έχουν διαβασθεί από μερικά μέλη της ομάδας, είναι πιθανό να ενδιαφέρει και τους υπόλοιπους χρήστες αυτής της ομάδας. Έτσι, αντί να βασιζόμαστε στις επιλογές μεμονωμένων χρηστών, η συστάδα ενσωματώνει και προσθέτει την απαραίτητη πληροφορία που χρειάζεται ένα CF σύστημα. Την λογική αυτή ακριβώς αξιοποιούμε και εμείς προκειμένου να αντιμετωπίσουμε το εν' λόγω πρόβλημα.

Οι δύο τεχνικές που παραδοσιακά αξιοποιούνται για εφαρμογή των παραπάνω προσεγγίσεων είναι η k Nearest Neighbors (k-NN) και η συσταδοποίηση.

3.3.2.2 Απαιτήσεις CF

Πολλές τεχνικές παραγοντοποίησης πινάκων έχουν εφαρμοστεί στο CF, όπως το Singular Value Decompossition (SVD), το probabilistic LSA, το probabilistic matrix factorization, κ.λπ. Παρόλα αυτά, ο συνδυασμός πολλαπλών αλγορίθμων φαίνεται να υπερτερεί των απλούστερων μεθοδολογιών [197]. Οι CF τεχνικές συχνά χρησιμοποιούν μία βάση δεδομένων για τις προτιμήσεις των χρηστών προς αντικείμενα. Σε ένα τυπικό σενάριο μίας λίστας m χρηστών $u_1, u_2, ..., u_m$ και μίας λίστας n αντικειμένων $i_1, i_2, ..., i_n$, όπου κάθε χρήστης u_i έχει μία λίστα από αντικείμενα I_{ui}, τα οποία ο χρήστης βαθμολόγησε άμεσα (π.χ. σε κλίματα 1-5) ή σε σχέση με τα οποία υπάρχει έμμεση ένδειξη ενδιαφέροντος με βάση τη συμπεριφορά του (π.χ. μέσω αγορών ή click-throughs).

Έτσι λοιπόν, οι αλγόριθμοι CF απαιτείται:

- να έχουν τη δυνατότητα να αντιμετωπίζουν τα αραιά διαθέσιμα δεδομένα
- να κλιμακώνονται με την αύξηση των χρηστών και των αντικειμένων
- να κάνουν ικανοποιητικές προτάσεις σε σύντομο χρονικό διάστημα (ικανοποιητική απόκριση)
- να μπορούν να αντιμετωπίσουν προβλήματα όπως η συνωνυμία (όπου παρόμοια αντικείμενα έχουν διαφορετικά ονόματα), shilling attacks [49], θόρυβο στα δεδομένα καθώς και θέματα προστασίας της ιδιωτικότητας [197]

3.3.2.3 Κατηγορίες CF

Οι τεχνικές CF έχουν χοντρικά τρεις κατηγορίες:

1. Memory-based, όπως για παράδειγμα τεχνικές που βασίζονται σε γείτονους (neighbor-based) [91] και item- based top-N τεχνικές [192][113]

2. Model-besed, για παράδειγμα Bayesian δίκτυα πεποίθησης (Bayesian belief nets) [202], λανθάνουσα σημασιοδότηση (latent semantic) [94] καθώς και περιορισμού διαστάσεων (dimensionality reduction) SVD [173]

3. Υβριδικά, τα οποία συνδυάζουν τα πλεονεκτήματα και των δύο παραπάνω κατηγοριών ενώ παράλληλα βελτιώνουν της απόδοση των προβλέψεων προτιμήσεων χρήστη [201]

Η αρχική γενιά CF συστημάτων χρησιμοποιούσε τα δεδομένα βαθμολόγησης των χρηστών προκειμένου να υπολογίσει την ομοιότητα ή το βάρος μεταξύ χρήστη και αντικειμένου, ώστε να κάνει προβλέψεις ή προτάσεις σύμφωνα με αυτές τις τιμές ομοιότητας. Τα memory-based CF συστήματα συχνά τα συναντούμε σε εμπορικές εφαρμογές [94] όπως το Amazon [10] και το Barnes and Noble [23] διότι είναι εύκολα ως προς την υλοποίηση τους και αρκετά αποδοτικά.

Για να επιτύχουν καλύτερα αποτελέσματα στις προβλέψεις τους και να αποφύγουν τα μειονεκτήματα των memory-based αλγορίθμων, οι model-based προσεγγίσεις κάνουν χρήση των πρωτογενών δεδομένων βαθμολόγησης προκειμένου να εκτιμήσουν και να εκμάθουν ένα μοντέλο το οποίο κάνει τις προβλέψεις. Το μοντέλο μπορεί να είναι κάποιος αλγόριθμος εξόρυξης δεδομένων ή μηχανικής εκμάθησης. Πολύ συχνές model-based CF τεχνικές είναι τα Bayesian δίκτυα πεποίθησης [145][195], τα CF μοντέλα συσταδοποίησης [203][46], καθώς και τα latent semantic CF μοντέλα [94]. Επίσης τα Markov decision process (MDP) μοντέλα CF [183] παράγουν αποτελέσματα με πολύ υψηλή απόδοση.

3.3.3 Φιλτράρισμα βάσει περιεχομένου

Πέρα από το συνεργατικό φιλτράρισμα, το φιλτράρισμα βάσει περιεχομένου (content-based filtering) είναι μια πολύ σημαντική κατηγορία συστημάτων προτάσεων. Τα συστήματα προτάσεων αυτού του είδους κάνουν προτάσεις αναλύοντας το περιεχόμενο της κειμενικής πληροφορίας και βρίσκοντας κανονικότητες στο περιεχόμενο, όπως π.χ. στο [87]. Η βασική διαφορά μεταξύ των CF και των content-based filtering συστημάτων προτάσεων είναι ότι τα πρώτα χρησιμοποιούν μόνο τις βαθμολογίες χρηστών-αντικειμένων για να κάνουν τις προβλέψεις και προτάσεις τους, ενώ τα δεύτερα βασίζονται στα χαρακτηριστικά των χρηστών και των αντικειμένων για αυτές [195].

Τόσο τα CF όσο και τα content-based filtering συστήματα όμως έχουν τους περιορισμούς τους: ενώ τα CF συστήματα δεν συμπεριλαμβάνουν άμεσα πληροφορία χαρακτηριστικών, τα content-based συστήματα δεν ενσωματώνουν απαραίτητα την πληροφορία για την ομοιότητα των προτιμήσεων μεταξύ των χρηστών [13]. Οι υβριδικές CF τεχνικές, όπως content-based CF αλγόριθμοι [141] και τεχνικές "διάγνωσης" προσωπικότητας (Personality Diagnosis (PD)) [176], συνδυάζουν το CF με το content-based με σκοπό την αποφυγή των περιορισμών των δύο κατηγοριών και

συνεπώς την βελτίωση της απόδοσης των προτάσεων. Η προσέγγιση αυτή αξιοποιείται για στην διδακτορική διατριβή για το σύστημα προτάσεων που αναπτύχθηκε.

3.4 Συστήματα προτάσεων

Όπως εξηγείται και στο [120], τα συστήματα προτάσεων έχουν μία ιστορία η οποία ξεκίνησε με τους εστιασμένους αλγόριθμους πρόβλεψης, οι οποίοι στην συνέχεια επεκτάθηκαν σε εμπορική χρήση και που πρόσφατα εστιάζουν σε πιο λεπτομερείς μεθοδολογίες ξεφεύγοντας από την λογική απλά και μόνο της ακρίβειας των προβλέψεων.

Στις αρχές της δεκαετίας 1990, καθώς η χρήση του διαδικτύου εξαπλωνόταν γρήγορα, συστήματα προτάσεων που βασίζονται σε συνεργατικό φιλτράρισμα εφευρέθηκαν για να βοηθήσουν τους χρήστες να αντιμετωπίσουν την υπερφόρτωση πληροφορίας με τη δημιουργία μοντέλων πρόβλεψης που εκτιμούν πόσο ο χρήστης θα ήθελε να έχει γνώση για τα εν' λόγω αντικείμενα. Το σύστημα GroupLens [181] βασιζόταν στην διαίσθηση ότι κάθε φορά που ένας χρήστης διάβαζε ένα άρθρο από το Usenet, σχημάτιζε και στη συνέχεια "πετούσε" μια πολύτιμη γνώμη. Αυτή η γνώμη καταγράφονταν από το σύστημα και έτσι, χρησιμοποιώντας τις αξιολογήσεις των "ομοϊδεατών" μπορούσε να παράγει τα προσωποποιημένες προβλέψεις που εμφανίζονταν ως μέρος της επικεφαλίδας του άρθρου.

Το σύστημα Ringo [196] προσέφερε προτάσεις για μουσικούς καλλιτέχνες χρησιμοποιώντας μια παρόμοια τεχνική που ονομάστηκε "κοινωνικό φιλτράρισμα των πληροφοριών". Ομοίως και για το πεδίο των προτάσεων πληροφορίας βίντεο [93], όπου χρησιμοποιήθηκαν παρόμοιοι αλγόριθμοι και ενημέρωση μέσω e-mail για τις εικονικές κοινότητες των ταινιόφιλων. Τα συστήματα προτάσεων γρήγορα έγιναν δημοφιλή, τόσο όσον αφορά την έρευνα, όσο και την εμπορική τους εκμετάλλευση και μέχρι το 1996, πολλές εταιρείες διαφήμιζαν και προωθούσαν τους μηχανισμούς προτάσεων τους.

Σε σχέση με την παραπάνω αρχή, το πεδίο έχει προχωρήσει τόσο μέσω της βασικής έρευνας και της εμπορικής ανάπτυξης, έως το σημείο όπου τα συστήματα συστάσεων σήμερα ενσωματώνονται σε ένα ευρύ φάσμα εφαρμογών περιεχομένου (online και offline). Παράλληλα, το πεδίο εφαρμογής των συστημάτων προτάσεων έχει διευρυνθεί, ενώ ο όρος, που αρχικά ήταν συνυφασμένος με το συνεργατικό φιλτράρισμα, γρήγορα επεκτάθηκε ώστε να συμπεριλάβει ένα ευρύτερο φάσμα από προσεγγίσεις που βασίζονται στο περιεχόμενο (content-based) αλλά και στη γνώση (knowledge-based).

Όλα τα πρώτα λοιπόν συστήματα προτάσεων χρησιμοποιούσαν παραλλαγές του ζυγισμένου k-NN αλγορίθμου. Διαισθητικά, αυτός ο αλγόριθμος προβλέπει πόσο ένα αντικείμενο i θα αρέσει σε έναν χρήστη u με το να επιλέγει μία γειτονία από άλλους χρήστες με ενδιαφέροντα όσο το δυνατόν κοντινότερα στον u. Η επιλογή γειτονικότητας γίνεται μέσω του υπολογισμού ενός μέτρου ομοιότητας μεταξύ των προηγούμενων επιλογών του u και επιλογών άλλων χρηστών (συχνά με βάση τη μετρική ομοιότητας του Pearson, ή ως ένα πίνακα ομοιότητας συνημιτόνου) και επιλέγοντας τα πιο όμοια αντικείμενα ως γείτονες [92].

Με την πρόβλεψη ενδιαφερόντων ως το βασικό έργο τους, δεν είναι περίεργο που οι πιο δημο-

φιλείς στρατηγικές αξιολόγησης των συστημάτων προτάσεων ήταν (και ακόμα και τώρα είναι σε μεγάλο βαθμό) η ακρίβεια των παραγόμενων προβλέψεων. Τα περισσότερα από τα πρώτα συστήματα προτάσεων αξιολογούνταν με βάσει κριτήρια όπως το σφάλμα ή η συσχέτιση. Στα παραπάνω περιλαμβάνονται το απόλυτο σφάλμα και το μέσο τετραγωνικό σφάλμα, προσφέροντας μία εκτίμηση του πόσο κοντά βρίσκονται οι προβλέψεις στα πραγματικά ενδιαφέροντα ή βαθμολογήσεις. Η συσχέτιση παρέχει ένα παρόμοιο μέτρο, αλλά εστιάζει στις σχετικές προβλέψεις, παρά στις απόλυτες τιμές πρόβλεψης. Σε κάθε περίπτωση, αυτές οι μετρικές εφαρμόζονται σε μέρος των δεδομένων (παρακρατημένα από το μηχανισμό προτάσεων) προκειμένου να εκτιμηθεί η ακρίβεια. Υπάρχει ένα σημαντικό μειονέκτημα όπως των παραπάνω μετρικών που πρέπει να αναφέρουμε. Μπορεί να κάνουν καλή δουλειά στο να εκτιμούν τα συστήματα προτάσεων ως προσεγγίσεις ανάκτησης ελλιπών δεδομένων, δεν κάνουν και τόσο καλή δουλειά όμως στο να αξιολογούν αν τα συστήματα προτάσεων προτείνουν αντικείμενα με αξία και προηγούμενος άγνωστα στον χρήστη (κάτι που είναι και ο βασικός στόχος άλλωστε των συστημάτων προτάσεων).

Σύντομα η λογική των συστημάτων προτάσεων μετατοπίστηκε στον τομέα της εμπειρίας χρήστη, μία δύσκολη γενικά πρόκληση. Η μέτρηση της εμπειρίας χρήστη θα αποτελούσε μελέτη διαφορετικού είδους. Το παραπάνω όμως απαιτεί χρήστες μακράς διαρκείας, οι οποίοι θα είναι πρόθυμοι να αξιολογήσουν το σύστημα - ο μόνος αξιόπιστος τρόπος δηλαδή μέτρησης συμπεριφορών σε πραγματική χρήση. Η έρευνα προς αυτή την κατεύθυνση διακρίνεται σε τρεις κατηγορίες:

- ανάπτυξη συστημάτων αποκλειστικά για πειραματική χρήση. Παραδείγματα σχετικών μελετών αποτελούν το [48], το TechLens το οποίο αξιοποιήθηκε από πολλές έρευνες ([111],[65], κ.α.)

- συνεργασία με χειριστές live συστημάτων για την εκτέλεση πειραμάτων πάνω σε συστήματα προτάσεων, όπως π.χ. με το BookCrossing.com στο [233] και την Wikipedia στο [55].

- ανάπτυξη και υποστήριξη ερευνητικών συστημάτων και κοινότητες χρηστών. Χαρακτηριστικός αντιπρόσωπος αποτελεί το ερευνητικό project GroupLens [181]

Τα παραπάνω δεν αποτελούν παρά μια σύνοψη της ιστορίας των συστημάτων προτάσεων. Για περισσότερες πληροφορίες σχετικά με το θέμα, παραπέμπουμε τον αναγνώστη στα [120] και [64], πηγές εξαιρετικά χρήσιμες και επίκαιρες.

3.5 Προεπεξεργασία κειμένου

Το να κρατήσουμε μία αναπαράσταση των κειμένων η οποία περιλαμβάνει κάθε keyoword (ή n-gram), είναι κάτι το απαγορευτικό για ένα πραγματικό σύστημα που αξιοποιεί την κειμενική πληροφορία. Ο λόγος είναι απλός και έχει να κάνει με την κλιμάκωση του χρόνου και χώρου υπολογισμού σε αυτή την περίπτωση. Αντιθέτως, είναι απαραίτητη μία διαδικασία προεπεξεργασίας κειμένου η οποία θα καταλήγει στον εντοπισμό των σημαντικών οντοτήτων αυτού, είτε αυτά είναι keywords, είτε n-grams, είτε κάποια άλλη (συνήθως στατιστική) πληροφορία.

Υπάρχει μία πληθώρα προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία σε ότι έχει να κάνει με την προεπεξεργασία κειμένου. Οι πιο γνωστές τεχνικές είναι τα Hidden Markov Models [53], η Naive Bayes [160] και τα Support Vector Machines [115]. Πέρα από τις παραπάνω τεχνικές μοντελοποίησης των δεδομένων, μία συχνά χρησιμοποιούμενη τεχνική, και δει αυτή που αξιοποιούμε και στα πλαίσια της διδακτορικής διατριβής, είναι η tf-idf (term frequency - inverse document frequency) [109]. Η μετρική αυτή είναι μία στατιστική μετρική η οποία στοχεύει να αναπαραστήσει πόσο σημαντικό είναι ένα keyword σε μία συλλογή. Αυξάνει δε αναλογικά σε σχέση με το πλήθος που εμφανίζεται το keyword στο κείμενο σε σύγκριση με την συχνότητα εμφάνισής του στη συνολική βάση δεδομένων. Η λογική πίσω από αυτή την αντιμετώπιση είναι σχετικά απλή: ενδιαφερόμαστε για κειμενικές 'μονάδες' (π.χ. keywords) τα οποία είναι συχνά στο κείμενο αλλά όμως δεν είναι το ίδιο συχνά σε μεγάλο μέρος των κειμένων της συλλογής. Άλλες τεχνικές, οι οποίες επίσης προταθεί στη βιβλιογραφία είναι το κέρδος πληροφορίας [224], odds ratio [146], κ.λπ.

3.5.1 Εξαγωγή λέξεων κλειδιών

Η αυτοματοποιημένη εξαγωγή λέξεων κλειδιών αποσκοπεί στον εντοπισμό ενός μικρού συνόλου λέξεων, φράσεων-κλειδιών ή πιο συγκεκριμένα, keywords από ένα κείμενο, τα οποία θα μπορούν να περιγράψουν το νόημα του κειμένου [97]. Θα πρέπει να γίνεται με συστηματικό τρόπο, είτε με ελάχιστη ή καθόλου ανθρώπινη παρεμβολή, ανάλογα το μοντέλο. Ο σκοπός της εξαγωγής λέξεων κλειδιών είναι η αναπαράσταση του κειμένου κατά τρόπο σύντομο, συγκεκριμένο και αποτελεσματικό με την μικρότερη δυνατή απώλεια νοηματικής πληροφορίας. Τα μοντέλα προεπεξεργασίας κειμένου που αναφέρθηκαν στην προηγούμενη ενότητα έχουν στον πυρήνα τους την διεργασία εξαγωγής λέξεων κλειδιών του κειμένου.

3.5.2 Εξαγωγή n-grams

Ένα n-gram αποτελεί την κειμενική ακολουθία μήκους n που βρίσκεται σε ένα κείμενο. Στην εργασία μας ασχολούμαστε με τα n-grams λέξεων (word n-grams) τα οποία μπορούν να ιδωθούν υπό την αναλογία τοποθέτησης ενός μικρού μεταβαλλόμενου παραθύρου πάνω από μία πρόταση του κειμένου, στο οποίο μόνο n λέξεις είναι "ορατές" κάθε στιγμή. Σε κάθε θέση του παραθύρου, η ακολουθία λέξεων μέσα του καταγράφεται. Σε ορισμένες περιπτώσεις, το παράθυρο μπορεί να μετακινείται περισσότερο από μία λέξη αφού κάθε n-gram έχει καταγραφεί. Η απλούστερη μορφή n-gram είναι το unigram, όπου $n = 1$, η οποία ανάγεται στην BOW αναπαράσταση των keywords του κειμένου. Τυπικά το n είναι ένας σταθερός αριθμός, υψηλά εξαρτώμενος από το συγκεκριμένο σύνολο δεδομένων (π.χ. τη γλώσσα, τον τομέα, κ.λπ.) καθώς και τα ερωτήματα προς αυτό.

Καθένα από τα n-grams είναι ένα σύνολο συντεταγμένων που αναπαριστά το κείμενο που μελετάται, και η συχνότητα εμφάνισης του n-gram μπορεί να είναι το βάρος του n-gram. Μπορούμε επομένως να χρησιμοποιήσουμε αυτή την αναπαράσταση σε εφαρμογές όπως η συμπίεση κειμένου, καθώς και πλήθος άλλων εφαρμογών στον τομέα του IR συμπεριλαμβανομένης και της συσταδοποίησης αντικειμένων όπως στην περίπτωση της διδακτορικής διατριβής.

Η χρήση της πιθανότητας κατανομής των n-grams και των n-grams μοντέλων στο NLP είναι

μία σχετικά απλή ιδέα, η οποία όμως έχει βρει τεράστια απήχηση. Για παράδειγμα μοντέλα n-grams σε επίπεδο χαρακτήρων κειμένου μπορούν να εφαρμοστούν σε κάθε γλώσσα, ή ακόμη και σε μη γλωσσικές ακολουθίες, όπως ακολουθίες DNA και μουσικής. Έχουν επίσης χρησιμοποιηθεί στην συμπίεση κειμένου, π.χ. το PPM μοντέλο [28], και έχουν επίσης αποδειχθεί αποτελεσματικά σε προβλήματα εξόρυξης δεδομένων [218]. Στον τομέα της κατηγοριοποίησης κειμένου, ανεξαρτήτου γλώσσας, n-grams μοντέλα σε επίπεδο λέξεων έχουν χρησιμοποιηθεί για την Αγγλική και Γερμανική γλώσσα με καλά αποτελέσματα [16]. Η ανάλυση των n-grams έχει επίσης αποδειχθεί μεγάλης σημασίας για πολλές περιοχές της φυσικής επεξεργασίας γλώσσας και εξόρυξης κειμένου, όπως το διαπέρασμα (parsing) κειμένου και IR εφαρμογές. Ορισμένα παραδείγματα συμπεριλαμβάνουν:

- αναζήτηση και κατηγοριοποίηση παρόμοιων κειμένων, όπως στο [152], όπου οι συγγραφείς παρουσιάζουν μία προσέγγιση n-grams χαρακτήρων για την περίπτωση της κατηγοριοποίησης κειμένων
- εντοπισμός επαναχρησιμοποιημένου, διπλότυπου ή κειμένου λογοκλοπής (plagiarized text) [24]
- εντοπισμός επιβλαβούς (malicious) κώδικα [3]
- πλήθος γλωσσολογικών διαδικασιών, όπως αναγνώριση γλώσσας [138]

Η διαίσθηση πίσω από τις προαναφερθείσες προσεγγίσεις είναι κοινή: οι φράσεις, ως σύνολο, μάλλον κουβαλάνε περισσότερη πληροφορία σε σχέση με το άθροισμα των αυτόνομων συστατικών τους. Έτσι, η εξαγωγή τους, μπορεί να οδηγήσει σε αποτελεσματικότερη κειμενική αναπαράσταση άρα και αποτελέσματα.

Ένα ακόμη θέμα που έχει να κάνει με την ανάλυση των n-grams και το οποίο θα πρέπει να αναφερθεί, είναι ότι τα εντελώς σπάνια εμφανιζόμενα n-grams είναι κατά κανόνα μη ενδιαφέροντα και έτσι χρειάζεται μόνο να μετράμε τα n-grams που εμφανίζονται στο σύνολο δεδομένων μας με συχνότητα από κάποιο όριο και πάνω. Δεν θα πρέπει η παραπάνω κατηγορία όμως να συγχέεται με τα μη συχνά n-grams, τα οποία και αποτελούν πιθανότητα σημαντικά (αντίστοιχη ζύγιση tf-idf).

Τέλος, ο καθορισμός της τιμής του n, δηλαδή του μεγέθους του μήκους παραθύρου που χρησιμοποιείται, όταν αναφερόμαστε σε n-grams λέξεων, είναι μια περιοχή πειραματισμού για την συγκεκριμένη περιοχή γνώσης των κειμένων. Για παράδειγμα, στο τομέα του εντοπισμού κειμένου λογοκλοπής, οι συγγραφείς του [24] εξηγούν ότι χαμηλές τιμές για το n φαίνεται να οδηγούν στα καλύτερα αποτελέσματα για συγκεκριμένες τιμές ακρίβειας-ανάκλησης. Τιμές πάνω από 4, μάλλον έχουν αρνητική επίπτωση στην αποτελεσματικότητα της προσέγγισης. Παρόμοιο αποτέλεσμα δίνεται και στο [73], όπου οι συγγραφείς καταλήγουν στο συμπέρασμα πως οι ακολουθίες λέξεων μεγέθους 2 ή 3 είναι πολύ πιο χρήσιμες σε σχέση με μεγαλύτερες ακολουθίες οι οποίες και μειώνουν την απόδοση της κατηγοριοποίησης.

Όσον αφορά τον τομέα της συσταδοποίησης, την επίδραση στον οποίο η χρήση των n-grams λέξεων μελετάται στην διδακτορική διατριβή, δεν βρήκαμε κάποια σχετική έρευνα στην βιβλιογραφία.

3.6 Ταξινόμηση κειμένων

Δεδομένου ενός συνόλου πινάκων κειμένων $\{d_1, d_2, ..., d_n\}$ και των συσχετιζόμενων με αυτά ετικετών $c(d_i) \in \{c_1, c_2, ..., c_l\}$, η διαδικασία της ταξινόμησης αφορά στον καθορισμό της σωστής ετικέτας του νέου κειμένου d. Η ταξινόμηση κειμένων (text classification) έχει μελετηθεί σε μεγάλο βαθμό, ιδιαίτερα ύστερα από την εμφάνιση του διαδικτύου. Οι περισσότεροι αλγόριθμοι βασίζονται στο μοντέλο "συνόλου λέξεων" του κειμένου [189]. Ένας απλός και συνάμα αποτελεσματικός αλγόριθμος είναι αυτός του Naive Bayes [144]. Για το πρόβλημα της ταξινόμησης κειμένων, διάφορες παραλλαγές του Naive Bayes έχουν χρησιμοποιηθεί αλλά έχει βρεθεί [139] ότι η παραλλαγή που βασίζεται στο πολυωνυμικό μοντέλο οδηγεί σε καλύτερα αποτελέσματα.

Η μέθοδος των Support Vector Machine (SVM) έχει επίσης χρησιμοποιηθεί επίσης με καλά αποτελέσματα [105][43]. Για ιεραρχικά δεδομένα κειμένων, όπως οι ιεραρχίες θεμάτων του Yahoo! [223] και το Open Directory Project [164], έχει μελετηθεί στα [119][45][61].

Για να αποφευχθούν οι πολλές διαστάσεις στην αναπαράσταση των κειμένων, πολλές μέθοδοι επιλογής χαρακτηριστικών έχουν προταθεί [224][119][45]. Επίσης συχνά επιζητείται η ιδιότητα της "ισχυρής" ταξινόμησης όπου η κάθε λέξη του κειμένου μπορεί να αντιπροσωπευθεί από τη μοναδική ομάδα που ανήκει. Τέτοια ιδιότητα αξιοποιείται στα [139][198]. Η επιλογή του μεγίστου πλήθους των λέξεων που θα απαρτίζουν ένα cluster είναι επίσης κάτι σημαντικό [216][185].

3.7 Συσταδοποίηση κειμένων

Η συσταδοποίηση δεδομένων γενικά έχει μελετηθεί σε βάθος στην υπάρχουσα βιβλιογραφία τα τελευταία 20 χρόνια. Η εξερεύνηση αυτής τις βιβλιογραφίας περιπλέκεται από το γεγονός ότι υπάρχουν πολλά πεδία γνώσης πάνω στα οποία η συσταδοποίηση μπορεί να εφαρμοστεί. Ειδικά για την περίπτωση της συσταδοποίησης κειμένων, μία τεράστια ποικιλία τεχνικών έχει προταθεί.

Σε αυτή στην ενότητα δεν θα προσπαθήσουμε να παρουσιάσουμε διεξοδικά όλους τους διαθέσιμους αλγορίθμους, αντίθετα θα ασχοληθούμε περισσότερο με τις γενικότερες κατηγορίες αυτών των αλγορίθμων καθώς και τους κυριότερους αντιπροσώπους αυτών. Παρότι έχουμε προσπαθήσει να επιλέξουμε προσεκτικά τους καλύτερους αντιπροσώπους κάθε ομάδας, υπάρχουν αναμφισβήτητα αλγόριθμοι οι οποίοι δεν αναφέρονται καθώς και πιθανά περισσότερες κατηγορίες αλγορίθμων.

Ένας βασικός στόχος της συσταδοποίησης κειμένων είναι η βελτίωση των αποτελεσμάτων των συστημάτων ανάκτησης πληροφορίας σε σχέση με τις μετρικές αυτών. Αυτό στη συνέχεια οδηγεί σε εξυπηρέτηση καλύτερων αποτελεσμάτων και φιλτραρισμένης πληροφορίας προς τους χρήστες διευκολύνοντας έτσι την διαδικασία λήψης αποφάσεων.

Οι αλγόριθμοι συσταδοποίησης έχουν αξιολογηθεί κατά καιρούς στην βιβλιογραφία με πολλούς τρόπους. Δυστυχώς όμως δεν υπάρχει ένας de-facto προ-συμφωνημένος τρόπος για αυτή τη διαδικασία. Επίσης, η επιλογή των μεθόδων αξιολόγησης συχνά εξαρτάται από το πεδίο γνώσης πάνω στο οποίο η έρευνα εφαρμόζεται. Για παράδειγμα στο πεδίο του AI, μπορεί να προτιμάται η αμοιβαία πληροφορία, ενώ στο πεδίο του IR προτιμάται η μετρική του F-measure.

Η συσταδοποίηση έχει επίσης αξιοποιηθεί και για το πεδίο της μηχανικής εκμάθησης (ML)

[163] όπως για εξόρυξη χρονοσειρών (time series clustering) [187] όποιο αξιοποιούνται συχνές λίστες αντικειμένων (κειμένων) προκειμένου να εντοπισθούν κανόνες συσχέτισης σε μεγάλες transactional databases.

Στα παρακάτω θα επιχειρήσουμε μια γενική κατηγοριοποίηση των τεχνικών συσταδοποίησης της βιλιογραφίας επιμένοντας λίγο παραπάνω στις τεχνικές που έχουν ιδιαίτερο ενδιαφέρον για την περίπτωση κειμενικής πληροφορίας (όπως τα άρθρα νέων).

3.7.1 Αλγόριθμοι συσταδοποίησης

Παραδοσιακά, οι ποικίλοι αλγόριθμοι συσταδοποίησης κατατάσσονται σε δύο γενικές κατηγορίες: ιεραρχικοί (agglomerative hierarchical) και μερισματικοί (partitional).

Οι τυπικοί ιεραρχικοί αλγόριθμοι συσταδοποίησης [86] παράγουν ένα σύνολο από διαμερίσματα πάνω στα δεδομένα, τα οποία μπορούν να ποικίλουν από μία συστάδα η οποία περιέχει όλα τα αντικείμενα, μέχρι και n συστάδες καθεμία από τις οποίες περιέχει ένα αντικείμενο, και τα οποία μπορούν να αναπαρασταθούν γραφικά ως ένα διαιρετικό (από την ρίζα προς τα φύλλα) ή συνδυαστικό (από τα φύλλα προς τη ρίζα) δέντρο. Από την άλλη μεριά, οι μερισματικοί αλγόριθμοι συσταδοποίησης τυπικά καθορίζουν όλες τις συστάδες μονομιάς, αλλά μπορούν να χρησιμοποιηθούν και ως διαμερισματικοί αλγόριθμοι στην περίπτωση της ιεραρχικής συσταδοποίησης (σε συνδυασμό των δύο μεθοδολογιών).

3.7.1.1 Ιεραρχικοί αλγόριθμοι

Η ιεραρχική συσταδοποίηση, συχνά αναφερόμενη και ως ανάλυση συστάδων ή Hierarchical Clustering Analysis (HCA) είναι μεθοδολογία η οποία αναζητεί την κατασκευή μίας ιεραρχίας συστάδων δεδομένων των δεδομένων προς συσταδοποίηση. Οι στρατηγικές για ιεραρχική συσταδοποίηση γενικά ταξινομούνται σε δύο κατηγορίες:

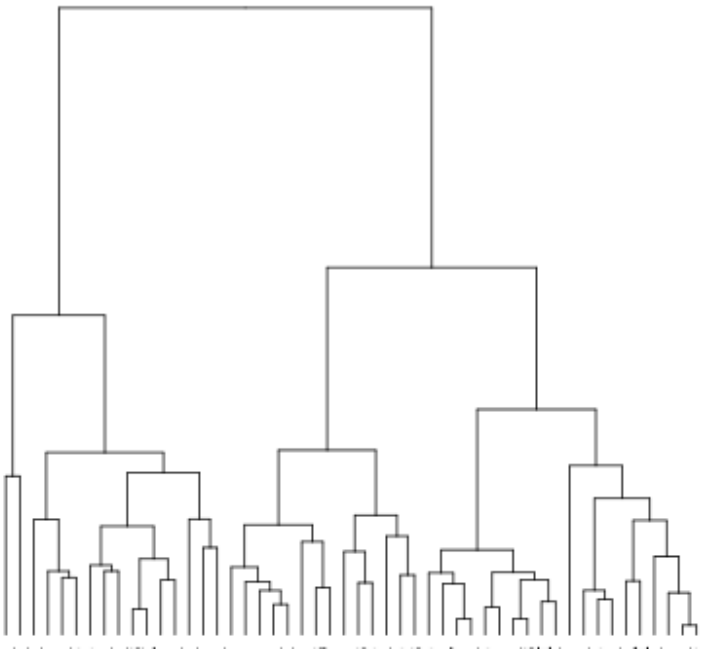
- Συνδυαστικές (Agglomerative): πρόκειται για μία "από κάτω προς τα πάνω" προσέγγιση όπου το κάθε αντικείμενο ξεκινάει ως μία συστάδα μόνο του και στη συνέχεια ζεύγη από συστάδες συνενώνονται συνεχώς όσο προχωράμε προς τα πάνω στην ιεραρχία. Οι agglomerative ιεραρχικοί αλγόριθμοι επομένως ξεκινούν θεωρώντας κάθε αντικείμενο ως μία συστάδα από μόνο του και συνδυάζοντας συστάδες μαζί παράγουν τους κόμβους του δέντρου οι οποίοι μοιράζονται ορισμένη ομοιότητα.

- Διαχωριστικές (Divisive): πρόκειται για μία "από πάνω προς τα κάτω' προσέγγιση όπου όλα τα αντικείμενα ξεκινούν ως μία συστάδα και στη συνέχεια οι συστάδες διαχωρίζονται αναδρομικά καθώς κατεβαίνουμε την ιεραρχία. Οι divisive ιεραρχικές μέθοδοι παράγουν επομένως μία εμφωλευμένη ακολουθία από διαμερίσεις των αντικειμένων με μία, όλα συμπεριλαμβανόμενη συστάδα στην κορυφή και μοναδιαίες συστάδες (singleton) με ατομικά αντικείμενα στη βάση [199]

Με τις παραπάνω έννοιες, οι ιεραρχικές τεχνικές απαιτούν έναν ορισμό ομοιότητας μεταξύ των συστάδων, ή αλλιώς μία μετρική απόστασης, προκειμένου σε διαδοχικά βήματα να μπορέσουν να

διαχωρίσουν ή να ενώσουν τις συστάδες. Είναι σύνηθες αυτή η μετρική να είναι ένας πίνακας ομοιοτήτων (αποστάσεων), το στοιχείο i, j του οποίου εκφράζει την απόσταση μεταξύ της i και j συστάδας. Αυτός ο πίνακας ανανεώνεται σε κάθε βήμα, όπου μετέπειτα κόμβοι δημιουργούνται με την ένωση τους σε ζεύγη (για agglomerative) ή διαχωρισμό (για divisive) έως ότου η διαδικασία ολοκληρωθεί.

Το αποτέλεσμα των παραπάνω αλγορίθμων είναι μία δεντρική δομή, ή αλλιώς δενδρόγραμμα (dendogram), το οποίο αποτυπώνει την διαδικασία συνένωσης (ή διαχωρισμού) των συστάδων κατά την διαδικασία της ιεραρχικής συσταδοποίησης. Οι ενδιάμεσες συστάδες που προκύπτουν στην την πορεία, μπορούν να συλλεχθούν "κόβοντας" το δέντρο σε επιθυμητό επίπεδο ακρίβειας. Ένα τυπικό παράδειγμα δενδρογράμματος φαίνεται στο σχήμα 4 με τις τομές να μπορούν να γίνουν σε οποιοδήποτε βάθος της ιεραρχίας κρατώντας τις επιθυμητές συστάδες.



Σχήμα 4: Τυπικό δενδρόγραμμα ιεραρχικής συσταδοποίησης

3.7.1.1.1 Τυπικές ιεραρχικές μέθοδοι συσταδοποίησης

Υπάρχουν πολλές διαφορετικές μέθοδοι ιεραρχικής συσταδοποίησης τις οποίες και αξιολογούμε στην διδακτορική διατριβή. Η διαφορά τους έγκειται στο πως ορίζεται η απόσταση μεταξύ των συστάδων σε σχέση με τα μέλη αυτών (άρθρα νέων). Οι μέθοδοι αυτοί και ο τρόπος ορισμού της απόστασης είναι οι εξής:

- pairwise single linkage, όπου η κοντινότερη απόσταση μεταξύ των όρων δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)

- pairwise maximum linkage, όπου η μακρινότερη απόσταση μεταξύ των όρων δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)
- pairwise average linkage, όπου η μέσος όρος όλων των αποστάσεων μεταξύ των όρων δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)
- centroid linkage, όπου κάθε συστάδα αναπαρίσταται από το κέντρο της, το οποίο και υπολογίζεται σε κάθε βήμα του αλγορίθμου. Η δια-συσταδική απόσταση (ομοιότητα) σε αυτή την περίπτωση είναι η απόσταση μεταξύ των κέντρων των συστάδων

Κάθε μία από τις προαναφερθείσες μεθοδολογίες ιεραρχικής συσταδοποίησης αξιολογήθηκε στα πλαίσια της διδακτορικής διατριβής και τα αποτελέσματα παρουσιάζονται στο κεφάλαιο 7.

3.7.1.1.2 Πολυπλοκότητα

Η προηγούμενη διαδικασία είναι ντετερμινιστική, παράγοντας κάθε φορά το ίδιο δενδρόγραμμα, επομένως και το ίδιο αποτέλεσμα συσταδοποίησης, κάτι που δεν ισχύει για τους μερισματικούς αλγορίθμους συσταδοποίησης που περιγράφονται στη συνέχεια. Παρόλα αυτά, όπως εξηγείται από τους Day και Edelsbrunner [58], οι σειριακοί agglomerative μη επικαλυπτόμενοι ιεραρχικοί αλγόριθμοι συσταδοποίησης (Sequential Agglomerative Hierarchical Non-overlapping (SAHN)) έχουν μέση πολυπλοκότητα $O(n^2)$ και πιο συχνά $O(n^3)$ ως προς το μέγεθος εισόδου (πλήθος αντικειμένων) n. Το παραπάνω στις περισσότερες περιπτώσεις είναι αποτρεπτικό για χρήση με πολλά αντικείμενα μιας και ο χρόνος εκτέλεσης κλιμακώνεται πολύ γρήγορα για πραγματικές εφαρμογές.

3.7.1.2 Μερισματικοί αλγόριθμοι

Στους μερισματικούς αλγορίθμους συσταδοποίησης χρησιμοποιείται ένα καθολικό κριτήριο, η βελτιστοποίηση του οποίου καθοδηγεί και την συνολική διαδικασία, παράγοντας επομένως έναν διαμερισμό των δεδομένων. Δοθέντος του πλήθους των επιθυμητών συστάδων, έστω k, οι μερισματικοί αλγόριθμοι βρίσκουν και τις k συστάδες μονομιάς, έτσι ώστε το άθροισμα των αποστάσεων όλων των στοιχείων από τις συστάδες τους να είναι ελάχιστο. Επιπλέον, για ένα αποτέλεσμα συσταδοποίησης να είναι ακριβές, εκτός από την χαμηλή εσω-συσταδική απόσταση, η υψηλή εξω-συσταδική απόσταση είναι επίσης επιθυμητή. Προκειμένου επομένως ένας αλγόριθμος συσταδοποίησης να είναι αποτελεσματικός, θα πρέπει να ικανοποιούνται όσο το δυνατόν καλύτερα οι δύο ακόλουθες συνθήκες:

- μικρή εσω-συσταδική απόσταση: τα μέλη της ίδιας συστάδας να είναι στενά συνδεδεμένα μεταξύ τους
- μεγάλη εξω-συσταδική απόσταση: τα μέλη διαφορετικών συστάδων να απέχουν αρκετά μεταξύ τους ώστε οι συστάδες να είναι καλά διακριτές

Μερικοί κλασικοί μερισματικοί αλγόριθμοι είναι οι: k-means, k-medians, και k-medoids. Οι αλγόριθμοι αυτοί βασίζονται στην λογική του κέντρου συστάδας (cluster center), ένα σημείο

δηλαδή στο χώρο των δεδομένων, συχνά μη φυσικά υπαρκτό μέσα στα ίδια τα δεδομένα, το οποίο αντιπροσωπεύει τη συστάδα. Η διαφορά των παραπάνω έγκειται στο πως το κέντρο συστάδας ορίζεται. Στα παρακάτω θα περιγράψουμε σύντομα καθεμία από τις πιο συνηθισμένες προσεγγίσεις μερισματικών αλγορίθμων, καθώς και παραλλαγές αυτών.

3.7.1.3 Οικογένεια k-means

Οι αλγόριθμοι της οικογένειας συσταδοποίησης k-means [89] στοχεύουν στον διαμερισμών n αντικειμένων σε k συστάδες όπου κάθε αντικείμενο ανήκει στην συστάδα με τον κοντινότερο μέσο (κέντρο της συστάδας). Το πρόβλημα της συσταδοποίησης είναι υπολογιστικά NP-hard [135][215], παρόλα αυτά υπάρχει πληθώρα αποδοτικών ευρετικών παραλλαγών που συχνά εφαρμόζονται και οδηγούν σχετικά γρήγορα σε τοπικό βέλτιστο.

Οι αλγόριθμοι της οικογένειας k-means χρησιμοποιούν τα κέντρα των συστάδων για να μοντελοποιήσουν τα δεδομένα που ανήκουν σε αυτές. Το κέντρο συστάδας ορίζεται ως το μέσο διάνυσμα δεδομένων βάσει του μέσου όρου όλων των στοιχείων της συστάδας. Στον αλγόριθμο k-medians, αντί για τον μέσο όρο, ο διάμεσος υπολογίζεται για κάθε διάσταση του διανύσματος δεδομένων. Παρόμοια, στον αλγόριθμο k-medoids το κέντρο συστάδας ορίζεται ως το αντικείμενο εκείνο το οποίο έχει το μικρότερο άθροισμα αποστάσεων από τα υπόλοιπα στοιχεία της συστάδας, πρόκειται επομένως για πραγματικό αντικείμενο στα δεδομένα. Ο k-medoids έχει το πλεονέκτημα της καλύτερης διαχείρισης των ακραίων τιμών (outliers) στα δεδομένα, ενώ παράλληλα δεν εξαρτάται από την σειρά με την οποία τα στοιχεία εξετάζονται.

Η οικογένεια των k-means αλγορίθμων [230] συχνά επιχειρεί να ελαχιστοποιήσει μία δεδομένη μετρική ομοιότητας, κατά κανόνα την Ευκλείδεια απόσταση, μεταξύ των στοιχείων της ίδιας συστάδας. Ένας πιο αυστηρός ορισμός είναι ο παρακάτω:

Ορισμός 3.7.1. Αν $d_1, d_2, ..., d_n$ είναι τα n κείμενα και $c_1, c_2, ..., c_k$ είναι τα k κέντρα συστάδων, ο αλγόριθμος k-means προσπαθεί να ελαχιστοποιήσει την καθολική συνάρτηση:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} sim(d_j, c_i)$$

Ένα μέτρο επομένως του πόσο καλά τα κέντρα των συστάδων αντιπροσωπεύουν τα αντικείμενα των συστάδων είναι υπολειπόμενο άθροισμα τετραγώνων ή αλλιώς Residual Sum of Squares (RSoS), η τετραγωνική απόσταση του κάθε αντικειμένου (που αναπαρίσταται φυσικά ως πίνακας στο πολυ-διάστατο χώρο των αντικειμένων) από το κέντρο του, αθροισμένη για όλα τα αντικείμενα:

$$RSoS_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2 \tag{10}$$

όπου $\vec{\mu}$ ο πίνακας αναπαράστασης του κέντρου της συστάδας ω. Άρα για όλες τις k συστάδες

μπορούμε αθροιστικά να υπολογίζουμε την καθολική συνάρτηση αξιολόγησης RSoS ως:

$$RSoS = \sum_{n=1}^{k} RSoS_k \tag{11}$$

3.7.1.3.1 Expectation Maximization

Ο αλγόριθμος EM [154] αποτελεί μία αποτελεσματική επαναληπτική διαδικασία για τον υπολογισμό μίας λύσης μέγιστης πιθανότητας (Maximum Likelihood (MaxL)) για το δεδομένο μοντέλο. Αποτελείται από δύο βήματα. Στο βήμα αναμονής (expectation step, E-step) τα ελλειπή δεδομένα υπολογίζονται βασιζόμενοι στα υπάρχοντα δεδομένα (τη συλλογή των κειμένων) καθώς και την τρέχουσα εκτίμηση του μοντέλου (για τις συστάδες). Στο βήμα μεγιστοποίησης (miximization step, M-step), η συνάρτηση πιθανότητας μεγιστοποιείται υπό την υπόθεση ότι τα ελλειπή δεδομένα είναι γνωστά. Για πιο πολλές πληροφορίες προτείνουμε το [30]. Μία επανάληψη του αλγορίθμου EM αποτελείται:

- από το βήμα αναμονής στο οποίο η πιθανότητα P υπολογίζεται για κάθε κείμενο δεδομένων των προβλέψεων για τις συστάδες ως:

$$P(\theta \mid d) = \frac{P(\theta)P(d \mid \theta)}{\sum_{\theta \in \Theta} P(d \mid \theta)} \tag{12}$$

$$P(\theta)^* = \sum_{d \in D} P(\theta \mid d) \tag{13}$$

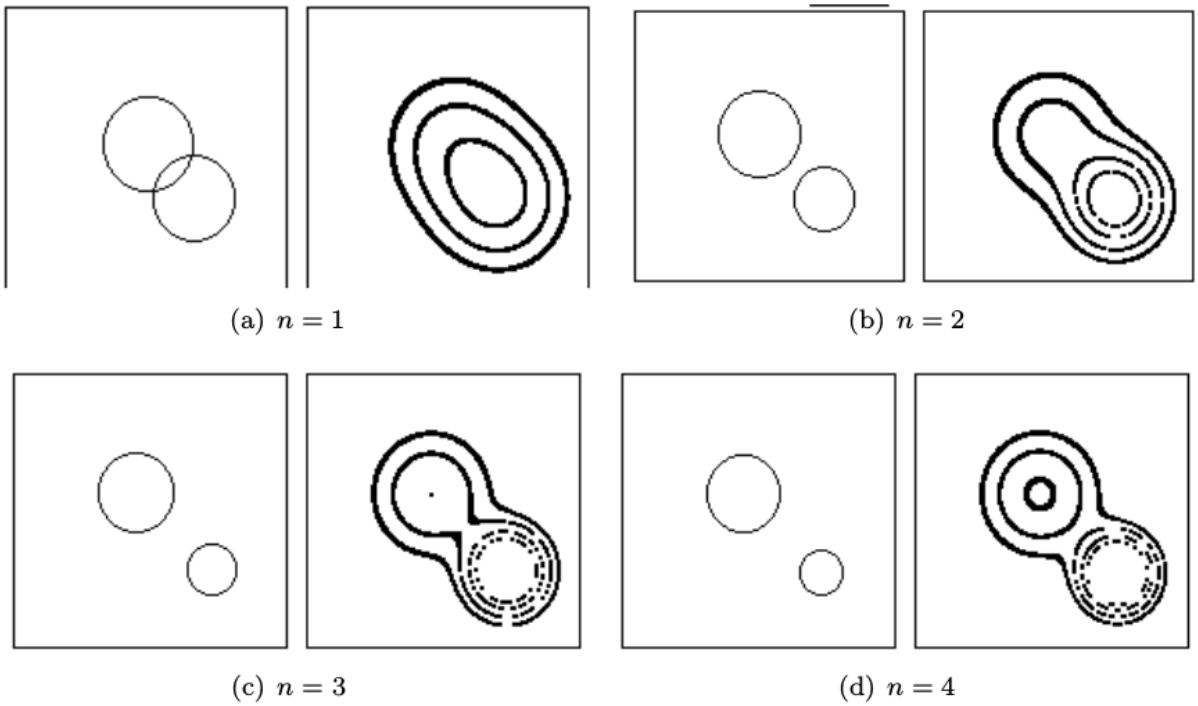
- από το βήμα μεγιστοποίησης, το οποίο ανανεώνει τις παραμέτρους του μοντέλου θ για μεγιστοποίηση της πιθανότητας δεδομένων των πιθανοτήτων που υπολογίστηκαν στο E-step:

$$\mu = \frac{\sum_{d \in D} P(\theta \mid d)d}{\sum_{d \in D} P(\theta \mid d)} \tag{14}$$

$$\mu = \Sigma = \frac{\sum_{d \in D} P(\theta \mid d)(d-\mu)(d-\mu)^T}{\sum_{d \in D} P(\theta \mid d)} \tag{15}$$

Έχει αποδειχθεί [140] ότι ο αλγόριθμος συγκλίνει σε τοπικό ελάχιστο με λογαριθμική πιθανότητα με το συνόλο των κειμένων D να παράγεται από το μοντέλο Θ ως συνθήκη τερματισμού. Το σχήμα 5 δείχνει τέσσερις επαναλήψεις του αλγορίθμου EM. Μία βοηθητική προσέγγιση ως προς την κατανόηση του αλγορίθμου είναι ως προς την εκτίμηση κάτω ορίου: σε κάθε επανάληψη, ένα πιο στενό κατώτερο όριο υπολογίζεται και οι εκτιμώμενες συστάδες "σκαρφαλώνουν" προς την άγνωστη τελική κατανομή.

Όπως σε κάθε περίπτωση μέγιστης πιθανότητας, το να υπάρχουν πολλές ελεύθερες μεταβλητές με ελλειπή δεδομένα μπορεί να οδηγήσει σε προβλήματα (π.χ. overfitting, μεγάλος χρόνος εκτέλεσης, κ.λπ.). Στο [131], αυτό το πρόβλημα αντιμετωπίζεται με χρήση SVD στο χώρο τον κειμένων.



Σχήμα 5: Ο αλγόριθμος EM σε τέσσερις επαναλήψεις του

Στη συνέχεια επιλέγονται ορισμένες διαστάσεις οι οποίες έχουν τις περισσότερες μοναδικές τιμές για τον σχηματισμό ενός μειωμένου χώρου πάνω στον οποίο διενεργείται η συσταδοποίηση.

Ένα πρόβλημα του τυπικού EM αλγόριθμου είναι ότι είναι τετραγωνικός ως προς τον αριθμό των συστάδων k, η αλλιώς $O(k^2n)$, δεδομένου ότι οι πιθανότητες επανυπολογίζονται για κάθε συστάδα. Για την περίπτωση της οικογένειας αλγορίθμων k-means όμως, μία πιο περιορισμένη (κομματιασμένη) έκδοση του EM αλγόριθμου είναι ο model-based k-means. Αυτός ο αλγόριθμος, μεταβαίνει μεταξύ του βήματος επανυπολογισμού του μοντέλου και του βήματος επανανάθεσης έχοντας ως αποτέλεσμα γραμμική πολυπλοκότητα. Επίσης, παρά τις σημαντικά θετικές ιδιότητές τους, ο αλγόριθμος αυτός δεν αποδίδει χειρότερα από τον πλήρη EM αλγόριθμο [231]. Η σημαντική διαφοροποίηση μεταξύ του κλασικού EM αλγορίθμου και της k-means παραλλαγής του είναι ότι ο δεύτερος, δεν επανεκπεδεύει το μοντέλο βασισμένος στην εκ των υστέρων πιθανότητα.

Τυπικά λοιπόν, όλοι οι αλγόριθμοι της οικογένειας k-means μοιράζονται τα EM βήματα που δίνονται στον αλγόριθμο 1 [19]. Ως αποτέλεσμα αυτού παράγεται ένας διαμοιρασμός των αντικειμέ-

νων σε ομάδες από τις οποίες η μετρική που θέλουμε να ελαχιστοποιείται μπορεί και υπολογίζεται.

Αλγόριθμος 1: Model-based k-means EM αλγόριθμος (τυπικός k-means αλγόριθμος)

Είσοδος: αντικείμενα προς συσταδοποίηση, k

1 Τυχαία επέλεξε k σημεία στον χώρο που αναπαρίσταται από τα αντικείμενα προς συσταδοποίηση (αυτά τα σημεία είναι τα αρχικά κέντρα των συστάδων)

2 Ανάθεσε κάθε αντικείμενο στην ομάδα που έχει το κοντινότερο κέντρο

3 Όταν όλα τα αντικείμενα έχουν ανατεθεί, επανυπολόγησε τις θέσεις των k κέντρων

4 Επανέλαβε τα βήματα 2 και 3 έως ότου δεν αλλάζουν οι αναθέσεις των κέντρων

Παρότι μπορεί να αποδειχθεί ότι η παραπάνω διαδικασία πάντα τερματίζει, ο αλγόριθμος EM δεν βρίσκει απαραίτητα και την βέλτιστη ανάθεση σε συστάδες. Επίσης ο αλγόριθμος EM συχνά πάσχει από σύγκλιση σε τοπικά ελάχιστα (ή μέγιστα) δεδομένης της τυχαιότητας της αρχικής επιλογής των κέντρων των συστάδων. Ο υπολογισμός επομένως μίας εξεζητημένης αρχικής συνθήκης μπορεί να επιφέρει σημαντικές βελτιώσεις όπως αποδείχθηκε στο [39]. Παραδείγματος χάριν, ο αλγόριθμος k-means++ [18], αφού επιλέξει τυχαία το πρώτο κέντρο συστάδας από τα δεδομένα, στη συνέχεια επιλέγει κάθε σημείο ως αρχικό κέντρο συστάδας χρησιμοποιώντας μία πιθανότητα η οποία είναι ανάλογη με το τετράγωνο της απόστασης μεταξύ κάθε διαδοχικής επιλογής κέντρου και της προηγούμενης. Τέλος προχωράει με τα βήματα του κλασικού k-means για να καταλήξει στις συστάδες. Αυτό το ευρετικό προσφέρει μία σημαντική ώθηση σε σύγκριση με τον τυπικό k-means όσον αφορά στο εύρος σφάλματος καθώς και στον χρόνο εκτέλεσης.

Μία ακόμη προσέγγιση είναι η χρήση πολλαπλών εκτελέσεων του αλγορίθμου k-means, με διαφορετικές αρχικές συνθήκες, και τελικά σύγκριση των αποτελεσμάτων ώστε να κρατηθεί μόνο το καλύτερο. Εάν μία συγκεκριμένη ανάθεση συστάδων εμφανίζεται να επαναλαμβάνεται, παρά τις διαφορετικές αρχικές συνθήκες, αυτό αποτελεί την καλύτερη ένδειξη ότι η συσταδοποίηση μάλλον είναι η βέλτιστη.

Ο bisecting k-means αλγόριθμος [126] εισάγει μία εναλλακτική προσέγγιση: αρχικά όλα τα δεδομένα αντιμετωπίζονται ως μία συστάδα. Μία συστάδα επιλέγεται για διαμερισμό σε δύο σε κάθε βήμα του αλγορίθμου χρησιμοποιώντας ένα κριτήριο, όπως το μέγεθος της συστάδας, ή η συνολική ομοιότητα. Ο διαμερισμός της επιλεγμένης συστάδας γίνεται με χρήση του κλασικού k-means και η διαδικασία ολοκληρώνεται όταν ο επιθυμητός αριθμός συστάδων έχει δημιουργηθεί. Κατά συνέπεια, σε αντίθεση με τον τυπικό k-means, ο οποίος διαχωρίζει τα συνολικά δεδομένα σε k συστάδες σε κάθε βήμα επανάληψης, η bisecting παραλλαγή του χωρίζει μόνο μία προ-υπάρχουσα συστάδα σε δύο υπο-συστάδες. Η επιλογή της συστάδας προς διαμερισμό μπορεί να βασίζεται στο μέγεθός της, ή στο δίκτυο γειτόνων του κέντρου της. Ενδιαφέρον αποτελεί ότι ο bisecting k-means αναφέρεται ως καλύτερος από άποψη απόδοσης σε σχέση με τον τυπικό k-means αλλά ακόμα και σε σχέση με ιεραρχικές προσεγγίσεις, ενώ παράλληλα κρατάει την πολυπλοκότητα γραμμική.

3.7.1.3.2 Spherical k-means

Ο κλασικός αλγόριθμος k-means χρησιμοποιεί την Ευκλείδεια απόσταση για τον καθορισμό της ομοιότητας μεταξύ των αντικειμένων καθώς και μεταξύ των συστάδων και των αντικειμένων.

Όμως αυτό το μέτρο απόστασης είναι συχνά αναποτελεσματικό για την συσταδοποίηση συλλογών κειμένων [200]. Ένα αποτελεσματικό μέτρο ομοιότητας μεταξύ κειμένων, και ένα που συχνά χρησιμοποιείται στον τομέα του IR είναι η ομοιότητα συνημιτόνου, η οποία χρησιμοποιεί το συνημίτονο της γωνίας μεταξύ πινάκων. Ο αλγόριθμος k-means μπορεί να προσαρμοστεί ώστε να χρησιμοποιεί το μέτρο ομοιότητας του συνημιτόνου, καταλήγοντας στον spherical k-means (S-kmeans) αλγόριθμο, ο οποίος ονομάζεται έτσι διότι δρα πάνω σε πίνακες οι οποίοι βρίσκονται πάνω στη μοναδιαία σφαίρα [60]. Δεδομένης της μετρικής του, ο (S-kmeans) εκμεταλλεύεται την αραιότητα των πινάκων των κειμένων και η εκτέλεσή του μπορεί να παραλληλοποιηθεί, κάτι που τον κάνει εξαιρετικά αποτελεσματικό [59], [123]. Τις ιδιότητες αυτές ακριβώς αξιοποιούμε στην διδακτορική διατριβή σε σχέση με τον προτεινόμενο W-kmeans αλγόριθμο.

3.7.1.3.3 Πολυπλοκότητα k-means

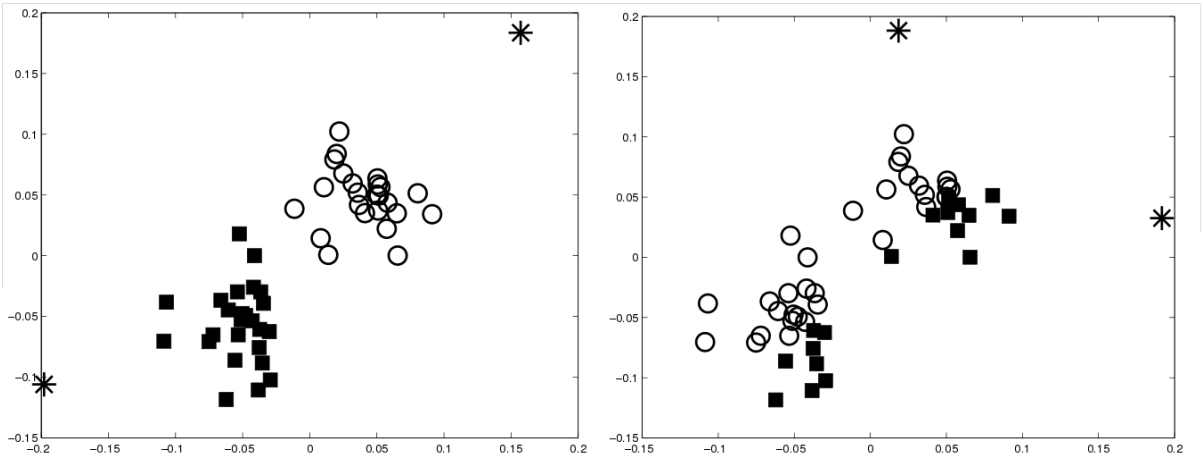
Παρότι το πρόβλημα της συσταδοποίησης είναι NP-hard στη γενική περίπτωσή του [9][57][135], η χαμηλή υπολογιστική πολυπλοκότητα είναι συνηθισμένη για όλους από τους προαναφερθέντες μερισματικούς αλγορίθμους. Ως αποτέλεσμα, αυτοί ταιριάζουν καλύτερα σε συσταδοποίηση μεγάλου όγκου δεδομένων, κάτι που μας ενδιαφέρει ιδιαίτερα και στην περίπτωσή μας (άρθρα νέων). Ειδικά για τον γενικό αλγόριθμο 1, η μέση πολυπλοκότητα είναι ουσιαστικά γραμμική, (nk) σε όλες τις σχετικές παραμέτρους: επαναλήψεις, πλήθος συστάδων καθώς και πλήθος κειμένων [19]. Παράλληλα, για την χειρότερη περίπτωση χρόνου εκτέλεσης, έχει υπολογιστεί από τους Arthur και Vassilvitskii [17] ως υπερ-πολυωνυμικός και συγκεκριμένα: $2^{\Omega\sqrt{n}}$.

3.7.1.3.4 Προβλήματα k-means

Παρότι ο k-means αλγόριθμος είναι διαισθητικά αποτελεσματικός σε αυτό που κάνει, παρουσιάζει ορισμένα μειονεκτήματα. Ένα από αυτά είναι ότι είναι εξαιρετικά ευαίσθητος στην αρχικοποίησή του, μιας και η επιλογή των αρχικών συστάδων παίζει μεγάλο ρόλο ως προς το αποτέλεσμα. Όπως εξηγείται στο [169] και φαίνεται στο σχήμα 6, δύο διαφορετικές αρχικοποιήσεις (με αστερίσκο στο σχήμα) μπορούν να οδηγήσουν σε σημαντικά διαφορετικά αποτελέσματα συσταδοποίησης.

Για την αντιμετώπιση του παραπάνω προβλήματος, ευρετικές μέθοδοι του k-means έχουν προταθεί στη βιβλιογραφία [117] [18] οι οποίες επιχειρούν να εντοπίσουν την καταλληλότερη αρχική ανάθεση. Η ευαισθησία αυτή στην αρχικοποίηση οφείλεται ουσιαστικά στο μη κυρτό πρόβλημα βελτιστοποίησης (non-convex optimization problem) στο οποίο ανάγεται ο k-means. Προς αυτή την κατεύθυνση (κυρτότητα) ένα πλήθος προσεγγίσεων συσταδοποίησης έχουν επίσης προταθεί [125] [162].

Ένα ακόμη πρόβλημα του αλγορίθμου k-means έχει να κάνει με την εκ' των προτέρων απαραίτητη γνώση του πλήθους των συστάδων των δεδομένων. Είναι πολύ συχνό το φαινόμενο τέτοια γνώση να μην υπάρχει για τα δεδομένα και επομένως η επιλογή είτε να γίνεται χωρίς κάποια γνώση των δεδομένων (τυχαία), είτε με μη αποτελεσματικό τρόπο. Προς αυτή την κατεύθυνση έχουν εφαρμοστεί μία σειρά από μεθόδους και ευρετικά στη βιβλιογραφία τα οποία και περιγράφονται στην ενότητα 3.7.3



Σχήμα 6: Ευαισθησία του k-means στις αρχικές συνθήκες

Τέλος, ένα εξαιρετικά σημαντικό πρόβλημα του αλγορίθμου k-means που θα πρέπει να αναφέρουμε είναι η φανερή του αδυναμία να διαχειριστεί τα outliers στα δεδομένα. Μία κατάσταση η οποία μπορεί να επιφέρει σημαντικές αλλοιώσεις και μειωμένη αποδοτικότητα στην όλη διαδικασία.

3.7.1.4 Άλλες προσεγγίσεις συσταδοποίησης

Πέρα από την παραπάνω γενική κατηγοριοποίηση σε ιεραρχικούς και διαιρετικούς αλγορίθμους, αρκετοί ακόμη αλγόριθμοι έχουν αναπτυχθεί που βασίζονται σε πληθώρα τεχνικών [12] μερικές από τις οποίες θα περιγραφούν και στη συνέχεια.

3.7.1.4.1 Ασαφής συσταδοποίηση

Όλες οι παραπάνω προσεγγίσεις προϋποθέτουν ότι τα αντικείμενα προς συσταδοποίηση ανήκουν έκαστο σε μία και μόνο συστάδα. Ενώ αυτό στις περισσότερες περιπτώσεις είναι αρκετό, υπάρχουν εφαρμογές στις οποίες το να ανήκουν τα αντικείμενα σε παραπάνω των μία συστάδων είναι επιθυμητό. Η συσταδοποίηση αυτού του είδους αναφέρεται ως ασαφής.

Στην ασαφή (fuzzy) συσταδοποίηση [161], σε αντιστοιχεία με την ασαφή λογική, κάθε σημείο έχει ένα βαθμό συμμετοχής στις συστάδες. Επομένως, τα αντικείμενα που βρίσκονται στις παρυφές των συστάδων μπορεί να ανήκουν σε μικρότερο βαθμό στη συστάδα τους σε σχέση με τα αντικείμενα που βρίσκονται εγγύτερα στο κέντρο της. Κάθε σημείο x λοιπόν έχει ένα σύνολο από συντελεστές που δίνουν τον βαθμό με τον οποίο αυτό ανήκει στην k συστάδα: $w_k(x)$. Με τον fuzzy c-means αλγόριθμο, το κέντρο της συστάδας είναι ο μέσος από όλα τα σημεία ζυγισμένα με τον βαθμό με τον οποίο αυτά ανήκουν στη συστάδα:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}. \tag{16}$$

Ο βαθμός $w_k(x)$ είναι σχετιζόμενος αντίστροφα με την απόσταση του x από το κέντρο της

συστάδας όπως υπολογίζεται από το προηγούμενο πέρασμα του αλγορίθμου. Εξαρτάται επίσης και από την παράμετρο m η οποία ελέγχει πόσο βάρος δίνεται στο κοντινότερο κέντρο. Ο fuzzy c-means αλγόριθμος είναι πολύ κοντά στον κλασικό k-means όσον αφορά στα βήματά του:

- Επέλεξε ένα πλήθος για τις συστάδες
- Ανάθεσε τυχαία κάθε αντικείμενο συντελεστές για συμμετοχή στης συστάδες
- Επανέλαβε έως ότου ο αλγόριθμος έχει συγκλίνει: οι συντελεστές ανάμεσα στα δύο τελευταία περάσματα δεν αλλάζουν παραπάνω από ϵ - το δοθέν όριο ευαισθησίας
- Υπολόγισε το κέντρο κάθε συστάδας με βάση την συνάρτηση 16
- Για κάθε σημείο, υπολόγισε τους συντελεστές του για συμμετοχή στις συστάδες με βάση την συνάρτηση 16

Ο αλγόριθμος c-means ελαχιστοποιεί την εσω-συσταδική απόσταση αλλά έχει τα ίδια προβλήματα όπως και ο k-means: το μέγιστο είναι συχνά τοπικό και τα αποτελέσματα εξαρτώνται σε μεγάλο βαθμό από τις αρχικές αναθέσεις βαρών. Ο αλγόριθμος c-means έχει χρησιμοποιηθεί ευρύτατα ως ένα σημαντικό εργαλείο για την επεξεργασία εικόνων και εύρεση συστάδων σε αυτές.. Μία ακόμη προσέγγιση συσχετιζόμενη με τον c-means είναι και ο Soft k-means.

3.7.1.4.2 Παραγωγικοί Αλγόριθμοι

Αλγόριθμοι όπως ο fuzzy c-means είναι ευαίσθητοι σε ακραίες τιμές (outliers). Σε ετερογενείς συλλογές κειμένων, οι ακραίες τιμές είναι ένα αρκετά σύνηθες φαινόμενο. Με το να κάνουμε ορισμένες υποθέσεις όμως για την κατανομή των δεδομένων, πιο ισχυρές και μη επιρρεπείς σε σφάλματα στατιστικές μέθοδοι μπορούν να εφαρμοστούν για την ανίχνευση συστάδων παρουσία θορύβου, λαμβάνοντας υπόψιν και τις αλληλεπικαλυπτόμενες συστάδες.

Μέθοδοι διακρίσεων (discriminative) που βασίζονται σε ζεύγη ομοιοτήτων κειμένων έχουν εξ' ορισμού $O(n^2)$ πολυπλοκότητα. Συχνά κιόλας αυτές οι ομοιότητες μπορούν να προ-υπολογιστούν και να αποθηκευθούν σε πίνακα. Τα παραγωγικά (generative) μοντέλα από την άλλη πλευρά, δεν απαιτούν κάποιον τέτοιο πίνακα και χρησιμοποιούν μία επαναληπτική διαδικασία η οποία μεταβαίνει μεταξύ των βημάτων εκτίμησης μοντέλου και ανάθεσης κειμένου.

3.7.1.4.3 Gaussian Μοντέλα

Τα Gaussian μοντέλα αναπαριστούν τα κείμενα ως ένα σύνολο από πίνακες μέσων τιμών (means) και συνδιακύμανσης (covariances). Σε αυτά τα μοντέλα, κάθε συστάδα βρίσκεται στο κέντρο της μέσης τιμής και περιγράφεται από το συσχετιζόμενο πίνακα. Το πρόβλημα συσταδοποίησης για αυτά τα μοντέλα ανάγεται στην εύρεση των παραπάνω πινάκων οι οποίοι ταιριάζουν καλύτερα στα κείμενα.

3.7.1.4.4 Μείωση διαστατικότητας

Στις περισσότερες των περιπτώσεων, η ανάλυση δεδομένων μπορεί να γίνει ευκολότερα και ακριβέστερα σε χώρο λιγότερων διαστάσεων. Η μείωση του πλήθους να διαστάσεων (dimensionality reduction) είναι η διαδικασία ελαχιστοποίησης του αριθμού των ανεξαρτήτων μεταβλητών ενός προβλήματος (σ.σ. συσταδοποίηση) και μπορεί χοντρικά να χωριστεί σε επιλογή χαρακτηριστικών και εξαγωγή χαρακτηριστικών.

Οι προσεγγίσεις επιλογής χαρακτηριστικών προσπαθούν να βρουν ένα υποσύνολο των αρχικών μεταβλητών χρησιμοποιώντας μία από τις δύο εξής στρατηγικές: φιλτράρισμα (κέρδος πληροφορίας) και αναζήτηση υποβοηθούμενη από την ακρίβεια.

Η εξαγωγή χαρακτηριστικών μετασχηματίζει τα δεδομένα από έναν χώρο υψηλού αριθμού διαστάσεων σε έναν με λιγότερες. Ο μετασχηματισμός αυτός μπορεί να είναι γραμμικός, όπως για παράδειγμα στην περίπτωση του Principal Component Analysis (PCA), όμως υπάρχουν και πολλές μη-γραμμικές τεχνικές μείωσης του αριθμού των διαστάσεων.

Η βασική γραμμική τεχνική μείωσης διαστατικότητας, PCA [108], εφαρμόζει μία γραμμική αντιστοίχηση των δεδομένων σε έναν χώρο λιγότερων διαστάσεων, με τέτοιο τρόπο ώστε η διακύμανση (διασπορά) των δεδομένων στον νέο χώρο να μεγιστοποιείται. Στην πράξη, ο πίνακας συσχετίσεων των δεδομένων κατασκευάζεται και οι ιδιοτιμές (eigenvalues) του πίνακα υπολογίζονται. Οι ιδιοπίνακες (eigenvectors) που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές, τα βασικά συστατικά δηλαδή, μπορούν εν' συνεχεία να χρησιμοποιηθούν για να ανακατασκευαστεί ένα μεγάλο ποσοστό της διακύμανσης των αρχικών δεδομένων. Επίσης, τα πρώτα λίγα ιδιοδιανύσματα μπορούν συχνά να ερμηνευτούν με όρους μεγάλης κλίμακας συμπεριφοράς τους συστήματος. Ο αρχικός χώρος έχει μειωθεί (με απώλεια δεδομένων αλλά συνήθως κρατώντας την πιο σημαντική διακύμανση) στο χώρο που καλύπτεται από τα λίγα ιδιοδιανύσματα.

Η PCA είναι μία στατιστική διαδικασία που χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό για να μετατρέψει ένα σύνολο από παρατηρήσεις από πιθανά εξαρτημένες μεταξύ τους μεταβλητές, σε ένα σύνολο από τιμές γραμμικών μη εξαρτημένων μεταβλητών οι οποίες αποκαλούνται πρωταρχικά χαρακτηριστικά (principal components). Το πλήθος των principal components είναι μικρότερο ή ίσο του πλήθους των αρχικών μεταβλητών. Αυτός ο μετασχηματισμός ορίζεται με τέτοιο τρόπο ώστε το πρώτο χαρακτηριστικό να έχει την μέγιστη δυνατή μεταβλητότητα (επομένως να ανταποκρίνεται σε όσο περισσότερη μεταβλητότητα των δεδομένων είναι αυτό εφικτό), και κάθε επόμενο χαρακτηριστικό έχει την επόμενη μέγιστη δυνατή μεταβλητότητα υπό την προϋπόθεση ότι είναι ορθογώνιο (δηλαδή μη συσχετιζόμενο) με τα προηγούμενα χαρακτηριστικά. Τα πρωταρχικά χαρακτηριστικά είναι ορθογώνια διότι είναι τα ιδιοδιανύσματα του πινάκα συνδιακύμανσης, ο οποίος είναι συμμετρικός. Το PCA είναι ευαίσθητο στην σχετική κλιμάκωση των αρχικών μεταβλητών.

3.7.1.4.5 Συσταδοποίηση δέντρου επιθεμάτων

Η συσταδοποίηση δέντρου επιθεμάτων (Suffix tree clustering) εξάγει συστάδες βασιζόμενος σε φράσεις που μοιράζονται μεταξύ τους τα κείμενα. Ο αλγόριθμος είναι γραμμικού χρόνου και

βασίζεται στον εντοπισμό των φράσεων εκείνων που είναι κοινές σε ομάδες κειμένων. Μία φράση είναι μία ακολουθία από λέξεις στη σειρά. Ορίζουμε λοιπόν μία βασική συστάδα ως το σύνολο κειμένων που μοιράζονται μία κοινή φράση. Το Suffix tree clustering έχει τρία λογικά βήματα:

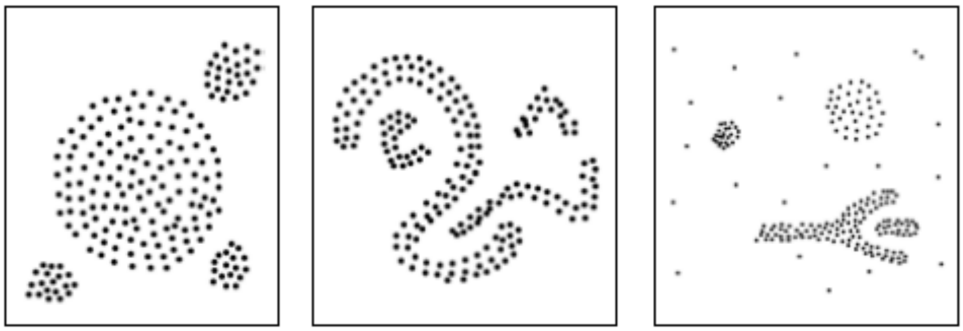
1. καθαρισμός κειμένου
2. εντοπισμός των βασικών συστάδων με χρήση δέντρου επιθεμάτων
3. συνδυασμός όλων αυτών των βασικών συστάδων σε μεγαλύτερες συστάδες

Περισσότερες πληροφορίες για το Suffix tree clustering είναι διαθέσιμες στα [228] [68] [212].

3.7.1.4.6 DBSCAN

Ο DBSCAN είναι ένας βασιζόμενος στην πυκνότητα αλγόριθμος ο οποίος βρίσκει ένα πλήθος από συστάδες ξεκινώντας από την εκτιμώμενη κατανομή πυκνότητας των κόμβων. Ο DBSCAN είναι ένας από τους πιο συνηθισμένους αλγορίθμους συσταδοποίησης μεγάλου όγκου δεδομένων. Ο DBSCAN μπορεί να εντοπίσει συστάδες σε μεγάλων χωρικών διαστάσεων δεδομένα ελέγχοντας την τοπική πυκνότητα των αντικειμένων, χρησιμοποιώντας μία μόνο παράμετρο εισόδου. Επίσης, ο χρήστης παίρνει μία πρόταση για την τιμή της παραμέτρου που θα ήταν η πιο ταιριαστή στα δεδομένα. Ως εκ' τούτου, απαιτείται ελάχιστη γνώση για τα ίδια τα δεδομένα. Ο αλγόριθμος μπορεί επίσης να καθορίσει ποια πληροφορία πρέπει να θεωρηθεί ως θόρυβος ή outliers. Είναι αρκετά γρήγορος και κλιμακώνεται σχεδόν γραμμικά με το μέγεθος των δεδομένων εισόδου.

Κάνοντας χρήση της κατανομής πυκνότητας των δεδομένων, ο DBSCAN μπορεί να κατηγοριοποιήσει αυτά σε χωριστές συστάδες οι οποίες μάλιστα, όπως φαίνεται και στο σχήμα 7, μπορούν να έχουν οποιοδήποτε σχήμα - κάτι που δεν ισχύει για τους προηγούμενους αλγορίθμους που παρουσιάστηκαν στην τρέχουσα ενότητα. Όμως, οι συστάδες που βρίσκονται κοντά μεταξύ τους συνήθως εν' τέλει ανήκουν στην ίδια κλάση δεδομένων.



Σχήμα 7: Τυπικές συστάδες του αλγορίθμου DBSCAN

Ο αλγόριθμος OPTICS μπορεί επίσης να ειδωθεί και ως μία γενίκευση του DBSCAN σε πολλαπλά εύρη τιμών, που επί της ουσίας αντικαθιστά την παράμετρο ϵ με μία μέγιστη ακτίνα αναζήτησης.

3.7.1.5 Μετρικές απόστασης (ομοιότητας)

Όλες οι μεθοδολογίες συσταδοποίησης οι οποίες περιγράφηκαν στο παρόν κεφάλαιο προϋποθέτουν την ύπαρξη ενός κατάλληλου χώρου ομοιότητας (similarity space) και επομένως απαιτούν την χρήση μίας μετρικής, ή αλλιώς ομοιότητας, μεταξύ δύο σημείων δεδομένων, δύο συστάδων ή ενός σημείου δεδομένων και μιας συστάδας. Όταν η μετρική ομοιότητας έχει καθοριστεί, καθένας από τους αλγορίθμους συσταδοποίησης μπορεί να υπολογίσει τον πίνακα ομοιότητας (distance matrix) ο οποίος περιλαμβάνει όλες τις αποστάσεις μεταξύ των αντικειμένων που συσταδοποιούνται.

Έστω λοιπόν δύο μεταβλητές, σημεία, ή κείμενα a και b. Παρακάτω περιγράφουμε ορισμένες από τις συνηθέστερες μετρικές απόστασης που αναφέρονται στη βιβλιογραφία.

3.7.1.5.1 Ευκλείδεια απόσταση

Η Ευκλείδεια απόσταση μεταξύ δύο σημείων αποτελεί την "κανονική" απόσταση τους - αυτή που κάποιος θεωρητικά θα μετρούσε με ένα χάρακα. Η απόσταση αυτή αποτελεί την στανταρ επιλογή σχεδόν για όλη την οικογένεια k-means αλγορίθμων. Ουσιαστικά μάλιστα ο k-means αλγόριθμος ορίζεται με βάσει την χρήση της Ευκλείδειας απόστασης ως μετρικής ομοιότητας.

Η Ευκλείδεια απόσταση μεταξύ των a και b ορίζεται βάσει του Πυθαγορείου θεωρήματος ως:

$$d(a,b) = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{17}$$

όπου a_i και b_i η αναπαράσταση του κειμένου a και b στην διάσταση i του n-διάστατου χώρου αναπαράστασης των κειμένων. Η Ευκλείδεια απόσταση λαμβάνει υπόψιν της και το μέγεθος της εισόδου (π.χ. κείμενο) και ως εκ' τούτου διατηρεί περισσότερη πληροφορία σχετικά με αυτή. Επίσης η Ευκλείδεια απόσταση είναι "πραγματική" μετρική μιας και ικανοποιεί την τριγωνική ανισότητα.

3.7.1.5.2 City-block / απόσταση Manhattan

Η απόσταση Manhattan μεταξύ δύο σημείων του n-διάστατου χώρου αναπαράστασης τους, είναι το άθροισμα των μηκών των προβολών αυτών πάνω στους άξονες συντεταγμένων. Πιο συγκεκριμένα:

$$d(a,b) = \frac{1}{n}\sum_{i=1}^{n} \mid a_i - b_i \mid \tag{18}$$

Η απόσταση Manhattan είναι επίσης "πραγματική" μετρική μιας και ικανοποιεί την τριγωνική ανισότητα.

3.7.1.5.3 Απόσταση Pearson

Ο συντελεστής συσχέτισης (correlation coefficient) Pearson μεταξύ δύο μεταβλητών ορίζεται ως η συνδιακύμανση (covariance) των δύο μεταβλητών διαιρεμένη με το γινόμενο της τυπικής τους

απόκλισης. Πιο συγκεκριμένα:

$$r(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i - \bar{a}}{\sigma_a}\right)\left(\frac{b_i - \bar{b}}{\sigma_b}\right) \tag{19}$$

όπου $\bar{a}$ και $\bar{b}$ είναι η μέση τιμή του a και b αντίστοιχα, ενώ σ_a και σ_b είναι η τυπική απόκλιση του a και b. Θα λέγαμε ότι ο συντελεστής συσχέτισης του Pearson, ως μετρική, αντιπροσωπεύει πόσο καλά μία ευθεία γραμμή μπορεί να ταιριάξει στο καρτεσιανό επίπεδο των a και b. Οι απόλυτες τιμές του συντελεστή συσχέτισης Pearson είναι μικρότερες ή ίσες του 1. Συγκεκριμένα, τιμές ίσες με +1 και -1 αντιστοιχούν σε σημεία δεδομένων του πέφτουν ακριβώς πάνω στη ευθεία γραμμή. Επίσης ο συντελεστής συσχέτισης Pearson είναι συμμετρικός για δύο σημεία: $r(a,b) = r(b,a)$.

Μία βασική μαθηματική ιδιότητα του συντελεστή συσχέτισης Pearson είναι ότι είναι αδιάφορος σε ξεχωριστές αλλαγές στην τοποθεσία και κλίμακα των δύο μεταβλητών. Ως εκ' τούτου, μπορούμε να μετασχηματίσουμε το a σε $\alpha + \beta a$ και το b σε $\gamma + \delta b$, όπου α, β, γ και δ σταθερές με $\beta, \delta > 0$, χωρίς να μεταβληθεί η τιμή του συντελεστή συσχέτισης.

Με βάση τα παραπάνω, η απόσταση Pearson ορίζεται ως:

$$d(a,b) = 1 - r \tag{20}$$

3.7.1.5.4 Ομοιότητα συνημιτόνου

Πρόκειται για ίσως την πιο χρησιμοποιούμενη μετρική σε συστήματα ανάκτησης πληροφορίας. Ορίζεται ως:

$$d(a,b) = cos(\theta) = \frac{a \cdot b}{\mid a \mid\mid b \mid} \in [0,1] \tag{21}$$

Η ομοιότητα συνημιτόνου μεταξύ δύο σημείων αντιστοιχεί στην γωνία που σχηματίζεται μεταξύ τους στον n-διάστατο χώρο αναπαράστασης. Βασίζεται στο εσωτερικό γινόμενο των διανυσμάτων που αποτελούνται από τις συντεταγμένες των a και b. Το συνημίτονο μηδενικής γωνίας είναι 1 και για οποιαδήποτε άλλη γωνία είναι μικρότερο του 1. Πρόκειται επομένως για μία μετρική που αποτυπώνει στην διάταξη στον n-διάστατο χώρο και όχι το μέτρο των παραπάνω διανυσμάτων.

3.7.1.5.5 Απόσταση Spearman-rank

Η απόσταση Spearman-rank είναι μία μη-παραμετρική μετρική η οποία αποδίδει καλά απέναντι σε ακραίες τιμές δεδομένων (outliers). Πηγάζει από τον συντελεστής συσχέτισης Pearson μέσω αντικατάστασης κάθε τιμής με την σειρά κατάταξης της αφού οι τιμές έχουν πρώτα ταξινομηθεί. Λόγω της απαλοιφής των τιμών δεδομένων, δεν υπάρχει πληροφορία βάρους η οποία να έχει ρόλο στον υπολογισμό της απόστασης (σε σχέση με τις προηγούμενες - παραμετρικές μετρικές ομοιότητας). Ο συντελεστής συσχέτισης Spearman-rank ορίζεται ως ακολούθως:

$$(a,b) = 1 - \frac{6\sum d_i^2}{n(n^2-1)}) \tag{22}$$

όπου $d_i = a_i - b_i$ η απόσταση μεταξύ της σειρά κατάταξης.

Η απόσταση Spearman-rank μεταξύ δύο σημείων a και b ορίζεται επομένως ως:

$$d(a, b) = 1 - \rho \tag{23}$$

3.7.1.5.6 Απόσταση Kendall's

Ο συντελεστής συσχέτιση Kendall's τ (Kendall's tau) είναι παρόμοιος με εκείνον του Spearman-rank, κάνοντας χρήση όμως σχετικών σειρών κατάταξης και όχι απολύτων. Πιο συγκεκριμένα:

Ορισμός 3.7.2. έστω $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ η λίστα από τις παρατηρήσεις (δεδομένα) των τυχαίων μεταβλητών a και b, τέτοιες ώστε όλες οι τιμές x_i και y_i να είναι μοναδικές. Κάθε ζεύγος παρατήρησης (x_i, y_i) και (x_j, y_j) είναι συγκλίνων, αν οι σειρές κατάταξης και για τα δύο στοιχεία συμφωνούν, δηλαδή: αν $x_i > x_j$, τότε και $y_i > y_j$, ή αν $x_i < x_j$, τότε και $y_i < y_j$. Αντίστοιχα το ζεύγος παρατήρησης είναι αποκλίνον αν $x_i > x_j$ και $y_i < y_j$ ή αν $x_i < x_j$ και $y_i > y_j$. Προφανώς αν $x_i = x_j$ or $y_i = y_j$ τότε το ζεύγος δεν είναι ούτε συγκλίνων ούτε αποκλίνον.

Με βάση τα παραπάνω ο συντελεστής συσχέτισης Kendall's τ ορίζεται ως:

$$\tau = \frac{(\text{πλήθος συγκλινόντων ζευγών}) - (\text{πλήθος αποκλινόντων ζευγών})}{\frac{1}{2}n(n-1)} \tag{24}$$

Τέλος, η απόσταση Kendall's ορίζεται ως

$$d(a, b) = 1 - \tau \tag{25}$$

3.7.1.6 Μετρικές αξιολόγησης συσταδοποίησης

Μία συνάρτηση αξιολόγησης της συσταδοποίησης κειμένων αποτελεί ένα ποσοτικό κριτήριο προκειμένου να αποκριθούμε αν και πόσο αποτελεσματικός είναι ένας αλγόριθμος συσταδοποίησης. Οι μέθοδοι αξιολόγησης που έχουν προταθεί στη βιβλιογραφία θα μπορούσαν να χωριστούν χοντρικά σε τρεις κατηγορίες:

- Οπτική αναπαράσταση των παραγόμενων συστάδων, π.χ. [95]. Ο τρόπος αξιολόγησης αυτός απλά παρουσιάζει τα αποτελέσματα της συσταδοποίησης σε ένα δισδιάστατο χώρο, παρέχοντας έτσι ένα οπτικό τρόπο για την κατανόηση των αποτελεσμάτων. Η αξιολόγηση αυτού του είδους όμως δεν είναι συνήθως αρκετή για την κρίση της απόδοσης των αλγορίθμων.

- Βασιζόμενοι σε IR κριτήρια αξιολόγησης. Η συσταδοποίηση, ως ένα κεντρικό IR task, συχνά μοιράζεται τις ίδιες μετρικές αξιολόγησης των αποτελεσμάτων της όπως και τα υπόλοιπα IR tasks. Οι μετρικές αυτές οποίες παρουσιάστηκαν αναλυτικά στην ενότητα 3.2.2. Για παράδειγμα στο [228] γίνεται αξιολόγηση της συσταδοποίησης με δέντρα επιθεμάτων χρησιμοποιώντας την μετρική της ακρίβειας. Η ίδια μετρική χρησιμοποιήθηκε και στα [118][114] για την αξιολόγηση των δικτύων Kohonen για συσταδοποίηση

- Ακρίβεια με βάση σύγκρισης της διαφορά μεταξύ των επιθυμητών και πραγματικών αποτελεσμάτων συσταδοποίησης. Για παράδειγμα [81][228]. Αυτή η μέθοδος αξιολόγησης απαιτεί τον ορισμό των επιθυμητών συστάδων ώστε να μπορούμε πράγματι να αξιολογήσουμε ένα μοντέλο συσταδοποίησης. Είναι λοιπόν δυνατή μόνο σε επίπεδα μοντέλα, όπου το πλήθος των συστάδων είναι γνωστό από πριν (δίνεται σαν παράμετρος), όπως για παράδειγμα ο αλγόριθμος k-means. Μία ακόμα μέθοδος που ανήκει σε αυτή την κατηγορία είναι και η βασιζόμενη στην εντροπία των κειμένων εντός και εκτός των συστάδων [103].

3.7.1.6.1 Δείκτης συσταδοποίησης (Clustering Index)

Η μετρική αξιολόγησης Clustering Index [104] βασίζεται στην παραδοχή ότι η καλύτερη συσταδοποίηση έχει να κάνει τόσο με την υψηλότερη δυνατή ενδο-συσταδική ομοιότητα, όσο και με τη χαμηλότερη δυνατή δια-συσταδική ομοιότητα. Μέσα σε μία συστάδα, τα κείμενα θα πρέπει να είναι όσο πιο όμοια γίνεται, ενώ αντίθετα μεταξύ των συστάδων, τα κείμενα θα πρέπει να είναι όσο πιο διαφορετικά γίνεται. Η μετρική Clustering Index επομένως ορίζεται ως ο λόγος της εσω-συσταδικής ομοιότητας, $\bar{\sigma}$, ως προς το άρθροισμα της εσω-συσταδικής και δια-συσταδικής ομοιότητας, $\bar{\delta}$. Επομένως:

$$CI = \frac{\bar{\sigma}^2}{\bar{\sigma} + \bar{\delta}} \tag{26}$$

Γενικά η τιμή του Clustering Index κανονικοποιείται μεταξύ 0 και 1. Τιμή 1 αντιστοιχεί στην απόλυτα επιθυμητή συσταδοποίηση, ενώ τιμή 0 το ακριβώς αντίθετο. Μεγιστοποίηση της τιμής CI σημαίνει μεγιστοποίηση της ενδο-συσταδικής ομοιότητας με παράλληλη ελαχιστοποίηση της δια-συσταδικής ομοιότητας. Ως εκ' τούτου ο δείκτης αυτός μπορεί να απεικονίσει την συνοχή των παραγόμενων συστάδων.

3.7.1.6.2 Μέσο απόλυτο σφάλμα

Το μέσο απόλυτο σφάλμα ή αλλιώς Mean Absolute Error (MAE), αποτελεί μία στατιστική μετρική η οποία χρησιμοποιείται για την μέτρηση του πόσο κοντά βρίσκονται οι προβλέψεις ενός συστήματος προτάσεων σε σχέση με τα πραγματικά αποτελέσματα. Το MAE ορίζεται ως:

$$MAE = \frac{\sum \mid r(u,i) - r'(u,i) \mid}{\mid R' \mid} \tag{27}$$

όπου $r(u,i) \in R$ η πραγματική τιμή της μεταβλητής i στο u και $r'(u,i) \in R'$ οι προβλέψεις που κάνει το σύστημα προτάσεων για την μεταβλητή i.

3.7.2 Αξιοποίηση Εξωτερικών Βάσεων Γνώσης

3.7.2.1 WordNet

Το WordNet[220] αποτελεί έναν από τους πιο χρησιμοποιημένους και αξιόπιστους θησαυρούς λέξεων της Αγγλικής γλώσσας, έτσι, μοντελοποιεί την λεξιλογική γνώση και χρήση των λέξεων

της Αγγλικής. Περιλαμβάνοντας πάνω από 150.000 όρους, ομαδοποιεί ουσιαστικά, ρήματα, επίθετα και επιρρήματα σε ομάδες συνωνύμων τα οποία και ονομάζονται Synonym sets (Synsets). Τα synsets οργανώνονται σε:

- ερμηνείες (senses) δίνοντας έτσι τα συνώνυμα από κάθε λέξη
- υπώνυμα / υπερώνυμα (δηλαδή, "είναι ένα..." (Is-A)) και μερώνυμα / ολόνυμα (δηλαδή, "μέρος από..." (Part-Of)) συσχετίσεις, παρέχοντας έτσι μία ιεραρχική δενδρική δομή για κάθε όρο.

3.7.2.1.1 Χρήση του WordNet στην συσταδοποίηση

Οι εφαρμογές του WordNet σε μία ποικιλία από IR τεχνικές έχουν μελετηθεί εκτενώς στην βιβλιογραφία σε σχέση με την εύρεση σημασιολογικής ομοιότητας των ανακτημένων αντικειμένων [214], ή σε σχέση με τις τεχνικές συσταδοποίσης. Για παράδειγμα, στο [47] οι συγγραφείς συνδυάζουν την γνώση από το WordNet με ασαφείς κανόνες συσχέτισης, ενώ στο [193] επεκτείνεται ο bisecting k-means αλγόριθμος με χρήση του WordNet, όμως, λόγω του ότι επιλέγονται τα υπερώνυμα / συνώνυμα σε "επίπεδα', οι συγγραφείς καταλήγουν στο συμπέρασμα ότι ο θόρυβος υποβιβάζει τα αποτελέσματα συσταδοποίησης.

Στο [76] ερευνάται η ιδέα χρήσης του WordNet σαν ένα εργαλείο αποσαφήνισης αναθέτοντας τις ρίζες των λέξεων κλειδιών στην λεξικολογική τους κατηγορία. Η παραπάνω προσέγγιση βελτιώνει την αποτελεσματικότητα του εφαρμοζόμενου αλγορίθμου συσταδοποίησης, όμως, φαίνεται να υπερ-γενικοποιεί τις αναφερόμενες λέξεις κλειδιά. Αυτό προκύπτει και από μία παρόμοια έρευνα στο [11], όπου οι συγγραφείς αποδέχονται των όρων σε έννοιες οντολογίας μπορεί να είναι εν' γένει διφορούμενη και να οδηγήσει σε απώλεια πληροφορίας στην προσπάθεια μείωσης των διαστάσεων του προβλήματος. Και οι δύο προαναφερθείσες προσεγγίσεις δεν λαμβάνουν υπόψιν τους τα υπερώνυμα του WordNet για την πραγματική ενίσχυση της λίστα των λέξεων κλειδιών, κάτι που εμείς προτείνουμε στην παρούσα διατριβή. Σε αντίθεση με τις παραπάνω προσεγγίσεις, πιστεύουμε ότι ένα αξιόπιστο σύστημα ζυγίσματος για τα υπερώνυμα του WordNet μπορεί να επιφέρει σημαντικά οφέλη στη διαδικασία συσταδοποίησης, όπως π.χ. στο [179].

3.7.3 Πλήθος συστάδων

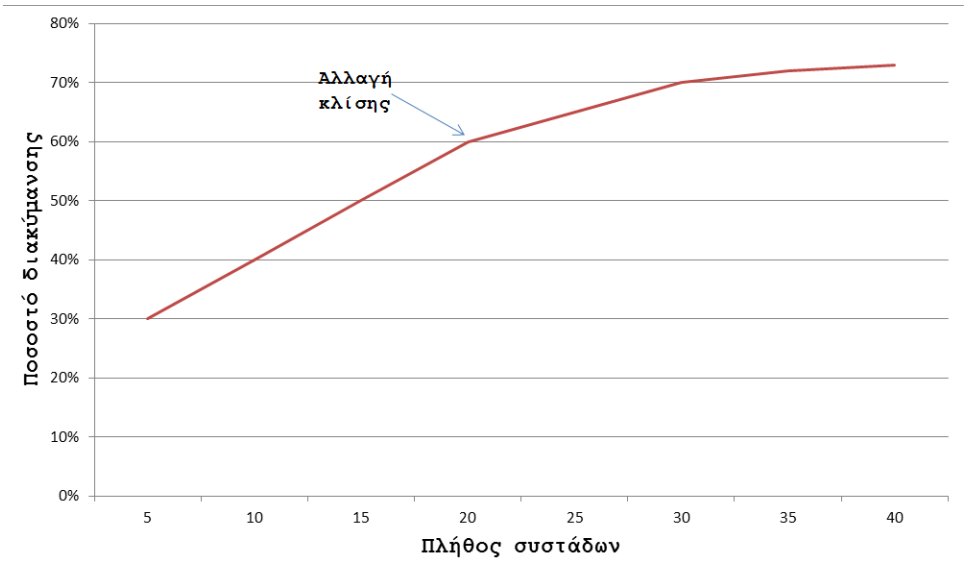
Ο αριθμός των συστάδων που τελικά αντιστοιχούν σε ένα σύνολο δεδομένων, είναι ένα πρόβλημα που αφορά σχεδόν όλους τους διαμερισματικούς αλγορίθμους - και δει της οικογένειας k-means. Μία ατυχής επιλογή για τον αριθμό των συστάδων συνήθως οδηγεί σε μη χρήσιμα και γενικά λανθασμένα αποτελέσματα. Αρκετές προσεγγίσεις έχουν προταθεί στην βιβλιογραφία:

- Εμπειρικός κανόνας: σε πολλές περιπτώσεις δεδομένων, μεγέθους n, έχει παρατηρηθεί ότι το πραγματικό πλήθος συστάδων, k, βρίσκεται κοντά στην τιμή:

$$k = \sqrt{n/2} \tag{28}$$

- Η μέθοδος του αγκώνα: η μέθοδος αυτή κοιτάζει στο ποσοστό διακύμανσης που δίνεται ως μια συνάρτηση του πλήθους των συστάδων. Κάποιος θα επιλέξει ένα πλήθος συστάδων έτσι ώστε η προσθήκη ακόμα μιας δεν δίνει καλύτερη μοντελοποίηση για τα δεδομένα. Πιο συγκεκριμένα, αν κάποιος σχεδιάσει το ποσοστό της διακύμανσης που δίνεται από τις συστάδες σε σχέση με το πλήθος των συστάδων, η πρώτες συστάδες γενικά θα προσθέσουν πολύ πληροφορία (υψηλή διακύμανση), όμως σε κάποιο σημείο το οριακό κέρδος (κλίση) θα αρχίσει να πέφτει αποτυπωνόμενο ουσιαστικά σαν μια γωνία στο γράφημα, όπως στο σχήμα 8 για παράδειγμα. Ο αριθμός των συστάδων επιλέγεται σε αυτό το σημείο. Συχνά όμως αυτό το σημείο δεν μπορεί να αποτυπωθεί εύκολα [116]. Το ποσοστό της διακύμανσης στην παραπάνω περίπτωση είναι ο λόγος μεταξύ της ενδο-συσταδικής διακύμανσης προς την συνολική διακύμανση (γνωστός και ως F-test). Μπορούμε επίσης αντί για την F-test μετρική να χρησιμοποιήσουμε την μετρική RSoS (11) η οποία θα μας δώσει τα ίδια αποτελέσματα όσον αφορά το σημείο οριακού κέρδους αλλά με ανεστραμμένη τη γραφική παράσταση (φθίνουσα RSoS όσο αυξάνεται το πλήθος των συστάδων).

  Εκτός από τα παραπάνω ο αλγόριθμος k-means είναι γενικά αποτελεσματικός όταν οι συστάδες είναι σχεδόν σφαιρικές σε σχέση με το μέτρο ομοιότητας που χρησιμοποιείται. Δεν υπάρχει όμως κάποιος λόγος να πιστεύουμε ότι τα κείμενα μιας συλλογής, υπό την τυπική αναπαράστασή τους ως ζυγισμένοι πίνακες λέξεων και κάποιας μορφής κανονικοποίησης μετρικής ομοιότητας εσωτερικού γινομένου, θα πρέπει να ανήκουν σε σχεδόν σφαιρικές συστάδες.



Σχήμα 8: Εκτιμώμενη αύξηση διακύμανσης με παράλληλη αύξηση του πλήθους των συστάδων

- Προσεγγίσεις κριτηρίου πληροφορίας: πρόκειται για μία κατηγορία μεθόδων οι οποίες ορίζουν κάποιο κριτήριο πληροφορίας η κάθε μία και αποφασίζουν για το πλήθος των συστάδων βάσει αυτού. Τυπικά παραδείγματα είναι τα: Akaike information criterion (AIC) [188], Bayesian

information criterion (BIC) [217] και Deviance information criterion (DIC) [29].

- Μέσω χρήσης silhouette: η silhouette κάποιων μονάδων δεδομένων είναι ένα μέτρο του πόσο κοντά ταιριάζει αυτή η μονάδα στα δεδομένα της συστάδας καθώς και πόσο χαλαρά ταιριάζει στα δεδομένα των γειτονικών συστάδων. Μία silhouette κοντά στο 1 υπονοεί ότι η μονάδα δεδομένων είναι στην σωστή συστάδα, ενώ μία τιμή κοντά στο -1 εκφράζει ότι η συστάδα του είναι λανθασμένη. Τεχνικές βελτιστοποίησης όπως οι γενετικοί αλγόριθμοι είναι χρήσιμοι στο να καθορίζουν το πλήθος των συστάδων το οποίο παράγει ουσιαστικά την μεγαλύτερη silhouette [132].

- Μέσω διασταυρωμένης επικύρωσης (cross-validation): σε αυτή την διαδικασία, τα δεδομένα χωρίζονται σε y μέρη. Κάθε μέρος τίθεται στην άκρη με την σειρά ως δεδομένα ελέγχου (test set) και ένα μοντέλο συσταδοποίησης υπολογίζεται χρησιμοποιώντας τα υπόλοιπα $y-1$ δεδομένα εκμάθησης (training set) και η τιμή της συνάρτησης στόχου (για παράδειγμα το άθροισμα του τετραγώνου των αποστάσεων των κέντρων για τον k-means) υπολογίζεται για τα δεδομένα ελέγχου. Ο μέσος όρων αυτών των y τιμών υπολογίζεται για κάθε εναλλακτικό πλήθος συστάδων και το πλήθος μου ελαχιστοποιεί το σφάλμα στα δεδομένα ελέγχου επιλέγεται [69].

- Για βάσεις κειμένων μεγέθους πίνακα όρων-κειμένων $D(mxn)$ όπου m το πλήθος των κειμένων και n το πλήθος των όρων, το πλήθος των συστάδων μπορεί χοντρικά να εκτιμηθεί ως:

$$k = \frac{mn}{t} \tag{29}$$

όπου t ο αριθμός των μη μηδενικών εγγραφών στον πίνακα D. Βασική προυπόθεση του παραπάνω αποτελεί ότι στον πίνακα D κάθε γραμμή και κάθε στήλη θα πρέπει να περιέχει τουλάχιστον ένα μη μηδενικό στοιχείο [42].

3.7.4 Ονοματοδοσία συστάδων

Η ονοματοδοσία συστάδων, μία διαδικασία που είναι ευρύτερα γνωστή ως "ετικετοποίηση" ή αλλιώς cluster labeling, αποτελεί ένα βήμα που τυπικά έπεται της ίδιας της συσταδοποίησης. Συνηθέστερα μάλιστα, στις περιπτώσεις που έχουμε να κάνουμε με συστάδες κειμενικής πληροφορίας ανθρωπίνου λόγου (π.χ. κείμενα στην αγγλική γλώσσα). Σκοπός του cluster labeling είναι η αντιστοίχιση νοηματικά κατανοητών λέξεων ή φράσεων στις συστάδες προκειμένου το περιεχόμενο αυτών να εύκολα αντιληπτό. Ο τελικός αποδέκτης βέβαια είναι ο άνθρωπος, είτε ο τελικός χρήστης του συστήματος, είτε κάποιος διαχειριστής αυτού, που μαζί με τις συστάδες παίρνει και τις ετικέτες αυτών για πληρέστερη κατανόηση του αποτελέσματος.

Οι τεχνικές του cluster labelling [211], συχνά αξιολογούν ετικέτες οι οποίες προέρχονται από τα ίδια τα δεδομένα, π.χ. λέξεις κλειδιά που ήδη εξάγονται από αυτά και ανήκουν στις συστάδες προς ονοματοδοσία [209]. Πρόσφατα στο [210], οι συγγραφείς προτείνουν μία αποτελεσματική Fuzzy Frequent Itemset-based προσέγγιση συσταδοποίησης κειμένων η οποία συνδυάζει εξόρυξη

ασαφών κανόνων συσχέτισης με την γνώση που εμπεριέχεται στα υπερώνυμα του WordNet για την δημιουργία των ετικετών. Παρόλα αυτά οι συγγραφείς τονίζουν ότι η διαδικασία εξόρυξης των ασαφών κανόνων συσχέτισης καθώς και η ίδια η συσταδοποίηση είναι δύο χρονοβόρα βήματα, κάτι που οδηγεί σε μεγάλους χρόνους εκτέλεσης των δεδομένων (παρότι αυτοί κλιμακώνονται γραμμικά με την είσοδο). Αντίθετα, στην περίπτωση ενός συστήματος προτάσεων άρθρων νέων, εστιάζουμε σε μία προσέγγιση η οποία θα παράγει τόσο τις συστάδες όσο και τις ετικέτες αυτών σχετικά γρήγορα ώστε να μπορεί να ανταπεξέρχεται στο ρυθμό παραγωγής των άρθρων από τις πηγές τους.

3.8 Προσωποποίηση στον Χρήστη

Το ζήτημα της προσωποποίησης του περιεχομένου στον χρήστη, αποτελεί ένα ερευνητικό πεδίο από μόνο του με πληθώρα διαστάσεων. Στη συνέχεια προσπαθούμε ουσιαστικά να εισαγάγουμε τον αναγνώστη σε διάφορες τεχνικές που έχουν προταθεί για το πρόβλημα όσον αφορά ορισμένες μόνο διαστάσεις του.

Η προσωποποιημένη αναζήτηση είναι μία σημαντική ερευνητική περιοχή η οποία αποσκοπεί στην επίλυση της ασάφειας των αποτελεσμάτων. Προσβλέποντας στην βελτίωση της σχετικότητας των αποτελεσμάτων αναζήτησης, οι μηχανές προσωποποιημένης αναζήτησης δημιουργούν προφίλ χρήστη για να καταγράψουν τις προσωπικές προτιμήσεις των χρηστών, και ως εκ' τούτου, να αναγνωρίσουν τον πραγματικό σκοπό ενός ερωτήματος. Δεδομένου όμως ότι οι χρήστες είναι συχνά διστακτικοί στην άμεση έκφραση των προτιμήσεών τους, κυρίως λόγω της επιπλέον δουλειάς που αυτό περιλαμβάνει, η πρόσφατη έρευνα έχει εστιάσει στην αυτοματοποιημένη εκμάθηση των προτιμήσεων του χρήστη κάνοντας χρήση των ιστορικών αναζήτησης και πλοήγησης. Τα προσωποποιημένα συστήματα γενικά σχεδιάζονται ώστε να βασίζονται στις προτιμήσεις χρηστών που έχουν ήδη εντοπιστεί με τον παραπάνω τρόπο. Οι περισσότερες προσεγγίσεις εφαρμόζουν ένα μοναδικό (και συνήθως μεγάλο) προφίλ για κάθε χρήστη που συμμετέχει στην διαδικασία. Στην πραγματικότητα όμως, οι θετικές προτιμήσεις δεν είναι αρκετές για να αποτυπώσουν πλήρως και εις βάθος τα ενδιαφέροντα ενός χρήστη.

Οι στρατηγικές δημιουργίας προφίλ χρηστών μπορούν να αντιστοιχιστούν σε δύο γενικές προσεγγίσεις: αυτές που βασίζονται στα κείμενα (document-based), και αυτές που βασίζονται στις έννοιες (concept-based).

Οι document-based μεθοδολογίες δημιουργίας προφίλ, στοχεύουν στην αποτύπωση της συμπεριφοράς του χρήστη σε ότι έχει να κάνει με τα "clicks" και γενικότερα τα μονοπάτια πλοήγησης που ακολουθεί. Οι προτιμήσεις σε κείμενα πρώτα εξάγονται από τα click-through δεδομένα και στη συνέχεια χρησιμοποιούνται για να παραγχθούν μοντέλα συμπεριφοράς χρήστη, τα οποία συνήθως αναπαρίστανται ως ένα σύνολο από ζυγισμένα χαρακτηριστικά.

Από την άλλη μεριά, οι concept-based μεθοδολογίες δημιουργίας προφίλ, στοχεύουν στην αποτύπωση των εννοιολογικών αναγκών των χρηστών. Τα κείμενα στα οποία οι χρήστες έχουν πλοηγηθεί, καθώς και τα ιστορικά αναζήτησής τους, αντιστοιχίζονται αυτομάτως σε ένα σύνολο από θεματικές κατηγορίες. Τα προφίλ χρηστών παράγονται βασιζόμενοι στις προτιμήσεις των χρηστών

όπως αυτές εξάγονται μέσα από τις θεματικές κατηγορίες.

Στο [106] μελετάται μία μέθοδος η οποία εφαρμόζει εξόρυξη προτιμήσεων και μηχανική εκμάθηση προκειμένου να μοντελοποιηθεί η συμπεριφορά από "clicks" και πλοήγησης. Η μέθοδος αυτή υποθέτει ότι ένας χρήστης θα διαβάσει τα αποτελέσματα από την λίστα που επιστρέφονται από την αρχή προς το τέλος. Εάν ο χρήστης προσπεράσει ένα κείμενο d_i στην θέση i, πριν κάνει click σε ένα κείμενο d_j στη θέση j, υποθέτει ότι μάλλον είδε για ποιο κείμενο πρόκειται και εσκεμμένα αποφάσισε να το αποφύγει. Κατά συνέπεια μπορούμε να υποθέσουμε ότι ο χρήστης προτιμά το κείμενο d_j περισσότερο από το d_i (δηλαδή $r_{d_i} < r_{d_j}$) όπου r είναι η σειρά προτίμησης των κειμένων στην λίστα που επιστράφηκε.

Στο [214] οι συγγραφείς εστιάζουν στην προσωποποιημένη παραγωγή προτάσεων από σελίδες Web οι οποίες προσαρμόζονται ανάλογα με τα πρότυπα πρόσβασης που κατασκευάζονται μέσω της ανάλυσης της πληροφορίας πλοήγησης των χρηστών. Δείχνουν ότι η μεθοδολογία που ενσωματώνει την συσταδοποίηση χρηστών μέσα στο πλαίσιο ενός συστήματος προτάσεων εντοπίζοντας ενδιαφέροντα μονοπάτια πλοήγησης χρηστών, μπορεί να είναι βοηθητική.

Στο [133] οι συγγραφείς προβλέπουν την προτίμηση του χρήστη για ένα αντικείμενο μέσω της ζύγισης των συνεισφορών παρόμοιων χρηστών, που ονομάζονται γείτονες, για αυτό το αντικείμενο. Η ομοιότητα μεταξύ των χρηστών υπολογίζεται μέσω σύγκρισης των τρόπων αξιολόγησης που αυτοί χρησιμοποιούν, π.χ. ένα σύνολο από βαθμολογήσεις που δόθηκαν για τα ίδια αντικείμενα, ή μέσω των συνηθειών πλοήγησής τους.

Σε αντίθεση με τις παραπάνω προσεγγίσεις, στην διδακτορική διατριβή προτείνουμε μία νέα μεθοδολογία η οποία ενσωματώνει τον αλγόριθμο συσταδοποίησης W-kmeans στο πλαίσιο της παραγωγής προσωποποιημένων προτάσεων προς τον χρήστη. Περισσότερα σχετικά με την προσέγγισή μας στα επόμενα κεφάλαια.

3.9 Το Πρόβλημα του νέου Χρήστη

Ένα βασικό πρόβλημα με το CF είναι ότι δεν δουλεύει πάντα καλά λόγω ελλιπών δεδομένων για τους χρήστες, κάτι που είναι επίσης γνωστό και ως πρόβλημα νέου χρήστη. Το πρόβλημα αυτό προκύπτει από το γεγονός ότι κάθε χρήστης έχει δει μόνο ένα μικρό μέρος από τα δεδομένα και επομένως ακριβείς προβλέψεις δεν μπορούν να γίνουν εύκολα, τουλάχιστον μέχρις ότου η κάλυψη χρήστη/δεδομένων έχει φτάσει σε κάποιο επίπεδο.

Οι προσεγγίσεις που περιγράφονται στη βιβλιογραφία για το πρόβλημα νέου χρήστη εστιάζουν κυρίως στα μετα-δεδομένα και στις ερωτήσεις προς τους χρήστες. Τα μετα-δεδομένα σχετικά με αντικείμενα μπορούν να χρησιμοποιηθούν για να παραχθούν προτάσεις από συστήματα προτάσεων που βασίζονται στο περιεχόμενο, όπως στο [22], ή σε υβριδικές προσεγγίσεις με συστήματα βασισμένα σε βαθμολογήσεις, π.χ. [110]. Τα filterbots [167] συνιστούν μία ακόμη προσέγγιση όπου ψευδο-χρήστες και αντικείμενα παράγονται αλγοριθμικά σε μία προσπάθεια να παρέχονται αναφορές βαθμολογήσεων στο σύστημα, έτσι ώστε κανείς χρήστης ή αντικείμενο να μην είναι χωρίς βαθμολόγηση. Η τεχνική αυτή, όπως αποτιμάται στο [82], μπορεί να λειτουργήσει καλύτερα όταν

χρησιμοποιείται σε συνδυασμό με τεχνικές CF, και πιο συγκεκριμένα, οι CF τεχνικές έχουν την μεγαλύτερη επίπτωση στα αποτελέσματα αυτού του συνδυαστικού σεναρίου χρήσης. Άλλες μέθοδοι οι οποίες συνδυάζουν δημογραφικά δεδομένα διαθέσιμα στο σύστημα έχουν επίσης προταθεί. Το πρόβλημα όμως αυτών των προσεγγίσεων είναι ότι η συλλογή τέτοιων δεδομένων συνήθως προσκρούει σε προβλήματα ιδιωτικότητας.

Τα συστήματα προτάσεων, εσωτερικά, έχουν επίσης χρησιμοποιηθεί για να αντιμετωπίσουν το πρόβλημα νέου χρήστη. Μερικές προσεγγίσεις, όπως περιγράφονται στο [159], παράγουν κατηγορίες χρηστών όπου νέοι χρήστες αντιστοιχίζονται γρήγορα αξιοποιώντας ένα σύνολο από προκαθορισμένες ερωτήσεις. Αυτές οι προσεγγίσεις εκκινούν το σύστημα χρησιμοποιώντας δημογραφικά χαρακτηριστικά, ή χαρακτηριστικά βασισμένα σε μοντέλα. Παρότι σχετικά περιορισμένα όσον αφορά τον τομέα γνώσης, μπορούν και παράγουν ακριβή αποτελέσματα.

3.9.1 Ερωτήσεις προς, και βαθμολογήσεις από τον χρήστη

Μία ακόμη μέθοδος αντιμετώπισης του προβλήματος νέου χρήστη είναι η απευθείας ερώτηση των χρηστών ώστε να παρέχουν βαθμολογήσεις σε αντικείμενα (άρθρα νέων για την περίπτωσή μας). Η προσέγγιση αυτή είναι σχετικά απλή: όταν ένας νέος χρήστης εγγράφεται στο σύστημα, του παρουσιάζονται αντικείμενα προς βαθμολόγηση. Τα αντικείμενα αυτά δεν είναι προτάσεις, αλλά επιλέγονται έτσι ώστε να συλλέγεται όσο τον δυνατόν περισσότερη πληροφορία για το προφίλ των χρηστών. Όσο ο χρήστης δίνει βαθμολογήσεις, το σύστημα αποφασίζει αν θα σταματήσει ή θα συνεχίσει τη διαδικασία, βελτιώνοντας στην δεύτερη περίπτωση όλο και περισσότερο το προφίλ του χρήστη. Παρόλα αυτά, τα μεγάλα ερωτηματολόγια έχουν και το αντίστοιχο κόστος: οι χρήστες ενοχλούνται σχετικά εύκολα και επομένως μπορεί να εγκαταλείψουν την διαδικασία βαθμολόγησης ή ακόμη χειρότερα, την διαδικασία εγγραφής. Ειδικά κιόλας αν οι ερωτήσεις έρχονται σε αντιπαράθεση με την ιδιωτικότητά τους. Όταν λοιπόν η παραπάνω διαδικασία τελειώσει, το σύστημα, έχοντας μία βασική γνώση για τις προτιμήσεις του χρήστη, ξεκινάει τις προτάσεις προς αυτόν. Η επιλογή ή μη των προτάσεων μπορεί να διαμορφώνει ένα βρόγχο ανάδρασης με το σύστημα το οποίο έτσι να ενημερώνει συνεχώς το προφίλ χρήστη.

Η παραπάνω διαδικασία ερωτήσεων και βαθμολογήσεων εισήχθηκε από τους Kohrs and Merialdo [143] οι οποίοι ερεύνησαν τη διάταξη των αντικειμένων σε σχέση με την διακύμανση και την εντροπία. Υπάρχουν δύο εξαιρετικά σημαντικές παράμετροι που καθορίζουν την πορεία της παραπάνω διαδικασίας: ποια αντικείμενα να επιλεχθούν για αξιολόγηση από το χρήστη και με ποια σειρά αυτά να προβληθούν. Πολλές προσεγγίσεις σχετικά με την διαδικασία επιλογής αντικειμένων έχουν προταθεί στη βιβλιογραφία. Κάθε μία από αυτές πρέπει να λάβει υπόψιν της συγκεκριμένες παραμέτρους, όπως η προσπάθεια που απαιτείται από τον χρήστη και η ικανοποίηση που λαμβάνει από την διαδικασία αξιολόγησης. Επίσης, η ακρίβεια προτάσεων, δηλαδή το πόσο καλές είναι οι επιλογές προς βαθμολόγηση.

Οι μεθοδολογίες σε σχέση με την διαδικασία ερωτήσεων και βαθμολογήσεων προς τον χρήστη χωρίζονται σε μη προσωποποιημένες και προσωποποιημένες [56]. Οι μη προσωποποιημένες περιλαμβάνουν:

- την τυχαία μέθοδο (random), όπου τα αντικείμενα προς βαθμολόγηση επιλέγονται με τυχαίο τρόπο με ομοιόμορφη πιθανότητα στο σύνολο των αντικειμένων. Αν η κατανομή των βαθμολογήσεων είναι κανονική, η συγκεκριμένη προσέγγιση έχει το πλεονέκτημα ότι καλύπτει το σύνολο των αντικειμένων

- την μέθοδο δημοφιλίας (popularity), όπου τα αντικείμενα διατάσσονται σε σειρά με βάση του πλήθους των αξιολογήσεων που τους έχουν δοθεί από όλους τους χρήστες. Παρότι εύκολη προς τους υπολογισμούς, η συγκεκριμένη προσέγγιση προάγει υπέρμετρα τα αντικείμενα τα οποία έχουν αξιολογηθεί από πολλούς χρήστες και ως εκ' τούτου φανερώνουν μικρή πληροφορία

- την μέθοδο εντροπίας (και παραλλαγές αυτής), οι οποίες βασίζονται στο γεγονός ότι συγκεκριμένα αντικείμενα μπορούν να φανερώσουν περισσότερη πληροφορία για τις προτιμήσεις του χρήστη. Γενικά ένα αντικείμενο που έχει ορισμένες αρνητικές και μερικές θετικές βαθμολογήσεις μπορεί να μας πει περισσότερα για τον χρήστη σε σχέση με ένα αντικείμενο που αρέσει σε όλους

- τις ζυγισμένες μεθόδους, οι οποίες αποτελούν συνδυασμό των μεθόδων δημοφιλίας και εντροπίας με την μορφή: $Popularity \times entropy$ ή $\log(Popularity \times entropy)$. Μια προσέγγιση αυτού του είδους, κάνοντας χρήση του θεωρήματος του Bayes, υποθέτει σιωπηλά ότι η δημοφιλία και η εντροπία είναι ανεξάρτητες μεταβλητές όσον αφορά στην επιλογή των αντικειμένων (κάτι που προφανώς δεν είναι πάντα σωστό)

- την "άπληστη" μέθοδο, όπου το επόμενο αντικείμενο επιλέγεται από εκείνα τα οποία ο χρήστης μπορεί να βαθμολογήσει, έτσι ώστε το σφάλμα πρόβλεψης για το σύνολο ελέγχου του να ελαχιστοποιείται. Εμφανώς αυτή η μέθοδος δεν έχει πρακτική αξία μίας και απαιτεί εκ' των προτέρων γνώση όχι μόνο για το τι ένας χρήστης μπορεί να βαθμολογήσει, αλλά και για το πως θα το βαθμολογήσει

- την "άπληστη" άλλων χρηστών μέθοδο - other people's greedy (και παραλλαγές αυτής), όπου τα αντικείμενα προς παρουσίαση στον χρήστη επιλέγονται από τα top-n της επιλεγμένης λίστας άλλων χρηστών.

Πρόσφατα, μία νέα μη προσωποποιημένη [79] και μία προσωποποιημένη [78] μεθοδολογία στοίχισης των αντικειμένων προτάθηκε από τους Golbandi et al. Επίσης στο [172] οι συγγραφείς κάνοντας χρήση μίας μεθόδου πρόβλεψης η οποία είναι μία παραλλαγή της παραγοντοποίησης πινάκων (matrix factorization), έδειξαν ότι πιο ακριβείς προβλέψεις μπορούν να γίνουν όταν ο χρήστης έχει δώσει ελάχιστες αξιολογήσεις, παρά όταν το σύστημα χρησιμοποιεί μετα-δεδομένα για τα αντικείμενα προκειμένου να κάνει προβλέψεις.

Οι προσωποποιημένες μεθοδολογίες από την άλλη μεριά, λαμβάνουν υπόψιν τις απαντήσεις τις οποίες ο χρήστης έχει δώσει στα αντικείμενα που ήδη έχουν παρουσιαστεί. Ορισμένες προσωποποιημένες μεθοδολογίες είναι οι εξής:

- αντικείμενο με αντικείμενο (item by item), όπου αρχικά τα αντικείμενα παρουσιάζονται με οποιαδήποτε άλλη μη προσωποποιημένη μεθοδολογία έως ότου μία βαθμολόγηση γίνει από τον χρήστη. Ύστερα από αυτό, οι προτάσεις για επόμενες βαθμολογήσεις γίνονται βασιζόμενοι σε κάποιο μέτρο ομοιότητας με το τι έχει ήδη αξιολογήσει ο χρήστης
- Naive Bayes, όπου με την γνώση για το αν ο χρήστης μπορεί να βαθμολογήσει ένα αντικείμενο, μπορούμε να υπολογίσουμε την Naive Bayes πιθανότητα να βαθμολογήσει τα υπόλοιπα αντικείμενα
- "διαταρασσόμενη άπληστη" άλλων χρηστών - perturbed other people's greedy, η οποία συνδυάζει την "άπληστη" άλλων χρηστών με την Naive Bayes μέθοδο.

Στο [177] παρουσιάζονται και αξιολογούνται αρκετές ακόμη προσωποποιημένες μεθοδολογίες για την βελτίωση της σειράς με την οποία παρουσιάζονται αντικείμενα στους χρήστες. Μία ακόμη προσέγγιση που έχει επιτυχώς χρησιμοποιηθεί για την αντιμετώπιση του προβλήματος νέου χρήστη είναι η παραγοντοποίηση πινάκων (matrix factorization) [122].

ΚΕΦΑΛΑΙΟ 4

ΑΡΧΙΤΕΚΤΟΝΙΚΗ

> In science, nothing is ever 100% proven.
>
> *Michio Kaku, American Physicist, 1947*

Στο παρόν κεφάλαιο παρουσιάζεται η αρχιτεκτονική του συστήματος προτάσεων (recommendation system) το οποίο αναπτύχθηκε κατά τη διάρκεια εκπόνησης της διδακτορικής διατριβής. Απεικονίζεται η ροή πληροφορίας των διαφόρων υποσυστημάτων, εξηγώντας πως αυτά αλληλεπιδρούν μεταξύ τους προκειμένου το τελικό αποτέλεσμα να είναι προτάσεις χρήσιμων άρθρων νέων προς τους χρήστες του συστήματος.

4.1 Στόχοι του συστήματος

Συχνά στις μέρες μας έχει παρατηρηθεί να μιλούμε για την ποιότητα στην ενημέρωση που παρέχει το διαδίκτυο. Ο κεντρικός στόχος του συστήματος που αναπτύχθηκε είναι να παρέχει ως έξοδο, στο χρήστη ή σε άλλα συστήματα, ποιοτική πληροφορία. Όπως έχει ήδη αναφερθεί στα προηγούμενα κεφάλαια, η πληροφορία του παγκοσμίου ιστού είναι σχεδόν χαοτική με αποτέλεσμα οι χρήστες να μην είναι εφικτό να προσεγγίσουν πληροφορία που τους είναι χρήσιμη και επιθυμητή. Σκοπός του συστήματός μας είναι να δημιουργήσουμε την κατάλληλη υποδομή ούτως ώστε να πραγματοποιείται φιλτράρισμα και να παράγονται προτάσεις για τα άρθρα νέων του διαδικτύου. Για να επιτευχθεί αυτό, αξιοποιούμε τεχνικές και αλγορίθμους από πολλά πεδία της επιστήμης των υπολογιστών και όχι μόνο.

Το σύστημά μας αντλεί και επεξεργάζεται περιεχόμενο που εντοπίζεται σε ειδησεογραφικούς δικτυακούς τόπους. Το περιεχόμενό τους παραλαμβάνεται σε συνεχή ρυθμό, και στη συνέχεια μπαίνει σε μία ακολουθιακή (pipelining) διαδικασία επεξεργασίας του, όπου: φιλτράρεται, αναλύεται, κατηγοριοποιείται, περιλήπτεται, συσταδοποιείται και στο τέλος προσωποποιείται στους χρήστες. Οι χρήστες επίσης συμμετέχουν στην διαδικασία μέσω συνεργατικού φιλτραρίσματος μιας και οι επιλογές τους οδηγούν το προτεινόμενο περιεχόμενο όχι μόνο προς αυτούς, αλλά και προς άλλους χρήστες που ανήκουν στις ίδιες συστάδες χρηστών.

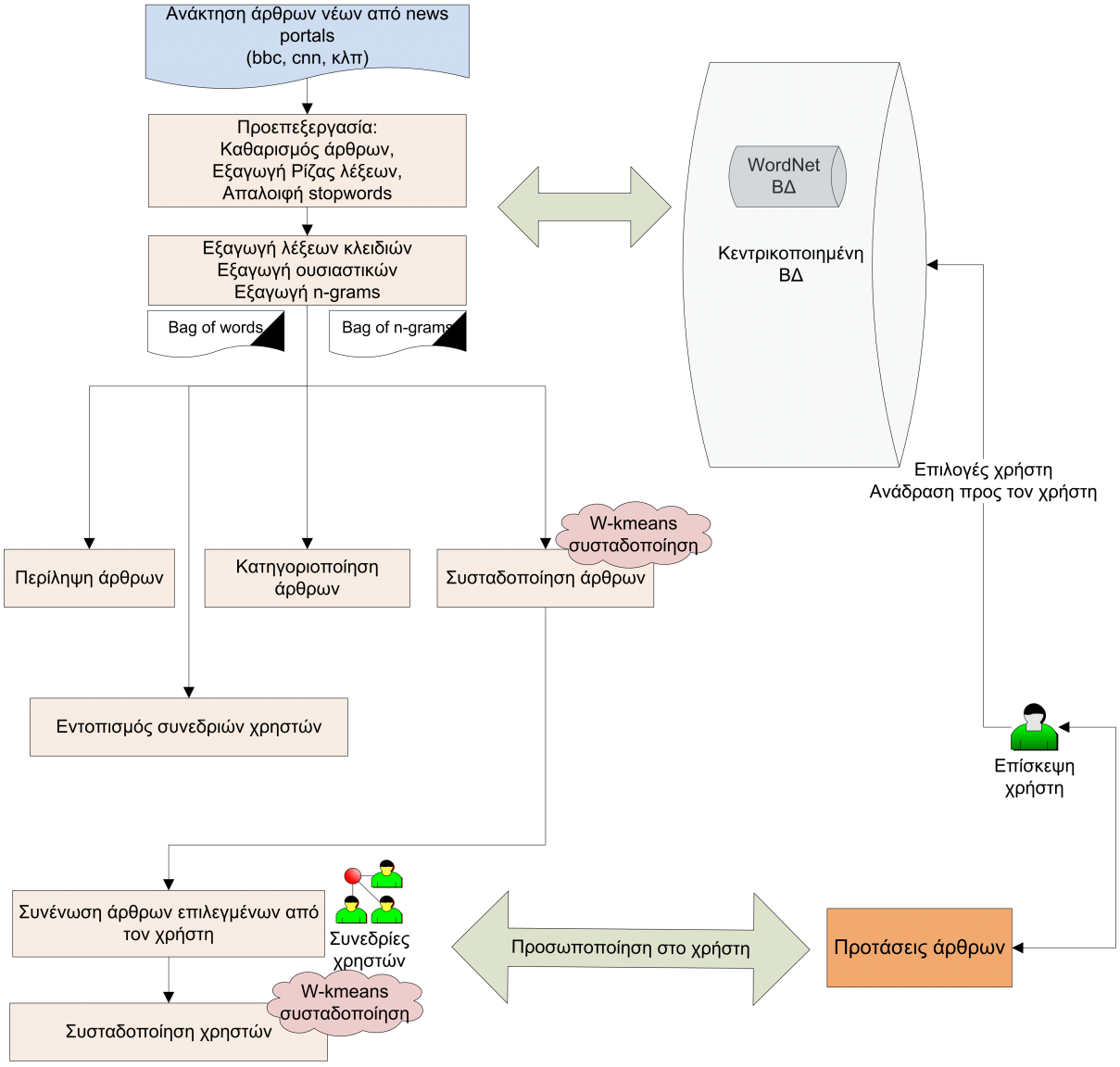
Ορισμένες από τις παραπάνω διεργασίες έχουν περιγραφεί διεξοδικά και στην μεταπτυχιακή διπλωματική εργασία μου [235], και ως εκ' τούτου, θα περιγραφούνε επιγραμματικά μόνο στο παρόν κεφάλαιο.

4.2 Γενική αρχιτεκτονική

Το σύστημα που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας είναι αρκετά πολύπλοκο και περιλαμβάνει αρκετά υποσυστήματα που επιτελούν τις επιμέρους λειτουργίες. Αποτελεί επομένως έναν τμηματοποιημένο μηχανισμό, κάθε κομμάτι του οποίου σχεδιάστηκε με σκοπό να μπορεί να λειτουργήσει και αυτόνομα ή, σε ορισμένες περιπτώσεις, ακόμα και να μπορεί να παρακαμφθεί (όπου αυτό απαιτείται). Η επιθυμητή αυτή ιδιότητα επιτυγχάνεται με τη χρήση της κοινής βάσης δεδομένων όπου αποθηκεύονται οι έξοδοι ενός συστήματος όπου αυτές αποτελούν εισόδους για κάποιο άλλο. Είναι επομένως εύκολο να αντικατασταθεί ένα τμήμα (module) του συστήματος από ένα νεότερο ή καλύτερο, όπως και να προστεθεί κάποιο ακόμα το οποίο θα χρησιμοποιεί υπάρχουσα πληροφορία από τη ΒΔ, δεδομένου φυσικά ότι θα χρησιμοποιεί την υπάρχουσα διεπαφή επικοινωνίας (communication interface). Η παραπάνω λογική σχεδίασης αναφέρεται συχνά ως modular και αποτελεί σημαντικό στοιχείο της αρχιτεκτονικής προσέγγισης κάθε συστήματος το οποίο σχεδιάζεται με την προοπτική επέκτασης του στο μέλλον.

4.3 Ροή Πληροφορίας

Η γενική αρχιτεκτονική τους συστήματος προτάσεων άρθρων νέων στο οποίο καταλήξαμε παρουσιάζεται στο σχήμα 9. Καθένα από αυτά τα υποσυστήματα που φαίνονται θα αναλυθεί στις ενότητες που ακολουθούν. Στην παρούσα ενότητα απλά αναφέρουμε συνολικά και επιγραμματικά τις λειτουργίες τους.



Σχήμα 9: Αρχιτεκτονική του συστήματος προτάσεων άρθρων νέων

Αρχικά, στο στάδιο εισόδου του, το σύστημά μας ανακτά άρθρα νέων που παράγονται από ειδησεογραφικά πρακτορεία του διαδικτύου. Αυτό αποτελεί μία offline διαδικασία η οποία επαναλαμβάνεται ανά τακτά χρονικά διαστήματα με χρήση ενός crawler. Ο συγκεκριμένος crawler, διαβάζει την λίστα από RSS feeds τα οποία υπάρχουν καταχωρημένα στη ΒΔ και στη συνέχεια ανακτά τα

άρθρα που αυτά αναφέρουν. Η συχνότητα αναζήτησης για ενημερώσεις στα RSS feeds, επομένως και η ανάκτηση των νέων άρθρων νέων, γίνεται κάθε 10 λεπτά. Η παραπάνω διαδικασία ανακτά σημαντικό όγκο ακατέργαστων δεδομένων τα οποία και αποθηκεύεται φυσικά στην ΒΔ προκειμένου να χρησιμοποιηθούν από τα υποσυστήματα που ακολουθούν. Κομμάτι της λειτουργικότητας του crawler είναι επίσης ο εντοπισμός του χρήσιμου κειμένου στις ανακτημένες ιστοσελίδες (π.χ. σώμα και τίτλος νέου, κ.λπ.).

Η προεπεξεργασία κειμένου αποτελεί μία κεντρική διαδικασία του συστήματος συνολικά, ίσης ή ίσως και μεγαλύτερης βαρύτητας των IR διαδικασιών που την ακολουθούν. Η προεπεξεργασία κειμένου εφαρμόζεται στο περιεχόμενο των ανακτημένων άρθρων και έχει ως αποτέλεσμα την εξαγωγή τόσο των λέξεων κλειδιών (keywords), όσο και των n-grams από τα οποία αποτελείται το κάθε άρθρο. Σε αυτό το επίπεδο ανάλυσης, εφαρμόζουμε ορισμένες τυπικές τεχνικές καθαρισμού κειμένου, στην οποίες περιλαμβάνονται:

- εύρεση ρίζας λέξεων (stemming)
- αφαίρεση stopwords

Παράλληλα με τα παραπάνω, χρησιμοποιούμε και ορισμένες τεχνικές που έχουν να κάνουν με:

- επιλογή/μείωση χαρακτηριστικών όπου επιχειρούμε να επιλέξουμε ένα υποσύνολο από τα χαρακτηριστικά τα οποία είναι πιο χρήσιμα για τις IR που ακολουθούν. Αυτό επιτυγχάνεται μέσω:
  - αντιστοίχιση μερών του λόγου (POS tagging) και πιο συγκεκριμένα, εύρεση των ουσιαστικών του κειμένου
  - "κλάδεμα" θορύβου ή ασήμαντων λέξεων οι οποίες εμφανίζονται με πολύ μικρή συχνότητα στο σύνολο των κειμένων (corpus). Οι λέξεις αυτές επομένως δεν εμπεριέχουν σημαντική νοηματική πληροφορία αναπαράστασης
- παραγωγή/εξαγωγή χαρακτηριστικών όπου νέα χαρακτηριστικά αναζητούνται για αναπαράσταση. Στην περίπτωσή μας αυτό επιτυγχάνεται με δύο τρόπους:
  - με την εξαγωγή των ουσιαστικών του κειμένου (POS tagging)
  - με την παραγωγή των δενδρικών δομών υπερωνύμων των λέξεων με χρήση της εξωτερικής βάσης γνώσης WordNet

Μετά τις παραπάνω τεχνικές προεπεξεργασίας κειμένου, ακολουθεί η εξαγωγή λέξεων κλειδιών, η οποία, κάνοντας χρήση του vector space μοντέλου, παράγει τον πίνακα όρων-συχνοτήτων του κειμένου (term-frequency vector). Ο πίνακας αυτός, ο οποίος περιγράφει το κάθε κείμενο σαν ένα σύνολο από λέξεις, ή αλλιώς "bag of words" (πίνακας λέξεων-συχνοτήτων) στις IR τεχνικές που ακολουθούν: κατηγοριοποίηση, περίληψη και συσταδοποίηση. Στην διδακτορική διατριβή ενισχύσαμε αυτή την αναπαράσταση με χρήση της εξωτερικής βάσης γνώσης WordNet, προκειμένου να βελτιώσουμε τα αποτελέσματα του αλγορίθμου συσταδοποίησης που ακολουθεί.

Παράλληλα, και κατ' αντίστοιχο τρόπο με αυτόν της εξαγωγής λέξεων κλειδιών, στην διδακτορική διατριβή προσθέσαμε μία νέα τεχνική παραγωγής χαρακτηριστικών η οποία κάνει χρήση των n-grams του κειμένου. Τα n-grams εξάγονται και δεικτοδοτούνται σε αυτό το σημείο ανάλυσης του κειμένου με τρόπο παρόμοιο με αυτόν της εξαγωγής λέξεων κλειδιών. Μάλιστα η εξαγωγή των keywords μπορεί να ιδωθεί ως η απλούστερη περίπτωση εξαγωγής n-grams, όπου $n = 1$.

Για κάθε άρθρο λοιπόν και για τιμές του n από 2 έως 6, εντοπίζουμε τα n-grams λέξεων του κειμένου και τα αποθηκεύουμε στη ΒΔ. Σε αυτή την περίπτωση, η συνολική ομοιότητα μεταξύ δύο άρθρων ή ενός άρθρου και μίας κατηγορίας ή συστάδας, δεν αποτυπώνεται μόνο σε σχέση με την μετρική συσχέτισης συχνότητας κειμένου/ανάστροφης συχνότητας σε όλα τα κείμενα, keyword frequency/inverse document frequency metric (kf-idf), αλλά πιο ακριβέστερα ως ο συνδυασμός της παραπάνω μετρικής και της αντίστοιχης n-grams μετρικής, έστω: gram frequency/inverse document frequency metric (gf-idf). Ο συνδυασμός των δύο αυτών μετρικών για ζύγιση της σημαντικότητας των λέξεων θα αναλυθεί στο επόμενο κεφάλαιο.

Ακολουθούν ορισμένα IR υποσυστήματα του μηχανισμού και τα οποία αφορούν στην κατηγοριοποίηση και εξαγωγή περίληψης του κειμένου. Τα υποσυστήματα αυτά δεν θα μας απασχολήσουν στα πλαίσια της διδακτορικής διατριβής και αναφέρονται απλά και μόνο διότι αποτελούν μέρος του συνολικού συστήματος. Σημαντικό ίσως εδώ είναι να αναφέρουμε ότι το υποσύστημα κατηγοριοποίησης αλληλεπιδρά με αυτό της εξαγωγής περίληψης προκειμένου να το υποβοηθήσει όσον αφορά στην βελτίωση της ποιότητας των εξαγόμενων περιλήψεων [235].

Η ενισχυμένη λίστα από χαρακτηριστικά που προκύπτει από την προεπεξεργασία κειμένου, τροφοδοτεί τον W-kmeans αλγόριθμο συσταδοποίησης που ακολουθεί. Είναι σημαντικό να αναφέρουμε όμως ότι η διαδικασία (αλγόριθμος) συσταδοποίησης είναι ανεξάρτητη από τα υπόλοιπα βήματα και επομένως θα μπορούσε εύκολα να αντικατασταθεί από μία άλλη διαδικασία στο μέλλον. Ο W-kmeans αποτελεί μία καινοτόμα προσέγγιση στο πρόβλημα της συσταδοποίησης επεκτείνοντας τον κλασικό αλγόριθμο συσταδοποίησης k-means. Ο W-kmeans κάνει χρήση της εξωτερικής γνώσης από τα υπερώνυμα του WordNet ενισχύοντας την "bag of words" αναπαράσταση των κειμένων.

Ακολουθώντας τις βασικές IR διεργασίες του μηχανισμού μας βρίσκεται ο αλγόριθμος προσωποποίησης. Ο αλγόριθμος μπορεί εύκολα να προσαρμοστεί σε λεπτές αλλαγές όσον αφορά στις προτιμήσεις των χρηστών. Αυτές οι αλλαγές, οι οποίες εκφράζονται μέσω της συμπεριφοράς πλοήγησης των χρηστών, εντοπίζονται και διαρκώς προσαρμόζουν το προφίλ του χρήστη όπου αυτό είναι απαραίτητο. Ο αλγόριθμος προσωποποίησης χρησιμοποιεί μία πληθώρα πληροφοριών που έχουν να κάνουν με τον χρήστη προκειμένου τελικά να φιλτράρει τα αποτελέσματα σε αυτόν, προτείνοντας τελικά μόνο ότι θεωρεί πως ταιριάζει καλύτερα στο προφίλ του. Επιπλέον, λαμβάνει υπόψιν του με έναν ζυγισμένο τρόπο την πληροφορία η οποία πηγάζει από τις προηγούμενες IR τεχνικές, την κατηγοριοποίηση, την περίληψη, καθώς και την συσταδοποίηση άρθρων νέων.

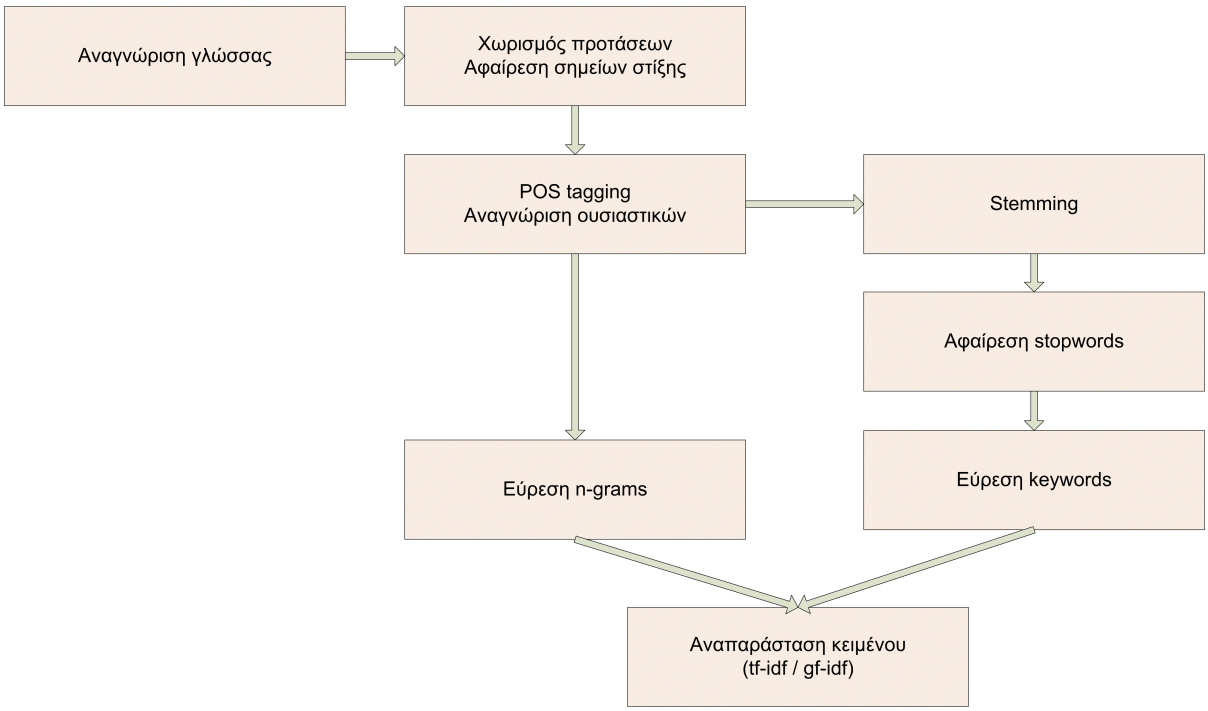
Τα προφίλ από πολλαπλούς χρήστες και χρονικά πλαίσια επίσης συσταδοποιούνται με χρήση του αλγορίθμου W-kmeans παράγοντας έτσι συστάδες χρηστών. Ο W-kmeans για την περίπτωση της συσταδοποίησης χρηστών ενισχύει τα προφίλ χρήστη με υπερώνυμα του εξάγονται από την

βάση γνώσης WordNet μέσω ενός ευρετικού τρόπου ο οποίος θα αναλυθεί στη συνέχεια. Αυτές οι συστάδες από προφίλ χρηστών επίσης χρησιμοποιούνται (παράλληλα με την παραπάνω πληροφορία) στη φάση παραγωγής προτάσεων προς τον χρήστη, προκειμένου να βελτιώσουν την ευχρηστία και αποτελεσματικότητα του συστήματος προτείνοντας έτσι πιο προσαρμοσμένα αποτελέσματα στους χρήστες που επανεπισκέπτονται το σύστημα.

Όταν λοιπόν ένας χρήστης επιστρέφει, το συσταδοποιημένο προφίλ του χρήστη ανακτάται και άρθρα τα οποία ταιριάζουν στο προφίλ αυτό εξάγονται και αξιολογούνται προς πρόταση για τον χρήστη.

4.3.1 Προεπεξεργασία κειμένου

Ο μηχανισμός προεπεξεργασίας κειμένου είναι ένα σημαντικό τμήμα του συνολικού μηχανισμού ο οποίος αναλαμβάνει το καθάρισμα του σώματος του κειμένου και καταλήγει στην εξαγωγή λέξεων κλειδιών και n-grams. Η διαδικασία της προεπεξεργασίας κειμένου φαίνεται στο Σχήμα 10. Η είσοδος στο υποσύστημα αυτό από τα δεδομένα της ΒΔ περιέχει τα απαραίτητα μόνο στοιχεία: τίτλος και σώμα κειμένου.



Σχήμα 10: Προεπεξεργασία κειμένου που οδηγεί στην εξαγωγή keywords και n-grams

Εκτός από τις παραπάνω εισόδους, ο μηχανισμός δέχεται ορισμένες παραμέτρους λειτουργίας, κάτι που μας επιτρέπει τόσο να μεταβάλλουμε εύκολα την λειτουργία του, όσο και να αξιολογήσουμε στη συνέχεια τις επιδόσεις για διάφορες τιμές των εισόδων αυτών. Οι παράμετροι του μηχανισμού προεπεξεργασίας κειμένου είναι:

- το ελάχιστο μήκος λέξης (οι λέξεις που είναι μικρότερες από αυτό το μήκος θα αφαιρεθούν)

- καθορισμός εάν τα αριθμητικά δεδομένα θα κρατηθούν ή θα αφαιρεθούν
- καθορισμός μιας λίστας από λέξεις τετριμμένες και συνηθισμένες οι οποίες δεν εκφράζουν κάποιο συγκεκριμένο νόημα και μπορούν να θεωρηθούν ως "σκουπίδια" (stopwords)
- καθορισμός του αλγορίθμου stemming που θα χρησιμοποιηθεί για τις λέξεις κλειδιά
- καθορισμός της βαρύτητας που δίνεται στα ουσιαστικά του κειμένου (αν αυτά ζυγίζουν περισσότερο)
- καθαρισμός των λέξεων που εμφανίζονται με μικρή συχνότητα ($<0.01\%$) στην ΒΔ (και ως εκ' τούτου πιθανότατα αποτελούν σκουπίδια)
- εύρος της τιμής n για τον καθορισμό των n-grams του κειμένου

Η διαδικασία που ακολουθείται από τον μηχανισμό προεπεξεργασίας κειμένου έχει ως εξής. Αρχικά, η γλώσσα του κειμένου αναγνωρίζεται κάτι που γίνεται είτε με ειδικό λογισμικό αναγνώρισης είτε έμμεσα χρησιμοποιώντας την προκαθορισμένη γλώσσα του RSS feed από το οποίο προέρχεται το άρθρο. Ακολουθεί η διαδικασία χωρισμού των προτάσεων, ο ορθογραφικός έλεγχος, και έπειτα η αφαίρεση των σημείων στίξης που υπάρχουν. Στη συνέχεια λαμβάνει χώρα η διεργασία αναγνώρισης των ουσιαστικών του κειμένου χρησιμοποιώντας τον POS SVM-based tagger από το [77] ο οποίος μπορεί να καθορίσει με μεγάλη ακρίβεια τα ουσιαστικά που περιέχει η κάθε πρόταση. Μερικές κοινότυπες τεχνικές εξαγωγής λέξεων κλειδιών ακολουθούν με σκοπό να περιοριστεί ο θόρυβος των αποτελεσμάτων: η αφαίρεση των stopwords και το stemming. Είναι σημαντικό να τονιστεί ότι η διαδικασία εύρεσης των ουσιαστικών του κειμένου πρέπει να προηγείται αυτών των διεργασιών αν επιθυμούμε να επιτύχει με μεγάλη πιθανότητα, μιας και οι λέξεις μπορούν εύκολα να αντιστοιχιστούν με μέρη του λόγου μέσα στην πρόταση στην οποία ανήκουν. Ένα εξίσου σημαντικό στοιχείο είναι ότι οι διαδικασίες της αναγνώρισης των ουσιαστικών, της αφαίρεσης των stopwords και του stemming είναι ισχυρά εξαρτώμενες από την γλώσσα του κειμένου. Γνωρίζοντας επομένως την γλώσσα του κειμένου (κάτι που γίνεται όπως είπαμε στα αρχικά στάδια), μπορούμε να λάβουμε τις σωστές αποφάσεις προεπεξεργασίας του: να αποφασίσουμε ποια θα πρέπει να είναι η λίστα με τα stopwords που θα πρέπει να αφαιρεθούν, ποιοι θα πρέπει να είναι οι κανόνες για το POS tagging που θα εφαρμόσει ο SVM tagger, ποιοι θα είναι οι κανόνες για την διαδικασία stemming που θα εφαρμοστεί και τελικά ποιο θα είναι το μέγεθος των αρχικών λέξεων που θα πρέπει να κρατηθούν, μιας και ορισμένες γλώσσες περιέχουν κατά κόρων μεγαλύτερες λέξεις από κάποιες άλλες.

Τα παραπάνω αφορούν το δεξί σκέλος του σχήματος 10. Παρόμοιες διαδικασίες ακολουθούνται και για την εύρεση των n-grams του κειμένου (αριστερό σκέλος του σχήματος 10) με την βασική διαφορά ότι η εξαγωγή της ρίζας των λέξεων (stemming) καθώς και η αφαίρεση των stopwords δεν προηγείται της εξαγωγής n-grams. Για την ακρίβεια, οι τεχνικές αυτές δεν έχουν εφαρμογή πέρα από συστήματα που βασίζονται μόνο σε εξαγωγή λέξεων. Και ο λόγος είναι απλός: σχεδόν όλα τα n-gram που μπορεί να εξαχθούν βασίζονται ακριβώς στα stopwords που συνδέουν ορισμένες

λέξεις, (για παράδειγμα: president of the United States) καθώς και στις καταλήξεις των λέξεων που απαρτίζουν τα n-grams.

Τα παραπάνω χαρακτηριστικά προσδίνουν την content-based φύση του συστήματος, μιας και η ανάλυση που περιγράφηκε μέχρις στιγμής γίνεται αποκλειστικά και μόνο με χρήση του κειμενικού περιεχομένου του ιδίου του κειμένου των άρθρων.

Η έξοδος του μηχανισμού προεπεξεργασίας κειμένου αποθηκεύεται στη βάση δεδομένων του συστήματος, και έπειτα διαβάζεται από τα υποσυστήματα που ακολουθούν. Στις εξόδους περιλαμβάνονται:

- οι λέξεις κλειδιά που προέκυψαν από την διαδικασία του keyword extraction
- τα n-grams που προέκυψαν από την διαδικασία του gram extraction
- τις θέσεις των keywords και των n-grams στο αρχικό κείμενο, σε ποιες προτάσεις δηλαδή εμφανίζονται
- το πλήθος με το οποίο εμφανίζονται τα keywords και τα n-grams κάτι που εκφράζεται είτε ως απόλυτη συχνότητα εμφάνισης (π. χ. ένα keyword εμφανίζεται 5 φορές στο κείμενο), είτε ως σχετική συχνότητα εμφάνισης (π. χ. ένα n-gram εμφανίζεται 5 φορές σε ένα κείμενο 50 n-grams, άρα με σχετική συχνότητα 0,1).
- την πληροφορία για το αν το keyword είναι ουσιαστικό ή όχι

Τα παραπάνω αναπαριστώνται μέσω πινάκων στο vector space μοντέλο: term frequency - inverse document frequency (tf-idf) για την περίπτωση των λέξεων κλειδιών, και gram frequency - inverse document frequency (gf-idf) για την περίπτωση των n-grams. Οι πίνακες αυτοί αποθηκεύονται στην βάση δεδομένων και αξιοποιούνται από τις διαδικασίες του επόμενου επιπέδου.

4.3.2 Συσταδοποίηση

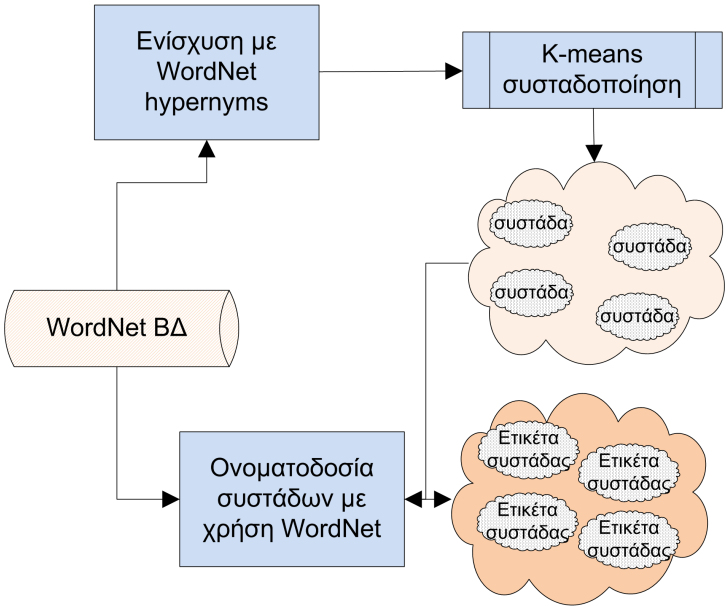
Η συσταδοποίηση αποτελεί μία από τις βασικές διεργασίες πυρήνα του συστήματος προτάσεων που αναπτύχθηκε. Ο αλγόριθμος συσταδοποίησης που αναπτύχθηκε ονομάζεται W-kmeans (WordNet-enabled k-means) ο οποίος και παρουσιάζεται στη συνέχεια.

4.3.2.1 Συσταδοποίηση W-kmeans

Ο αλγόριθμος συσταδοποίησης W-kmeans εξερευνά την υπόθεση ότι η ενσωμάτωση λεξικολογικής πληροφορίας στην αναπαράσταση κειμένου, μπορεί να οδηγήσει σε βελτιώσεις σχετικά με την ακρίβεια συσταδοποίησης. Αυτό ισχύει είτε έχουμε να κάνουμε με άρθρα νέων, είτε με χρήστες προς συσταδοποίηση, κάτι που κάνει τον αλγόριθμο να δρα με τον ίδιο τρόπο, ανεξάρτητα από την είσοδο (πίνακες λέξεων κλειδιών άρθρων και πίνακες λέξεων κλειδιών προφίλ χρηστών αντίστοιχα).

Στον πυρήνα του W-kmeans βρίσκεται ο αλγόριθμος k-means ο οποίος ενισχύεται ώστε να κάνει χρήση ενός ευρετικού που βασίζεται στη βάση γνώσης WordNet. Πιο συγκεκριμένα, κάνει

χρήση της εξωτερικής βάσης γνώσης υπερωνύμων του WordNet προκειμένου να ενισχύσει την αναπαράσταση "bag of words" που προκύπτει από το υποσύστημα προεπεξεργασίας κειμένου στο στάδιο εισαγωγής του. Η ενισχυμένη λίστα χαρακτηριστικών που προκύπτει οδηγεί τον αλγόριθμο k-means ο οποίος κάνοντας χρήση της μετρικής ομοιότητας συνημιτόνου παράγει τις συστάδες των αντικειμένων (άρα τυπικά, πρόκειται για τον αλγόριθμο spherical k-means (s-kmeans)).



Σχήμα 11: Συσταδοποίηση άρθρων νέων και χρηστών

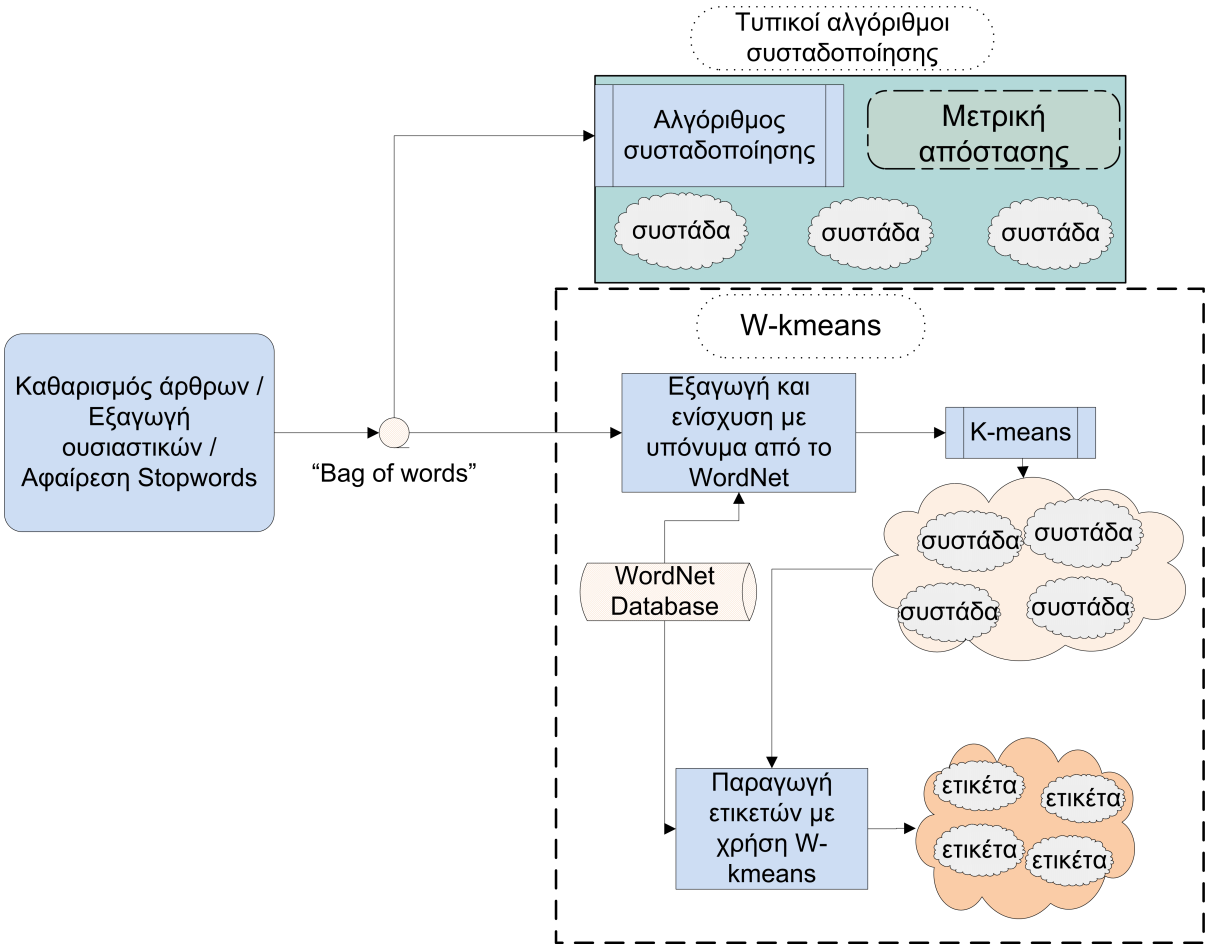
Όπως φαίνεται και στο χρήμα 11, η συσταδοποίηση άρθρων νέων και η συσταδοποίηση χρηστών αποτελούν δύο διαφορετικές διεργασίες του συστήματος που όμως χρησιμοποιούν τον ίδιο πυρήνα (αλγόριθμο) προκειμένου να παράγουν την έξοδό τους (τις συστάδες τους). Οι παραγόμενες συστάδες στο τέλος οδηγούνται προς την διαδικασία εξαγωγής ετικετών / ονοματοδοσίας συστάδων (labelling), η οποία και αντιστοιχίζει μία ή περισσότερες λέξεις κλειδιά σε κάθε συστάδα. Αυτές οι λέξεις, εν' τέλει, αντιπροσωπεύουν διαισθητικά και σε ανθρώπινη γλώσσα την κάθε μία συστάδα και αποτελούν έναν φυσικό τρόπο κατανόησης των περιεχομένων των συστάδων που προκύπτουν.

4.3.2.2 Συσταδοποίηση άρθρων νέων

Η διαδικασία συσταδοποίησης άρθρων νέων απεικονίζεται στο σχήμα 12, με τις διεργασίες στο τετραγωνισμένο κουτί να αποτελούν τα θεμελιώδη βήματα του W-kmeans αλγορίθμου (όπως παρουσιάστηκε και στο σχήμα 11). Αρχικά, ένας τυπικός αλγόριθμος συσταδοποίησης δέχεται την έξοδο του συστήματος προεπεξεργασίας και βάσει της δεδομένης μετρικής απόστασής του, προχωράει στην εξαγωγή συστάδων από τα κείμενα. Βάσει της γενικής αυτής ροής αξιολογούνται διάφοροι αλγόριθμοι συσταδοποίησης στην ενότητα 7.2.1.

Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος W-kmeans για την περίπτωση της συσταδοποίησης άρθρων νέων, δέχεται ως είσοδο την έξοδο του μηχανισμού προεπεξεργασίας και συγκεκριμένα τις λέξεις κλειδιά του κειμένου καθώς και τις σχετικές συχνότητες εμφάνισης αυτών

στα κείμενα προς συσταδοποίηση, σε σχέση πάντα με τη συνολική συχνότητα εμφάνισης τους στα κείμενα της ΒΔ (BOW αναπαράσταση). Έχοντας αυτές τις πληροφορίες, εξάγει για κάθε μία από τις λέξεις κλειδιά του κάθε κειμένου προς συσταδοποίηση το δέντρο υπερώνυμων, όπως αυτό δίνεται από το WordNet. Τα αυτόνομα δέντρα υπερωνύμων έπειτα προστίθενται, παράγοντας έτσι ένα αθροιστικό δέντρο για κάθε κείμενο. Ακολουθεί η εφαρμογή πάνω στο σύνολο των keywords και των υπερωνύμων του αλγορίθμου k-means, απ' όπου εξάγονται οι συστάδες των άρθρων νέων. Η ενίσχυση των χαρακτηριστικών των κειμένων κατ' αυτόν τον τρόπο βελτιώνει την ποιότητα της συσταδοποίησης αισθητά, όπως θα δούμε και σε επόμενα κεφάλαια. Παράλληλα, εξάγονται και οι ετικέτες που χαρακτηρίζουν την κάθε συστάδα - πάλι με χρήση των υπερωνύμων του Wordnet. Οι αναθέσεις άρθρων σε συστάδες, καθώς και ετικετών στις συστάδες, αποτελούν επομένως τις εξόδους του υποσυστήματος συσταδοποίησης άρθρων νέων οι οποίες και αποθηκεύονται στη ΒΔ.



Σχήμα 12: Συσταδοποίηση άρθρων νέων - τυπικοί αλγόριθμοι και W-kmeans

4.3.2.3 Μοντελοποίηση και συσταδοποίηση χρηστών

Για κάθε χρήστη που διαβάζει άρθρα νέων από το σύστημα, κρατάμε τις ενέργειες του οι οποίες χαρακτηρίζουν μία συνεδρία χρήστη. Προκειμένου να συνδέσουμε το υποσύστημα συσταδοποί-

ησης χρηστών με τον αλγόριθμο προσωποποίησης στο χρήστη, χρησιμοποιούμε την λογική των συνεδριών χρήστη (user sessions). Μία συνεδρία λοιπόν, ορίζεται ως η λίστα από επιλεγμένα άρθρα τα οποία ο χρήστης αποφάσισε να δει για μία ελάχιστη χρονική περίοδο και μέσα σε ένα περιορισμένο χρονικό παράθυρο συνεχής διάρκειας, παράμετροι οι οποίες προσαρμόζονται και αποτιμούνται κατάλληλα με βάσει τη πειραματική αξιολόγηση του υποσυστήματος. Τα επιλεγμένα άρθρα που συμπεριλαμβάνονται σε αυτές τις συνεδρίες ενώνονται εν' συνεχεία σε επίπεδο λέξεων κλειδιών, παράγοντας έτσι ένα χρονικά φραγμένο προφίλ χρήστη. Τα προφίλ από πολλαπλούς χρήστες και χρονικές περιόδους συσταδοποιούνται συνεχώς από το σύστημα με χρήση του W-kmeans αλγορίθμου, παράγοντας έτσι συστάδες από προφίλ οι οποίες και αποθηκεύονται στη ΒΔ.

Όπως είναι σαφές από τα παραπάνω, η συσταδοποίηση των χρηστών ανάγεται στο πρόβλημα της συσταδοποίησης επιλεγμένων άρθρων που ανήκουν σε συνεδρίες χρηστών και όχι στην αυστηρή αντιστοίχιση keywords με χρήστες όπως προτείνεται από πολλές τεχνικές της βιβλιογραφίας. Η προσέγγιση αυτή προσφέρει μεγαλύτερη ευελιξία όσον αφορά στα μεταβαλλόμενα ενδιαφέροντα των χρηστών του συστήματος, τα οποία και έχουν με αυτό τον τρόπο άμεση απεικόνιση στις επίκαιρες συστάδες χρηστών που εξάγονται.

4.3.2.4 Υπολογισμός πλήθους συστάδων

Όπως έχει αναφερθεί, ένα βασικό πρόβλημα της οικογένειας αλγορίθμων k-means είναι η εκ' των προτέρων ανάγκη καθορισμού του πλήθους των υποκείμενων συστάδων που βρίσκονται στα προς συσταδοποίηση δεδομένα. Κάτι τέτοιο όμως σπάνια είναι γνωστό, ειδικά κιόλας στην περίπτωση συσταδοποίησης άρθρων νέων ή μεταβαλλόμενων προφίλ χρηστών που μας αφορά: τα μεν άρθρα νέων καταφθάνουν με γοργούς ρυθμούς χωρίς κάποια πρότερη γνώση για το τι περιέχουν, οι δε συστάδες των προφίλ χρηστών μεταβάλλονται επίσης συχνά αναλόγως με τα ενδιαφέροντα του χρήστη στην συγκεκριμένη φραγμένη χρονική περίοδο. Ως εκ' τούτου, και δεδομένου ότι μία τυχαία επιλογή πλήθους συστάδων έχει συνήθως αρνητικά αποτελέσματα (εκτός βέβαια και αν βρίσκεται εξαιρετικά κοντά στο πραγματικό πλήθος συστάδων των δεδομένων), χρησιμοποιούμε έναν συνδυασμό της μεθόδου του εμπειρικού κανόνα και της μεθόδου του αγκώνα για την εκτίμηση του πλήθους των συστάδων. Οι μέθοδοι αυτοί περιγράφηκαν στην ενότητα 3.7.3.

Πιο συγκεκριμένα, δεδομένου ότι μία γενική εκτίμηση, και μάλιστα αρκετά κοντά στις περισσότερες των περιπτώσεων, μπορεί να γίνει με τον εμπειρικό κανόνα, αρχίζουμε βασιζόμενοι με αυτόν. Στη συνέχεια, κρατώντας τις μισές συστάδες αυτής της προσέγγισης ξεκινάμε εκτελέσεις του αλγορίθμου W-kmeans αυξάνοντας σταδιακά το πλήθος των συστάδων και μετρώντας την μετρική RSoS (11). Η οριακή κλίση της φθίνουσας τιμής RSoS όσο το πλήθος συστάδων αυξάνεται είναι η τιμή που κρατάμε όσον αφορά το πλήθος των συστάδων. Η απόφασή μας να χρησιμοποιήσουμε τον συνδυασμό των δύο τεχνικών έχει να κάνει με τις εξής παρατηρήσεις:

- ο εμπειρικός κανόνας συχνά δίνει μια καλή εκτίμηση του πλήθους των συστάδων. Δεν είναι όμως σπάνιο το φαινόμενο η εκτίμηση αυτή να απέχει λίγο έως πολύ από το πραγματικό πλήθος

- σε μεγάλο πλήθος δεδομένων, το να χρησιμοποιήσουμε απλά την μέθοδο του αγκώνα αρχίζοντας από $k = 2$ συστάδες (η περίπτωση $k = 1$ ποτέ δεν αποτελεί λύση) θα σήμαινε συχνά πάρα πολλές εκτελέσεις του αλγορίθμου W-kmeans μιας και τα δεδομένα μας είναι πολλά. Κάτι τέτοιο θα ήταν πρακτικά αδύνατο για ένα σύστημα που θέλει σχετικά γρήγορα να καταναλώνει την πληροφορία που παράγεται στο διαδίκτυο.

Ως εκ' τούτου καταλήξαμε στην επιλογή του να ξεκινάμε με τις μισές συστάδες από αυτές που μας προτείνει ο εμπειρικός κανόνας και συνεχίζοντας αυξητικά να προσπαθήσουμε να εκτιμήσουμε το πραγματικό πλήθος. Προφανώς η συσταδοποίηση η οποία θα δώσει και την οριακή κλίση στην μείωση της τιμής RSoS για τα δεδομένα μας είναι και το αποτέλεσμα που εν' τέλει ψάχνουμε.

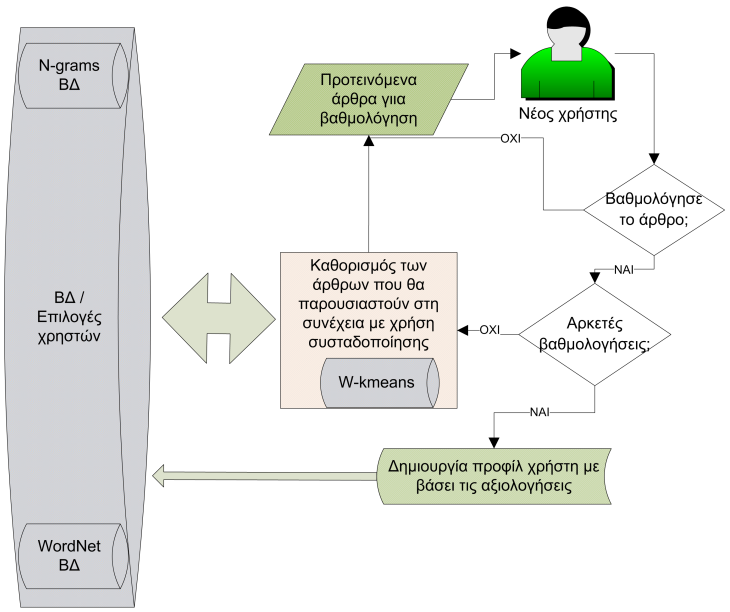
4.3.3 Πρόβλημα νέου χρήστη

Η προσέγγισή μας προκειμένου να αντιμετωπίσουμε το πρόβλημα του νέου χρήστη για το σύστημα προτάσεων που αναπτύχθηκε, βασίζεται στην οργάνωση των άρθρων σε συστάδες παρόμοιων ενδιαφερόντων. Με αυτόν τον τρόπο, συμμετρικά, μπορεί κάποιος να οργανώσει άρθρα με βάση του ποιος τα έχει δει και να αντιστοιχίσει έτσι ομάδες άρθρων σε χρήστες. Για παράδειγμα, έστω μία ομάδα από άρθρα τα οποία έχουν να κάνουν με πολιτική και πιο συγκεκριμένα, με τον Obama. Ένας χρήστης που έχει προηγούμενα εκδηλώσει κάποιο ενδιαφέρον προς την ανάγνωση άρθρων σχετικών με τον Obama ή το δημοκρατικό κόμμα, μπορεί δυνητικά να ενδιαφέρεται για άρθρα της παραπάνω ομάδας. Η αντίστροφη προσέγγιση είναι επίσης πιθανή: έστω μία ομάδα χρηστών οι οποίοι έχουν προηγουμένως εκφράσει το ενδιαφέρον τους για αυτό το θέμα. Ένα πρόσφατο άρθρο με ομοιότητες με κάποια από τα άρθρα που έχουν προηγουμένως διαβάσει κάποια από τα μέλη αυτής της ομάδας μπορεί δυνητικά να είναι ενδιαφέροντα και για τους υπόλοιπους χρήστες αυτής της ομάδας. Η παραπάνω προσέγγιση υποθέτει ότι αντί να βασιζόμαστε στις αποφάσεις μεμονωμένων χρηστών, η συστάδα ενσωματώνει συνολικά την απαραίτητη πληροφορία. Δύο ευρύτερα χρησιμοποιούμενες τεχνικές για το παραπάνω σενάριο παραδοσιακά είναι η k-NN και οι συσταδοποίηση, ή κάποιος συνδυασμός των δυο αυτών [166] [136].

Όταν ένας νέος χρήστης εγγράφεται στο σύστημα, ακολουθεί μία διαδικασία κατά την οποία το σύστημα προσπαθεί, μέσω προβολής άρθρων νέων προς βαθμολόγηση, να εντοπίσει τα ενδιαφέροντα του χρήστη. Η ροή πληροφορίας για την διαδικασία εγγραφής παρουσιάζεται στο σχήμα 13.

Όπως φαίνεται και στο σχήμα, όσο ο χρήστης δεν έχει δώσει αρκετές αξιολογήσεις σχετικά με τα ενδιαφέροντά του, προκειμένου να σύστημα να "τον γνωρίσει" επαρκώς, η διαδικασία επιλογής άρθρων και παρουσίασή τους προς αξιολόγηση συνεχίζεται. Όταν έχουν συλλεχθεί τουλάχιστον 5 βαθμολογήσεις άρθρων, η διαδικασία εγγραφής (registration - user prompting) στο σύστημα ολοκληρώνεται με ένα αρχικό προφίλ χρήστη να έχει διαμορφωθεί.

Όσον αφορά στον τρόπο επιλογής των προς παρουσίαση άρθρων στον χρήση για βαθμολόγηση, εστιάζουμε σε μία προσωποποιημένη μεθοδολογία, παρόμοια με την item by item στρατηγική που περιγράφηκε στην ενότητα 3.9.1. Η προσέγγισή μας εκμεταλλεύεται την συσταδοποίση μέσω του αλγορίθμου W-kmeans τόσο των άρθρων όσο και των χρηστών του συστήματος, προκειμένου να επιλεγούν τα άρθρα νέων τα οποία πρόκειται να παρουσιαστούν. Στο κεφάλαιο 7 αξιολογούμε επίσης



Σχήμα 13: Ροή πληροφορίας κατά την εγγραφή νέου χρήστη

και την αποτελεσματικότητα της συγκεκριμένης στρατηγικής σε σύγκριση με τις βασικότερες από τις προαναφερθείσες της ενότητας 3.9.1.

Κατ' αρχάς, επιθυμούμε ο συνολικός αριθμός άρθρων που θα παρουσιαστούν στον χρήστη να ελαχιστοποιηθεί, ενώ παράλληλα να συλλέξουμε όσο περισσότερη πληροφορία για τον χρήστη μπορούμε. Όπως αναφέρθηκε ήδη, ο καθορισμός των άρθρων που θα παρουσιαστούν καθώς και η σειρά τους είναι καίριας σημασίας για την αποτελεσματικότητα της διαδικασίας. Αρχικά λοιπόν, επιλέγουμε προς παρουσίαση άρθρα με βάσει οποιαδήποτε στρατηγική. Παρότι θα περιμέναμε η επιλεγμένη στρατηγική να παίζει κάποιο ρόλο, αυτό δεν φαίνεται να κάνει κάποια διαφορά, όπως δείχθηκε και στο [178]. Ως εκ' τούτου, επιλέγουμε άρθρα τα οποία έρχονται από τα πιο συχνά αξιολογημένα του συστήματος - εκείνα δηλαδή που έχουν τις περισσότερες, αριθμητικά, αξιολογήσεις. Όσο τα άρθρα που παρουσιάζονται στον χρήστη δεν αξιολογούνται, συνεχίζουμε τις προτάσεις βάση αυτής της προαναφερθείσας λίστας. Μόλις ο χρήστης βαθμολογήσει ένα άρθρο, κάνουμε χρήση των δεδομένων συσταδοποίησης προκειμένου να εντοπίσουμε και να προτείνουμε στη συνέχεια άλλα άρθρα που:

- ανήκουν στην ίδια συστάδα άρθρων με αυτό που αξιολογήθηκε
- έχουν επιλεχθεί από χρήστες οι οποίοι έχουν επίσης αξιολογήσει το ίδιο άρθρο με παρόμοιο τρόπο στο παρελθόν

Η παραπάνω διαδικασία συνεχίζεται έως ότου αρκετές αξιολογήσεις άρθρων έχουν συγκεντρωθεί για το χρήστη. Με το πέρας της, το αρχικό προφίλ χρήστη έχει δημιουργηθεί και το σύστημα είναι έτοιμο να μπει στην τυπική λειτουργία προτάσεών του.

4.3.4 Προσωποποίηση στο χρήστη

Ακολουθώντας τις βασικές IR διεργασίες του μηχανισμού μας, το υποσύστημα προσωποποίησης και γενικότερα παραγωγής προτάσεων ακολουθεί. Είναι προφανές πως το υποσύστημα αυτό αποτελεί ένα καίριο τμήμα του συστήματος προτάσεων που αναπτύχθηκε. Το σύστημα προσωποποίησης προσαρμόζεται εύκολα στον τελικό χρήστη, μιας και μικρές αλλαγές στις προτιμήσεις του, όπως εκφράζονται από την συμπεριφορά πλοήγησής του εντοπίζονται και προσαρμόζουν συνεχώς το προφίλ του στο σύστημα. Ο αλγόριθμος προσωποποίησης χρησιμοποιεί μία πληθώρα από συσχετιζόμενες με τον χρήστη πληροφορίες προκειμένου να φιλτράρει τα αποτελέσματα που παρουσιάζονται προς αυτόν. Μεταξύ αυτών, οι χρόνοι που ο χρήστης ξοδεύει διαβάζοντας τα άρθρα ή τις περιλήψεις τους, η συστάδα χρηστών στην οποία ανήκει, κ. α. Επιπλέον, λαμβάνει υπόψιν του, με ζυγισμένο τρόπο, την πληροφορία η οποία πηγάζει από τα προηγούμενα επίπεδα ανάλυσης του συστήματος σε σχέση με την κατηγοριοποίηση/συσταδοποίηση καθώς και την συσταδοποίηση άρθρων/χρηστών.

Ο αλγόριθμος W-kmeans ενισχύει τα προφίλ των χρηστών με υπερώνυμα που εξάγονται από το WordNet μέσω ενός ευρετικού τρόπου παρόμοιου με διαδικασία συσταδοποίησης άρθρων νέων. Αυτά τα προφίλ χρηστών, που επί της ουσίας αποτελούν συστάδες χρηστών, χρησιμοποιούνται στο στάδιο των προτάσεων προκειμένου να βελτιώσουν την εμπειρία χρήσης του συστήματος με το να παράγουν αποτελέσματα που ταιριάζουν καλύτερα στον χρήστη που επισκέπτεται το σύστημα. Ακολουθώντας την διαδικασία συσταδοποίησης συνεδριών, οι συστάδες που προκύπτουν ονοματίζονται με χρήση της διαδικασίας του WordNet cluster labeling που αναφέρθηκε και νωρίτερα.

Όταν ένας χρήστης επιστρέφει, το συσταδοποιημένο του προφίλ ανακτάται από την ΒΔ του συστήματος. Τα άρθρα που ταιριάζουν στο προφίλ του εξάγονται και αξιολογούνται προς πρόταση. Τα προτεινόμενα άρθρα δεν θα πρέπει να ανήκουν σε εκείνα που έχει προηγουμένως ήδη δει ο χρήστης και επίσης, δεν θα πρέπει να βρίσκονται πολύ κοντά σε άρθρα που έχει ο χρήστης αξιολογήσει αρνητικά στο παρελθόν. Η παραπάνω προσέγγιση προσδίδει ουσιαστικά και την φύση συνεργατικού φιλτραρίσματος στο σύστημα προτάσεων μας, μια και πρακτικά εμπεριέχει τους συσχετιζόμενους χρήστες στην διαδικασία λήψης απόφασης για τις προτάσεις. Αναμένουμε ότι ο συνδυασμός αυτού του είδους της προσέγγισης συνεργατικού φιλτραρίσματος μαζί με την τυπική λειτουργία που βασίζεται στο περιεχόμενο του κειμένου (εξαγωγή λέξεων κλειδιών και n-grams) θα βελτιώσει σημαντικά τις προτάσεις που το σύστημα μπορεί να προσφέρει προς τους χρήστες.

ΚΕΦΑΛΑΙΟ 5

ΑΝΑΛΥΣΗ ΚΑΙ ΑΛΓΟΡΙΘΜΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

> In order to succeed, we must
> first believe that we can.
>
> *Nikos Kazantzakis, Greek*
> *Writer, 1883*

Στο παρόν κεφάλαιο αναλύονται εκτενώς οι αλγόριθμοι που υλοποιήθηκαν στα πλαίσια της διδακτορικής διατριβής καθώς και πως αυτοί ενσωματώνονται στο σύστημα προτάσεων. Για κάθε υποσύστημα το οποίο υλοποιήθηκε ή ενισχύθηκε κατά την διδακτορική διατριβή γίνεται εκτενής αναφορά στον αλγοριθμικό τρόπο λειτουργίας του.

Οι αλγόριθμοι αυτοί αξιολογούνται μέσω των πειραματικών διαδικασιών που ακολουθούν στο επόμενο κεφάλαιο. Κάποιες σχεδιαστικές αποφάσεις όμως, όπως επιλογή παραμέτρων ζυγίσματος αναφέρονται και εδώ.

5.1 Υποσύστημα προεπεξεργασίας κειμένου

Στην παρούσα ενότητα αναλύουμε τις αλλαγές που έγιναν σε ότι έχει να κάνει με το αλγοριθμικό κομμάτι του μηχανισμού προεπεξεργασίας κειμένου του συστήματος προτάσεων νέων. Όπως προείπαμε σε προηγούμενα κεφάλαια, η προεπεξεργασία κειμένου έχει τον ρόλο της αξιολόγησης των δομικών μονάδων του κειμένου (π.χ. keywords, n-grams) όσον αφορά στην χρησιμότητά τους αποτύπωσης του νοήματος του κειμένου.

Το εν' λόγω υποσύστημα, όπως περιγράφεται στην μεταπτυχιακή μου εργασία [235], λειτουργούσε μόνο με keywords. Πλέον ενισχύεται με δύο τρόπους:

- με την εξαγωγή n-grams από τα άρθρα νέων
- με την αξιοποίηση της εξωτερικής γνώσης των υπερωνύμων από το WordNet

Από τα παραπάνω αναλύουμε μόνο το πρώτο σε αυτή την ενότητα μιας και το δεύτερο ταιριάζει περισσότερο στην ενότητα συσταδοποίσης όπου και αξιοποιείται μέσω του αλγορίθμου W-kmeans. Στην συνέχεια, αναλύουμε επίσης την διαδικασία ζύγισης των keywords και n-grams από το σύστημά μας.

5.1.1 Αξιοποίηση n-grams

Η διαδικασία εξαγωγής keywords, κάνοντας χρήση του vector space μοντέλου, παράγει τον πίνακα όρων-συχνοτήτων ο οποίος και περιγράφει κάθε άρθρο σαν μία BOW αναπαράσταση στις τεχνικές IR που ακολουθούν: κατηγοριοποίηση, προσωποποίηση και συσταδοποποίηση. Όταν λοιπόν η διαδικασία εξαγωγής keywords ολοκληρώνει την λειτουργία της για κάθε άρθρο που έχει προηγουμένως ανακτηθεί από το διαδίκτυο, μία λίστα από stemmed keywords (ρίζες λέξεων μόνο) παράγεται και αποθηκεύεται στη ΒΔ. Για παράδειγμα, έστω ένα άρθρο που ανήκει στον τομέα της κοσμολογίας (κατηγοριοποιημένο ως "επιστήμη') και για το οποίο το υποσύστημα εξαγωγής keywords εντοπίζει τα 18 keywords που (ορισμένα από τα οποία) φαίνονται στον πίνακα 1.

Από την λίστα με keywords του κειμένου κρατάμε τις ρίζες των ουσιαστικών μόνο, κάτι που φαίνεται στον πίνακα 1. Αυτά ταξινομούνται σε φθίνουσα κατάταξη με βάση την απόλυτη συχνότητα εμφάνισής τους στο κείμενο. με βάση τα παραπάνω, καθώς και τα δεδομένα των keywords για όλα τα άρθρα της ΒΔ, μπορούμε εύκολα να βρούμε τα tf-idf βάρη για τα keywords που μας ενδιαφέρουν.

Παράλληλα με τα εξαγόμενα keywords, η προσέγγισή μας επίσης εξάγει τα n-grams των άρθρων, όπου $2 < n <= 6$ και με συχνότητα εμφάνισης $kw_{fr} > 1$. Σε αυτή την περίπτωση, η συνολική ομοιότητα είτε μεταξύ δύο άρθρων είτε μεταξύ ενός άρθρου και μίας κατηγορίας ή συστάδας, δεν απεικονίζεται μόνο σε σχέση με τα tf-idf στατιστικά των keywords αλλά και σε σχέση με την αντίστοιχη μετρική για τα n-grams.

Για παράδειγμα στο ίδιο άρθρο που αναφέραμε προηγουμένως, τα εξαγόμενα n-grams μαζί με τις συχνότητες εμφάνισης τους φαίνονται στον πίνακα 2.

ID	Keyword	Συχνότητα
1	year	6
2	cosm	4
3	radiat	4
4	profess	4
5	mass	3
6	intens	3
7	event	3
8	Neuhauser	2
...	...	...
18	burst	2

Πίνακας 1: Stemmed keywords με τις συχνότητες εμφάνισής τους όπως εξάγονται από ένα τυχαίο άρθρο

ID	Keyword	Συχνότητα
1	light years	4
2	the most	3
3	the past 3000	3
4	to have	3
5	light years away	3
...	...	...
24	Professor Neuhauser	2

Πίνακας 2: Τα πιο συχνά εμφανιζόμενα n-grams όπως εξάγονται από το ίδιο άρθρο

Από την λίστα με n-grams του πίνακα 2, θα μπορούσαμε να συμπεράνουμε ότι κάποια απ' αυτά, όπως "light years away" και "Professor Neuhauset" μπορούν να θεωρηθούν ως καλοί εκπρόσωποι του νοήματος καθώς και του συγκεκριμένου πεδίου στο οποίο ανήκει το παραπάνω άρθρο (κοσμολογία). Έτσι για παράδειγμα θα μπορούσαμε να πούμε ότι η συστάδα στην οποία ανήκει αυτό το άρθρο θα πρέπει να έχει αυξημένο το βάρος αυτών των n-grams.

Αν λοιπόν έχουμε τέτοιους πίνακες π.χ. για δύο άρθρα, μπορούμε εύκολα να υπολογίσουμε την συσχέτισή τους χρησιμοποιώντας οποιαδήποτε μετρική ομοιότητας μας ενδιαφέρει, π.χ. Ευκλείδεια απόσταση στην περίπτωσή μας. Να σημειώσουμε τέλος ότι δεδομένης της φύσης των n-grams σε σχέση με το νόημά τους και την χρήση τους στην φυσική γλώσσα, δεν μπορούμε να αγνοήσουμε τα stopwords όπως κάνουμε για παράδειγμα στην περίπτωση εξαγωγής των keywords για μείωση διαστατικότητας. Τα stopwords αποτελούν δομικό στοιχείο των n-grams του κειμένου.

5.1.2 Ζύγιση άρθρων

Όταν έχουμε τους παραπάνω πίνακες για κάποιο άρθρο λοιπόν, μπορούμε να υιοθετήσουμε ένα σχήμα ζυγίσματος που θα αξιοποιεί την πληροφορία συχνότητας εμφάνισης τόσο των keywords, όσο και των n-grams. Για το βάρος του κάθε keyword i, W_{kw_i}, ξεκινούμε από την κλασική tf-idf

μετρική που δίνεται στη σχέση 30,

$$W_{kw_i} = tf - idf_i = freq_i * log\frac{N}{M_i} \tag{30}$$

όπου $freq_i$ η συχνότητα εμφάνισης του keyword i στο κείμενο, N το συνολικό πλήθος των άρθρων στη ΒΔ και M_i το πλήθος των άρθρων που περιέχουν το keyword i. Ενισχύοντας την παραπάνω λογική ζυγίσματος κάθε άρθρου, μπορούμε να αναθέσουμε βάρη στα n-grams που έχει το κάθε άρθρο χρησιμοποιώντας τα αντίστοιχα tf-idf στατιστικά. Πιο συγκεκριμένα, για κάθε n-gram j, το βάρος του, W_{ng_j}, μπορεί να εκφραστεί από την tf-idf συχνότητά του, την οποία και στο εξής αποκαλούμε gf-idf (gram frequency / inverse document frequency). Το βάρος αυτό θα μπορούσε να γραφεί όπως φαίνεται στη εξίσωση 31.

$$W_{ng_j} = gf - idf_j = freq_j * log\frac{N}{M_j} \tag{31}$$

όπου $freq_j$ η συχνότητα εμφάνισης του n-gram j στο κείμενο, N το συνολικό πλήθος των άρθρων στη ΒΔ και M_j το πλήθος των άρθρων που περιέχουν το n-gram j.

Παρόλα αυτά, όπως αναλύθηκε στα πλαίσια της μεταπτυχιακής μου εργασίας, το βάρος των keywords δεν μπορεί να βασίζεται απλά και μόνο στην συχνότητα εμφάνισης. Αντίθετα επεκτείνουμε το σχήμα ζύγισης χρησιμοποιώντας τις λογικές υποθέσεις που ακολουθούν:
Συνήθως, ένα keyword που ανήκει στον τίτλο ενός άρθρου, είναι πιο σημαντικό μιας και αποτυπώνει περισσότερο νόημα από το κείμενο. Επιπλέον, μιας και η διαδικασία περίληψης του συστήματός μας βασίζεται στην επιλογή των πιο σημαντικών προτάσεων από το ίδιο το κείμενο μέσω του ζυγίσματος αυτών, τα αποτελέσματα της κατηγοριοποίησης μπορούν να είναι βοηθητικά στο να προσαρμόσουν πιο αποτελεσματικά το ζύγισμα των προτάσεων. Η κοινή λογική λέει ότι ένα keyword που έχει πολύ υψηλή συχνότητα εμφάνισης σε μία κατηγορία θα δίνει περισσότερο βάρος σε ένα κείμενο ή πρόταση για ένα κείμενο που γνωρίζουμε ότι ανήκει σε αυτή την κατηγορία. Αντίστοιχα, ένα keyword με μικρή ή μηδενική συχνότητα εμφάνισης σε μία κατηγορία μπορεί να έχει λιγότερο βάρος για μία πρόταση. Για μία πιο βαθιά ανάλυση των παραπάνω παραπέμπουμε τον αναγνώστη στο [34].

Έτσι, το βάρος (σκορ) μίας πρότασης ή ενός κείμενου a δίνεται από την ευρετική γραμμική σχέση 32:

$$S_a = \sum_i W_{kw_i,a} * \gamma_{i,a} \tag{32}$$

όπου $W_{kw_i,a}$ η tf-idf μετρική του keyword i στο κείμενο a, και $\gamma_{i,a}$ ο παράγοντας ζυγίσματος κάθε keyword i που εξαρτάται από:

- την σχετική συχνότητα εμφάνισης[1] του i στο σώμα του άρθρου (σε σχέση με τον συνολικό αριθμό εμφανίσεών του στη ΒΔ)

[1] η σχετική συχνότητα ορίζεται ως το πηλίκο του πλήθους των εμφανίσεων ως προς το πλήθος όλων των keywords του κειμένου

- την σχετική συχνότητα εμφάνισης του i στον τίτλο του άρθρου (σε σχέση με τον συνολικό αριθμό εμφανίσεών του στη ΒΔ)
- την επίδραση που έχει η διαδικασία κατηγοριοποίησης και η διαδικασία περίληψης κειμένου

Συμπεριλαμβάνοντας την πληροφορία των n-grams, η 32 γίνεται όπως η 33:

$$S_a = A * (\sum_i W_{kw_i,a} * \gamma_{i,a}) + B * \sum_j W_{ng_j,a} \tag{33}$$

όπου $W_{ng_j,a}$ η gf-idf μετρική του n-gram j στο κείμενο a, A και B οι παράγοντες ζυγίσματος των keywords και n-grams αντίστοιχα (οι οποίοι και θα αναλυθούν στη συνέχεια).

Ο παράγοντας ζυγίσματος των keywords της σχέσης 33, $\gamma_{i,a}$, για κάθε keyword i, καθορίζεται όπως φαίνεται στην σχέση 34

$$\gamma_{i,a} = \alpha * F_{i,body} + \beta * F_{i,title} + \delta * F_{i,category} \tag{34}$$

όπου:

- α και β είναι οι παράγοντες ζυγίσματος της σημαντικότητας του keyword i που ανήκει στο σώμα ή στον τίτλο του άρθρου αντίστοιχα. Στην έρευνά μας: $\alpha = 0.1$ και $\beta = 0.7$
- $F_{i,body}$ είναι η σχετική συχνότητα εμφάνισης του keyword i στο σώμα του άρθρου
- $F_{i,title}$ είναι η σχετική συχνότητα εμφάνισης του keyword i στον τίτλο του άρθρου
- $F_{i,category}$ είναι η σχετική συχνότητα εμφάνισης του keyword i στην κατηγορία την οποία γνωρίζουμε ότι ανήκει το άρθρο
- δ είναι ο παράγοντας ζύγισης της επίπτωσης που έχει η κατηγοριοποίηση/συσταδοποίηση. Στην έρευνά μας $\delta = 0.2$

Με βάση τα παραπάνω, η σχέση 34 απλοποιείται στην 35

$$\gamma_{i,a} = 0.1 * F_{i,body} + 0.7 * F_{i,title} + 0.2 * F_{i,category} \tag{35}$$

5.1.2.1 Ζύγιση keywords για την συσταδοποίηση

Αξιοποιώντας την επιπλέον πληροφορία που περιγράφηκε για κάθε keyword, το βάρος του keyword i ενός άρθρου a γίνεται:

$$\bar{W}_{kw_i} = W_{kw_i} * \gamma_{i,a} = freq_i * log\frac{N}{M_i} * \gamma_{i,a} \tag{36}$$

Η σχέση 36 είναι και αυτή που χρησιμοποιείται από τον αλγόριθμο συσταδοποίησης προκειμένου να αποδοθεί το βάρος του κάθε keyword των άρθρων προς συσταδοποίηση ώστε έπειτα να εκτιμηθούν οι συστάδες των άρθρων.

Όπως φαίνεται στην σχέση 33 μπορούμε να ελέγξουμε και να κανονικοποιήσουμε την επίπτωση που έχει η ζύγιση keywords ή n-grams χρησιμοποιώντας μία γραμμική συσχέτιση μεταξύ τους βάσει των παραμέτρων A και B, έτσι ώστε:

$$W'_{kw_i} = \bar{W}_{kw_i} * A \tag{37}$$

$$W'_{ng_j} = W_{ng_j} * B \tag{38}$$

όπου:

$$A + B = 1 \tag{39}$$

Άρα η σχέση 33 μπορεί να γραφεί με γραμμικό τρόπο ως την ακόλουθη σχέση ο οποία και αποτελεί τον πυρήνα του ζυγίσματος που επιτελεί το σύστημα προεπεξεργασίας:

$$S_a = \sum_i W'_{kw_i} + \sum_j W'_{ng_j} \tag{40}$$

Ο καθορισμός των βαρών A και B στις παραπάνω εξισώσεις είναι αποτέλεσμα της πειραματικής διαδικασίας για το σύνολο δεδομένων που μας ενδιαφέρει. Πειραματικά εντοπίσαμε ότι για τον τομέα των άρθρων νέων που μας ενδιαφέρει, τα καλύτερα αποτελέσματα ήταν όταν:

$$A = 0.7, \quad B = 0.3 \tag{41}$$

5.2 Υποσύστημα συσταδοποίησης

Στην ενότητα αυτή περιγράφονται οι αλγόριθμοι που αφορούν την συσταδοποίηση τόσο άρθρων νέων όσο και χρηστών του συστήματος προτάσεων που αναπτύχθηκε.

5.2.1 Αλγόριθμος W-kmeans

Στον πυρήνα του αλγόριθμου W-kmeans υπάρχουν τα τυπικά βήματα του k-means αλγορίθμου. Προπαρασκευαστικό βήμα όμως πριν εφαρμοστεί ο k-means αλγόριθμος είναι η ενίσχυση της εισόδου (είτε άρθρα νέων, είτε συνεδρίες χρηστών) με υπερώνυμα που εξάγονται από το WordNet. Επομένως, σε ένα πολύ υψηλό επίπεδο θα μπορούσαμε να περιγράψουμε τα αλγοριθμικά βήματα του W-kmeans όπως δίνονται στον αλγόριθμο 2. Η ανάλυση αυτών των βημάτων για κάθε περίπτωση ακολουθεί στα επόμενα.

Αλγόριθμος 2: Αλγόριθμος W-kmeans

Είσοδος: articles A, number of clusters, X

// η λίστα με τα άρθρα προς συσταδοποίηση

Έξοδος: cluster assignments C, cluster labels L

// λίστες με τις συστάδες που προκύπτουν καθώς και οι ετικέτες τους

```
foreach article a ∈ A do
    keywords = 20% * restore_keywords(a);
    // κρατάμε μόνο το 20% των συχνότερων keywords
    enriched_keywords = wordnet_enrich(a);
    total_keywords = keywords + enriched_keywords;
    ngrams = n_grams_extract(a, 1 < n < X);
    // εξαγωγή των n-grams που μας ενδιαφέρουν
C[] = kmeans(total_keywords + ngrams);
L[] = wordnet_cluster_labeling(C, enriched_keywords);
// η ζύγιση των keywords και n-grams βασίζεται στην εξίσωση 33
return C[], L[]
```

5.2.2 Συσταδοποίηση άρθρων νέων

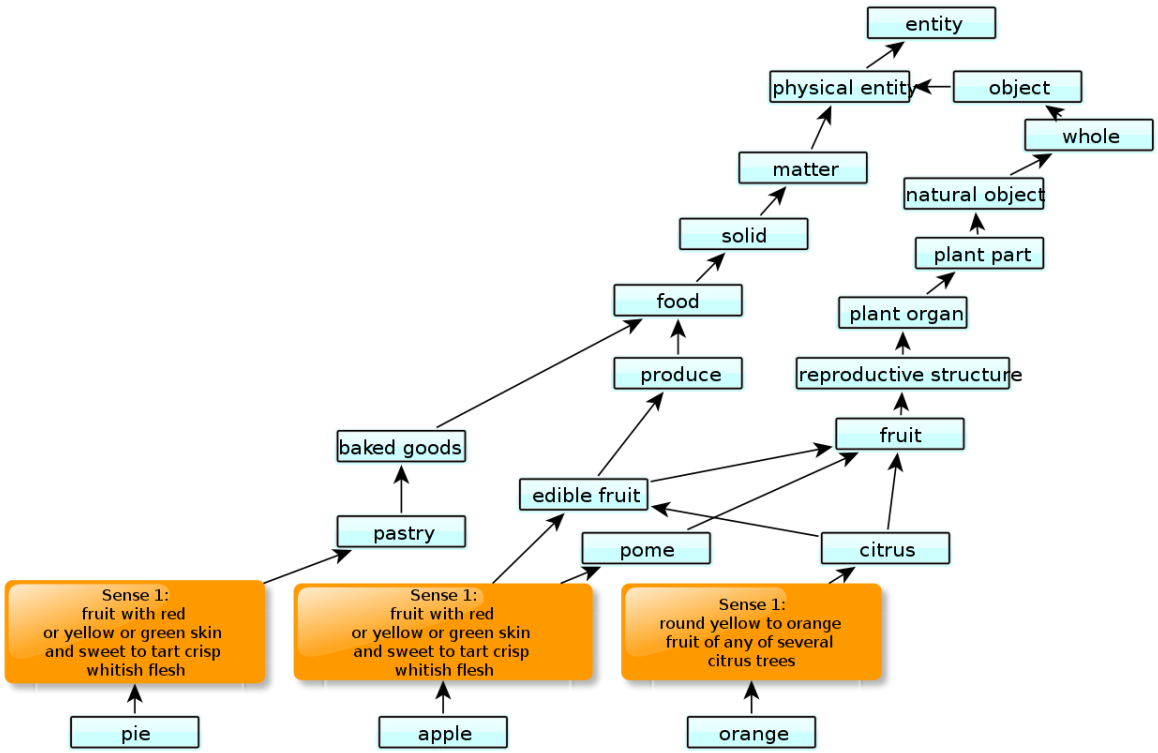
Στην παρούσα ενότητα αναλύουμε την αλγοριθμική μας προσέγγιση για την αξιοποίηση της εξωτερικής βάσης γνώσης WordNet στο σύστημά μας. Όπως προείπαμε, το σύστημα WordNet οργανώνει διαφορετικές λεξικολογικές συσχετίσεις σε ιεραρχίες. Έτσι, δεδομένου ενός ουσιαστικού, ρήματος, επιθέτου και επιρρήματος, το WordNet μπορεί να παρέχει αποτελέσματα σε σχέση με τα υπερώνυμα, υπώνυμα μερόνυμα και ολόνυμά του. Χρησιμοποιώντας αυτές τις δενδρικές δομές που παράγονται από το WordNet, μπορούμε να αναζητήσουμε σε αυτό για όλα τα υπερώνυμα ενός συνόλου λέξεων που μας ενδιαφέρουν και έπειτα να τα ζυγίσουμε κατάλληλα, επιλέγοντας τελικά αντιπροσωπευτικά υπερώνυμα που φαίνεται να ενισχύουν το συνολικό νόημα του κειμένου

Αυτή η διαισθητική προσέγγιση όμως, εξαρτάται αποκλειστικά από την εξίσωση ζύγισης των υπερωνύμων που θα επιλεχθεί. Είναι φυσικά σημαντικό η ζύγιση αυτή να εισάγει "νέα γνώση" σε σχέση με την υπάρχουσα λίστα από keywords και όχι να χειροτερεύει τα ήδη υπάρχοντα αποτελέσματα. Τελικός σκοπός είναι φυσικά η συσταδοποίηση να είναι λιγότερο ασαφής και περισσότερο ακριβής ως προς τα αποτελέσματα που παράγει.

5.2.2.1 Εξαγωγή και ζύγιση υπερωνύμων

Όσον αφορά την εξαγωγή υπερωνύμων και την πρόσθεση τους στα συνολικά keywords, η διαδικασία έχει ως εξής: για κάθε keyword ενός άρθρου, παράγουμε το γράφο/δέντρο υπερωνύμων το οποίο καταλήγει στο υπερώνυμο-ρίζα το οποίο το WordNet ονομάζει ως "οντότητα" για τα ουσιαστικά. Στη συνέχεια συνδυάζουμε κάθε ξεχωριστό γράφο σε έναν αθροιστικό. Για παράδειγμα,

το σχήμα 14 αθροιστικά απεικονίζει τους γράφους υπερωνύμων του WordNet για τις λέξεις: 'pie', 'apple' και 'orange'.



Σχήμα 14: Αθροιστικό δέντρο υπερωνύμων για τρεις λέξεις: 'pie', 'apple' και 'orange'

Βλέπουμε έτσι ότι η λέξη 'apple' έχει τρία διαφορετικά μονοπάτια προς τη ρίζα:

- apple → edible fruit → fruit → ...
- apple → edible fruit → produce → ...
- apple → pome → fruit → ...

Έχοντας τα παραπάνω υπόψιν μας, μπορούμε να αναζητήσουμε στο σύνολο των υπερωνύμων τα "καλύτερα" χρησιμοποιώντας μία ευρετική συνάρτηση. Υπάρχουν πρακτικά δύο παράμετροι τις οποίες πρέπει να λάβουμε υπόψιν για κάθε υπερώνυμο της παραπάνω ανεστραμμένης δενδρικής δομής και οι οποίες καθορίζουν τη σημαντικότητά του:

1. το βάθος στο δέντρο
2. η συχνότητα εμφάνισης στα διάφορα μονοπάτια από τα φύλλα (χαμηλά) ως τη ρίζα (στην κορυφή)

Μπορούμε να παρατηρήσουμε ότι όσο πιο ψηλά (δηλαδή λιγότερο βαθιά όπως προχωράμε από την ρίζα προς τα κάτω) είναι το υπερώνυμο στο γράφο, τόσο πιο γενικό είναι. Όμως, όσο πιο χαμηλά

είναι το υπερώνυμο στο γράφο, τόσο λιγότερες πιθανότητες έχει να εμφανίζεται σε πολλά μονοπάτια (δηλαδή η συχνότητα εμφάνισής του είναι μικρή). Να σημειώσουμε επίσης ότι σε περιπτώσεις όπου ένα υπερώνυμο έχει πολλαπλά μονοπάτια που οδηγούν στη ρίζα, το κοντινότερο από αυτά κρατείται για την μέτρηση του βάθους του στο γράφο.

Στην προσέγγισή μας και στα πλαίσια του αλγορίθμου W-kmeans, οι δύο αυτές αντικρουόμενες παράμετροι ζυγίζονται όπως φαίνεται στην συνάρτηση 42.

$$W(d,f) = \frac{1}{2} * (\frac{1}{1+e^{-0.125(d^3 \frac{f}{TW})}} - 0.5) \tag{42}$$

όπου d είναι το βάθος του κόμβου (μετρώντας από πάνω προς τα κάτω στο σχήμα 14), f είναι η συχνότητα εμφάνισης του υπερωνύμου (κόμβου) στα πολλαπλά μονοπάτια (υπο-γράφοι) και TW είναι το πλήθος των συνολικών λέξεων που χρησιμοποιήθηκαν για να παραχθεί το δέντρο, δηλαδή τα συνολικά keywords και υπερώνυμα των άρθρων για την περίπτωση της συσταδοποίησης άρθρων νέων, ή τα συνολικά keywords και υπερώνυμα των άρθρων των συνεδριών για την περίπτωση της συσταδοποίησης χρηστών.

Η συνάρτηση 42 είναι "σιγμοείδης" (sigmoid) της γενικής μορφής:

$$S(x) = \frac{1}{1+e^{-x}} \tag{43}$$

όπου:

$$x = d^3 * \frac{f}{TW} \tag{44}$$

και ζυγίζεται ως:

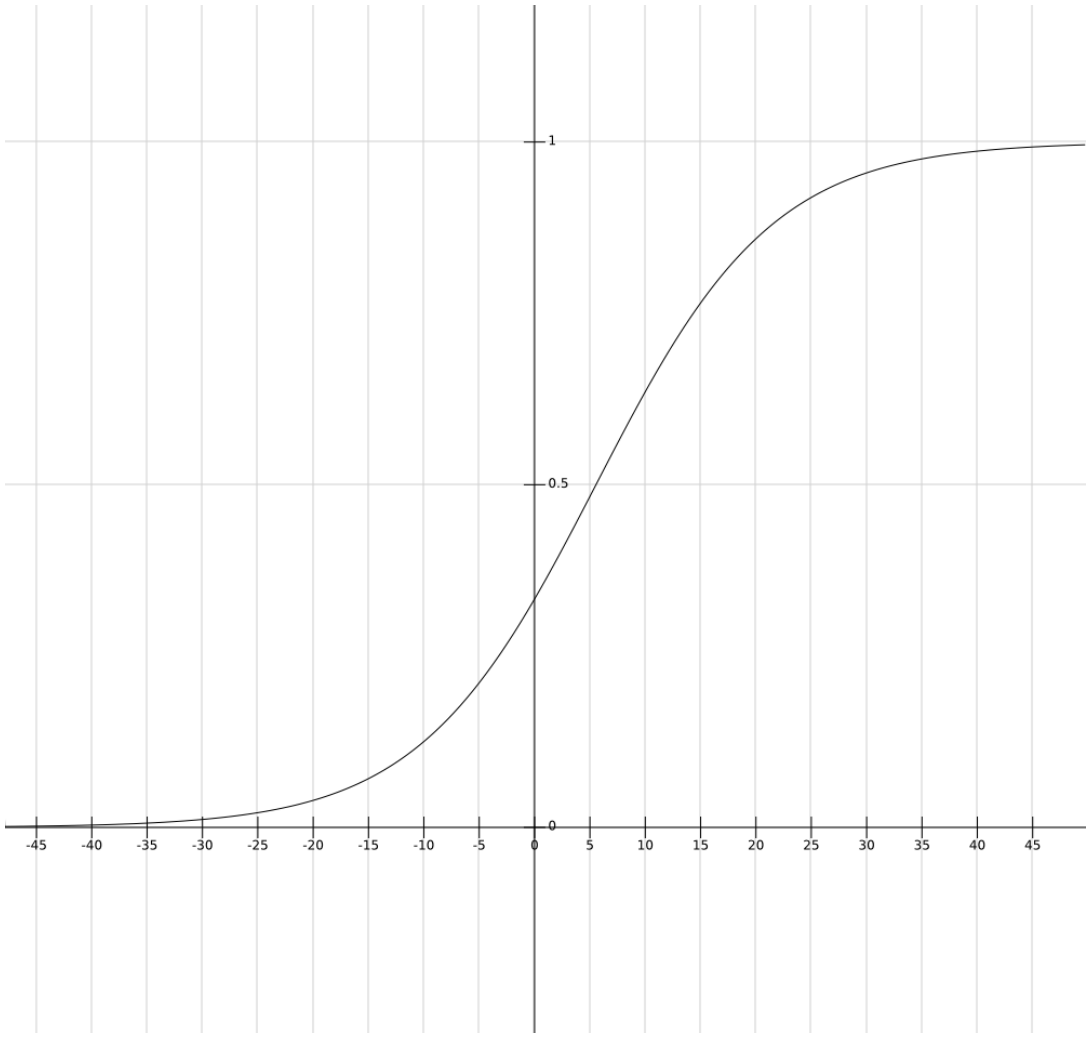
$$S(d,f) = a * Sig(d,f) - b \tag{45}$$

με $a = b = 0.5$. Η γραφική αναπαράσταση της συνάρτησης 42 φαίνεται στο σχήμα 15.

Το πόσο "απότομη" είναι η συνάρτηση 42 καθορίζεται από τον εκθέτη της βάσης του φυσικού λογαρίθμου e που έχει (-0.125 στην περίπτωσή μας). Επιλέξαμε sigmoid συνάρτηση αφού παρατηρήσαμε το πως τόσο η συχνότητα όσο και το βάθος των υπερωνύμων στο δέντρο επηρεάζουν τα παραγόμενα αποτελέσματα συσταδοποίησης η οποία και περιγράφεται στη συνέχεια.

Η σημαντικότητα (βάρος - weight) κάθε υπερωνύμου εμφανίζει μία εξέλιξη από χαμηλή αρχική τιμή η οποία κορυφώνεται απότομα σε κάποιο σημείο. Για συνδυασμούς βάθους και συχνότητας με υψηλές τιμές, το βάρος των υπερωνύμων πλησιάζει πολύ πιο γρήγορα στην μέγιστη τιμή 1 (μίας και καμία από τις δύο παραμέτρους δεν μπορεί να είναι αρνητική). Αντίθετα για συνδυασμούς βάθους και συχνότητας με χαμηλές τιμές το βάρος προσεγγίζει όλο και περισσότερο στην ελάχιστη τιμή της συνάρτησης, δηλαδή $W = 1/3$. Η ελάχιστη αυτή τιμή επιλέχθηκε προκειμένου κάθε υπερώνυμο που έχει εν' τέλει συμμετοχή στην διαδικασία ζύγισης να έχει κάποια ουσιαστική τιμή που μπορεί να επηρεάσει την όλη διαδικασία. Με άλλα λόγια, τιμές πιο κοντά στο μηδέν δεν έχουν κάποια πρακτική αξία στην διαδικασία ζύγισης.

Ένα keyword που δεν έχει υπερώνυμα ή που δεν υπάρχει στο WordNet αφαιρείται τόσο από τον συνολικό γράφο όσο και από το άθροισμα TW. Επιπλέον, ένα υπερώνυμο μπορεί να έχει πολλαπλά



Σχήμα 15: Γραφική αναπαράσταση της sigmoid συνάρτησης 42 που χρησιμοποιείται από τον αλγόριθμο W-kmeans

μονοπάτια προς την ρίζα, όμως μετριέται μόνο μία φορά για κάθε δεδομένο keyword και κρατείται το ελάχιστο πάντα βάθος. Να σημειωθεί επίσης ότι το βάθος παίζει αρκετά πιο σημαντικό ρόλο στη διαδικασία ζύγισης σε σχέση με την συχνότητα εμφάνισης. Η συχνότητα εμφάνισης όμως, δρα όλο και περισσότερο ως επιλεκτικός παράγοντας όσο ένα δέντρο επεκτείνεται και περισσότερα keywords με τα υπερώνυμά τους προστίθενται. Καταλήξαμε στο παραπάνω σύστημα ζύγισης μετά από παρατηρήσεις σε αθροιστικούς γράφους υπερωνύμων που περιέχουν εκατοντάδες κόμβους και βλέποντας το ζύγισμα αυτό κλιμακώνεται ικανοποιητικά.

Έτσι, από το δέντρο του σχήματος 14 μπορούμε να υπολογίσουμε το βάρος κάθε υπερωνύμου. Για παράδειγμα οι τιμές βάρους για ορισμένα υπερώνυμα φαίνονται στον πίνακα 3 (με $TW = 21$).

Υπερώνυμο	**Βάθος (d)**	**Συχνότητα (f)**	**Βάρος (W)**
fruit	9	2	0.57
edible fruit	7	2	0.47
food	5	3	0.44

Πίνακας 3: Βάρος ορισμένων υπερωνύμων του σχήματος 14

Οι παραπάνω τιμές a και b καθορίστηκαν πειραματικά ως εξής: χρησιμοποιώντας ένα σύνολο 1000 προ-κατηγοριοποιημένων άρθρων νέων, προσπαθήσαμε να αξιολογήσουμε την αποτελεσματικότητα του προτεινόμενου W-kmeans αλγορίθμου με το να συσταδοποιήσουμε αυτά τα άρθρα στο προκαθορισμένο σύνολο από τις κατηγορίες του συστήματος. Με αυτό το σενάριο, η προσέγγιση συσταδοποίησης θα πρέπει να επιτυγχάνει συστάδες όσο πιο κοντά γίνεται στις υπάρχουσες κατηγορίες άρθρων. Προφανώς το πλήθος των συστάδων k είναι ίσο με το πλήθος των κατηγοριών. Μια μεγάλη ποικιλία από συνδυασμούς των a και b χρησιμοποιήθηκαν και το καλύτερο συνολικά αποτέλεσμα παρατηρήθηκε για $a = b = 0.5$.

5.2.2.2 Αλγόριθμος ενίσχυσης άρθρων νέων με υπερώνυμα

Ο αλγόριθμος ενίσχυσης των άρθρων με χρήση των υπερωνύμων του WordNet, όπως περιγράφεται στο 3, λειτουργεί πάνω στα keywords του άρθρου παράγοντας το γράφο υπερωνύμων για καθένα ξεχωριστά. Χρησιμοποιούμε μόνο το 20% των πιο σημαντικών keywords, μειώνοντας έτσι τη διαστατικότητα και το θόρυβο. Στη συνέχεια, ένας αθροιστικός γράφος παράγεται από τον οποίο το βάρος του κάθε υπερωνύμου υπολογίζεται βάσει της συνάρτησης 42. Οι κόμβοι του γράφου ύστερα ταξινομούνται με βάση το βάρος τους και η λίστα από τα πιο σημαντικά keywords και υπερώνυμα επιστρέφεται ως η προτεινόμενη για ενίσχυση του άρθρου. Προκειμένου να αποφύγουμε την αύξηση της διαστατικότητας και υπεργενίκευσης των αποτελεσμάτων, λαμβάνουμε υπόψιν μας το ένα τέταρτο (25%) από τα συνολικά keywords και υπερώνυμα που επιστρέφονται από την παραπάνω διαδικασία. Παρατηρήσαμε ότι η επιλογή αυτή παράγει καλύτερα αποτελέσματα με ελάχιστο επιπλέον κόστος στον χρόνο εκτέλεσης.

Αλγόριθμος 3: Εμπλουτισμός άρθρων νέων με χρήση των υπερωνύμων του WordNet

Είσοδος: article a

// το άρθρο προς ενίσχυση

Έξοδος: enrichedKeywords

// λίστα από keywords ενισχυμένα με υπερώνυμα του WordNet

```
1  total_hypen_tree = NULL;
2  kws = fetch 20% most frequent k/ws for a;
3  foreach keyword kw in kws do
4      htree = wordnet_hypen_tree(kw);
       // εξαγωγή του δέντρου υπερωνύμου για αυτό το keyword
5      foreach hypen h in htree do
6          if (h not in total_hypen_tree) then
7              h.frequency=1;
8              total_hypen_tree ->append(h);
9          else
10             total_hypen_tree ->at(h)->freq++;
11 foreach h in total_hypen_tree do
12     calculate_depth(h);
13     weight = 1/2 * ((1/(1 + exp(−0.0125 * (h− > depth3 * h− > freq/kws_in_wn− >
       size)))) − 0.5));
14 sort_weights(total_hypen_tree);
15 important_hypens = (kws -> size / 4) * top(total_hypen_tree);
16 return_kws += important_hypens;
17 return return_kws//
```

5.2.3 Ονοματοδοσία συστάδων

Προκειμένου να παραχθούν οι προτεινόμενες ετικέτες για κάθε μία από τις συστάδες που προκύπτουν (είτε συστάδες άρθρων νέων, είτε συστάδες συνεδριών χρηστών), κάνουμε επίσης χρήση των υπερωνύμων του WordNet. Τα βήματα που ακολουθούνται παρουσιάζονται στον αλγόριθμο 4.

Αλγόριθμος 4: Ονοματοδοσία συστάδων με χρήση των υπερωνύμων του WordNet

Είσοδος: clusters C

Έξοδος: cluster labels L

```
1  total_hypen_tree = NULL;
2  foreach cluster c ∈ C do
3      kws += fetch 10% most frequent k/ws for c;
4  foreach keyword kw ∈ kws do
5      hypens_tree = wordnet_hypen_tree(kw);
6      foreach hypen h in hypens_tree do
7          if h not in total_hypen_tree then
8              h.frequency = 1;
9              total_hypen_tree->append_child(h);
10         else
11             total_hypen_tree->at(h)->frequency++;
12 foreach hypen h in total_hypen_tree do
13     calculate_depth(h);
14 sort_weights(total_hypen_tree);
15 cluster_labels[]-> append(5 top(total_hypen_tree));
16 return cluster_labels[]
```

Η διαδικασία ονοματοδοσίας λειτουργεί πάνω σε κάθε συστάδα ανακτώντας αρχικά μόνο το 10% των πιο σημαντικών keywords που ανήκουν στη συστάδα. Η μελέτη σε άρθρα νέων που υπάρχουν στο σύστημά μας έδειξε ότι το παραπάνω ποσοστό είναι αρκετό προκειμένου η διαδικασία να παράξει ετικέτες υψηλού επιπέδου και χωρίς την εισαγωγή θορύβου. Για κάθε συστάδα λοιπόν, ανακτούμε και αθροίζουμε το δέντρο υπερωνύμων του WordNet σε μία διαδικασία παρόμοια με αυτή του σχήματος 14. Οι κόμβοι που προκύπτουν ζυγίζονται βάσει της συνάρτησης 42, ταξινομούνται και τελικά τα 5 σημαντικότερα υπερώνυμα επιστρέφονται ως ετικέτες της συστάδας. Μιας και πρόκειται ουσιαστικά για ανάθεση ετικετών (tagging) θεωρούμε ότι 5 ετικέτες είναι αρκετές για να καλύψουν με συντομία το περιεχόμενο των εκάστοτε συστάδων.

5.3 Προσωποποίηση στο χρήστη

Η προσέγγιση που προτείνεται για την προσωποποίηση στον χρήστη αποτελείται από τρία βασικά αλγοριθμικά τμήματα που χρησιμοποιούνται για:

1. την offline διαδικασία εντοπισμού των συνεδριών χρηστών, όπως αυτοί αξιοποιούν την πληροφορία που παρέχει το σύστημα προτάσεων

2. την offline διαδικασία συσταδοποίησης των εντοπισμένων συνεδριών

3. την online διαδικασία παραγωγής προτάσεων άρθρων νέων από το σύστημα βασιζόμενοι σε πληθώρα πληροφοριών

Τα αλγοριθμικά βήματα για τις παραπάνω διαδικασίες δίνονται στη συνέχεια.

5.3.1 Εύρεση συνεδρίων χρηστών

Ο εντοπισμός των συνεδριών μέσα στο ιστορικό πλοήγησης ενός χρήστη επιτυγχάνεται ακολουθώντας τα βήματα που περιγράφονται στον αλγόριθμο 5. Ο εν' λόγω αλγόριθμος χρησιμοποιεί δύο σημαντικές τιμές κατωφλίου:

1. το κατώφλι προβολής, δηλαδή τον ελάχιστο χρόνο που ο χρήστης αναμένεται να ξοδέψει σε ένα άρθρο που τον ενδιαφέρει

2. το κατώφλι συνεδρίας, δηλαδή τον μέγιστο χρόνο τον οποίο κατά μέσο όρο ξοδεύει ο χρήστης πλοηγημένος συνεχόμενα σε άρθρα νέων - αξιοποιώντας επομένως την πληροφορία που του παρέχει το σύστημα

Για τον καθορισμό των παραπάνω τιμών, αναλύσαμε τις συνήθειες πλοήγησης των χρηστών που χρησιμοποιούσαν το σύστημα προτάσεων. Κατά την ανάλυση παρατηρήθηκε ότι, κατά μέσο όρο, ένα άρθρο νέου θα διαβαζόταν για όχι λιγότερο από 30 δευτερόλεπτα από τους χρήστες αφού το έχουν επιλέξει πρώτα προς ανάγνωση και άρα αποτελεί ένα ενδιαφέρον άρθρο για αυτούς. Να σημειωθεί παρόλα αυτά ότι αυτό το κατώφλι πραγματικά εξαρτάται από το μέγεθος του άρθρου και η παραπάνω τιμή αποτελεί απλά έναν οδηγό (μέσος όρος) βασιζόμενοι στην ανάλυση που έγινε για το συγκεκριμένο σύνολο δεδομένων. Παράλληλα, από τα ίδια δεδομένα πλοήγησης παρατηρήθηκε ότι ένας χρήστης, στις περισσότερες των περιπτώσεων, δεν ξοδεύει πάνω από 10 λεπτά συνεχόμενα διαβάζοντας άρθρα στο σύστημα προτάσεων προτού αποχωρήσει από το σύστημα. Είναι αυτή η συνεχής ροή από άρθρα νέων κατά μία τέτοια περίοδο χρήσης που πραγματικά μας ενδιαφέρει προκειμένου να αποτυπωθεί σωστά το προφίλ του χρήστη.

Η έξοδος του αλγορίθμου 5 είναι μία λίστα από συνεδρίες για κάθε χρήστη η οποία και αποθηκεύεται στη ΒΔ για περαιτέρω χρήση.

Αλγόριθμος 5: Εύρεση συνεδριών στα μονοπάτια πλοήγησης των χρηστών

Είσοδος: history/* το παράθυρο χρόνου που χρησιμοποιείται για εξαγωγή συνεδριών χρήστη */

Έξοδος: Sessions[]// οι εντοπισμένες συνεδρίες

```
1  viewing_threshold = 30 // τουλάχιστον 30 δευτερόλεπτα
2  session_threshold = 10 * 60 // το πολύ 10 λεπτά
3  previous_user = NULL;
4  current_session = NULL;
5  while (fetch from DB (user, viewed article, timestamp, viewing_time)) do
6      if (viewing_time < viewing_threshold || timestamp < history) then
7          continue;
8      if (current_session.username != user) then
           /* μιας και αυτό είναι ταξινομημένο ως προς το username, όταν ένας νέος χρήστης
              βρεθεί, τότε αρχίζει και μία νέα συνεδρία */
9          if (current_session.username !="" && current_session.articles != "") then
10             Sessions[]+=current_session;
11         current_session.username = user;
12         current_session.user_id = user_id;
13         current_session.start = timestamp;
14         current_session.articles.add(article_id);
15     else
           /* εάν ο χρήστης είναι ο ίδιος με πριν, αλλά ο χρόνος προσπέλασης για αυτό το
              άρθρο ξεπερνά το όριο, μία νέα συνεδρία αρχίζει */
16         if (timestamp – current_session.start) > session_threshold ) then
17             if (current_session.username!="" && current_session.articles=!empty) then
18                 Sessions[]+=current_session;
19             current_session.username = user;
20             current_session.user_id = user_id;
21             current_session.start = timestamp;
22             current_session.end = timestamp;
23             current_session.articles.add(article_id);
24         else
               // ο χρόνος προσπέλασης για αυτό το άρθρο δεν ξεπερνά το όριο
25             current_session.articles.add(article_id);
26             current_session.end = timestamp;
27 return Sessions[]
```

5.3.2 Συσταδοποίηση Χρηστών με χρήση του W-kmeans

Όταν οι συνεδρίες χρηστών έχουν εξαχθεί, για κάθε μία από αυτές, μπορούμε να προσθέσουμε τα άρθρα νέων που την απαρτίζουν σε μία λίστα. Στο επόμενο βήμα, εμπλουτίζουμε τα keywords που αποτελούν τα άρθρα νέων της συνεδρίας χρησιμοποιώντας υπερώνυμα από την εξωτερική βάση γνώσης WordNet και στη συνέχεια προχωρούμε στην συσταδοποίηση των συνεδριών, εμμέσως επομένως και των χρηστών.

Αλγόριθμος 6: Συσταδοποίηση συνεδριών χρηστών με χρήση του αλγορίθμου W-kmeans

Είσοδος: sessions, number of clusters

Έξοδος: session to cluster assignments[]// οι συστάδες των συνεδριών

```
1 foreach s in sessions do
2     foreach article a belonging to s do
3         s.kws += fetch 20% most frequent k/ws for a;
4     wordnet_enrich(s) // δες τον αλγόριθμο 3
5 clusters[] = kmeans(sessions);
6 return clusters[]
```

Όσον αφορά την εξαγωγή υπερωνύμων και την πρόσθεση τους στα συνολικά keywords, η διαδικασία είναι παρόμοια με εκείνη που περιγράφηκε για τα άρθρα νέων στην ενότητα 5.2.2.1, μόνο που τώρα το αθροιστικό δέντρο αποτελείται από keywords και τα υπερώνυμά τους που απαρτίζουν τα άρθρα των συνεδριών χρηστών προς συσταδοποίηση. Επίσης, ο τρόπος ζύγισης των εξαγόμενων υπερωνύμων είναι ίδιος με την περίπτωση της συσταδοποίησης άρθρων νέων όπως παρουσιάστηκε στη σχέση 42.

Τα παραπάνω βήματα συνοψίζονται στους αλγορίθμους 6 και 7. Η διαδικασία συσταδοποίησης των συνεδριών των χρηστών τρέχει διαρκώς στο σύστημα και επομένως οι συνεδρίες που προκύπτουν συσταδοποιούνται παραπάνω από μία φορά σε διαφορετικά περάσματα συσταδοποίησης. Ύστερα από αυτό, για να αποφανθούμε ποια συστάδα συνεδριών θα συσχετίσουμε με έναν συγκεκριμένο χρήστη, χρησιμοποιούμε μόνο αυτές που προέκυψαν από τα πιο πρόσφατα περάσματα συσταδοποίησης. Παρά τις όποιες ομοιότητες που μπορεί να έχει η παραπάνω προσέγγιση με αυτή του fuzzy clustering, δεν πρέπει να συγχέεται μίας και αποτελεί κάτι το διαφορετικό αφού σε κάθε πέρασμα συσταδοποίησης η κάθε συνεδρία εξακολουθεί να ανήκει σε μία και μόνο συστάδα.

Αλγόριθμος 7: Εμπλουτισμός συνεδριών χρηστών με χρήση των υπερωνύμων του WordNet

Είσοδος: session s
// η συνεδρία προς ενίσχυση
Έξοδος: enrichedSession
// συνεδρία ενισχυμένων keywords με υπερώνυμα του WordNet

```
1  total_hypen_tree = NULL;
2  kws = fetch 20% most frequent k/ws for s;
3  foreach keyword kw in kws do
4      htree = wordnet_hypen_tree(kw);
       // εξαγωγή του δέντρου υπερωνύμου για αυτό το keyword
5      foreach hypen h in htree do
6          if (h not in total_hypen_tree) then
7              h.frequency=1;
8              total_hypen_tree ->append(h);
9          else
10             total_hypen_tree ->at(h)->freq++;
11 foreach h in total_hypen_tree do
12     calculate_depth(h);
13     weight = 12 * ((1(1 + exp(-0.0125 * (h->depth 3 * h->freq  kws_in_wn->size)))) -
       0.5));
14 sort_weights(total_hypen_tree);
15 important_hypens = (kws -> size  4) * top(total_hypen_tree);
16 return_kws += important_hypens return return kws[]
```

Έχοντας τα αποτελέσματα της συσταδοποίησης συνεδριών χρηστών προχωρήσαμε σε ένα πρώτο επίπεδο παράγωγης προτάσεων για τον χρήστη. Το επίπεδο αυτό δεν αποτελεί το τελικό του συστήματος προτάσεων, κάτι που θα περιγραφεί σε επόμενη ενότητα.

Πιο συγκεκριμένα λοιπόν, όταν ο χρήστης επιστρέφει στο σύστημα, η συστάδα του, με βάση τις τελευταίες συνεδρίες του, είναι ήδη γνωστή. Μπορούμε επομένως να υποθέσουμε ότι επιλογές που έχουν γίνει από άλλους χρήστες της ίδιας συστάδας είναι πολύ πιθανό να τον ενδιαφέρουν. Βασιζόμενοι σε αυτή την απλή παραδοχή (που δεν αποτελεί την ολοκληρωμένη προσέγγιση του μηχανισμού παραγωγής προτάσεων ο οποίος αναλύεται στα επόμενα), μπορούμε να επιλέξουμε άρθρα προς πρόταση με χρήση των βημάτων που περιγράφονται στον αλγόριθμο 8. Γενικά, σε αυτή την φάση κρατούσαμε 10 από τα πιο συχνά εμφανιζόμενα άρθρα στις επιλογές άλλων χρηστών της

συστάδας του χρήστη.

Αλγόριθμος 8: Παραγωγή προτάσεων άρθρων νέων βασιζόμενοι (μόνο) στην συσταδοποίηση χρηστών

Είσοδος: user u, cluster c
// ο χρήστης και η συστάδα που ανήκει
Έξοδος: suggestions
// προτάσεις άρθρων νέων βασιζόμενοι στην πληροφορία συσταδοποίησης χρηστών

```
1 suggestions [] = NULL;
2 num_sug = 10;
  // πλήθος προτάσεων προς παραγωγή
3 sessions = recover_user_clustering_info(u, c);
  // ανάκτηση πληροφορίας συσταδοποίησης χρηστών από τη ΒΔ
4 foreach s in sessions do
      // για τους χρήστες που ανήκουν στην ίδια συστάδα
5     foreach article a in s do
          // εντοπισμός άρθρων με την μεγαλύτερη συχνότητα εμφάνισης στη συστάδα
6         if freq(a) > min(freq(suggestions)) then
7             suggestions [] += article;
8 return suggestions
```

5.3.3 Προφίλ χρηστών και προσωποποίηση με χρήση συσταδοποίησης

Δεδομένου ενός χρήστη u και ενός συνόλου από άρθρα νέων R στα οποία ο u έδωσε, είτε άμεσα είτε έμμεσα, θετική ή αρνητική ανατροφοδότηση (feedback) σε σχέση με το πόσο τον ενδιαφέρουν ή όχι, ένα προφίλ χρήστη U_p διατηρείται στο σύστημα προτάσεων. Το U_p είναι ένα σύνολο από keywords, το οποίο αναλύεται σε δύο μέρη:

- το θετικό μέρος U_p^+ το οποίο αποτελείται από keywords άρθρων νέων που αξιολογήθηκαν θετικά από τον u
- το αρνητικό μέρος U_p^- το οποίο αποτελείται από keywords άρθρων νέων που αξιολογήθηκαν αρνητικά από τον u

Επιπλέον, κάθε keyword ζυγίζεται με βάρος W_{kw_i} το οποίο εξαρτάται από την ικανότητά του να αντιπροσωπεύσει την θετική ή αρνητική προτίμηση του u.

Πιο τυπικά:

$$U_p = \{U_p^+ \cup U_p^-\} \tag{46}$$

με:

$$U_p^+ = \{\cup kw_i\}, i = 1 \cdots q \quad \text{όπου} \quad q \leq \mid R \mid \tag{47}$$

$$U_p^- = \{\cup kw_j\}, j = 1 \cdots m \quad \text{όπου} \quad m \leq \mid R \mid \tag{48}$$

όπου: kw_i το θετικό keyword που εξέταση, kw_j το αρνητικό keyword που εξέταση.

Τα βήματα που ακολουθούνται από την διαδικασία προσωποποίησης προκειμένου να ποσοτικοποιηθεί η σχέση 46, παρουσιάζονται στον αλγόριθμο 9. Όταν ένας νέος χρήστης εγγράφεται στο σύστημα, δηλώνει (έμμεσα) τις προτιμήσεις του (με τον τρόπο που περιγράφεται στην ενότητα 5.4). Στην συνέχεια, και για κάθε επίσκεψη του χρήστη, ο μηχανισμός προσωποποίησης διατηρεί τις προαναφερθείσες λίστες από θετικά και αρνητικά keywords προσθαφαιρώντας στοιχεία με κατάλληλο τρόπο ώστε αυτές να ανταποκρίνονται στις πιο πρόσφατες επιλογές του χρήστη. με βάση αυτές τις λίστες μπορούμε έπειτα να προσωποποιήσουμε τα άρθρα νέων και περιλήψεις που εν' τέλει προτείνονται από το σύστημα.

Η διαδικασία ενημέρωσης του προφίλ χρήστη που περιγράφεται στον αλγόριθμο 9, τρέχοντας διαρκώς σε κάθε επίσκεψη χρήστη, λαμβάνει υπ' όψιν του τις ακόλουθες παραμέτρους:

(α′) τα άρθρα τα οποία ο χρήστης έχει επιλέξει να διαβάσει

(β′) τον χρόνο που ο χρήστης ξοδεύει διαβάζοντας την περίληψη ή το πλήρες κείμενο ενός άρθρου

(γ′) τα άρθρα που ο χρήστης αποφεύγει να διαβάσει (είτε την περίληψη ή το πλήρες κείμενο ενός άρθρου)

Τα παραπάνω πηγάζουν από τις εξής λογικές υποθέσεις:

- ένας χρήστης μάλλον θα ξοδέψει ένα χρονικό διάστημα από ένα συγκεκριμένο όριο και πάνω διαβάζοντας το πλήρες κείμενο ή την περίληψη ενός άρθρου που τον ενδιαφέρει (παράγοντας α στον αλγόριθμο 9). Τα κάτω όρια αυτά ορίζονται ως: Rar_{thr_1} και $Rsum_{thr_1}$ αντιστοίχως.
- παρόλα αυτά, ένα πάνω όριο Rar_{thr_2} και $Rsum_{thr_2}$ θα πρέπει να τεθεί για τα παραπάνω μιας και δεν θέλουμε ο μηχανισμός μας λανθασμένα να λάβει υπόψιν άρθρα που έχει ξεχάσει ο χρήστης "ανοιχτά", συγχέοντας τα έτσι με ενδιαφέροντα.

Θέτουμε τα παραπάνω όρια σε ότι έχει να κάνει με την ανάγνωση πλήρους άρθρου σε λεπτά:

$$Rar_{thr_1} = 1/2 \tag{49}$$

$$Rar_{thr_2} = 3 \tag{50}$$

αντίστοιχα, καθορίζοντας έτσι το αρχικό σύνολο από keywords άρθρων που θα πρέπει να συμπεριληφθούν στην θετική λίστα με keywords του χρήστη. Καταλήξαμε στις τιμές αυτές για τα παραπάνω όρια μετά από μελέτη των επιλογών των χρηστών στη ΒΔ του συστήματος: στις περισσότερες των περιπτώσεων, όταν ένας χρήστης ξόδευε χρόνο ανάμεσα στα παραπάνω όρια διαβάζοντας το πλήρες άρθρο, θα δρούσε σε σχέση με αυτό το άρθρο, π.χ. ακολουθώντας το σύνδεσμο προς την πηγή του άρθρου ή διαβάζοντας άρθρα σχετικά με αυτό.

Τα όρια προβολής περίληψης άρθρων είναι πιο δυναμικά σε σχέση με εκείνα του πλήρους άρθρου και υπολογίζονται ως:

$$Rsum_{thr_1} = Rar_{thr_1} * S_{ratio} \tag{51}$$

$$Rsum_{thr_2} = Rar_{thr_2} * S_{ratio} \tag{52}$$

όπου το S_{ratio} εκφράζει το "ποσοστό συμπίεσης" επί του συνολικού κειμένου που επιτυγχάνει η περίληψη του εκάστοτε κειμένου:

$$S_{ratio} = \frac{\#words(summary)}{\#words(fulltext)} \tag{53}$$

με βάση τα παραπάνω όρια λοιπόν, μπορούμε να αποφανθούμε αν κάποιο άρθρο του οποίου το πλήρες κείμενο ή την περίληψη επέλεξε προς ανάγνωση ο χρήστης, είναι ενδιαφέρον ή όχι για τον ίδιο, αρχικοποιώντας έτσι τις προαναφερθείσες λίστες από keywords (θετική και αρνητική). Έτσι θα λέγαμε πως, στις περισσότερες των περιπτώσεων, είναι αναμενόμενο ένας χρήστης να επιλέξει να διαβάσει άρθρα από ένα θέμα που βρίσκει ελκυστικό (παράγοντας b στον αλγόριθμο 9) όπως αυτό διαφημίζεται από τον τίτλο ή την περίληψή του. Επιπλέον, ένας χρήστης μάλλον θα αποφύγει να διαβάζει άρθρα που δεν του άρεσαν στο παρελθόν ή γενικά τα βρίσκει μη ενδιαφέροντα και έτσι, σε αυτή την περίπτωση, τα keywords που αναπαριστούν αυτά τα άρθρα θα πρέπει να λαμβάνουν μειωμένο ή αρνητικό βάρος (παράγοντας c στον αλγόριθμο 9).

Εκτός από την παραπάνω αρχική αποτύπωση των προτιμήσεων των χρηστών, μπορούμε να αντλήσουμε αρκετά περισσότερη πληροφορία με βάση τα δεδομένα που υπάρχουν ήδη στο σύστημα. Έτσι, επιπλέον των παραπάνω παραγόντων (a-c) του αλγορίθμου 9, γνωρίζοντας ήδη την συστάδα στην οποία ανήκει ο χρήστης, μπορούμε να αξιοποιήσουμε και αυτή την πληροφορία. Πιο συγκεκριμένα, από την συστάδα στην οποία ανήκει ο χρήστης, μπορούμε να εξάγουμε λίστες από keywords που ανήκουν σε άρθρα που πρόσφατα έχουν επισκεφθεί αρκετοί άλλοι χρήστες της συστάδας (τουλάχιστον 20% των χρηστών), προκειμένου να ενισχύσουμε την λίστα με τα θετικά keywords. Για αυτά τα άρθρα, κρατάμε τα 5 από τα πιο σημαντικά keywords ή υπερώνυμα τα οποία έχουν εξαχθεί προηγουμένως από το WordNet. Το ευρετικό αυτό, το οποίο αξιοποιεί τα αποτελέσματα της συσταδοποίησης χρηστών, αναφέρεται ως παράγοντας d στον αλγόριθμο 9.

Με χρήση των παραπάνω παραγόντων, δημιουργούνται δύο λίστες από keywords, μία θετική και μία αρνητική, τις οποίες ο αλγόριθμος προσωποποίησης λαμβάνει υπόψιν του για τις αποφάσεις του. Οι λίστες αυτές εμπεριέχουν:

- keywords προς τα οποία ο χρήστης έχει εκφράσει θετική ή αρνητική προτίμηση στο παρελθόν
- keywords από παρόμοια ενδιαφέροντα των χρηστών της ίδιας συστάδας

Η παράμετρος που καταγράφει την προτίμηση του χρήστη για ένα keyword i με βάση τους προαναφερόμενους παράγοντες a-d είναι η U_i και βασίζεται στις σχετικές συχνότητες των keywords στις λίστες, συχνότητες που συνεχώς μεταβάλλονται καθώς αποτυπώνουν όλο και καλύτερα τις

επιλογές του χρήστη. Η U_i δίνεται από την εξίσωση 54.

$$U_i = rel(fr(i)) * (1 + T_i) \tag{54}$$

όπου $rel(fr(i))$ είναι η σχετική συχνότητα του i στην θετική ή αρνητική λίστα, δηλαδή το βάρος του, T_i είναι ο κανονικοποιημένος συνολικός χρόνος που ξόδεψε ο χρήστης στο i, αν αυτό ανήκει στην θετική λίστα. Αν αντίθετα το keyword ανήκει στην αρνητική λίστα, τότε $T_i = 0$ μίας και καθόλου χρόνος δεν ξοδεύτηκε από τον χρήστη για τα εκάστοτε keywords που ανήκουν σε αυτή τη λίστα. Για την περίπτωση που ένα keyword προέρχεται από την διαδικασία συσταδοποίησης χρηστών, και επομένως δεν έχει κάποια ρητή συσχέτιση με τον χρήστη (είτε θετική είτε αρνητική), για το T_i κρατάμε τον μέσο όρο του χρόνου που οι χρήστες οι οποίοι ανήκουν στην συστάδα του χρήστη ξόδεψαν στο άρθρο από το οποίο προέρχεται αυτό το keyword.

Επιπλέον, αναμένουμε ότι όταν το προφίλ χρήστη φτάνει σε μία σταθερή κατάσταση ύστερα από αρκετές επιλογές του χρήστη, οι μέσοι χρόνοι των προτιμήσεων ως προς τα keywords θα είναι σωστοί, απεικονίζοντας επομένως τις ολικές προτιμήσεις του χρήστη. Ο συνολικός παράγοντας προσωποποίησης λοιπόν για κάθε keyword i, U_{pi} είναι:

$$U_{pi} = B * U_i \tag{55}$$

όπου για την παράμετρο B:

- > 1 όταν το keyword ανήκει στην θετική λίστα
- < 1 όταν το keyword ανήκει στην αρνητική λίστα

Η απόλυτη τιμή της παραμέτρου B μπορεί να πάρει όποια τιμή επιθυμούμε, αυξάνοντας ή μειώνοντας έτσι την επίπτωση που έχει η προσωποποίηση και η δυναμική παραγωγή προφίλ στην διαδικασία ζύγισης προτάσεων. Επομένως, το U_{pi} μπορεί να είναι θετικό, αρνητικό ή και μηδέν αν δεν υπάρχει πληροφορία για την προτίμηση του χρήστη προς το συγκεκριμένο keyword.

Έχοντας υπολογίσει το βάρος του κάθε keyword i προς τον χρήστη U_i, εξάγουμε τα υποψήφια άρθρα για τα οποία η λογική λέει ότι θα περιλαμβάνουν πολλά από τα θετικά και καθόλου ή λίγα από τα αρνητικά keywords. Πιο συγκεκριμένα, για κάθε άρθρο a, κρατούμε το αλγεβρικό άθροισμα των προτιμήσεων του χρήστη ως προς τα keywords του a, το οποίο και αποκαλούμε U_{p_a}:

$$U_{p_a} = \sum_i U_{pi} \tag{56}$$

Τα άρθρα έπειτα ταξινομούνται σε φθίνουσα κατάταξη με βάση το ολικό βάρος τους (U_{p_a}) και η λίστα δίνεται ως αποτέλεσμα του συστήματος προτάσεων.

Αλγόριθμος 9: Αλγόριθμος προσωποποίησης που ενσωματώνει την πληροφορία ανάδρασης από τον χρήστη

```
Συνάρτηση update_profile
    Είσοδος: a, b, c, d // μεταβλητές ζύγισης των διαφόρων επιλογών χρήστη
    Έξοδος: updated user profile // το ενημερωμένο προφίλ χρήστη
1   get_articles(a,b,d) // ανάκτηση άρθρων από την ΒΔ βάσει των παραμέτρων a,b,d
2   foreach article do
3       if (full article) then
4           if (time_viewed > Rar_thr1 && time_viewed < Rar_thr2) then
5               Keywords_positive = top 5 frequent keywords;
6               update_list(Positive, Keywords_positive);
7       else
8           if (time_viewed> Rsum_thr1 && time_viewed < Rsum_thr2) then
9               Keywords_positive = top 5 frequent keywords;
10              update_list(Positive, Keywords_positive);
11      get_articles(c) // για την παράμετρο c
12      foreach article do
13          Keywords_negative = top 5 frequent keywords;
14          update_list(Negative, Keywords_negative);

Συνάρτηση get_article
    Είσοδος: lists // ανά περίπτωση a,b ή c ή d (δες στη συνέχεια)
    Έξοδος: Articles // τα ανακτημένα άρθρα από τη ΒΔ βάσει της εισόδου
    /* 1) τα άρθρα στα οποία έχει πλοηγηθεί ο χρήστης καθώς και το πόσο χρόνο ξόδεψε
       διαβάζοντας το πλήρες άρθρο ή την περίληψή του (περίπτωση εισόδου a,b)   */
    /* 2) τα άρθρα με αρνητική προτίμηση (περίπτωση εισόδου c)                   */
    /* 3) τα πιο συχνά αναγνωσμένα άρθρα από τους χρήστες της συστάδας ενός χρήστη
       (περίπτωση εισόδου d)                                                     */
15  return Articles

Συνάρτηση update_list
    Είσοδος: list, keywords // keywords και η λίστα προς ενημέρωση
    /* προσθέτει τα δεδομένα keywords στην λίστα που ορίζεται                   */
16  foreach (keyword in keywords) do
17      if (keyword not in list[]) then
18          list.add(keywords[keyword])
19      else
20          list.update_freq(keywords[keyword])
```

5.4 Πρόβλημα νέου χρήστη

Στην παρούσα ενότητα θα περιγράψουμε τα διάφορα αλγοριθμικά βήματα που ακολουθούνται προκειμένου το σύστημα προτάσεων που αναπτύχθηκε να μπορεί να αντιμετωπίσει το πρόβλημα του νέου χρήστη. Ο τρόπος που επιλέξαμε να αντιμετωπίσουμε το εν' λόγω πρόβλημα ήταν μέσω της τεχνικής των προτάσεων αντικειμένων για αξιολόγηση από τον νέο χρήστη (user prompting).

Ο αλγόριθμος 10 παρουσιάζει την διαδικασία επιλογής άρθρων νέων προς συλλογή των πιθανών βαθμολογήσεών τους από τον χρήστη ο οποίος έρχεται για πρώτη φορά στο σύστημα. Τα συγκεκριμένα αλγοριθμικά βήματα εκτελούνται κατά την διάρκεια εγγραφής ενός νέου χρήστη. Επίσης, ο αλγόριθμος 12 απαριθμεί τα βήματα τα οποία χρησιμοποιούνται για την ανάκτηση άρθρων βασιζόμενοι είτε στην συσταδοποίηση άρθρων νέων είτε στη συσταδοποίηση χρηστών. Παρότι οι συναρτήσεις αυτές δεν παρουσιάζονται σε μεγάλη τεχνική λεπτομέρεια, η λειτουργία τους θα πρέπει να είναι εύκολα κατανοητή από τον αναγνώστη.

Για την επιλογή των άρθρων προς παρουσίαση στο χρήστη προκειμένου να λάβουμε κάποιες αξιολογήσεις, κάνουμε χρήση της προσωποποιημένης στοιχείο προς στοιχείο (item by item) στρατηγικής που αναφέρθηκε και στην ενότητα 3.9.1 η οποία είναι παρόμοια με εκείνη που χρησιμοποιείται στο [178]. Αρχικά, όταν ένας χρήστης εγγράφεται στο σύστημα και μπαίνει στην διαδικασία προβολής προτάσεων, του παρουσιάζουμε άρθρα από την λίστα με τα πιο δημοφιλή όπως καταγράφονται στη ΒΔ από τις επιλογές άλλων χρηστών. Έστω L_1 η λίστα αυτή.

Η παρουσίαση των άρθρων από την L_1 συνεχίζεται έως ότου ένα άρθρο, έστω A_1, αξιολογηθεί από τον χρήστη με σκορ S_1. Χρησιμοποιούμε αυτή την πληροφορία προκειμένου να καθορίσουμε την συστάδα στην οποία ανήκει αυτό το άρθρο. Έπειτα, μπορούμε να προτείνουμε για αξιολόγηση προς τον χρήστη M από τα πιο συχνά αξιολογημένα άρθρα από αυτή τη συστάδα, τα οποία και διαμορφώνουν την λίστα L_2 Η L_2 επομένως περιέχει άρθρα βασισμένα στην πληροφορία συσταδοποίησης που έρχεται από τη ΒΔ. Η επιλογή της κατάλληλης τιμής για το M είναι αποτέλεσμα πειραματισμού μιας και υπάρχει ένα συγκεκριμένο trade-off. Μεγάλες τιμές M δίνουν πολλά σχετικά άρθρα σε σχέση με το A_1 και επομένως, μία επιτυχής αξιολόγηση, στη συνέχεια πιθανά θα αντλήσει πολλές αξιολογήσεις χρήστη στα επόμενα άρθρα που θα παρουσιαστούν. Παρόλα αυτά, εάν το αξιολογημένο άρθρο A_1 δεν καλύπτει αρκετά καλά τα ενδιαφέροντα του χρήστη (π.χ. αξιολογήθηκε επιπόλαια ή τυχαία), πολλά άρθρα θα προταθούν για αξιολόγηση στην συνέχεια από το σύστημα τα οποία με μεγάλη πιθανότητα δεν θα αξιολογηθούν, χωρίς κιόλας τη δυνατότητα να παραληφθούν. Το τελευταίο μπορεί να έχει αρνητική επίπτωση στην αποδοτικότητα του συστήματος προτάσεων, ενώ παράλληλα θα προκαλέσει δυσφορία στον νέο χρήστη (κάτι που προφανώς πρέπει να αποφεύγεται). Από την άλλη μεριά βέβαια, μικρές τιμές M μπορεί να οδηγήσουν σε μία παρόμοια αρνητική επίπτωση αλλά από διαφορετικό μονοπάτι: ένας χρήστης περιμένει από ένα σύστημα προτάσεων γρήγορα να αντιλαμβάνεται τις προτιμήσεις του και να μην ολισθαίνει σε άρθρα που δεν τον ενδιαφέρουν.

Συνοπτικά, δεν θέλουμε να υπερφορτώσουμε τον χρήστη προτείνοντας για αξιολόγηση πολλά άρθρα μόνο από μία συστάδα, αλλά όμως θέλουμε να αποφαινόμαστε σχετικά γρήγορα εάν τα άρ-

θρα που ανήκουν στην εν' λόγω συστάδα είναι πράγματι ενδιαφέροντα για τον χρήστη. Επιπλέον, θέλουμε να καλύπτουμε όσο πιο ευρύτερα γίνεται τις σχετικές συστάδες οι οποίες μπορεί να ενδιαφέρουν τον χρήστη μιας και μπορούμε να αντλήσουμε γρήγορα αξιολογήσεις με αυτό τον τρόπο. Κατά συνέπεια, μία μικρή προς μέση τιμή για το M φαντάζει πιο λογική.

Καθώς ο αλγόριθμός μας προχωράει, εάν δεν ληφθούν αξιολογήσεις για κανένα από τα M άρθρα στην L_2, αναζητούμε για συστάδες χρηστών οι οποίοι προηγουμένως βαθμολόγησαν το A_1 με σκορ S_1. Αξιοποιούμε αυτές τις συστάδες για να σχηματίσουμε μία λίστα από άρθρα, έστω L_3, η οποία αποτελείται από $M*$ το πλήθος των συστάδων των πιο συχνά αξιολογημένων άρθρων. Ξανά, επιλέγουμε να κρατήσουμε M άρθρα από καθεμία από τις παραπάνω συστάδες και όπως και πριν ισχύουν τα ίδια trade-offs για την επιλογή του M. Η λίστα $L3$, περιέχοντας πληροφορία συσταδοποίησης χρηστών στην συνέχεια προτείνεται (άρθρα ένα προς ένα) στον χρήστη προς αξιολόγηση. Όσες αξιολογήσεις ληφθούν, χρησιμοποιούνται αναδρομικά για την επαναδημιουργία της λίστας L_3 με παρόμοιο τρόπο όπως και πριν. Η διαδικασία αυτή συνεχίζεται έως ότου το πλήθος των αξιολογήσεων φτάσει το όριο που έχει οριστεί, έστω R_{min} όσον αφορά στο πλήθος των αξιολογήσεων.

Αντιθέτως, εάν ο χρήστης έχει αξιολογήσει τουλάχιστον ένα άρθρο από τα M της L_2 λίστας, αναζητούμε για συστάδες χρηστών που περιέχουν επιλογές με τα περισσότερα από τα προηγούμενα βαθμολογημένα άρθρα και ξανά επιλέγουμε τα $M*$ πλήθος συστάδων από τα πιο αξιολογημένα άρθρα τα οποία και αναθέτουμε στην λίστα L_4. Παρότι οι λίστες L_3 και L_4 μοιάζουν, δεν είναι ίδιες. Η διαφορά έγκειται στο ότι η L_3 βασίζεται αποκλειστικά στην πληροφορία συσταδοποίησης χρηστών, ενώ η L_4 αρχικοποιείται από την συσταδοποίηση άρθρων αρχικά και έπειτα ενισχύεται από την συσταδοποίηση χρηστών αξιοποιώντας έτσι την συνεργατική γνώση που βρίσκεται στη ΒΔ.

Τελικά, όταν τα R_{min} άρθρα έχουν αξιολογηθεί από τον χρήστη, η διαδικασία εγγραφής ολοκληρώνεται και ο χρήστης μπορεί να πλοηγηθεί στις προσωποποιημένες προτάσεις που το σύστημα πλέον μπορεί να του παρέχει.

Αλγόριθμος 10: Καθορισμός των άρθρων νέων που θα παρουσιαστούν στο χρήστη προς αξιολόγηση

Είσοδος: NULL

Έξοδος: user_ratings[]// βαθμολογημένα άρθρα από τον χρήστη

```
1  rated_article = NULL // πρώτο βαθμολογημένο άρθρο - A1
2  article_cluster = NULL;
3  articles_next [] = NULL;
4  rated_articles[] = NULL;
5  while (!rated_article and rated_article < average_rate (article)) do
6      rated_article = rate(present_next_most_rated_article());
       // συνέχισε να παρουσιάζεις άρθρα από την L1 λίστα
       // έως ότου ο χρήστης βαθμολογήσει 1 άρθρο
7  user_ratings[] += rated_article // Το άρθρο A1 βαθμολογείται με σκορ S1 >
       average_rate
8  article_cluster = find_article_cluster(rated_article);
9  articles_next[] = find_most_rated_articles(article_cluster, M);
   // το articles_next[] είναι τώρα η L2 λίστα που περιλαμβάνει M άρθρα
10 while (has_next(articles_next)) do
11     rated_articles[] = rate(present_next_article(articles_next))
12 if (!rated_articles[]) then
       // ο χρήστης δεν έχει βαθμολογήσει κανένα από τα M άρθρα της λίστας L1
13     articles_next[]=find_most_rated_articles_from_user_clusters (rated_article,M)
           // το articles_next[] είναι τώρα η λίστα L3
14     rated_articles[] = rate(articles_next[]);
15     user_ratings [] += rated_articles[];
16     GOTO: T ;
       // συνέχισε με πιθανές προτάσεις από τα αποτελέσματα της συσταδοποίησης χρηστών
17 else
       // ο χρήστης έχει βαθμολογήσει ορισμένα από τα M άρθρα
18     user_ratings [] += rated_articles[];
19     T:;
20     while (user_ratings.size() < Rmin) do
           // έχουμε αρκετές βαθμολογήσεις;
21         articles_next[]=find_most_rated_articles_from_user_clusters(rated_articles,M)
               // το articles_next[] είναι τώρα η λίστα L4
22         rated_articles[] = rate(articles_next[]);
23         user_ratings [] += rated_articles[];
24 return user_ratings[]
```

Ορισμένες βοηθητικές συναρτήσεις του αλγορίθμου 10 ακολουθούν δίνονται στον αλγόριθμο 11.

Αλγόριθμος 11: Συναρτήσεις που χρησιμοποιούνται στον αλγόριθμο 10

Συνάρτηση *average_rate*
Είσοδος: article
Έξοδος: average rating
/* Ανάκτηση από τη ΒΔ της μέσης βαθμολογίας για αυτό το άρθρο από οποιονδήποτε χρήστη το βαθμολόγησε */
1 **return** *user_ratings[]*

Συνάρτηση *rate*
Είσοδος: articles[]
Έξοδος: rated_articles[]
/* Παρουσιάζει για βαθμολόγηση τα επιλεγμένα άρθρα και επιστρέφει τις βαθμολογίες (σκορ) ή null αν ένα άρθρο δεν βαθμολογήθηκε */
2 rated_articles[]=NULL;
3 **return** *rated_articles[]*

Συνάρτηση *find_article_cluster*
Είσοδος: rated_article
Έξοδος: article_cluster
/* Ανάκτηση από τη ΒΔ της συστάδας στην οποία ανήκει το άρθρο */
4 **return** *article_cluster*

Αλγόριθμος 12: Ανάκτηση άρθρων βασιζόμενοι σε συστάδες άρθρων ή χρηστών συστήματος

```
Συνάρτηση find_most_rated_articles
    Είσοδος: cluster, M
    Έξοδος: articles[M]
    /* Ανακτά τα M πιο βαθμολογημένα άρθρα τα οποία ανήκουν στην συστάδα cluster   */
    /* Χρησιμοποιεί τα αποτελέσματα της συσταδοποίησης από τη ΒΔ. Η συστάδα μπορεί
       να είναι είτε άρθρων νέων, είτε χρηστών                                      */

Συνάρτηση find_most_rated_articles_from_user_clusters
    Είσοδος: article/articles[], size M
    Έξοδος: rated_articles[]
    /* Ανακτά άρθρα από τις συστάδες χρηστών τα οποία περιέχουν χρήστες που
       προηγουμένως βαθμολόγησαν την δεδομένη λίστα από άρθρα. Χρησιμοποιεί τα
       αποτελέσματα της συσταδοποίησης χρηστών                                      */
    rated_articles[]=NULL;
    clusters[]= find_user_clusters(article);
    // βρες τις συστάδες χρηστών που έχουν βαθμολογήσει τα άρθρα
    foreach cluster in clusters[] do
        rated_articles[]+=find_most_rated_articles (cluster, M);
    return rated_articles[]
```

ΚΕΦΑΛΑΙΟ 6

ΤΕΧΝΟΛΟΓΙΕΣ ΥΛΟΠΟΙΗΣΗΣ ΚΑΙ ΠΡΟΔΙΑΓΡΑΦΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

> Ignorance, the root and the stem of every evil.
>
> *Plato, Greek Philosopher, 347 BC*

Στο παρόν κεφάλαιο περιγράφονται οι διαθέσιμες τεχνολογίες που υπάρχουν και που αφορούν στα διάφορα υποσυστήματα του μηχανισμού που υλοποιήθηκε. Παράλληλα αναφέρονται οι αποφάσεις που πάρθηκαν όσον αφορά στο ποιες τεχνολογίες τελικά επιλέχθηκαν και τέλος δίνεται και η αντίστοιχη αιτιολόγηση για τις αποφάσεις αυτές. Στο παρόν κεφάλαιο δεν αναφέρουμε διεξοδικά όλες οι διαθέσιμες τεχνολογίες, παρά περιγράφονται οι κατά τη γνώμη μας πιο σημαντικές. Για μία διεξοδική αναφορά ο αναγνώστης μπορεί φυσικά να ανατρέξει στις εκάστοτε αναφορές ή και στην μεταπτυχιακή μου εργασία.

Παράλληλα, δίνονται οι προδιαγραφές του συστήματος προτάσεων ώστε αυτό να είναι σε θέση να λειτουργεί σωστά και να παράγει αποτελέσματα που έχουν αξία. Τέλος αναφέρονται και ορισμένα στοιχεία που έχουν να κάνουν με τις απαιτήσεις του μηχανισμού σε υλικό και λογισμικό ώστε να μπορεί να λειτουργεί αποτελεσματικά.

6.1 Γλώσσα υλοποίησης βασικών υποσυστημάτων

Όλα τα βασικά υποσυστήματα του συστήματος προτάσεων που υλοποιήθηκε, είναι γραμμένα είτε σε γλώσσα C, είτε σε C++. Με τον όρο "βασικά" αναφερόμαστε στα υποσυστήματα που εκτελούν διεργασίες πυρήνα, δηλαδή: ανάκτηση άρθρων από τον παγκόσμιο ιστό, προεπεξεργασία κειμένων, κατηγοριοποίηση, παραγωγή περίληψης και συσταδοποίηση άρθρων νέων και χρηστών. Υποσυστήματα όπως η παρουσίαση της πληροφορίας μπορούν να είναι σε οποιαδήποτε γλώσσα αυτό επιθυμείται, μιας και δεν επιτελούν υπολογισμούς παρά μόνο διαβάζουν πληροφορία από την ΒΔ του συστήματος και την παρουσιάζουν. Ακολουθεί μία συνοπτική περιγραφή των δύο γλωσσών και τεκμηρίωση της παραπάνω απόφασης σχετικά με τις επιλεγμένες γλώσσες προγραμματισμού.

6.2 Προεπεξεργασία

6.2.1 Εξαγωγή n-grams

Για την εξαγωγή των n-grams από το σώμα των άρθρων νέων που ανακτούνται από το διαδίκτυο, αξιοποιήσαμε το πακέτο λογισμικού εξαγωγής n-grams από το [229]. Το πακέτο λογισμικού αυτό παρέχει μία σειρά από εργαλεία για την εξαγωγή και διαχείριση n-grams από ακατέργαστο κείμενο. Το βασικό πρόγραμμα υλοποιεί τον αλγόριθμο εξαγωγής που περιγράφεται στο [151] και μπορεί να εξάγει τόσο n-grams λέξεων (για δυτικές γλώσσες), όσο και χαρακτήρων (π.χ. για Κινέζικα).

Πρόκειται για μία εξαιρετικά αποδοτική υλοποίηση των εν' λόγω τεχνικών και δεν προβλημάτισε ο ελάχιστος χρόνος που απαιτήθηκε προκειμένου να εξάγονται τα n-grams του κειμένου. Παράλληλα βρέθηκε ακριβής ως προς τα αποτελέσματά του, της τάξης του 95% κατά μέσο όρο, και επομένως αποτέλεσε μία αυτονόητη επιλογή προς χρήση.

Άλλα εργαλεία εξαγωγής n-grams, τα οποία όμως δεν είναι σε γλώσσα C/C++, είναι τα εξής:

- TextToolbox NGramCounter - A Free Web API for N-Gram Generation [158]
- A Simple Ruby N-Gram Generator [2]
- A Java Example for N-Gram Generation [1]

6.2.2 Υπερώνυμα του WordNet

Για την εξαγωγή των υπερωνύμων του WordNet αξιοποιήσαμε την δυνατότητα που μας δίνει το ίδιο το εργαλείο του WordNet, σε μορφή βιβλιοθήκης της C, προκειμένου να αλληλεπιδράσει με το σύστημα προεπεξεργασίας μας.

Όπως έχουμε αναφέρει ήδη, το WordNet είναι μια λεξικολοφική βάση γνώσης για την Αγγλική γλώσσα. Ουσιαστικά, ρήματα, επίθετα και επιρρήματα οργανώνονται σε σύνολα από συνώνυμα (synset) με το καθένα να εκφράζει ένα διακριτό νόημα. Τα synsets διασυνδέονται μέσω

εννοιολογικών-σημασιολογικών καθώς και λεξικολογικών συσχετίσεων. Το δίκτυο που προκύπτει από συσχετιζόμενες λέξεις και νοήματα μπορεί να είναι διαθέσιμο στον χρήστη ή σε άλλα προγράμματα προς αξιοποίηση. Το WordNet είναι επίσης ελεύθερα διαθέσιμο για κατέβασμα και χρήση, βρίσκοντας εφαρμογή σε πολλά θέματα υπολογιστικής και λεξικολογικής επεξεργασίας φυσικής γλώσσας.

Η έκδοση του WordNet που χρησιμοποιήθηκε στα πλαίσια της διδακτορικής διατριβής είναι η 3.0 ή οποία αποτελεί και την τελευταία διαθέσιμη για Linux.

6.3 Συσταδοποίηση

6.3.1 Υλοποιήσεις αλγορίθμων συσταδοποίησης

Πολλές από τις προαναφερθείσες μεθοδολογίες συσταδοποίησης έχουν ενσωματωθεί σε πακέτα συσταδοποίησης, όπως το CLUTO [112], το SenseClusters [175], κ. α. Ακολουθεί μία σύντομη περιγραφή των πιο σημαντικών από αυτά.

6.3.1.1 CLUTO

Το CLUTO clustering toolkit [112] είναι ένα πακέτο λογισμικού για την συσταδοποίηση τόσο χαμηλής όσο και υψηλής διαστατικότητας δεδομένων, το οποίο δίνει τη δυνατότητα ανάλυσης των χαρακτηριστικών των διαφόρων συστάδων. Αποτελεί μία εξαιρετική λύση για συσταδοποίηση δεδομένων που πηγάζουν από διαφορετικές περιοχές εφαρμογών όπως ανάκτηση πληροφορίας, ιστός, GIS και βιολογία. Το CLUTO αποτελείται τόσο από αυτόνομα προγράμματα που υλοποιούν τους αλγορίθμους συσταδοποίησης, όσο και από μία βιβλιοθήκη μέσω της οποίας μπορούμε να αξιοποιήσουμε απευθείας τους διαφόρους αλγορίθμους είτε συσταδοποίησης είτε ανάλυσης.

Το CLUTO παρέχει τρεις διαφορετικές κλάσεις αλγορίθμων συσταδοποίησης οι οποίοι δρουν είτε απ' ευθείας πάνω στο χώρο των αντικειμένων (feature space), είτε στο χώρο ομοιότητας αυτών (similarity space). Ένα σημαντικό χαρακτηριστικό στους περισσότερους αλγορίθμους συσταδοποίησης του CLUTO είναι ότι αντιμετωπίζουν το πρόβλημα της συσταδοποίησης ως μία διαδικασία βελτιστοποίησης η οποία προσπαθεί να μεγιστοποιήσει ή να ελαχιστοποιήσει μία συγκεκριμένη συνάρτηση κριτηρίου η οποία ορίζεται είτε συνολικά, είτε τοπικά σε σχέση με το χώρο λύσεων του προβλήματος. Το CLUTO έχει δύο τρόπους εκτέλεσης, έναν που αντιμετωπίζει τα αντικείμενα ως διανύσματα σε έναν πολλαπλών διαστάσεων χώρο, και έναν ο οποίο ενεργεί πάνω στο χώρο ομοιότητας μεταξύ των αντικειμένων. Και οι δύο τρόποι υπολογίζουν την λύση στο πρόβλημα της συσταδοποίησης χρησιμοποιώντας μία από τις πέντε παρακάτω διαφορετικές προσεγγίσεις. Οι τέσσερις από αυτές τις προσεγγίσεις είναι μερισματικές στη φύση τους, ενώ η πέμπτη προσέγγιση είναι ιεραρχική (agglomerative).

Ορισμένα από τα προτερήματα του CLUTO είναι τα εξής:

- πολλαπλές κλάσεις αλγορίθμων συσταδοποίησης: διαμερισματικοί, ιεραρχικοί, καθώς και γραφο-μερισματικοί

- πολλαπλές μετρικές ομοιότητας/συναρτήσεις απόστασης για αξιοποίηση στους εν' λόγω αλγορίθμους:
  - Ευκλείδεια απόσταση
  - ομοιότητα συνημιτόνου
  - correlation coefficient
  - extended Jaccard
  - καθώς και δυνατότητα ορισμού από τον χρήστη
- πολλαπλά state of the art κριτήρια συσταδοποίησης καθώς και σχήματα συγχώνευσης ιεραρχικών αλγορίθμων
- παραδοσιακοί ιεραρχικοί αλγόριθμοι single-link, complete-link, UPGMA
- εκτεταμένες δυνατότητες οπτικής απεικόνισης συστάδων καθώς και εξόδου σε αρχείο: postscript, SVG, gif, xfig, κ.λπ.
- εύκολη κλιμάκωση για χιλιάδες αντικείμενα και δεκάδες χιλιάδες διαστάσεις

Περισσότερες πληροφορίες για το CLUTO clustering toolkit είναι διαθέσιμες στο [51].

6.3.1.2 SenseClusters

Το SenseClusters [175] είναι χοντρικά ένα σύστημα διακρίσεων ερμηνειών λέξεων. Παράγει συστάδες οι οποίες σχηματίζονται από τα συμφραζόμενα στα οποία μία δεδομένη λέξη εμφανίζεται. Δεν χρησιμοποιεί άλλη γνώση πέρα από αυτή που είναι διαθέσιμη σε ένα μη δομημένο corpus, ενώ οι συστάδες για μία δεδομένη λέξη-στόχος βασίζονται μόνο στις αμοιβαίες ομοιότητες από τα συμφραζόμενα.

Επί της ουσίας, το SenseClusters αποτελεί ένα πακέτο από (κυρίως) Perl προγράμματα το οποίο παρέχει αλγορίθμους συσταδοποίησης δεδομένων με χρήση μη εποπτευόμενων τεχνικών εκμάθησης. Το SenseClusters αξιοποιεί διαφορετικά άλλα προγράμματα (όπως το CLUTO) προκειμένου να παράγει τις επιθυμητές συσταδοποιήσεις. Βασίζεται αυστηρά σε λεξικολογικά χαρακτηριστικά των κειμένων και όχι στην εκπαίδευση συστήματος ή χρήση εξωτερικής βάσης γνώσης. Το SenseClusters έχει την δυνατότητα αυτόματου καθορισμού του πλήθους των συστάδων στα δεδομένα βασιζόμενο σε ένα πλήθος από κριτήρια αυτόματου τερματισμού της εκτέλεσης. Περισσότερες πληροφορίες για το SenseClusters είναι διαθέσιμες στο [194]

6.3.1.3 Συσταδοποίηση στη MATLAB

Υπάρχει πληθώρα από βιβλιοθήκες συσταδοποίησης που έχουν υλοποιηθεί στην MATLAB, και συγκεκριμένα το [50] αποτελεί μία καλή περίληψη των διαθέσιμων επιλογών. Εκτός από τις λύσεις που προτείνονται στο παραπάνω, διάφορα άλλα toolkits έχουν δημιουργηθεί τα οποία καλύπτουν αρκετούς αλγορίθμους συσταδοποίησης, όπως για παράδειγμα το TMG.

6.3.1.3.1 Text to Matrix Generator

Το Text to Matrix Generator (TMG)[234][205] αποτελεί ένα MATLAB toolbox το οποίο μπορεί να χρησιμοποιηθεί για πολλές εργασίες που έχουν να κάνουν με text mining. Το μεγαλύτερο μέρος του TMG είναι γραμμένο σε MATLAB, παρότι ένα μεγάλο τμήμα της φάσης δεικτοδότησης στην τελευταία έκδοση είναι γραμμένο σε Perl. Παράλληλα, το TMG συνεργάζεται εύκολα με MySQL προσφέροντας έτσι ευελιξία στη χρήση.

Το TMG ταιριάζει ιδιαίτερα σε text mining εφαρμογές με δεδομένα υψηλής διαστατικότητας αλλά εξαιρετικά αραιά, μίας και χρησιμοποιεί την υποδομή αραιών πινάκων (sparce matrices) της MATLAB. Αρχικά φτιάχτηκε στο Πανεπιστήμιο Πατρών ως ένα εργαλείο προ-επεξεργασίας κειμένου προκειμένου να παράγει τους πίνακες όρων-κειμένων από αδόμητο κείμενο. Η τελευταία έκδοση προσφέρει πολλά περισσότερα, όπως: δεικτοδότηση, ανάκτηση, μείωση διαστατικότητας, μη-αρνητική παραγοντοποίηση πινάκων, συσταδοποίηση και κατηγοριοποίηση.

Η λειτουργικότητα του TMG είναι διαθέσιμη στον χρήστη με διάφορους τρόπους. Είτε απευθείας μέσω του GUI που βασίζεται σε MATLAB (έκδοση 7.0 η μεταγενέστερη), είτε απευθείας μέσω του command line interface της MATLAB απ' όπου μπορούν να κληθούν οι επιθυμητές συναρτήσεις.

Η διαδικασία προεπεξεργασίας του TMG παρέχει παρόμοια λειτουργικότητα με αυτή που περιγράφηκε στα πλαίσια της διδακτορικής διατριβής, περιλαμβάνοντας διάφορα βήματα που έχουν να κάνουν με μείωση διαστατικότητας, όπως αφαίρεση κοινότυπων λέξεων (stopwords), αφαίρεση πολύ σύντομων ή πολύ μεγάλων όρων, αφαίρεση πολύ συχνών ή πολύ σπάνιων όρων. Επίσης δίνει τη δυνατότητα εξαγωγής ζυγίσματος όρων όπως και σχημάτων κανονικοποίησης και εξαγωγής ρίζας λέξεων (stemming).

6.3.1.4 C Clustering Library

Η βιβλιοθήκη C Clustering Library [40] αποτελεί μία ελαφριά (lightweight) και open source υλοποίηση των διαφόρων αλγορίθμων συσταδοποίησης που χρησιμοποιήθηκαν κατά την διάρκεια εκπόνησης της διδακτορικής διατριβής. Η βιβλιοθήκη περιλαμβάνει τόσο τους αλγορίθμους συσταδοποίησης τους οποίους αξιολογήσαμε απ' ευθείας (ενότητα 7.2.1.1), όσο και τον k-means πυρήνα του αλγορίθμου W-kmeans που αναπτύχθηκε.

Οι μέθοδοι συσταδοποίησης που προσφέρει η εν' λόγω βιβλιοθήκη μπορούν να αξιοποιηθούν με πολλούς τρόπους. Η έκδοση Cluster 3.0 παρέχει μία γραφική διεπαφή για την πρόσβαση σε ρουτίνες διεπαφής. Η βιβλιοθήκη είναι διαθέσιμη για όλες τις πλατφόρμες και παρέχει interfaces σε πολλαπλές γλώσσες προγραμματισμού. Το βασικότερο πλεονέκτημα όμως της βιβλιοθήκη είναι ότι είναι γρήγορη στις κλήσεις της (σε γενικές γραμμές βρήκαμε κάθε κλήση γρηγορότερη σε σχέση με τα υπόλοιπα toolkits) και εξαιρετικά αποτελεσματική όσον αφορά τις παραγόμενες συστάδες. Επιπλέον, η αξιοποίησή της και διασύνδεση με το σύστημα προτάσεων ήταν αρκετά εύκολη μιας και είναι υλοποιημένη σε γλώσσα προγραμματισμού C όπως και τα υπόλοιπα βασικά υποσυστήματα του συστήματος προτάσεων μας.

Οι αλγόριθμοι συσταδοποίησης που υποστηρίζει η C Clustering Library είναι οι εξής:

- ιεραρχικοί αλγόριθμοι: pairwise centroid linkage, single linkage, complete linkage, και average linkage
- k-means
- Self-organizing maps
- PCA

Οι μετρικές ομοιότητας/συναρτήσεις απόστασης που υποστηρίζει η C Clustering Library είναι οι εξής:

- συσχέτιση Pearson, απόλυτη τιμή της
- ομοιότητα συνημιτόνου
- συσχέτιση Spearman's rank
- Kendall's
- Ευκλείδεια απόσταση
- city-block απόσταση

Περισσότερες πληροφορίες σχετικά με την C Clustering Library μπορούν να βρεθούν και στο user manual της βιβλιοθήκης[1]. Τέλος, η έκδοση της βιβλιοθήκης C Clustering Library που αξιοποιήθηκε στα πλαίσια της διδακτορικής διατριβής ήταν η cluster-1.52a η οποία είναι και η τελευταία διαθέσιμη κατά την συγγραφή του παρόντος.

6.4 Βάση δεδομένων

6.4.1 MySQL

Η MySQL είναι η δημοφιλέστερη Βάση Δεδομένων ανοιχτού κώδικα που προσφέρεται από το Δίκτυο MySQL. Η αρχιτεκτονική της την κάνει να είναι εξαιρετικά γρήγορη και πολύ εύκολη σε αλλαγές και αναβαθμίσεις. Επιτρέπει επαναχρησιμοποίηση κώδικα όπου αυτό είναι αναγκαίο και παρέχει ένα μινιμαλιστικό τρόπο δημιουργίας στοιχείων διαχείρισης βάσης δεδομένων τέτοιο ώστε να κάνει τη MySQL ασύγκριτη σε ταχύτητα, σε κατάληψη χώρου, σταθερότητα και ευκολία. Ο μοναδικός στο είδος του διαχωρισμός του κεντρικού πυρήνα του server από το μηχανισμό αποθήκευσης κάνει δυνατή την ύπαρξη αυστηρού ελέγχου σε συναλλαγές και μείωση ταχύτητας ή ύπαρξη θεαματικά μεγάλης ταχύτητας με απευθείας προσπέλαση των δεδομένων, στοιχεία που μπορούν να χρησιμοποιηθούν ανάλογα με τις ανάγκες των χρηστών. Η MySQL περιλαμβάνει αποθήκευση σε μηχανή InnoDB, η οποία υποστηρίζει ασφάλεια στις συναλλαγές και ACID-συμβατή μηχανή αποθήκευσης με commit, rollback, crash recovery και low-level locking δυνατότητες. Η έκδοση της

[1] http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster.pdf

MySQL που βρίσκεται αυτή τη στιγμή σε σταθερή κατάσταση είναι η 5.5.40 και υποστηρίζει πολλά στοιχεία που αφορούν την απόδοση, τη διεθνοποίηση και τη δυνατότητα ένταξης του MySQL server σε άλλα στοιχεία υλικού και λογισμικού. Τα πιο βασικά στοιχεία που χαρακτηρίζουν τη MySQL είναι:

- Υποερωτήματα, που επιτρέπουν στους χρήστες να κάνουν σύνθετα ερωτήματα με μεγάλη ευκολία και αποδοτικά.
- Γρήγορη επικοινωνία μεταξύ server και client μέσα από ένα καινούριο πρωτόκολλο.
- Μικρότερη κατανάλωση πόρων από το server μέσα από βελτιστοποίηση στις βιβλιοθήκες.
- Υποστήριξη Unicode, διεθνείς χαρακτήρες και υποστήριξη αποθήκευσης στην πλειοψηφία των συνόλων χαρακτήρων.
- Υποστήριξη τύπων GIS για ερωτήματα που αφορούν χάρτες και γεωγραφικά δεδομένα.

Τα παραπάνω στοιχεία κάνουν τη MySQL ένα υπερ-πολύτιμο εργαλείο στα χέρια κάποιου χρήστη και τη θέτουν στην 1η θέση για επιλογή ως βάση δεδομένων του συστήματός μας [150]

6.4.2 Βάση δεδομένων του συστήματος

Το σύστημα μας λοιπόν στην παρούσα έκδοσή του χρησιμοποιεί την έκδοση 5.5.40 της MySQL και η οποία αποτελεί και το ουσιαστικό επίπεδο διασύνδεσης μεταξύ των διαφορετικών υποσυστημάτων που έχουν υλοποιηθεί.

Μία εκτενής ανάλυση των πινάκων του συστήματος ξεφεύγει από το σκοπό της παρούσας διατριβής, εξάλλου κάτι τέτοιο έχει ήδη γίνει στην μεταπτυχιακή μου εργασία. Παρόλα αυτά, κάποια γενικά στοιχεία για τη βάση δεδομένων αναφέρονται στη συνέχεια. Η ΒΔ του συστήματος προτάσεων είναι ουσιαστικά το βασικό επίπεδο συντονισμού και επικοινωνίας των διαφόρων υποσυστημάτων από τα οποία απαρτίζεται. Ορισμένοι πίνακες που αξίζει να αναφέρουμε είναι οι εξής:

- αυτοί που αφορούν τα άρθρα νέων και οι οποίοι είναι και οι μεγαλύτεροι σε όγκο πληροφορίας. Επίσης ο πίνακας των άρθρων έχει πολλές πληροφορίες ανάλογα με το στάδιο στο οποίο βρίσκεται το σύστημα μέσα από συγκεκριμένα flags τα οποία αποθηκεύονται σε πεδία της ΒΔ για κάθε άρθρο
- αυτοί που αφορούν την κατηγοριοποίηση που πραγματοποιεί το αντίστοιχο υποσύστημα, καθώς και οι πίνακες που κρατάνε την πληροφορία εκπαίδευσης

Η βάση δεδομένων του συστήματος έχει δεχθεί πολλές σχεδιαστικές αλλαγές σε σχέση με αυτή που χρησιμοποιήθηκε στα πλαίσια της διπλωματικής ή μεταπτυχιακής εργασίας. Αυτό είναι ένα στοιχείο θετικό για το σύστημα καθώς με αυτό τον τρόπο έχουμε μία πιο συνοπτική αναπαράσταση των δεδομένων που αποθηκεύει το σύστημά μας και επομένως καλύτερη απόδοση όσον αφορά στην εκτέλεση των ερωτημάτων.

Μία γενική εικόνα της βάσης δεδομένων φαίνεται στο σχήμα 16. Από το σχήμα αυτό βέβαια λείπουν οι νέοι πίνακες που προσθέθηκαν στα πλαίσια της διδακτορικής διατριβής και οι οποίοι αναλύονται στην ενότητα που ακολουθεί.

6.4.2.1 Νέοι πίνακες

Το E-R διάγραμμα των νέων πινάκων της ΒΔ που υλοποιήθηκαν φαίνεται στο σχήμα 17. Μπορούμε να δούμε ότι οι νέοι πίνακες χωρίζονται σε τρεις κατηγορίες:

- πίνακες συσταδοποίησης άρθρων νέων
- πίνακες συσταδοποίησης χρηστών
- πίνακες εξαγωγής πληροφορίας n-grams

Στη συνέχεια δίνουμε ορισμένα στοιχεία για καθέναν από τους νέους (συνολικά 12) πίνακες

6.4.2.1.1 Πίνακες συσταδοποίησης άρθρων νέων

Ακολουθεί μία σύντομη περιγραφή των πινάκων που έχουν να κάνουν με την διαδικασία συσταδοποίησης άρθρων νέων από τον αλγόριθμο W-kmeans.

6.4.2.1.1.1 clustering_passes Ο πίνακας clustering_passes την απαραίτητη πληροφορία που αφορά κάθε πέρασμα συσταδοποίησης άρθρων νέων. Πιο συγκεκριμένα, περιλαμβάνει:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

sum_of_dist Άθροισμα των αποστάσεων που καταγράφηκαν σε αυτό το πέρασμα. Η πληροφορία είναι χρήσιμη για τον μετέπειτα υπολογισμό μετρικών όπως η Clustering Index (CI)

avg_intra_sim Η μέση εσω-συσταδική απόσταση των συστάδων του περάσματος

avg_inter_sim Η μέση δια-συσταδική απόσταση των συστάδων του περάσματος

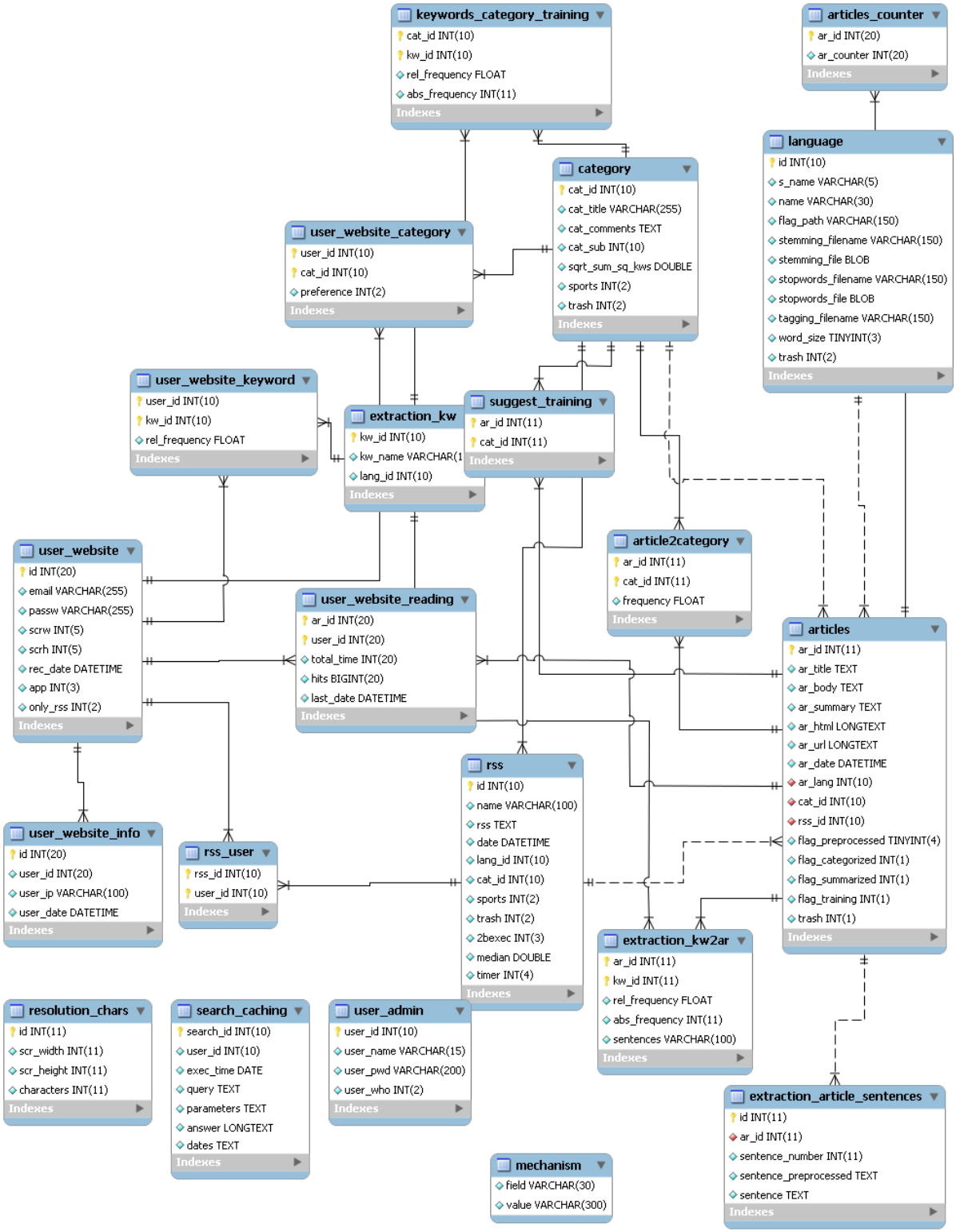
singletons Το πλήθος των μοναδιαίων συστάδων του περάσματος (συστάδες με ένα μόνο αντικείμενο)

clusterinig_index Δείκτης συσταδοποίησης για το συγκεκριμένο πέρασμα. Ουσιαστικά αποτελεί το βασικό κριτήριο αξιολόγησης του πόσο αποτελεσματικό ήταν το πέρασμα

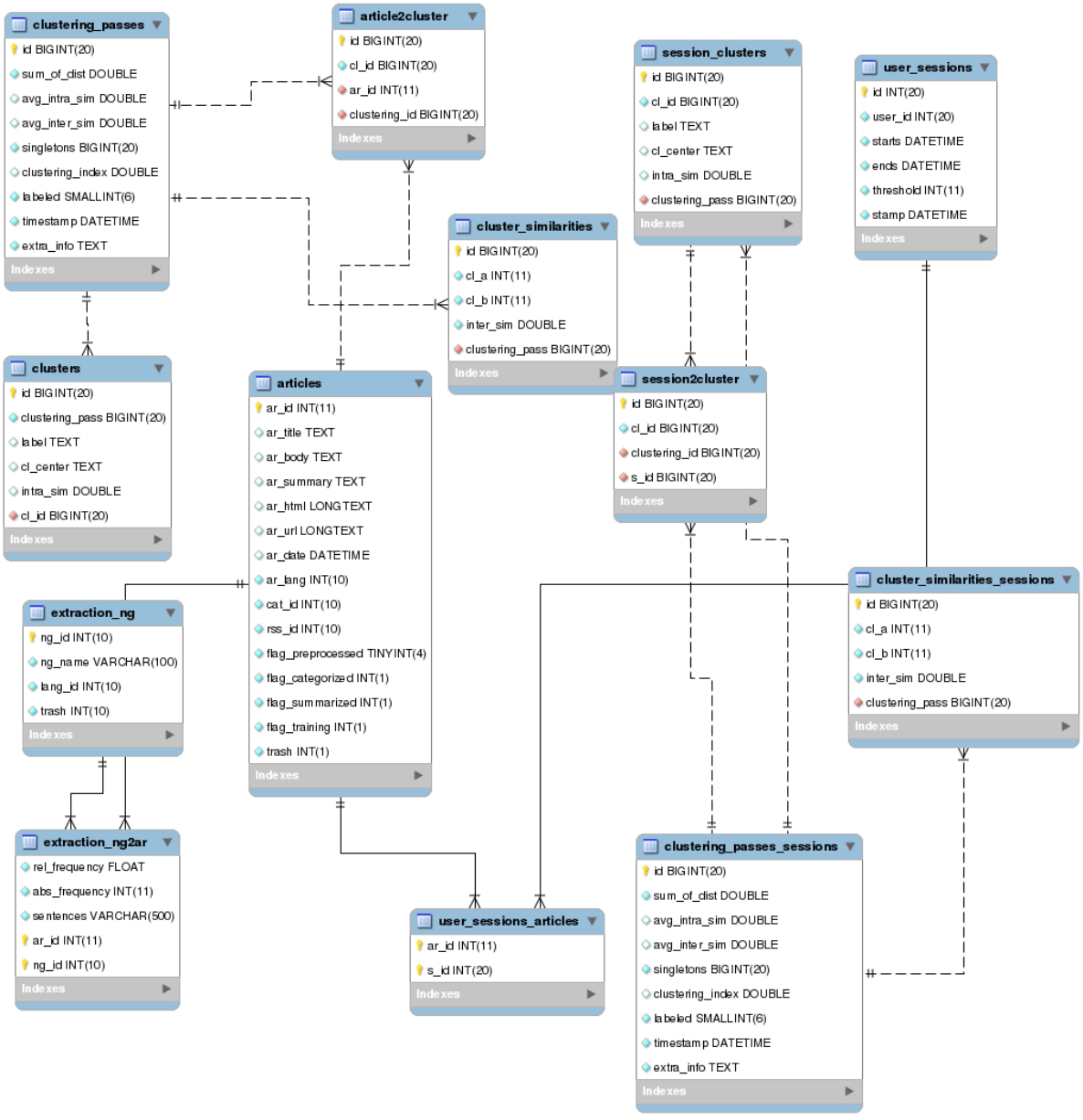
labeled flag που χαρακτηρίζει εάν έχει γίνει ονοματοδοσία συστάδων για το συγκεκριμένο πέρασμα. Αν ναι, οι επιλεγμένες συστάδες θα είναι διαθέσιμες στον πίνακα clusters

timestamp Το χρονικό σημείο όπου ξεκίνησε το εν' λόγω πέρασμα

exta_info Επιπλέον πληροφορίες για αυτό το πέρασμα. Κυρίως χρησιμοποιείται για να αποθηκεύουμε πληροφορίες εκτέλεσης γραμμής εντολών που χρησιμοποιήθηκε



Σχήμα 16: Διάγραμμα E-R της ΒΔ χωρίς τους νέους πίνακες



Σχήμα 17: Διάγραμμα E-R των νέων πινάκων της ΒΔ

6.4.2.1.1.2 clusters Ο πίνακας clusters αποθηκεύει τις εντοπισμένες συστάδες μετά την εκτέλεση του αλγορίθμου W-kmeans πάνω σε άρθρα νέων. Πιο συγκεκριμένα, τα πεδία που περιλαμβάνει είναι:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

clustering_pass Το πέρασμα συσταδοποίησης της εν' λόγω αποθηκευμένης συστάδας άρθρων νέων (foreign key στον πίνακα clustering_passes)

label Η ετικέτα που έδωσε ο μηχανισμός ονοματοδοσίας συστάδων, αν και εφόσον έχει τρέξει για το συγκεκριμένο πέρασμα συσταδοποίησης

cl_center Οι συντεταγμένες του κέντρου της συστάδας στον διανυσματικό χώρο των άρθρων που συμμετείχαν στην συγκεκριμένη διαδικασία συσταδοποίησης

intra_sim Η εσω-συσταδική ομοιότητα της συστάδας αυτή με τις υπόλοιπες της συγκεκριμένης συσταδοποίησης (μέση τιμή)

6.4.2.1.1.3 article2cluster Ο πίνακας article2cluster αποθηκεύει την συστάδα στην οποία ανήκει το κάθε άρθρο που συσταδοποιήθηκε. Περιλαμβάνει τα εξής πεδία:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

cl_id Η συστάδα στην οποία ανήκει το άρθρο νέων (foreign key στον πίνακα clusters)

ar_id Το άρθρο νέων (foreign key στον πίνακα articles (σχήμα 16)

clustering_id Η διαδικασία συσταδοποίησης που έκανε την εν' λόγω ανάθεση του άρθρου στη συστάδα (foreign key στον πίνακα clustering_passes)

6.4.2.1.1.4 cluster_similarities Ο πίνακας cluster_similarities αποθηκεύει τις ομοιότητες μεταξύ των συστάδων που προέκυψαν από τα διάφορα περάσματα συσταδοποίησης. Περιλαμβάνει τα εξής πεδία:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

cl_a Το αναγνωριστικό της πρώτης συστάδας άρθρων νέων (foreign key στον πίνακα clusters)

cl_b Το αναγνωριστικό της δεύτερης συστάδας άρθρων νέων (foreign key στον πίνακα clusters)

inter_sim Η δια-συσταδική ομοιότητα των δύο συστάδων της κάθε εγγραφής

clustering_id Η διαδικασία συσταδοποίησης που έκανε την εν' λόγω ανάθεση του άρθρου στη συστάδα (foreign key στον πίνακα clustering_passes)

6.4.2.1.2 Πίνακες συσταδοποίησης χρηστών

Ακολουθεί μία σύντομη περιγραφή των πινάκων που έχουν να κάνουν με την διαδικασία συσταδοποίησης συνεδριών (χρηστών) του συστήματος από τον αλγόριθμο W-kmeans.

6.4.2.1.2.1 clustering_passes_sesions Ο πίνακας clustering_passes_sesions την απαραίτητη πληροφορία που αφορά κάθε πέρασμα συσταδοποίησης συνεδριών χρηστών που έχουν καταγραφεί στο σύστημα. Πιο συγκεκριμένα, περιλαμβάνει:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

sum_of_dist Άθροισμα των αποστάσεων που καταγράφηκαν σε αυτό το πέρασμα. Η πληροφορία είναι χρήσιμη για τον μετέπειτα υπολογισμό μετρικών όπως η CI

avg_intra_sim Η μέση εσω-συσταδική απόσταση των συστάδων του περάσματος

avg_inter_sim Η μέση δια-συσταδική απόσταση των συστάδων του περάσματος

singletons Το πλήθος των μοναδιαίων συστάδων του περάσματος (συστάδες με ένα μόνο αντικείμενο)

clusterinig_index Δείκτης συσταδοποίησης για το συγκεκριμένο πέρασμα. Ουσιαστικά αποτελεί το βασικό κριτήριο αξιολόγησης του πόσο αποτελεσματικό ήταν το πέρασμα

labeled flag που χαρακτηρίζει εάν έχει γίνει ονοματοδοσία συστάδων για το συγκεκριμένο πέρασμα. Αν ναι, οι επιλεγμένες συστάδες θα είναι διαθέσιμες στον πίνακα clusters

timestamp Το χρονικό σημείο όπου ξεκίνησε το εν' λόγω πέρασμα

exta_info Επιπλέον πληροφορίες για αυτό το πέρασμα. Κυρίως χρησιμοποιείται για να αποθηκεύουμε πληροφορίες εκτέλεσης γραμμής εντολών που χρησιμοποιήθηκε

6.4.2.1.2.2 session_clusters Ο πίνακας session_clusters αποθηκεύει τις εντοπισμένες συστάδες συνεδριών μετά την εκτέλεση του αλγορίθμου W-kmeans. Πιο συγκεκριμένα, τα πεδία που περιλαμβάνει είναι:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

clustering_pass Το πέρασμα συσταδοποίησης της εν' λόγω αποθηκευμένης συστάδας χρηστών (foreign key στον πίνακα clustering_passes_sessions)

label Η ετικέτα που έδωσε ο μηχανισμός ονοματοδοσίας συστάδων, αν και εφόσον έχει τρέξει για το συγκεκριμένο πέρασμα συσταδοποίησης

cl_center Οι συντεταγμένες του κέντρου της συστάδας στον διανυσματικό χώρο των άρθρων που συμμετείχαν στην συγκεκριμένη διαδικασία συσταδοποίησης

intra_sim Η εσω-συσταδική ομοιότητα της συστάδας αυτή με τις υπόλοιπες της συγκεκριμένης συσταδοποίησης (μέση τιμή)

6.4.2.1.2.3 session2cluster Ο πίνακας session2cluster αποθηκεύει την συστάδα στην οποία ανήκει η κάθε συνεδρία χρήστη που συσταδοποιήθηκε. Τα πεδία που περιλαμβάνει είναι:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

cl_id Η συστάδα στην οποία ανήκει η συνεδρία (foreign key στον πίνακα session_clusters)

s_id Η συνεδρία χρήστη (foreign key στον πίνακα sessions)

clustering_id Η διαδικασία συσταδοποίησης που έκανε την εν' λόγω ανάθεση της συνεδρίας χρήστη στη συστάδα (foreign key στον πίνακα clustering_passes_sessions)

6.4.2.1.2.4 cluster_similarities_sessions Ο πίνακας cluster_similarities_sessions αποθηκεύει τις ομοιότητες μεταξύ των συστάδων που προέκυψαν από τα διάφορα περάσματα συσταδοποίησης. Περιλαμβάνει τα εξής πεδία:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

cl_a Το αναγνωριστικό της πρώτης συνεδρίας χρηστών (foreign key στον πίνακα session_clusters)

cl_b Το αναγνωριστικό της δεύτερης συνεδρίας χρηστών (foreign key στον πίνακα session_clusters)

inter_sim Η δια-συσταδική ομοιότητα των δύο συστάδων της κάθε εγγραφής

clustering_id Η διαδικασία συσταδοποίησης που έκανε την εν' λόγω ανάθεση του άρθρου στη συστάδα (foreign key στον πίνακα clustering_passes_sessions)

6.4.2.1.2.5 user_sessions Ο πίνακας user_sessions αποθηκεύει τις συνεδρίες χρηστών οι οποίες έχουν εξαχθεί από τις πλοηγήσεις που έχουν καταγραφεί στο σύστημα. Περιλαμβάνει τα εξής πεδία:

id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

user_id Το αναγνωριστικό του χρήστη (foreign key στον πίνακα website_users)

starts Timestamp που αποθηκεύει το χρονικό σημείο έναρξης της συνεδρίας

ends Timestamp που αποθηκεύει το χρονικό σημείο λήξης της συνεδρίας

threshold Το χρονικό όριο (threshold) που χρησιμοποιήθηκε κατά την εξαγωγή της συγκεκριμένης συνεδρίας. Αφορά την μεταβλητή session_threshold στον αλγόριθμο 5. Τυπικά το όριο εύρεσης συνεδριών είναι 10 λεπτά.

stamp Timestamp που αποθηκεύει το χρονικό σημείο που εισήχθηκε η συγκεκριμένη εγγραφή

6.4.2.1.2.6 user_sessions_articles Ο πίνακας user_sessions_articles αποθηκεύει τις εμφανίσεις άρθρων νέων σε συνεδρίες χρηστών. Τα πεδία που περιλαμβάνει είναι τα ακόλουθα:

ar_id Το αναγνωριστικό του άρθρου

s_id Το αναγνωριστικό της συνεδρίας

6.4.2.1.3 Πίνακες n-grams

Ακολουθεί μία συνοπτική παρουσίαση των πινάκων της ΒΔ οι οποίοι κρατάνε τις πληροφορίες για τα εξαγόμενα n-grams από τα άρθρα νέων.

6.4.2.1.3.1 extraction_ng Ο πίνακας extraction_ng αποθηκεύει τα εξαγόμενα n-grams σε αντιστοιχία του πίνακα keywords (σχήμα 16). Περιλαμβάνει τα εξής πεδία:

ng_id Μοναδικό αναγνωριστικό πρωτεύον κλειδί για το συγκεκριμένο πίνακα

ng_name Το όνομα (λεκτικό) του συγκεκριμένου n-gram - μπορεί να περιλαμβάνει 2 η περισσότερες λέξεις

lang_id Το αναγνωριστικό της γλώσσας στην οποία ανήκει το εν' λόγω n-gram. (foreign key στον πίνακα language του σχήματος 16)

trash Πεδίο για την επισήμανση μη χρήσιμων n-grams. Τα n-grams που έχουν την τιμή "1" στο εν' λόγω πεδίο δεν λαμβάνονται υπόψιν από τις διαδικασίες συστήματος

6.4.2.1.3.2 extraction_ng2ar Ο πίνακας extraction_ng2ar τη συσχέτιση μεταξύ των εξαγόμενων n-grams και άρθρων του συστήματος. Περιλαμβάνει τα εξής πεδία:

ng_id Το n-gram στο οποίο αναφέρεται η εν' λόγω εγγραφή (foreign key στον πίνακα extraction_ng)

ar_id Το άρθρο στο οποίο περιέχεται το εν' λόγω n-gram (foreign key στον πίνακα articles του σχήματος 16)

sentences Οι προτάσεις στις οποίες εμφανίζεται το n-gram στο συγκεκριμένο άρθρο. Οι προτάσεις αριθμούνται και καταγράφοντα με τη σειρά

abs_frequency Η απόλυτη συχνότητα εμφάνισης του n-gram μέσα στο άρθρο

rel_frequency Η σχετική συχνότητα εμφάνισης του n-gram μέσα στο άρθρο

6.5 Διασύνδεση μηχανισμών

Η διασύνδεση των μηχανισμών βασίζεται αποκλειστικά στο επίπεδο βάσης δεδομένων αλλά και στη σειριακή εκτέλεση των διαδικασιών που προσφέρει το σύστημα. Το γεγονός ότι χρησιμοποιούνται πολλαπλά επίπεδα στην υλοποίηση είναι σημαντικό για ένα τέτοιο σύστημα καθότι υπάρχει

ένα επίπεδο το οποίο είναι κοινό για όλα τα υποσυστήματα και συνεπώς είναι εφικτή η ανταλλαγή δεδομένων.

Παράλληλα, όλοι οι μηχανισμοί του συστήματος έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να δέχονται δεδομένα από δύο διαφορετικά κανάλια και αντίστοιχα να εξάγουν τα δεδομένα σε δύο διαφορετικά κανάλια, το ένα αυτό της βάσης δεδομένων και το άλλο σε μορφή XML. Μιλούμε για το κλασσικό πρότυπο μίας n-tier αρχιτεκτονικής η οποία επιτυγχάνει διασύνδεση των αυτόνομων μηχανισμών που την αποτελούν στο επίπεδο καναλιού επικοινωνίας. Με αυτό τον τρόπο έχουν μηχανισμούς που αποδεσμεύονται όσο αφορά το κομμάτι της υλοποίησης και δεν έχουν κανένα περιορισμό αρκεί να μπορούν να "διαβάσουν" δεδομένα από βάση δεδομένων ή από XML αρχεία και αντίστοιχα να είναι σε θέση να "γράψουν" σε βάση δεδομένων ή σε XML αρχεία.

6.6 Προδιαγραφές

6.6.1 Συλλογή άρθρων και εξαγωγή χρήσιμου κειμένου

Το σύστημα προτάσεων ξεκινά την διαδικασία δεικτοδότησης άρθρων με τον μηχανισμό συλλογής άρθρων από το διαδίκτυο ο οποίος τρέχει ανεξάρτητα από τα υπόλοιπα υποσυστήματα που έχουν αλληλεπίδραση με τον χρήστη. Σε αυτόν περιλαμβάνονται η συλλογή άρθρων από τον ιστό και η εξαγωγή του χρήσιμου κειμένου από αυτά. Η λειτουργία είναι αυτοματοποιημένη ώστε να αλληλεπιδρά με τη βάση δεδομένων και η ανθρώπινη επίδραση μπορεί να είναι μόνο έμμεση. Το συγκεκριμένο υποσύστημα δέχεται σαν είσοδο τα RSS Feeds που καταγράφονται στη βάση δεδομένων και για την ακρίβεια τα urls των RSS feeds των news portals τα οποία πρέπει να διαπεράσει ο crawler. Είναι εύλογο πως υπόκειται στον διαχειριστή του συστήματος ο καθορισμός έγκυρων RSS Feeds για την τροφοδότηση του μηχανισμού με άρθρα, κάτι που είναι εφικτό μέσω της διεπαφής διαχείρισης του συστήματος. Ο μηχανισμός εξαγωγής χρήσιμου κειμένου είναι σχεδιασμένος ώστε να εξάγει κείμενα άρθρων από τη σελίδα· δεν έχει επομένως νόημα, και για την ακρίβεια γεμίζει τη βάση δεδομένων με "σκουπίδια", η εισαγωγή urls από RSS feeds που δεν περιέχουν σώμα. Παρόμοια, πρέπει να αποφεύγεται η χρήση urls που δεν υπάρχουν (dead links) καθώς οδηγούν τον crawler και όλο συνολικά το σύστημα σε χάσιμο χρόνου. Περισσότερες πληροφορίες για το υποσύστημα εξαγωγής χρήσιμου κειμένου είναι διαθέσιμες στα [5] [6] [8] [4].

6.6.2 Προεπεξεργασία κειμένου

Ως γνωστών, ο μηχανισμός προεπεξεργασίας κειμένου είναι αυτοματοποιημένος ώστε να αλληλεπιδρά με τα κείμενα της βάσης δεδομένων. Η ορθή λειτουργία του επομένως εναπόκειται στην ορθή κατάσταση της βάσης δεδομένων και τις συναλλαγές που γίνονται με αυτή. Δεδομένου ότι όλα τα απαραίτητα πεδία των πινάκων της βάσης δεδομένων περιέχουν ορθές πληροφορίες, η εξαγωγή keywords προχωράει βάσει αυτών. Πρέπει να σημειωθεί επίσης ότι η διαδικασία της προεπεξεργασίας κειμένου (αφαίρεση στίξης και αριθμών, ανάκτηση ουσιαστικών, αφαίρεση stopwords, stemming) εκτελείται σειριακά και πριν τις διαδικασίες κατηγοριοποίησης περίληψης και συσταδοποίησης για το κάθε άρθρο. Τα αποτελέσματα του υποσυστήματος εξαγωγής keywords, όπως

έχουμε πει, αποθηκεύονται στους κατάλληλους πίνακες της βάσης δεδομένων του συστήματος για να είναι διαθέσιμα στα υποσυστήματα που ακολουθούν. Περισσότερες πληροφορίες για το υποσύστημα προεπεξεργασίας κειμένου είναι διαθέσιμες στα [38] [34].

6.6.3 Κατηγοριοποίηση εξαγωγή περίληψης και συσταδοποίησης

Τα υποσυστήματα κατηγοριοποίησης εξαγωγής περίληψης και συσταδοποίησης, που αποτελούν και τον πυρήνα του συστήματος μαζί με αυτό της προσωποποίησης, είναι σχεδιασμένα ώστε να δέχονται ως είσοδο τα δεδομένα της προεπεξεργασίας κειμένου. Όπως περιγράφεται και στο [34], η διαδικασία που ακολουθείται μετά την προεπεξεργασία κειμένου είναι: προσπάθεια για κατηγοριοποίηση του κειμένου βάσει κάποιων κριτηρίων και της βάσης γνώσης που έχουμε, αν η κατηγοριοποίηση είναι επιτυχής (το κείμενο είναι πολύ σχετικό με μία κατηγορία), προχωρούμε σε εξαγωγή γενικής περίληψης υποβοηθούμενη από την κατηγορία του κειμένου. Αν η κατηγοριοποίηση δεν είναι εφικτή, προχωρούμε σε εξαγωγή γενικής περίληψης και επιχειρούμε την κατηγοριοποίηση αυτής. Αν η δεύτερη απόπειρα κατηγοριοποίσης δώσει καλύτερα αποτελέσματα, αποθηκεύουμε αυτά στη βάση δεδομένων, αλλιώς τα πρώτα. Φυσικά τα υποσυστήματα μπορούν να κληθούν και αυτόνομα, π. χ. να ζητήσουμε περίληψη ή κατηγοριοποίηση ενός άρθρου που έχουμε στην κατοχή μας.

Η προσπάθεια κατηγοριοποίησης ενός άρθρου μοιάζει με την Linear Least Squares Fit - LLSF τεχνική και προχωράει ως εξής: η κατηγοριοποίηση των άρθρων γίνεται χρησιμοποιώντας την λίστα με τα πιο αντιπροσωπευτικά (stemmed) keywords του κειμένου μαζί με τις συχνότητες εμφάνισής τους. Έχοντας ήδη στη διάθεσή μας παρόμοιες λίστες που αφορούν στα πιο αντιπροσωπευτικά keywords της κάθε κατηγορίας, συγκρίνουμε τις λίστες χρησιμοποιώντας την ομοιότητα συνημιτόνου. Ένα επιπλέον σημαντικό χαρακτηριστικό είναι ότι η ανάλυση είναι διαφορετική για τους διαφορετικούς χρήστες. Όσο μεγαλύτερη είναι η αφαίρεση πληροφορίας σε τόσο λιγότερες προτάσεις ενός κειμένου πραγματοποιείται κατηγοριοποίηση του κειμένου και συνεπώς η κατηγορία στην οποία εντάσσεται ένα κείμενο είναι πιο γενική. Η παραπάνω διαδικασία έχει σαν αποτέλεσμα να δημιουργηθεί πολλαπλού είδους κατηγοριοποίηση στα κείμενα τα οποία θα διαθέτει το σύστημα με αποτέλεσμα να είναι διαφορετικά τα αποτελέσματα για κάθε χρήστη ανάλογα με τη λεπτομέρεια της αναζήτησης που πραγματοποιούν. Το ένα είδος κατηγοριοποίησης θα είναι καθαρά αλγοριθμικό ενώ το δεύτερο κομμάτι θα βασίζεται κυρίως στις προσωπικές επιλογές του χρήστη, οι οποίες δημιουργούν κατηγορίες αφαίρεσης πληροφορίας.

Αξίζει να τονίσουμε κάποια βασικά στοιχεία της λειτουργίας αυτού του μηχανισμού. Ο μηχανισμός αυτός από τη στιγμή που θα αρχικοποιηθεί με ένα σύνολο πρότυπων κειμένων για τη δημιουργία μίας κατηγορίας μπορεί να λειτουργεί ανεξάρτητα από το υπόλοιπο σύστημα κατηγοριοποιώντας συνεχώς κείμενα. Είναι πολύ βασικό για την καλή λειτουργία του συστήματος να ανανεώνεται συχνά η βάση γνώσης με επικαιροποιημένα κείμενα χρησιμοποιώντας το τμήμα της ανανέωσης της βάσης γνώσης του μηχανισμού (suggest training).

6.6.4 Προσωποποίηση

Η λειτουργία του υποσυστήματος προσωποποίησης αφορά στο προσωποποιημένο περιεχόμενο που παρουσιάζεται στο χρήστη. Προκειμένου η πληροφορία να καλύπτει κατά το καλύτερο δυνατό τις προτιμήσεις του χρήστη, είναι σημαντικό το σύστημα να αντιλαμβάνεται εγκαίρως αλλαγές στο προφίλ του. Οι χρήστες σπάνια ξοδεύουν χρόνο για να δηλώσουν ρητά τι επιθυμούν, πολλές φορές λόγω του ότι δεν εμπιστεύονται τις προτιμήσεις που έχουν σε ένα απρόσωπο σύστημα που ζητάει υπερβολικά πολλά στοιχεία γι' αυτούς. Ο μόνος δρόμος επομένως είναι οι πληροφορίες αυτές να συλλέγονται (όπου αυτό είναι δυνατό) έμμεσα καταγράφοντας τις επιλογές που κάνει ο χρήστης κατά την διάρκεια παραμονής του στο σύστημα. Η ερμηνεία όμως αυτών των συμπεριφορών που φαίνεται να παρουσιάζουν οι χρήστες πρέπει να ερμηνεύονται και κατάλληλα από το σύστημα βάσει σωστών παραμέτρων και μετρικών. Ήδη αναφέραμε τις παραμέτρους που αξιοποιεί το σύστημα προτάσεων προκειμένου να εξάγει το μητρώο με τα keywords και τις προτιμήσεις για καθένα που έχει ο χρήστης. Η διαδικασία όμως αυτή είναι επιρρεπής σε λάθη μεσοπρόθεσμα: ο χρήστης αρχικά πιθανών να μην βλέπει όλα τα νέα άρθρα που επιθυμεί ή πιθανών να βλέπει και κάποια που θεωρεί ότι αντιτίθενται στο προφίλ του. Μακροπρόθεσμα όμως, έχοντας αρκετά στοιχεία για την συμπεριφορά του χρήστη το σύστημα φαίνεται να προσαρμόζεται αρκετά καλά στις προτιμήσεις, κάτι που θα γίνει ορατό και στο επόμενο κεφάλαιο μέσα από την πειραματική διαδικασία.

6.7 Απαιτήσεις του συστήματος

Στην ενότητα αυτή παρουσιάζονται οι απαιτήσεις του συστήματος από άποψη λογισμικού και υλικού.

6.7.1 Λογισμικό και βιβλιοθήκες

Για την ανάπτυξη του συστήματος χρησιμοποιήθηκαν πακέτα λογισμικού και βιβλιοθήκες που αναφέρονται στον πίνακα 4

Η ανάπτυξη του συστήματος έγινε εξ' ολοκλήρου σε open source λογισμικό και λειτουργικό σύστημα Gentoo Linux [75].

6.7.2 Υλικό

Το σύστημα που αναπτύχθηκε δεν έχει υψηλές απαιτήσεις υλικού. Μπορεί να στηθεί σε κάποιον υπολογιστή γενιάς Pentium IV και νεότερο. Φυσικά εάν οι απαιτήσεις μας έχουν να κάνουν με ένα σύστημα που θα πραγματοποιεί real time κατηγοριοποίηση και εξαγωγή προσωποποιημένης περίληψης κειμένων είναι εύλογο να χρησιμοποιηθεί ένα πιο σύγχρονο σύστημα στο οποίο η βάση δεδομένων (η οποία και αποτελεί το bottleneck του συστήματος λόγω των πολλών συναλλαγών) θα έχει καλύτερους χρόνους εξυπηρέτησης. Ο server ο οποίος εξυπηρετεί το σύστημα προτάσεων βρίσκεται στο url `http://perssonal.cti.gr/` και έχει την παρακάτω σύνθεση υλικού (Πίνακας 5):

gcc-4.8.3 [207]
MySQL-5.5.40 [150]
apache-2.2.22 [206]
php-5.3.10 [208]
boost-1.49.0 [33]
cgicc-3.2.9 [44]
mysql++-2.3.2 [149]
libcurl-7.18.2 [127]
expat-2.0.1 [67]
xerces-2.7.0 [222]
libstemmer [129]
gd-2.0.35-r3 [74]
htmltidy-20090325-r1 [96]
icu-20090325-r1 [98]
libpng-1.5.15 [128]
openssl-1.0.1c [165]
aspell-0.60.6.1 [85]
wordnet-3.0-r3 [220]

Πίνακας 4: Σύνθεση υλικού για ανάπτυξη του συστήματος

CPU	Intel(R) Xeon(R) CPU E5-2407 0 @ 2.20GHz
RAM	6GB
Hard Disk	300GB, 7200rpm

Πίνακας 5: Σύνθεση υλικού του εξυπηρετητή του συστήματος προτάσεων άρθρων νέων

ΚΕΦΑΛΑΙΟ 7

ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΚΑΙ ΥΠΟΣΥΣΤΗΜΑΤΩΝ

> Wise men speak because they have something to say; fools because they have to say something.
>
> *Plato, Greek Philosopher, 428 BC*

Στο κεφάλαιο αυτό παρουσιάζεται η πειραματική διαδικασία που πραγματοποιήθηκε σε σχέση με τα διάφορα υποσυστήματα του μηχανισμού που αναπτύχθηκαν ή βελτιώθηκαν στη διδακτορική διατριβή. Κάθε βελτίωση αξιολογείται αυτόνομα καθώς και συνολικά με το σύστημα προτάσεων σε πλήρη λειτουργία. Σημαντικό είναι ίσως να αναφερθεί ότι τα πειραματικά αποτελέσματα του παρόντος κεφαλαίου παρουσιάζονται με νοηματική σειρά σε σχέση με το υποσύστημα που αγγίζουν, όχι με τη σειρά που εκτελέστηκαν.

7.1 Υποσύστημα Προεπεξεργασίας κειμένου

Στην παρούσα ενότητα παρουσιάζουμε την πειραματική διαδικασία που αφορά στην αξιολόγηση του υποσυστήματος προεπεξεργασίας κειμένου, σε σχέση με τις αλλαγές που έγιναν σε αυτό στα πλαίσια της διδακτορικής διατριβής.

7.1.1 Αξιοποίηση n-grams

Για την αξιολόγηση της επίπτωσης που έχει η αξιοποίηση των n-grams λέξεων από τα άρθρα νέων στις διαδικασίες του συστήματος, εκτελέσαμε ορισμένα offline πειράματα με βάση πληροφορία που υπήρχε ήδη στη ΒΔ του συστήματος. Πιο συγκεκριμένα, προσπαθούμε να εντοπίσουμε τυχόν βελτίωση στην αποτελεσματικότητα του W-kmeans αλγορίθμου συσταδοποίησης όταν χρησιμοποιούμε τα εξαγόμενα n-grams.

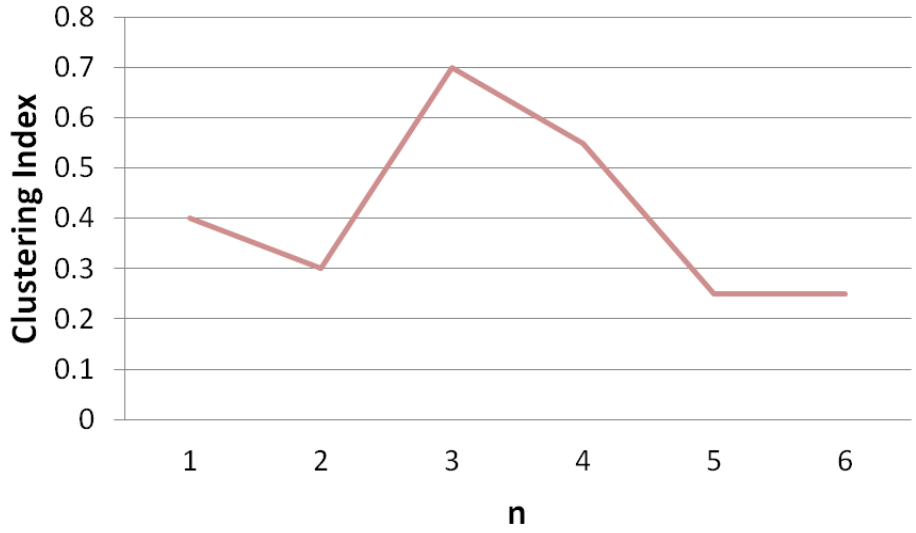
7.1.1.1 Σύνολο δεδομένων

Το σύνολο δεδομένων μας αποτελείται από 10.000 άρθρα νέων που συλλέχθηκαν σε διάστημα 5 μηνών από διάφορα news portals του διαδικτύου (BBC, CNN, κ.λπ.). Τα άρθρα αυτά επιλέχθηκαν ώστε να είναι ομοιόμορφα μοιρασμένα μεταξύ των 8 κατηγοριών του συστήματος προτάσεων: *business, politics, health, education, science, sports, technology* και *entertainment*. Ο ομοιόμορφος διαμοιρασμός των άρθρων έγινε προκειμένου να αποφύγουμε τυχών προκαταλήψεις ή στατιστικές ανωμαλίες που έχουν να κάνουν πιθανά με συγκεκριμένες κατηγορίες άρθρων. Ως μετρική αξιολόγησης χρησιμοποιήσαμε τον δείκτη συσταδοποίησης, CI, όπως αυτός περιγράφηκε στην ενότητα 3.7.1.6.1.

7.1.1.2 Αποτελέσματα και ανάλυση

Για το πρώτο μας πείραμα, προσπαθήσαμε να εντοπίσουμε την καλύτερη τιμή n για τα δεδομένα μας, δηλαδή μέχρι ποιο μέγεθος παράθυρου λέξεων θα πρέπει να κρατάμε κατά τον εντοπισμό n-grams για να έχουμε τα καλύτερα αποτελέσματα σε σχέση με τις CI τιμές. Για την περίπτωση αυτή, αυθαίρετα θέσαμε στις σχέσεις 37 και 38 $A = B = 0.5$ δίνοντας έτσι την ίδια βαρύτητα τόσο στην στα εξαγόμενα keywords όσο και στα n-grams (η επιλογή καταλληλότερων τιμών συζητείται σε επόμενο πείραμα), και στη συνέχεια δοκιμάσαμε διάφορες τιμές n όπου $2 \leq n \leq 6$. Για κάθε τιμή n, τρέξαμε την διαδικασία συσταδοποίησης του αλγορίθμου W-kmeans (αλγόριθμος 2) 10 φορές, με διαφορετικές αρχικές αναθέσεις κάθε φορά, πραγματοποιώντας έτσι ένα πείραμα 10 περασμάτων (10-pass experiment). Τα αποτελέσματα για τις διάφορες τιμές n φαίνονται στο σχήμα 18.

Από τις τιμές CI που απεικονίζονται στο σχήμα 18 μπορούμε να δούμε ότι για $n = 3$, δηλαδή όταν κρατάμε 3-grams για την ζύγιση των άρθρων, η απόδοση του αλγόριθμου W-kmeans αυξάνεται κατά μέσο όρο 0.3 όσον αφορά τις παραγόμενες συστάδες. Η σύγκριση αυτή γίνεται σε σχέση με την περίπτωση αξιοποίησης μόνο των εξαγόμενων keywords (περίπτωση $n = 1$ στο σχήμα

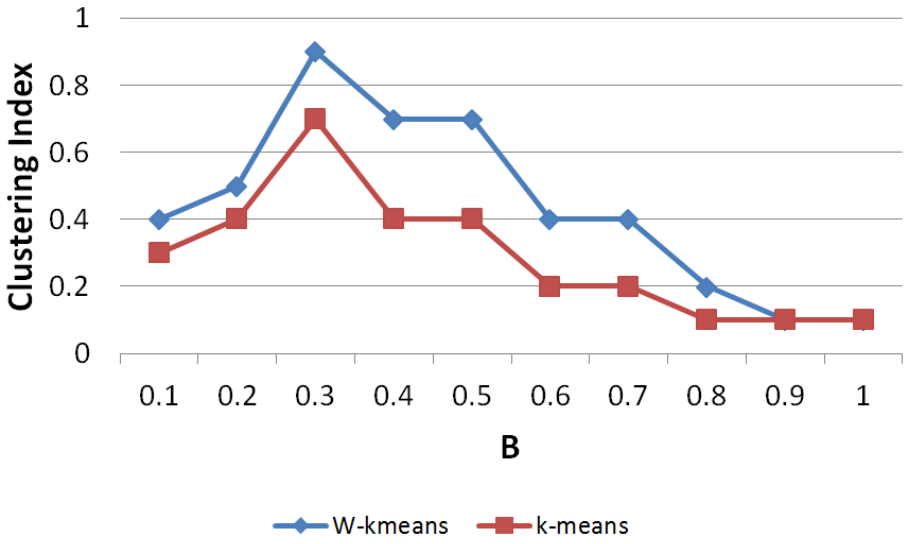


Σχήμα 18: Η επίδραση της αξιοποίησης των n-grams στην διαδικασία συσταδοποίησης για διάφορες τιμές του n

18). Το αποτέλεσμα αυτό είναι μάλιστα σε συμφωνία με ότι έχει παρατηρηθεί σε προηγούμενες εργασίες, π.χ. [72]. Για $n = 4$ βλέπουμε ακόμη μία αύξηση στην απόδοση του αλγορίθμου σε σχέση με την περίπτωση που δεν λαμβάνονται καθόλου υπόψιν τα n-grams. Η περαιτέρω αύξηση του παραθύρου n φαίνεται να έχει αρνητική επίπτωση στην απόδοση του W-kmeans, κάτι που μπορεί να ερμηνευθεί ως εξής: μεγαλύτερα παράθυρα κατά την εξαγωγή n-grams σημαίνουν ότι n-grams που τυχαία εμφανίζονται μαζί σε μεγαλύτερες ακολουθίες ζυγίζονται περισσότερο απ' ότι πρέπει, μία κατάσταση που πιθανά έχει αρνητική επίπτωση στην συνολική ζύγιση (δεδομένης της τυχαιότητας αυτών). Μία ακόμη ενδιαφέρουσα παρατήρηση από το παραπάνω πείραμα είναι ότι για $n = 2$, τα αποτελέσματα είναι ελαφρώς χειρότερα από την περίπτωση που δεν χρησιμοποιούμε καθόλου n-grams στο ζύγισμα. Η εκτίμηση μας για αυτό είναι ότι οφείλεται στην εξαγωγή μη χρήσιμων n-grams όταν $n = 2$, αφού πράγματι παρατηρήσαμε ότι για πληθώρα περιπτώσεων, τα 2-grams που εξάγονταν, δεν είχαν κάτι παραπάνω να προσφέρουν νοηματικά σε σχέση με την αναπαράσταση του κείμενου.

Για το δεύτερο πείραμα αξιολόγησης της επίδρασης των n-grams, προσπαθήσαμε να καθορίσουμε τις καλύτερες τιμές ζύγισης A και B όπως αυτές περιγράφονται στις σχέσεις 37 και 38. Έτσι, θέτοντας $n = 3$ (δεδομένου του αποτελέσματος του προηγούμενου πειράματος), τρέξαμε 10 φορές τον αλγόριθμο W-kmeans καθώς και τον k-means για όλα τα δεδομένα και για κάθε μία από τις αυξανόμενες τιμές, με βήμα 0.1, του B (και επομένως μειούμενες τιμές A μιας και $A = 1 - B$), ενώ παράλληλα καταγράψαμε τις CI τιμές για όλα τα περάσματα συσταδοποίησης. Οι μέσοι όροι των τιμών CI απεικονίζονται στο σχήμα 19.

Όπως φαίνεται και στο σχήμα 19, οι καλύτερες CI τιμές προκύπτουν για $B = 0.3$ και $A = 0.7$, δηλαδή όταν τα n-grams συμμετέχουν στη διαδικασία ζύγισης κατά 30% ενώ το υπόλοιπο 70% ανήκει στην ζύγιση BOW. Είναι ενδιαφέρον επίσης ότι η απόδοση των αλγορίθμων γρήγορα χει-



Σχήμα 19: Αποτελέσματα απόδοσης των αλγορίθμων W-kmeans και k-means για διάφορες τιμές ζυγίσματος των εξαγόμενων n-grams

ροτερεύει όσο το B αυξάνει, φτάνοντας την χειρότερη τιμή όταν $B = 1$, δηλαδή όταν λαμβάνονται υπόψιν μόνο τα εξαγόμενα n-grams. Το προηγούμενο μπορεί να εξηγηθεί από το γεγονός ότι δεν έχουν όλα τα άρθρα ξεχωριστά ή συχνά εμφανιζόμενα n-grams, είδικά τα μικρότερα σε μέγεθος. Κατά συνέπεια σε αυτή την περίπτωση όσο περισσότερο λαμβάνουμε υπόψιν τα n-grams σε βάρος των keywords, τόσο χειροτερεύει η αναπαράσταση των κειμένων από την εξαγόμενη πληροφορία (είτε keywords, είτε n-grams) και επομένως και οι CI τιμές των αποτελεσμάτων.

Μία ακόμα σημαντική παρατήρηση που μπορούμε εύκολα να κάνουμε είναι ότι ο W-kmeans εύκολα ξεπερνά σε απόδοση τον τυπικό k-means, ακόμη και για την περίπτωση που τα n-grams δεν λαμβάνονται υπόψιν στην διαδικασία ζύγισης. Για την ακρίβεια, τα αποτελέσματα ήταν συνεχώς υπερ του W-kmeans για κάθε τιμή της παραμέτρου B που δοκιμάσαμε. Αυτό αποτελεί μία καλή ένδειξη ότι το ευρετικό της χρήσης των υπερωνύμων του WordNet γρήγορα αποδίδει σε κάθε συνδυασμό ζύγισης μεταξύ των χαρακτηριστικών του κειμένου (περισσότερα για αυτό στα πειράματα που ακολουθούν). Παρόλα αυτά, οφείλουμε να ομολογήσουμε ότι δεν έχουμε κάποια εξήγηση για τα σχεδόν ίδια αποτελέσματα όταν $B = 0.9$. Στην παρούσα φάση το αποδίδουμε στο ότι η εφαρμογή καθενός από τους δύο αλγορίθμους δεν έχει ιδιαίτερη αξία όταν όταν σχεδόν μόνο τα n-grams των κειμένων συμμετέχουν στη ζύγιση. Πρόκειται δηλαδή για μη αποτελεσματική αναπαράσταση των δεδομένων για κάθε περίπτωση.

Συμπερασματικά, θα λέγαμε ότι η εξαγωγή και η αξιοποίηση n-grams κατά την διαδικασία ζύγισης των άρθρων έχει αξιοσημείωτα οφέλη σε ότι έχει να κάνει με την συσταδοποίησης άρθρων νέων. Ως εκ' τούτου, αναμένουμε και η απόδοση του συστήματος προτάσεων να ενισχυθεί από αυτή την εξέλιξη (περισσότερα για αυτό σε επόμενη ενότητα του παρόντος κεφαλαίου).

7.2 Συσταδοποίηση

Στην παρούσα ενότητα παρουσιάζουμε και αναλύουμε τα πειραματικά αποτελέσματα που προέκυψαν από την διαδικασία αξιολόγησης του υποσυστήματος συσταδοποίησης του συστήματος προτάσεων που αναπτύχθηκε. Όπως έχει αναφερθεί η συσταδοποίηση επιτυγχάνεται σε δύο διαστάσεις: άρθρων νέων και χρηστών.

7.2.1 Συσταδοποίηση άρθρων νέων

Ακολουθούν τα πειράματα και τα αποτελέσματα που αφορούν στη συσταδοποίηση άρθρων νέων. Πιο συγκεκριμένα, έγινε αξιολόγηση ορισμένων αλγορίθμων συσταδοποίσης για τον τομέα των άρθρων νέων και στη συνέχεια αξιολογήθηκε και συγκρίθηκε με τους παραπάνω ο αλγόριθμος W-kmeans.

7.2.1.1 Αξιολόγηση βασικών αλγορίθμων βιβλιογραφίας

Στα αρχικά βήματα της διδακτορικής διατριβής, αξιολογήθηκαν αρκετοί αλγόριθμοι συσταδοποίησης καθώς και μετρικές ομοιότητας που υπάρχουν στη βιβλιογραφία. Στόχος μας ήταν να εντοπίσουμε ποιος συνδυασμός αλγορίθμου/μετρικής ομοιότητας έδινε τα καλύτερα αποτελέσματα για την περίπτωση που μας ενδιαφέρει: συσταδοποίηση άρθρων νέων. Ένας σημαντικός παράγοντας που έχει να κάνει με άρθρα νέων γενικά, είναι τόσο η ποικιλομορφία όσο και η ομοιότητά τους ταυτόχρονα. Όταν ανακτούμε άρθρα νέων από πολλαπλά news portals, είναι αναμενόμενο να περιμένουμε, σε κάποιο βαθμό, ομοιότητα μεταξύ τους όσον αφορά το περιεχόμενο, μιας και ο μεγαλύτερος όγκος δημοσιευμένων άρθρων αποτελούν αναδημοσιεύσεις άλλων πηγών. Παρόλα αυτά, είναι σημαντικό να μπορούμε να αντιλαμβανόμαστε προκαταλήψεις (biases) στα άρθρα οι οποίες και φανερώνουν συνήθως διαφορετικές απόψεις. Επιπλέον, όταν έχουμε να κάνουμε με κείμενα σε φυσική γλώσσα, το πλήθος των όρων τους οποίους μπορούμε να συναντήσουμε είναι πρακτικά απεριόριστοι, συγκριτικά π.χ. με την περίπτωση της συσταδοποίησης γονιδίων. Με άλλα λόγια, σε ότι έχει να κάνει με άρθρα νέων από το διαδίκτυο, πρόκειται για δεδομένα υψηλής διαστατικότητας και αραιή αναπαράσταση στο vector space μοντέλο. Οι αλγόριθμοι και οι μετρικές ομοιότητας θα πρέπει επομένως να ανταποκρίνονται αποτελεσματικά στα παραπάνω προβλήματα.

7.2.1.1.1 Σύνολο δεδομένων

Για τα πειράματα που ακολουθούν χρησιμοποιήθηκε ένα σύνολο από 10.000 άρθρα νέων, τυχαία επιλεγμένα, με προέλευση από 50 διαφορετικά news portals και χρονικό εύρος δημοσίευσης 6 μηνών. Τα άρθρα αυτά ανήκουν αποκλειστικά σε μία από τις 8 βασικές κατηγορίες του συστήματός μας. Τυχών διπλά άρθρα με διαφορετική προέλευση αφαιρέθηκαν ήδη από την λίστα με βάση τόσο τον τίτλο όσο και το περιεχόμενο του κειμένου.

7.2.1.1.2 Αποτελέσματα και ανάλυση

Σε αυτό το σύνολο δεδομένων, εφαρμόσαμε τους εξής αλγορίθμους συσταδοποίησης που περιγράφηκαν στην ενότητα 3.7.1:

- ιεραρχικοί:
  - pairwise single linkage, όπου η κοντινότερη απόσταση μεταξύ δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)
  - pairwise maximum linkage, όπου η μακρινότερη απόσταση μεταξύ δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)
  - pairwise average linkage, όπου ο μέσος όρος όλων των αποστάσεων μεταξύ δύο συστάδων λαμβάνεται υπόψιν ως η δια-συσταδική απόσταση (ομοιότητα)
  - centroid linkage, όπου κάθε συστάδα αναπαρίσταται από το κέντρο της το οποίο υπολογίζεται σε κάθε βήμα του αλγορίθμου. Η δια-συσταδική απόσταση (ομοιότητα) σε αυτή την περίπτωση είναι η απόσταση μεταξύ των κέντρων των συστάδων
- διαμερισματικοί:
  - k-means
  - k-medians
  - k-means++

Επιπλέον, για κάθε έναν από τους παραπάνω αλγορίθμους, εκτός από τον k-means++ (ο οποίος υποστηρίζει μόνο Ευκλείδεια απόσταση), χρησιμοποιήσαμε τις ακόλουθες μετρικές ομοιότητας:

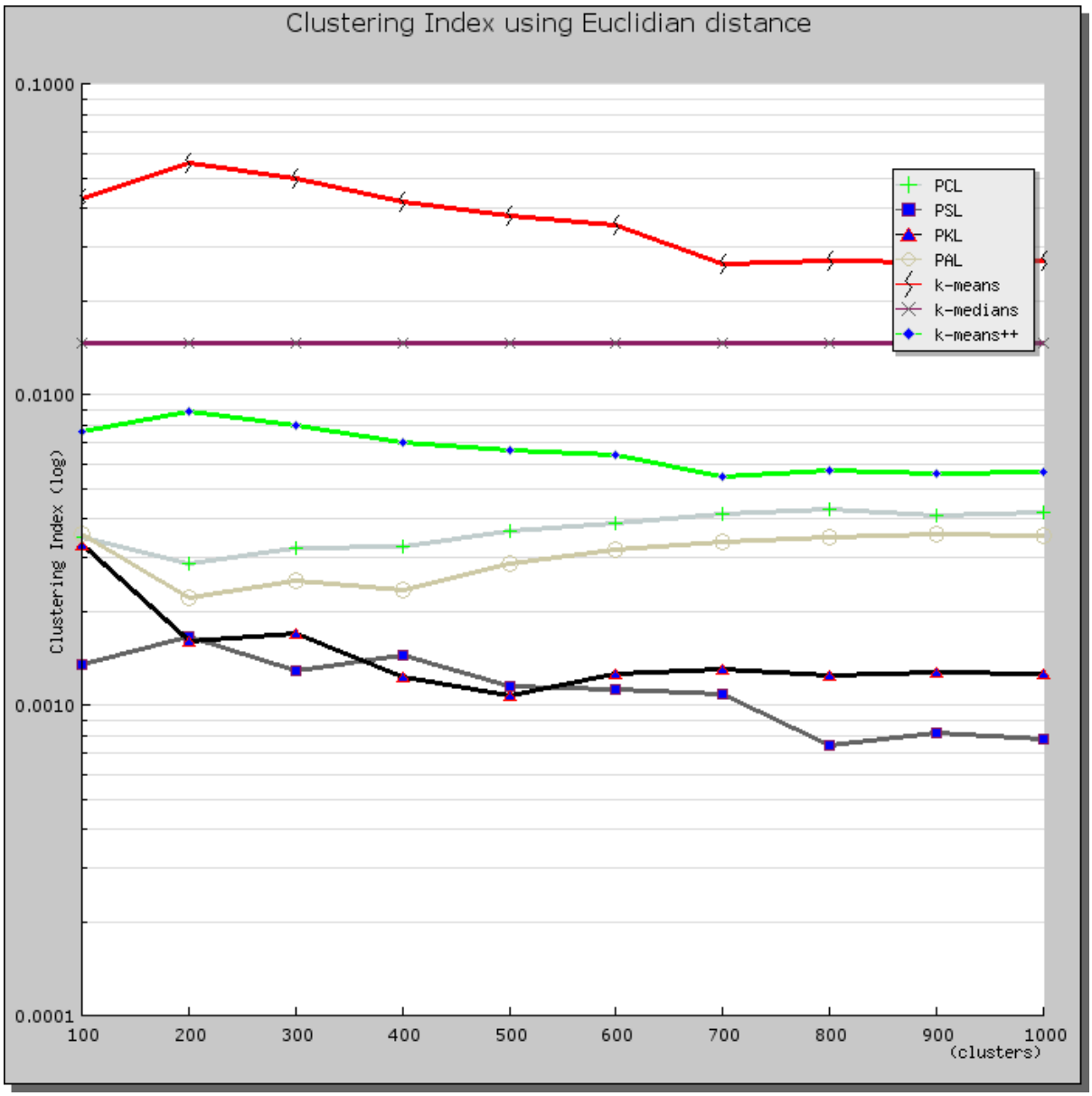
- Ευκλείδεια απόσταση
- City-block / Manhattan απόσταση
- Απόσταση Pearson
- Ομοιότητα συνημιτόνου
- Απόσταση Spearman-rank
- Απόσταση Kendall's τ

Για τους διαμερισματικούς αλγορίθμους χρησιμοποιήσαμε σχήμα 10 περασμάτων με διαφορετικές αρχικές συνθήκες κάθε φορά προκειμένου να αποφευχθούν τοπικά ελάχιστα/μέγιστα λόγω ανομοιογένειας των δεδομένων. Για την αξιολόγηση της αποτελεσματικότητας της κάθε μεθοδολογίας συσταδοποίησης, χρησιμοποιήσαμε την μετρική αξιολόγησης CI (βλέπε ενότητα 3.7.1.6.1). Επιπλέον, για τον καθορισμό της ομοιότητας μεταξύ δύο άρθρων, χρησιμοποιήσαμε τον πίνακα αποστάσεων που παράγεται από την εκάστοτε μετρική ομοιότητας.

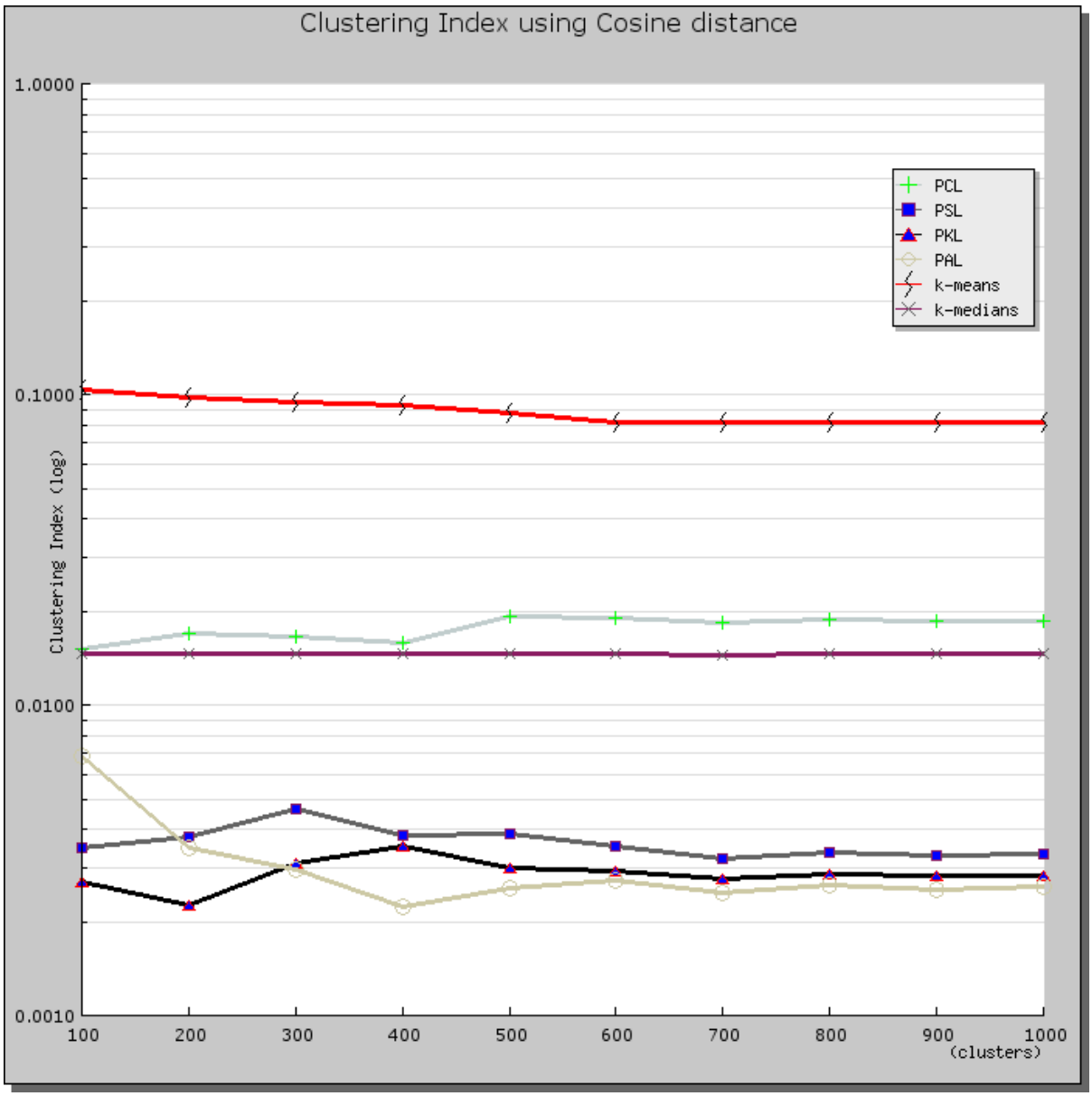
Τα αποτελέσματα για κάθε μεθοδολογία συσταδοποίησης και μετρικής ομοιότητας για πλήθη συστάδων από 100 έως 1000, αποτυπώνονται στα σχήματα 20 - 25. Στο πείραμα αυτό δεν χρησιμοποιήθηκαν όλα τα keywords των κειμένων παρά μόνο οι ρίζες (stemmed) των ουσιαστικών αυτών. Οι συμβολισμοί που χρησιμοποιούνται στις γραφικές παραστάσεις των εν' λόγω σχημάτων για τις ιεραρχικές μεθόδους φαίνονται στον πίνακα 6

Είδος απόστασης	**Συμβολισμός**
Pairwise Maximum (complete) linkage	PCL
Pairwise Single linkage	PSL
Pairwise Centroid linkage	PKL
Pairwise Average linkage	PAL

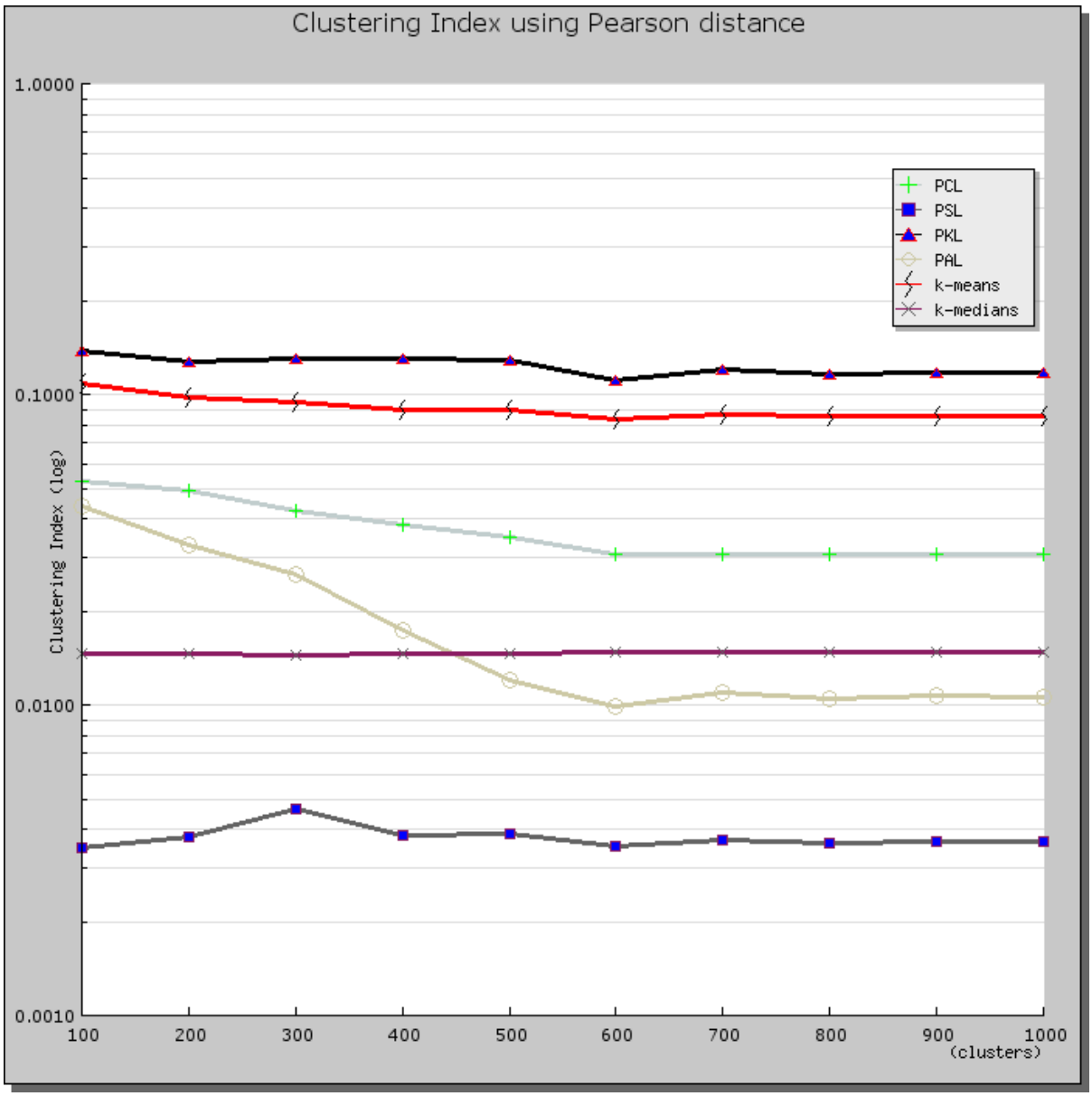
Πίνακας 6: Σημειογραφία ιεραρχικής συσταδοποίησης



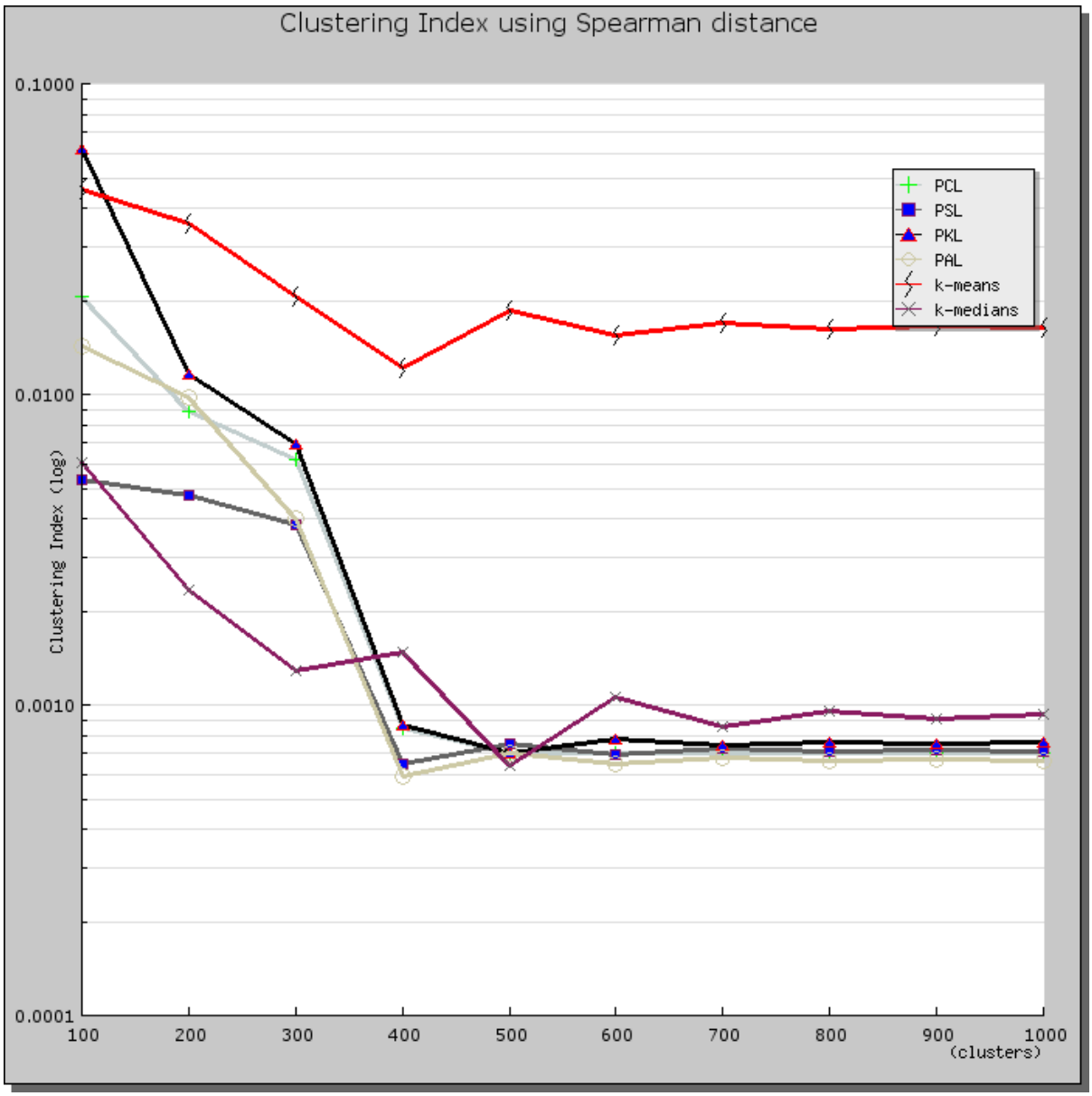
Σχήμα 20: Αποτελέσματα συσταδοποίησης με χρήση της Ευκλείδειας απόστασης



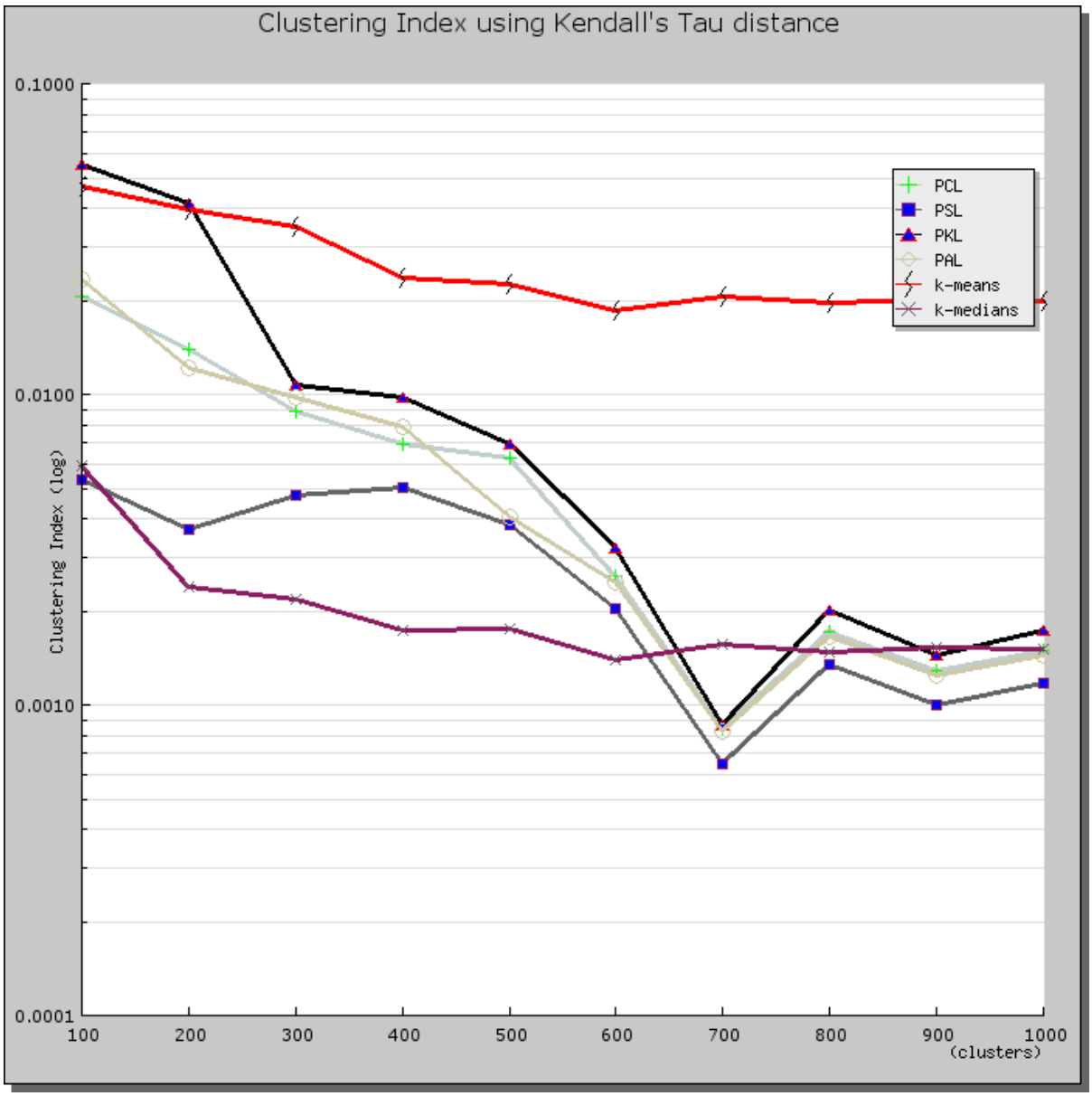
Σχήμα 21: Αποτελέσματα συσταδοποίησης με χρήση της απόστασης συνημιτόνου



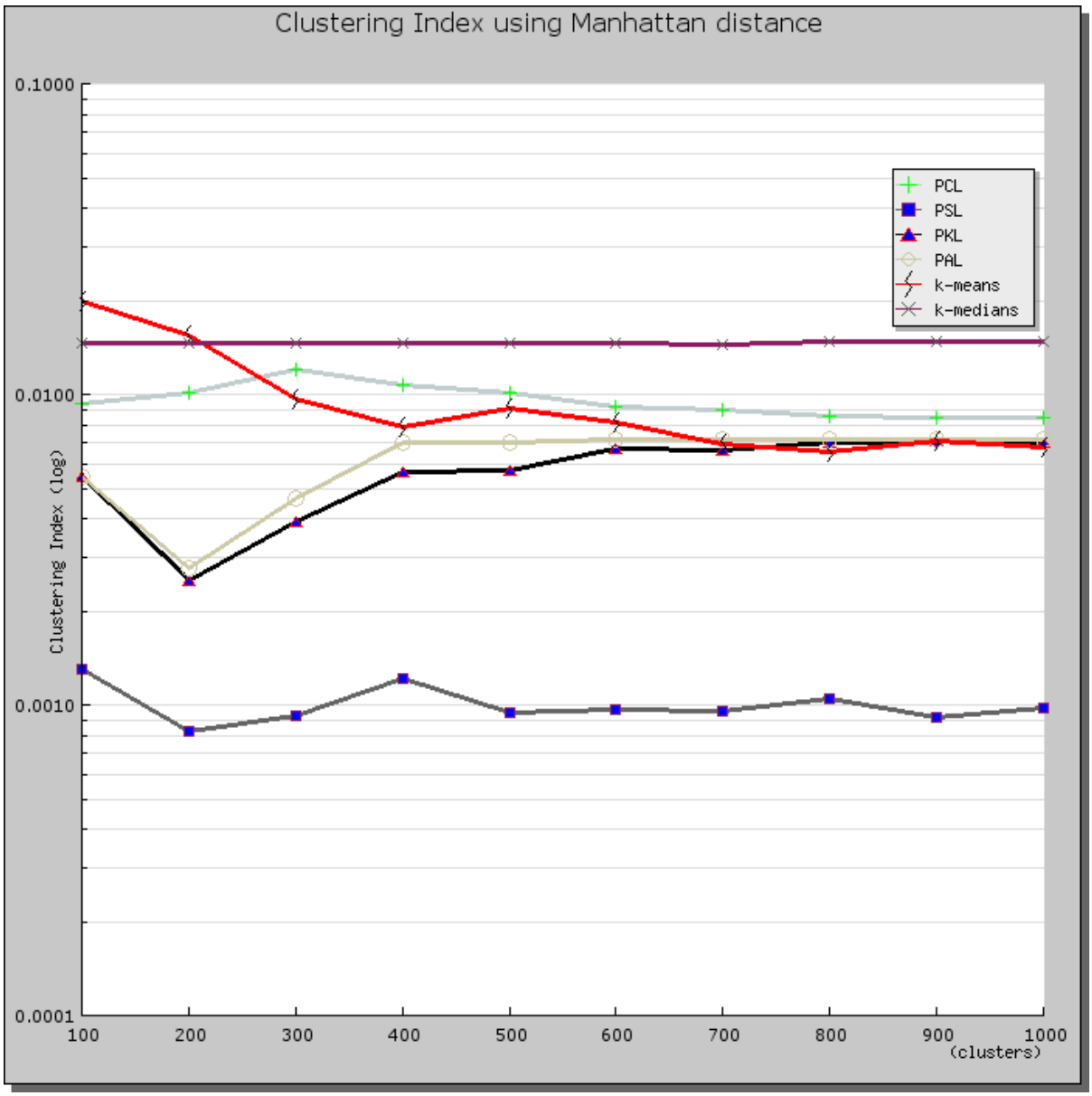
Σχήμα 22: Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Pearson



Σχήμα 23: Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Spearman



Σχήμα 24: Αποτελέσματα συσταδοποίησης με χρήση της απόστασης Kendals' τ



Σχήμα 25: Αποτελέσματα συσταδοποίησης με χρήση της απόστασης City-block

Από τις γραφικές των σχημάτων 20 - 25, φαίνεται ότι ο αλγόριθμος k-means σχεδόν πάντα ξεπερνά κάθε άλλη προσέγγιση συσταδοποίησης.

Επιπλέον, η ομοιότητα συνημιτόνου και η Ευκλείδεια απόσταση αποδεικνύονται καλύτερες για τον k-means μιας και οι συστάδες σε αυτή την περίπτωση είναι καλύτερα συνδεδεμένες, σε σχέση π.χ. με την απόσταση City-block η οποία φαίνεται να ταιριάζει καλύτερα στον αλγόριθμο k-medians. Για την ακρίβεια, ο k-means ξεπερνά τις άλλες προσεγγίσεις για την Ευκλείδεια απόσταση, την ομοιότητα συνημιτόνου, την απόσταση Spearman και την απόσταση Kendall's. Αντίθετα ο k-medians είναι καλύτερος για την απόσταση city-block και ο ιεραρχικός Pairwise Centroid linkage είναι καλύτερος για την απόσταση Pearson.

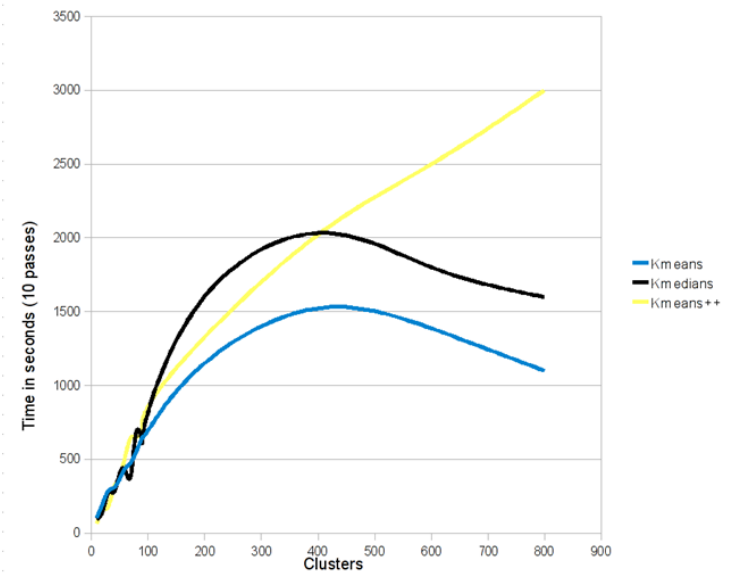
Άλλη μία παρατήρηση που μπορούμε να κάνουμε, είναι ότι το πλήθος των συστάδων επηρεάζει άμεσα την μετρική CI και μετά από ένα συγκεκριμένο όριο συστάδων, κάθε αλγόριθμος φαίνεται να χειροτερεύει σε ότι έχει να κάνει με την CI. Για παράδειγμα, η καλύτερη τιμή CI για την περίπτωση των διαμερισματικών αλγορίθμων παρατηρείται για τον k-means και την ομοιότητα συνημιτόνου και 100 συστάδες, ακολουθούμενη από τον συνδυασμό k-means, Ευκλείδειας απόστασης και 200 συστάδων.

Οι καλύτερες CI τιμές για τους ιεραρχικούς αλγορίθμους παρατηρούνται για τον Pairwise Centroid linkage αλγόριθμο και την απόσταση Pearson. Για τις περισσότερες μετρικές ομοιότητας πάντως παρατηρήσαμε χαμηλότερες τιμές CI για τους ιεραρχικούς αλγορίθμους σε σχέση με τους διαμερισματικούς. Αυτό μάλλον προκύπτει από τον συνήθως προβληματικό τρόπο με τον οποίο οι αλγόριθμοι αυτοί ενεργούν ενώ "κόβουν" το δενδρόγραμμα: παρατηρήσαμε πολλές συστάδες με μόνο 1 στοιχείο (singletons) και λίγες συστάδες με πολλά στοιχεία, κάτι που γενικά οδηγούσε σε πολύ χαμηλές τιμές CI.

Σε ότι έχει να κάνει με τις διαμερισματικές μεθοδολογίες, ο k-means φάνηκε καλύτερος από τον k-medians αλλά ακόμη και από τον k-means++ (Ευκλείδεια απόσταση μόνο), ο οποίος φαίνεται να χειροτερεύει σύντομα καθώς το πλήθος των συστάδων αυξάνει. Επιπλέον, όπως φαίνεται και στο σχήμα 26, ο k-means++ είναι σημαντικά πιο αργός σε σχέση με τους υπολοίπους της ίδιας οικογένειας αλγορίθμων για δεδομένο πλήθος συστάδων.

Για το επόμενο πείραμά μας, προσπαθήσαμε να αξιολογήσουμε την επίδραση της εξαγωγής των ουσιαστικών των κειμένων στην διαδικασία συσταδοποίησης. Για το λόγο αυτό, επαναλάβαμε την παραπάνω πειραματική διαδικασία χωρίς να χρησιμοποιήσουμε αυτή τη φορά μόνο τις ρίζες (stemmed) των ουσιαστικών του κειμένου, όπως αυτά προκύπτουν από την διαδικασία προεπεξεργασίας (χρησιμοποιούμε δηλαδή όλα τα keywords του κειμένου). Η μέση μεταβολή των CI τιμών που προέκυψαν παρουσιάζονται στον πίνακα 7

Εμφανώς, η διαδικασία του stemming και εξαγωγής των ουσιαστικών των άρθρων έχει ένα πολύ ωφέλιμο αποτέλεσμα για όλες τις εφαρμοζόμενες μεθοδολογίες, ειδικά δε για τον αλγόριθμο k-means, κάτι που εν' μέρει εξηγεί και τα αποτελέσματα των προηγούμενων γραφικών παραστάσεων (όπου εφαρμόζεται η εξαγωγή ουσιαστικών).



Σχήμα 26: Χρόνοι εκτέλεσης διαμερισματικών αλγορίθμων σε σχέση με τα πλήθη συστάδων

Μεθοδολογία συσταδοποίησης	**Ποσοστιαία μεταβολή CI**
PCL	+5%
PSL	+6%
PKL	+5%
PAL	+5%
k-means	+18%
k-medians	+16%
k-means++	+15%

Πίνακας 7: Επίδραση της εξαγωγής ουσιαστικών και stemming στις μεθοδολογίες συσταδοποίησης

Παρότι οι μετρικές αξιολόγησης όπως το CI μπορεί να είναι ικανές να δείξουν μία γενική τάση της απόδοσης των τεχνικών συσταδοποίησης που αξιολογήσαμε, δεν μπορούν πολλές φορές όμως να αποτυπώσουν την πρακτική αξία ή την ικανοποίηση των χρηστών. Κατά συνέπεια, για το επόμενο πείραμά μας, εφαρμόσαμε μία εναλλακτική προσέγγιση η οποία βασίζεται σε αξιολόγηση των παραγόμενων αποτελεσμάτων από τους ίδιους τους χρήστες. Προκειμένου λοιπόν να αξιολογήσουμε την ποιότητα των παραγόμενων συστάδων, ζητήσαμε από ένα σύνολο 10 χρηστών του συστήματος να επιτελέσουν το έργο της "χειροκίνητης συσταδοποίησης" σε ένα μικρό υποσύνολο από τα αρχικά μας δεδομένα. Πιο συγκεκριμένα, το ζητούμενο από τους χρήστες ήταν να τοποθετήσουν 50 τυχαία επιλεγμένα άρθρα σε 10 συστάδες με βάση την προσωπική τους και μόνο άποψη. Στη συνέχεια, βγάλαμε τον μέσο όρο των επιλογών τους και συγκρίναμε τα αποτελέσματα με τα περάσματα συσταδοποίησης για κάθε μία από τις προηγούμενες μεθοδολογίες με χρήση της ομοιότητας Ευκλείδειας απόστασης. Το κριτήριο αξιολόγησης σε αυτή την περίπτωση είναι το F-measure (βλέπε ενότητα 3.2.2.3), δηλαδή ο ζυγισμένος αρμονικός μέσος της ακρίβειας

και ανάκλησης μεταξύ των επιλογών των χρηστών και των αποτελεσμάτων των αλγορίθμων.

Τα αποτελέσματα που φαίνονται στον πίνακα 8 φανερώνουν ότι ακόμα και από την μεριά των χρηστών, οι συστάδες που προκύπτουν από τον k-means είναι εγγύτερα στις επιλογές που οι περισσότεροι χρήστες έκαναν για το επιλεγμένο υποσύνολο από άρθρα. Μάλιστα η μέση τιμή 0.61 είναι αρκετά ικανοποιητική δεδομένης της απλότητας του k-means αλγορίθμου.

Μεθοδολογία συσταδοποίησης	**F-measure**
PCL	0.42
PSL	0.42
PKL	0.43
PAL	0.41
k-means	0.61
k-medians	0.57
k-means++	0.51

Πίνακας 8: Αξιολόγηση των μεθοδολογιών συσταδοποίησης σε σχέση με την συσταδοποίηση των ίδιων των χρηστών

Συνοπτικά θα λέγαμε ότι από τα αποτελέσματα του σύνολου των προαναφερθέντων πειραμάτων, ο συνδυασμός του αλγορίθμου k-means με την μετρική ομοιότητας συνημιτόνου, ουσιαστικά ο αλγόριθμος S-kmeans (ενότητα 3.2.2.3), αποδείχθηκε ως η καλύτερη επιλογή για το σύνολο των δεδομένων πάνω στα οποία έτρεξαν. Το αποτέλεσμα αυτό αξιοποιήθηκε για την συνέχεια της ερευνητικής δραστηριότητας της διδακτορικής διατριβής.

7.2.1.2 Αξιολόγηση W-kmeans

Ο αλγόριθμος W-kmeans, όπως προείπαμε, βασίζεται στον κλασικό αλγόριθμο k-means (S-kmeans για την ακρίβεια), αξιοποιώντας την επιπλέον γνώση υπερωνύμων του WordNet. Έχει επιπλέον νόημα η απευθείας σύγκριση των αποτελεσμάτων που παράγουν οι δύο αλγόριθμοι προκειμένου να κατανοήσουμε την βελτίωση που επιφέρουν οι εφαρμοζόμενες τεχνικές.

7.2.1.2.1 Σύνολο δεδομένων

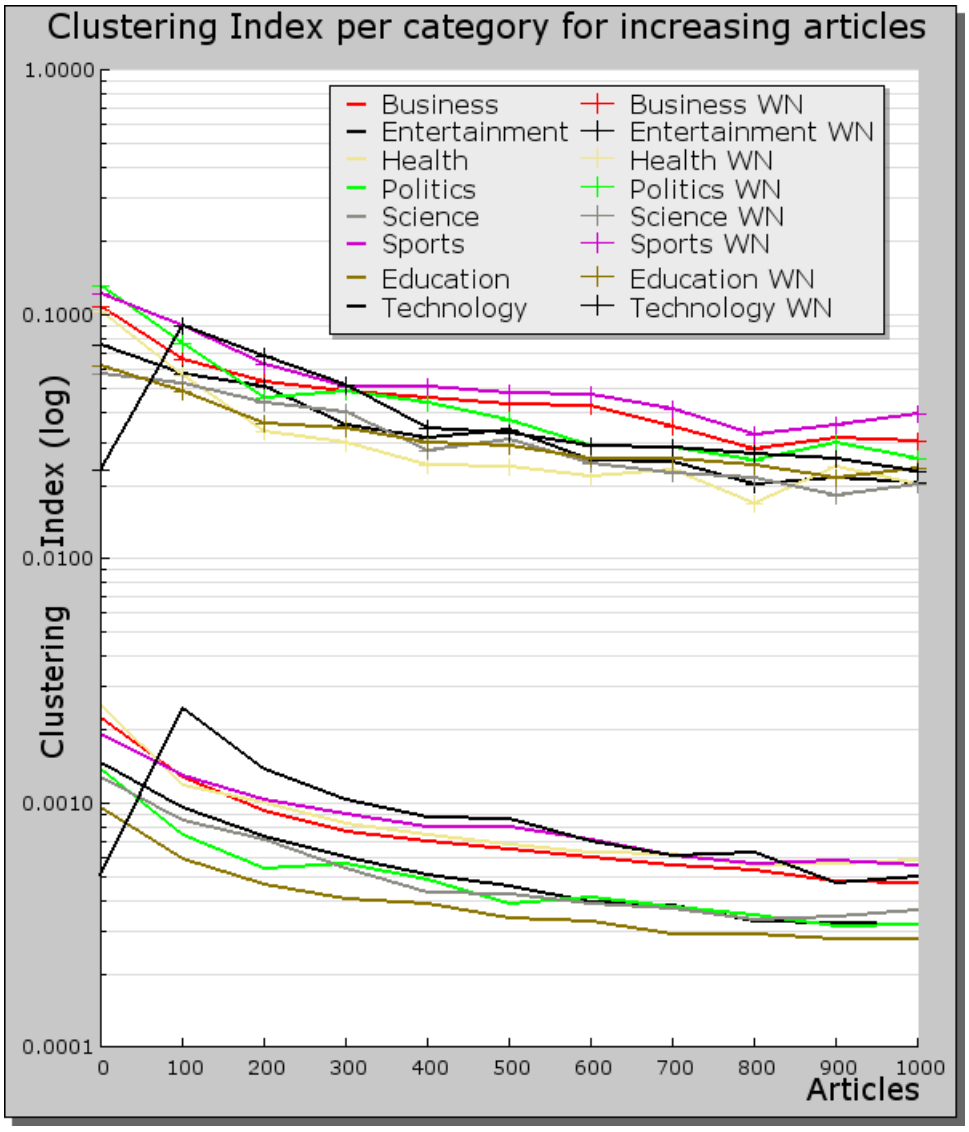
Για την αξιολόγηση του αλγορίθμου χρησιμοποιήσαμε ένα σύνολο από 8.000 άρθρα νέων που το σύστημά μας ανέκτησε από news portals όπως το bbc.com, cnn.com, κλπ. σε ένα χρονικό εύρος 2 μηνών. Τα άρθρα αυτά ήταν ομοιόμορφα κατανεμημένα στις 8 βασικές κατηγορίες του συστήματος. Ως μετρική αξιολόγησης χρησιμοποιήθηκε η CI (βλέπε ενότητα 3.7.1.6.1).

7.2.1.2.2 Αποτελέσματα και ανάλυση

Για το πρώτο μας πείραμα, τρέξαμε τόσο τον αλγόριθμο k-means όσο και τον W-kmeans για το σύνολο δεδομένων μας και υπολογίσαμε τις CI τιμές των παραγόμενων αποτελεσμάτων με

ελεύθερες μεταβλητές τις μεταβαλλόμενες κατηγορίες, το πλήθος των άρθρων καθώς και το πλήθος των συστάδων.

Για τα αποτελέσματα που φαίνονται στις γραφικές παραστάσεις του σχήματος 27, το πάνω σύνολο γραμμών δίνει τις CI τιμές για την περίπτωση των εκτελέσεων με αξιοποίηση του WordNet (W-kmeans), ενώ το κάτω σύνολο γραμμών αφορά τον k-means αλγόριθμο.



Σχήμα 27: Σύγκριση W-kmeans και k-means για διάφορες κατηγορίες και πλήθη άρθρων

Όπως φαίνεται ξεκάθαρα από το σχήμα 27, η ποιότητα της συσταδοποίησης (όσον αφορά την μετρική CI) του W-kmeans είναι αισθητά βελτιωμένη σε σχέση με τον απλό k-means αλγόριθμο. Κάτι που μάλιστα παρατηρείται ανεξάρτητα από το πλήθος των άρθρων ή της κατηγορίας που αυτά ανήκουν. Το παραπάνω αποτελεί μία επιβεβαίωση της αρχικής μας υπόθεσης ότι η χρήση εξωτερικής γνώσης (χαρακτηριστικών) της Αγγλικής γλώσσας, μπορεί να είναι εξαιρετικά χρήσιμη όσον αφορά την συσταδοποίηση.

Μία ακόμη παρατήρηση που μπορούμε να κάνουμε από το σχήμα 27, είναι ότι όσο το πλήθος των άρθρων αυξάνει, η διαφορά των τιμών CI μεταξύ W-kmeans και k-means αυξάνει επίσης. Θεωρούμε ότι αυτό συμβαίνει διότι όσο μεγαλώνει το εύρος των δεδομένων, τόσο αυξάνει και η πιθανότητα εμφάνισης υπερωνύμων μεταξύ τους. Επομένως, όσο μεγαλύτερο το dataset, τόσο μεγαλύτερη και η πιθανότητα του αλγορίθμου W-kmeans να παράγει συστάδες με καλύτερη συνεκτικότητα και περισσότερο διακριτές μεταξύ τους.

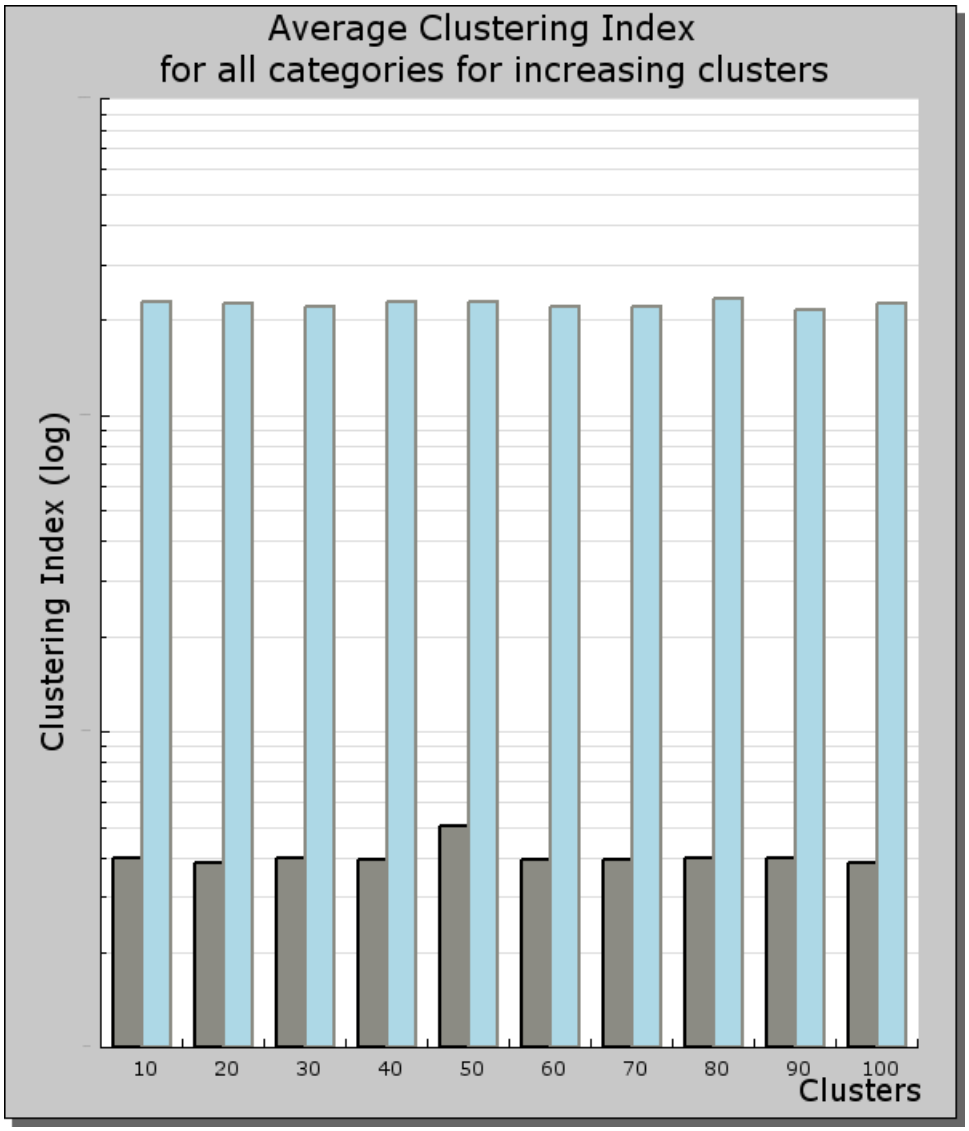
Επιπλέον, στο σχήμα 28 απεικονίζονται οι τιμές CI (μέσος όρος) για αυξανόμενο πλήθος συστάδων σε όλες τις κατηγορίες (για τα ίδια 8.000 άρθρα). Η βελτίωση, όπως και πριν, είναι περίπου 10 φορές καλύτερες CI τιμές σε σχέση με τον κλασικό k-means (οι κλίμακες στον άξονα y των σχημάτων 27 και 28 είναι λογαριθμικές). Επίσης για το συγκεκριμένο dataset παρατηρήσαμε ότι για την περίπτωση των 50 συστάδων, οι τιμές CI είναι σχετικά βελτιωμένες σε σχέση με τις υπόλοιπες τιμές πλήθους συστάδων. Το παραπάνω αποτέλεσε μία σαφή ένδειξη για το πραγματικό πλήθος συστάδων των δεδομένων το οποίο πράγματι στη συνέχεια επιβεβαιώσαμε ότι ήταν 51 συστάδες.

Στη συνέχεια, προχωρήσαμε σε ένα ακόμη πείραμα προκειμένου να συγκρίνουμε τον W-kmeans με state of the art εργαλεία συσταδοποίησης που υπάρχουν διαθέσιμα online, πιο συγκεκριμένα, τα CLUTO [112] και SenseClusters [175]. Χρησιμοποιήσαμε το ίδιο dataset με πριν και το κριτήριο αξιολόγησης CI. Το πλήθος των συστάδων στα δεδομένα μας τέθηκε ως 50. Για την περίπτωση του CLUTO, ο μέσος όρος του CI για όλες τις 5 προσεγγίσεις που προσφέρει (4 διαμερισματικές και 1 ιεραρχική) υπολογίστηκε. Τα αποτελέσματα της πειραματικής διαδικασίας φαίνονται στον πίνακα 9 όπου επίσης καταγράφονται και οι χρόνοι εκτέλεσης που απαιτήθηκαν για κάθε μία από τις προαναφερθείσες προσεγγίσεις.

Προσέγγιση συσταδοποίησης	**CI**	**Χρόνος εκτέλεσης (δευτερόλεπτα)**
CLUTO	0.80	204 (μέσος όρος των 5 προσεγγίσεων)
SenseCluster	0.56	302
W-kmeans	0.84	198

Πίνακας 9: Σύγκριση του W-kmeans με CLUTO και SenseCluster σε σχέση με CI και χρόνο εκτέλεσης.

Από τα αποτελέσματα του πίνακα 9 βλέπουμε ότι ο W-kmeans παρέχει αποτελέσματα συσταδοποίησης που υπερτερούν των υπολοίπων προσεγγίσεων όσον αφορά τις τιμές CI, ειδικά σε σχέση με το SenseCluster. Παράλληλα οι χρόνοι εκτέλεσης είναι σημαντικά μικρότεροι σε σχέση με όλες τις προαναφερθείσες μεθοδολογίες, κάτι που μάλλον έχει να κάνει με την απλή φύση του αλγορίθμου.



Σχήμα 28: Σύγκριση W-kmeans και k-means για συσταδοποίηση άρθρων νέων και για διάφορα πλήθη συστάδων

7.2.1.3 Αξιολόγηση ονοματοδοσίας συστάδων

Στην συνέχεια, προχωρήσαμε σε πειραματική αξιολογήση της αποτελεσματικότητας της διαδικασίας ονοματοδοσίας συστάδων που επιτελεί ο αλγόριθμος W-kmeans.

7.2.1.3.1 Σύνολο δεδομένων

Προς αυτή την κατεύθυνση αξιοποιήσαμε το προηγούμενο dataset (8.000 άρθρα νέων) το οποίο και περάσαμε από τον αλγόριθμο W-kmeans με ζητούμενο πλήθος συστάδων ίσο με το πλήθος των κατηγοριών του συστήματος (8). Μιας και γνωρίζουμε εκ' των προτέρων ότι τα άρθρα ανήκουν σε κάποια από τις 8 κατηγορίες, συγκρίναμε τις ετικέτες που προέκυψαν με λίστες από keywords που παράξαμε από τις κατηγορίες αυτές καθ' αυτές. Οι λίστες αυτές περιείχαν:

- τα 10 πιο συχνά keywords της κάθε κατηγορίας
- το όνομα της εκάστοτε κατηγορίας

Οι ετικέτες που "έπεφταν κοντά" (π.χ. συνώνυμα ή παράγωγα) στα περιεχόμενα αυτών των λιστών, αξιολογούνταν ως αντιπροσωπευτικές (επιτυχής ονοματοδοσία). Όλες οι υπόλοιπες, αξιολογούνταν ως ανεπιτυχείς επιλογές ονοματοδοσίας (miss).

7.2.1.3.2 Αποτελέσματα και ανάλυση

Με βάση τα παραπάνω, αξιολογήσαμε την ακρίβεια των παραγόμενων ετικετών σε σχέση με την εκάστοτε κατηγορία των άρθρων. Η μετρική επομένως που χρησιμοποιήσαμε είναι αυτή της ακρίβειας που αναλύθηκε στην ενότητα 3.2.2. Η ακρίβεια λοιπόν της ονοματοδοσίας i και της αναφερόμενης κατηγορίας j ορίζεται ως:

$$Pr_{i,j} = avg\_rank(i,j) * \frac{a}{a+b} \tag{57}$$

όπου $avg\_rank(i,j)$ η μέση κατάταξη που έχει η ετικέτα i στην συνολική λίστα της κατηγορίας j, a το πλήθος των όρων που η διαδικασία ονοματοδοσίας i έχει για την κατηγορία j και b το πλήθος των όρων που η διαδικασία ονοματοδοσίας i έχει αλλά δεν είναι στην j.

Τα αποτελέσματα ανά κατηγορία παρουσιάζονται στον πίνακα 10. Από αυτά βλέπουμε ότι το συνολικό ποσοστό ακρίβειας των παραγόμενων συστάδων αγγίζει κατά μέσο όρο το 75%. Το παραπάνω ποσοστό μάλιστα ίσως να ήταν ακόμη καλύτερο, αν οι κατηγορίες "επιστήμη" και "τεχνολογία" ενώνονταν μιας και παρατηρήθηκε ότι εμφάνιζαν σχετικά κοντινές ετικέτες.

Συνολικά, από τα πειραματικά αποτελέσματα που παρουσιάστηκαν στην παρούσα ενότητα, θα λέγαμε ότι ξεκινώντας από τον S-kmeans αλγόριθμο, ο W-kmeans, αξιοποιώντας την εξωτερική γνώση από το WordNet, παρουσιάζει σημαντικά βελτιωμένα αποτελέσματα σε ότι έχει να κάνει με τις παραγόμενες συστάδες, ενώ παράλληλα υπερτερεί από άποψη χρόνου εκτέλεσης σε σχέση με τα συγκρινόμενα εργαλεία συσταδοποίησης. Τα αποτελέσματα αυτά ήταν άκρως ενθαρρυντικά για την πορεία της διδακτορικής έρευνας και μας οδήγησαν ουσιαστικά στην ενσωμάτωση του αλγορίθμου στο συνολικό σύστημα προτάσεων.

Κατηγορία	Ακρίβεια
Business	85%
Entertainment	78%
Health	90%
Politics	88%
Science	65%
Technology	60%
Education	75%
Sports	90%

Πίνακας 10: Αποτελέσματα ακρίβειας της ονοματοδοσίας συστάδων του W-kmeans ανά κατηγορία

7.2.2 Συσταδοποίηση χρηστών

Για την αξιολόγηση του αλγορίθμου W-kmeans όσον αφορά την συσταδοποίηση χρήστων, εκτελέστηκαν ορισμένα πειράματα τα οποία και περιγράφονται στα επόμενα.

7.2.2.1 Σύνολο δεδομένων

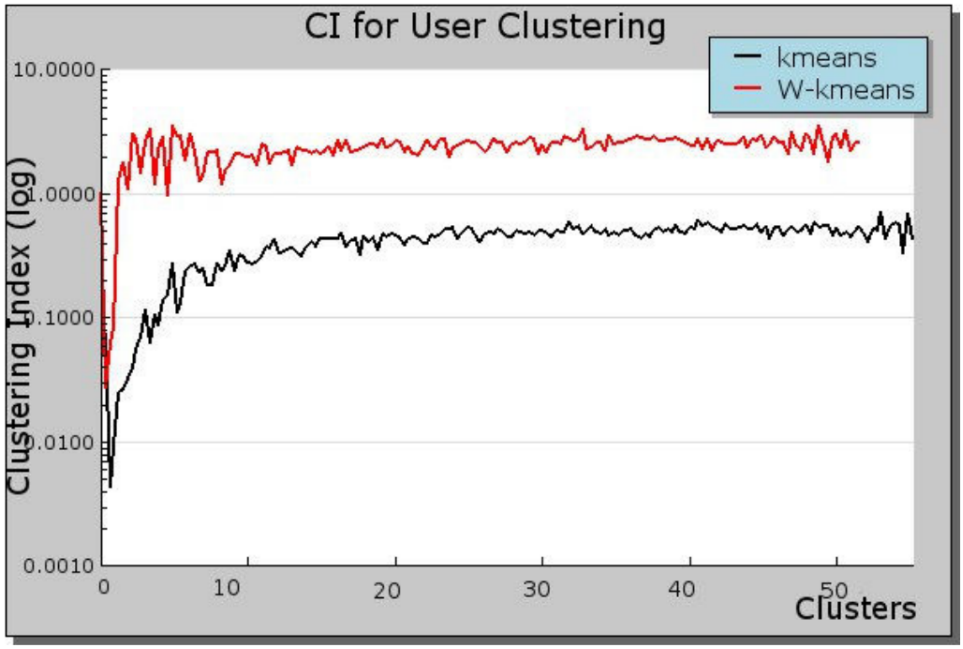
Το σύνολο δεδομένων για τα πειράματα που ακολουθούν αποτελείται από 10.000 άρθρα νέων από news portals τα οποία ανακτήθηκαν σε εύρος 6 μηνών. Αυτά τα άρθρα ήταν ομοιόμορφα κατανεμημένα ανάμεσα στις 8 βασικές κατηγορίες του συστήματός μας. Για καθένα από αυτά τα άρθρα, το αποτέλεσμα της προεπεξεργασίας ήταν stemmed ουσιαστικά. Εκτός από τα παραπάνω, αξιοποιήσαμε και τα πρότυπα πλοήγησης τα οποία καταγράφηκαν για 50 εγγεγραμμένους χρήστες του συστήματος την ίδια χρονική περίοδο. Για κάθε χρήστη κρατήσαμε τα επιλεγμένα άρθρα καθώς και τον χρόνο που ξόδεψαν διαβάζοντας το καθένα, όπως ακριβώς περιγράφεται στον αλγόριθμο 5, εξάγοντας έτσι τις συνεδρίες από τα ιστορικά πλοήγησής τους. Ως μετρικές αξιολόγησης χρησιμοποιήθηκαν η CI (ενότητα 3.7.1.6.1) και το F-measure (ενότητα 3.2.2.3).

7.2.2.2 Αποτελέσματα και ανάλυση

Για το πρώτο μας πείραμα, συγκρίναμε τους αλγόριθμους W-kmeans και k-means όσον αφορά την εφαρμογή τους στην συσταδοποίηση χρηστών. Πιο συγκεκριμένα, τρέξαμε τον κάθε αλγόριθμο για όλα τα δεδομένα των συνεδριών χρηστών, καθώς και πολλαπλές τιμές πλήθους συστάδων.

Τα αποτελέσματα συσταδοποίησης των εξαγόμενων συνεδριών, που φαίνονται στο σχήμα 29, δείχνουν ότι ο W-kmeans είναι σαφώς αποτελεσματικότερος του k-means, παρέχοντας έτσι, τουλάχιστον σε ότι έχει να κάνει το CI, συστάδες πολύ καλά συνδεδεμένες μεταξύ τους. Ως λογική συνέπεια του παραπάνω, οι παραγόμενες συστάδες μπορούν να αποτυπώσουν με μεγαλύτερη ακρίβεια χρήστες με παρόμοια ενδιαφέροντα, ενώ παράλληλα διαχωρίζουν επιτυχώς χρήστες με αντικρουόμενα ενδιαφέροντα.

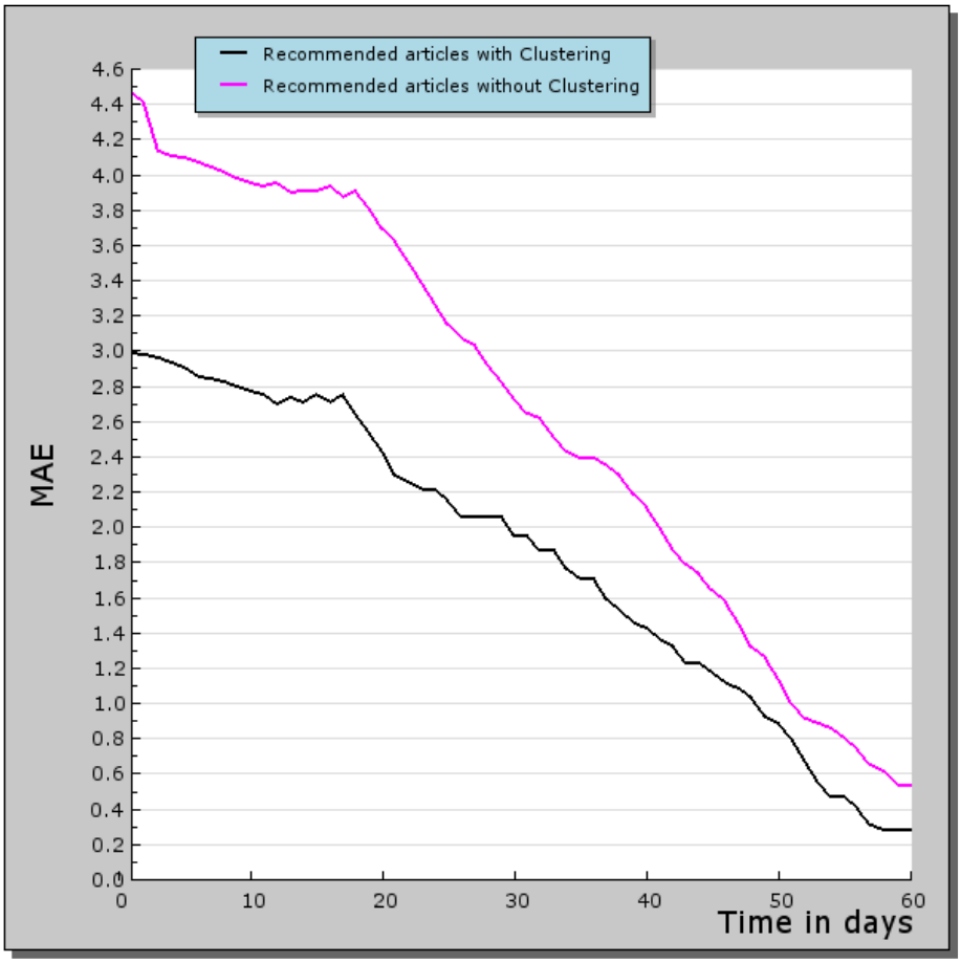
Στη συνέχεια προσπαθήσαμε να μετρήσουμε την επίδραση της συσταδοποίησης χρηστών σε ότι έχει να κάνει με τις προτάσεις του συστήματος. Έτσι, το σχήμα 30, απεικονίζει τα MAE αποτελέ-



Σχήμα 29: Σύγκριση W-kmeans και k-means για συσταδοποίηση συνεδριών χρηστών και διάφορα πλήθη συστάδων

σματα που αποκομίσαμε καθόσον πέρναγαν οι μέρες, όταν τόσο η συσταδοποίηση άρθρων νέων όσο και η συσταδοποίηση χρηστών εφαρμόζεται για την διαδικασία παραγωγής προτάσεων. Μπορούμε να παρατηρήσουμε ότι η εφαρμογή της συσταδοποίησης χρηστών οδηγεί σε σημαντική μείωση των τιμών MAE των προτάσεων. Πιο συγκεκριμένα, βλέπουμε ότι όσο οι χρήστες διάβαζαν ολοένα και περισσότερα άρθρα και το προφίλ τους διαμορφωνόταν, οι τιμές MAE μειώνονται. Το παραπάνω είναι αληθές τόσο όταν η συσταδοποίηση χρηστών εφαρμόζεται, όσο και όταν δεν εφαρμόζεται. Η πρακτική αξία της προαναφερθείσας παρατήρησης είναι ότι οι προτάσεις που παρέχονταν στους χρήστες ήταν, με αυξανόμενη τάση, ακριβείς δεδομένου ότι οι χρήστες επέλεγαν να τις διαβάσουν. Επίσης, με το να λαμβάνεται υπόψιν και η πληροφορία συσταδοποίησης χρηστών, οι τιμές MAE των προτάσεων άρθρων νέων σε σχέση με τις πραγματικές επιλογές των χρηστών μειώθηκαν κατά μέσο όρο 15%, συγκρινόμενες με την περίπτωση που η συσταδοποίηση χρηστών δεν εφαρμοζόταν. Το παραπάνω είναι πιο σαφές από την γραφική παράσταση ιδίως τις πρώτες μέρες του πειράματος, όταν τα προφίλ χρηστών δεν ήταν ακόμη σαφή. Παρόλα αυτά όμως, ακόμα και όταν τα προφίλ των χρηστών έφτασαν μία σταθερή κατάσταση, περί τις 45 ημέρες, οι MAE τιμές ήταν επίσης πιο χαμηλές όταν η συσταδοποίηση χρηστών λαμβάνονταν υπόψιν.

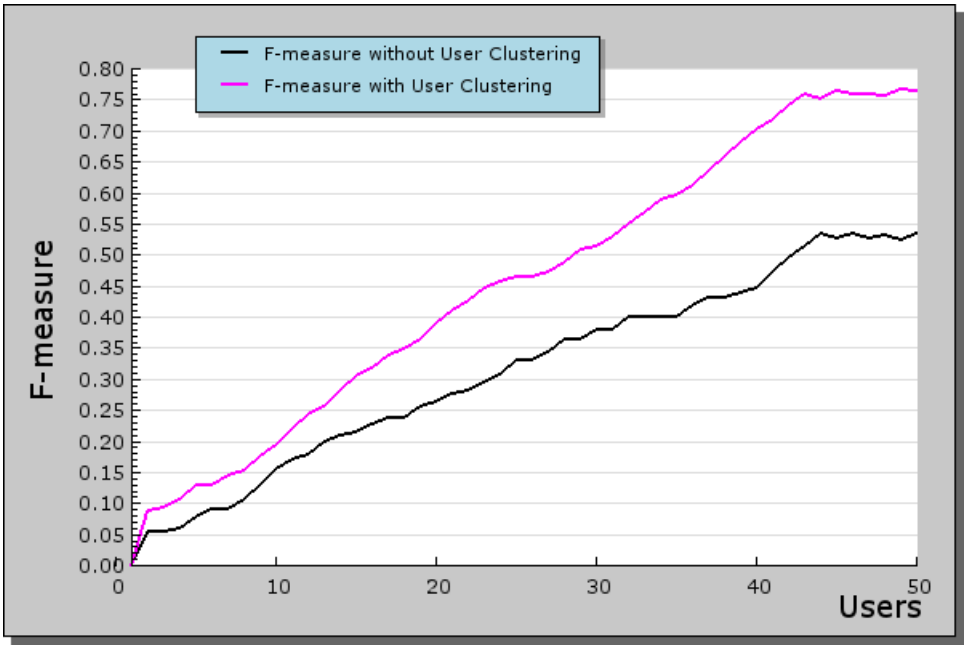
Για το επόμενο πείραμά μας, προσπαθήσαμε να εκτιμήσουμε την συνολική βελτίωση του έχει στον μηχανισμό παραγωγής προτάσεων η αξιοποίηση της πληροφορίας συσταδοποίησης χρηστών. Για το λόγω αυτό, όσον αφορά την παραγωγή προτάσεων χρησιμοποιήσαμε τα βήματα που περιγράφονται στον αλγόριθμο 8. Για χρήστες που επιστρέφουν προτείναμε 10 από τα πιο συχνά αναγνωσμένα άρθρα από τους χρήστες που ανήκουν στην ίδια συστάδα του χρήστη. Στη συνέχεια καταγράψαμε ποια από τα προτεινόμενα άρθρα διαβάστηκαν από τον χρήστη μέσα σε ένα χρονικό



Σχήμα 30: Τιμές MAE των προτάσεων του συστήματος με και χωρίς την χρήση του W-kmeans

ορίζοντα 30 λεπτών από την είσοδό του στο σύστημα. Η διαδικασία επαναλήφθηκε χωρίς την χρήστη της πληροφορίας συσταδοποίησης χρηστών.

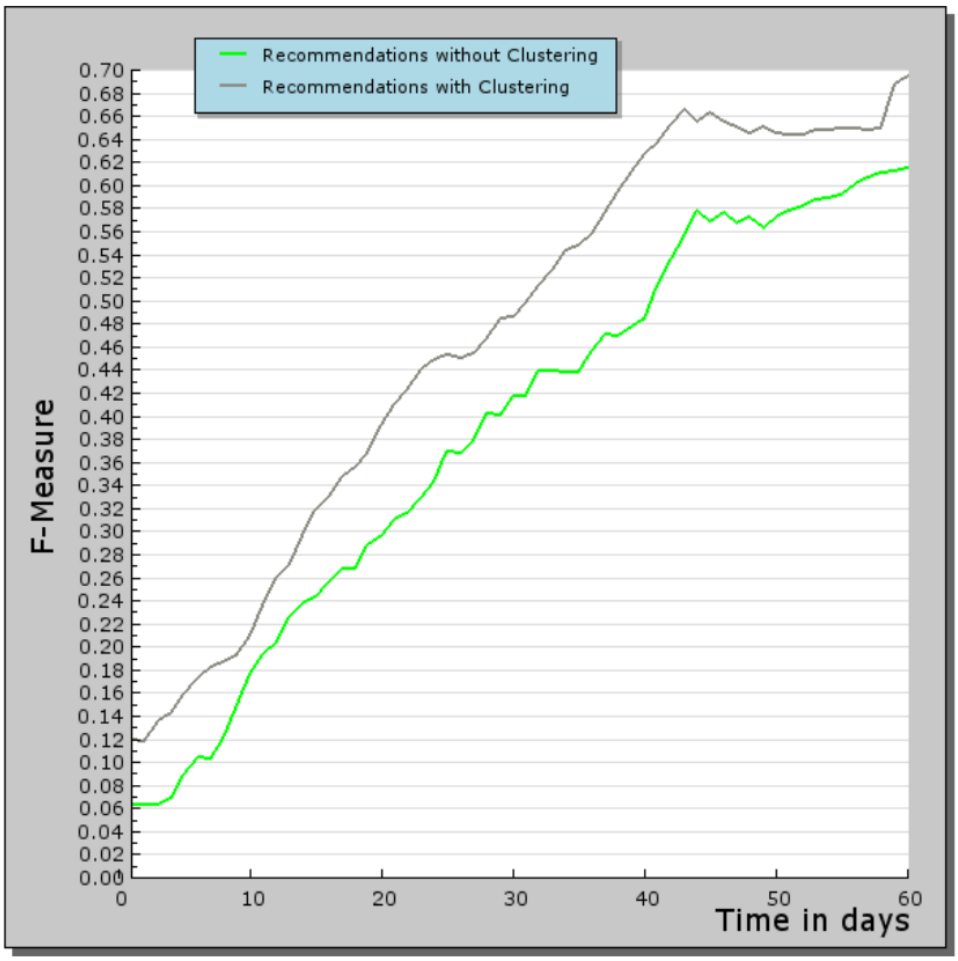
Τα αποτελέσματα που παρουσιάζονται στο σχήμα 31 δείχνουν τους μέσους όρους των τιμών F-measure για κάθε περίπτωση και για αυξανόμενο πλήθος χρηστών. Παρατηρούμε ότι η αποδοτικότητα των προτάσεων αυξάνει συνεχώς καθώς όλο και περισσότεροι χρήστες λαμβάνονται υπόψιν από το σύστημα, κάτι που είναι αναμενόμενο δεδομένων των χαρακτηριστικών προσωποποίησης του συστήματός μας. Στατιστικά, είδαμε ότι οι παραγόμενες προτάσεις ταιριάζουν στις επιλογές χρήστη 7 στις 10 φορές, κάτι που κατά τη γνώμη μας αποδεικνύει ότι η αξιοποίηση της πληροφορίας συσταδοποίησης χρηστών μπορεί να επιφέρει σημαντικά οφέλη για το σύστημα προτάσεων και σε μεγαλύτερη κλίμακα δεδομένων και χρηστών.



Σχήμα 31: Σύγκριση της απόδοσης του συστήματος προτάσεων με χρήστη της πληροφορίας συσταδοποίησης χρηστών και μη

Όπως και πριν, χρησιμοποιώντας τα ίδια δεδομένα μετρήσαμε τις τιμές του F-measure με το πέρασμα του χρόνου είτε όταν μόνο η συσταδοποίηση άρθρων νέων εφαρμοζόταν, είτε όταν και η συσταδοποίηση άρθρων νέων και χρηστών εφαρμοζόταν. Από τα αποτελέσματα, τα οποία παρουσιάζονται στο σχήμα 32, μπορούμε να παρατηρήσουμε ότι οι προτάσεις που κάνουν χρήση των παραγόμενων συστάδων άρθρων και χρηστών παράγουν κατά μέσο όρο 0.1 καλύτερες τιμές σε σχέση με το F-measure. Όπως και πριν, η βελτίωση γίνεται ακόμη καλύτερη ύστερα από μερικές μέρες χρήσης του συστήματος. Αυτό έχει δύο εξηγήσεις. Πρώτον, το σύστημα έχει περισσότερα δεδομένα σχετικά με τις προτιμήσεις και επιλογές των χρηστών, και δεύτερον, το σύστημα έχει περισσότερο χρόνο να παράξει συστάδες με καλύτερη συνοχή και γενικά πιο σωστές. Από το σχήμα 32 μπορούμε επίσης να δούμε ότι περίπου στις 45 ημέρες οι προτάσεις φτάνουν στην καλύτερη απόδοσή τους, αποκαλύπτοντας έτσι ότι, κατά μέσο όρο, τα προφίλ χρηστών έχουν φτάσει σε μία

σταθερή φάση.



Σχήμα 32: F-measure τιμές των προτάσεων του συστήματος με και χωρίς την χρήση του W-kmeans

Στη συνέχεια και για το επόμενο πείραμά μας, προσπαθήσαμε να εκτιμήσουμε την απόδοση της προτεινόμενης μεθοδολογίας συσταδοποίσης χρηστών όσον αφορά τις παραγόμενες προτάσεις προς τον χρήστη, σε σύγκριση με state of the art μεθοδολογίες που αξιοποιούνται στον τομέα του CF, όπως latent semantic CF, neighbor-based CF, καθώς και τεχνικές μείωσης διαστατικότητας (SVD). Τα αποτελέσματα για το ίδιο σύνολο δεδομένων με πριν, παρουσιάζονται στον πίνακα 11 και δείχνουν ότι η προσέγγιση συσταδοποίησης W-kmeans είναι σχεδόν τόσο αποτελεσματική όσο και τεχνικές μείωσης διαστατικότητας (SVD), ενώ υπερτερεί των Latent semantic CF και Neighbor-based CF.

Μεθοδολογία CF	**Μέσος όρος F-measure για όλους τους χρήστες**
W-kmeans	0.45
Latent semantic CF	0.4
Neighbor-based CF	0.35
SVD	0.5

Πίνακας 11: Σύγκριση μεθοδολογιών CF

7.3 Πρόβλημα νέου χρήστη

Για την αξιολόγηση της προτεινόμενης μεθοδολογίας αντιμετώπισης του προβλήματος νέου χρήστη, προχωρήσαμε στην πειραματική διαδικασία η οποία αξιολογεί ουσιαστικά τον αλγόριθμο 10.

7.3.1 Σύνολο δεδομένων

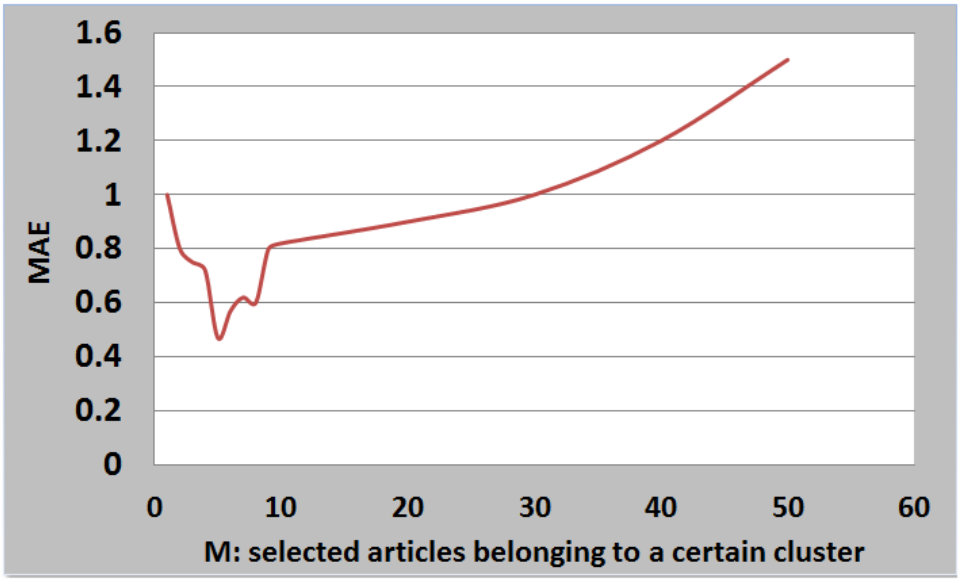
Το σύνολο δεδομένων που χρησιμοποιήθηκε αποτελείται από τα μοτίβα βαθμολόγησης άρθρων νέων των χρηστών του συστήματός μας. Πιο συγκεκριμένα, αφού αφαιρέθηκαν οι χρήστες του συστήματος με λιγότερες από 50 καταγεγραμμένες βαθμολογήσεις άρθρων νέων, κρατήσαμε τις βαθμολογήσεις από 60 χρήστες οι οποίοι είχαν αξιολογήσει 500 άρθρα νέων με πάνω από 1.000 αξιολογήσεις. Το όριο των 50 συνολικών αξιολογήσεων ανά χρήστη είναι σημαντικό, μιας και θέλουμε να αποφύγουμε χρήστες οι οποίοι δεν έχουν χρησιμοποιήσει εκτενώς το σύστημα από το να επηρεάσουν την διαδικασία αξιολόγησης του μηχανισμού. Σε γενικές γραμμές, επιθυμούμε πολλές αξιολογήσεις από κάθε χρήστη προκειμένου να έχουμε ένα καλό δείγμα από άρθρα τα οποία οι χρήστες ήταν πρόθυμοι να αξιολογήσουν - κάτι που ουσιαστικά απεικονίζει εμμέσως τις προτιμήσεις τους.

Δεδομένου ότι ένας νέος χρήστης στην πειραματική διαδικασία που ακολουθεί είναι πρακτικά κάθε ένας από τους 60 χρήστες που προαναφέραμε, για κάθε εκτέλεση που αφορούσε τον κάθε χρήστη, δεν υπολογίζαμε τις προηγούμενες αξιολογήσεις που ήταν καταγεγραμμένες από το σύστημα. Καθώς παρουσιάζαμε τα άρθρα βασισμένοι σε κάθε μία από τις στρατηγικές που παρουσιάστηκαν στην ενότητα 3.9.1 και που αξιολογούμε εδώ, κάναμε την εξής παραδοχή: οι χρήστες "αξιολόγησαν" τα άρθρα εκείνα για τα οποία έχουμε καταγραφή ανάγνωσης στη ΒΔ. Έτσι, εάν ένα άρθρο που παρουσιάζεται στον χρήστη έχει βρεθεί ως αναγνωσμένο ή βαθμολογημένο στη ΒΔ, θεωρούμε ότι ήταν μία επιτυχής πρόταση για αξιολόγηση προς τον χρήστη (hit). Σε ότι έχει να κάνει με τα σκορ βαθμολόγησης, το σκορ εκείνο που βρίσκεται στη ΒΔ χρησιμοποιείται για την περίπτωση που γενικά ο χρήστης έχει αξιολογήσει το συγκεκριμένο άρθρο. Εάν ωστόσο το άρθρο έχει επιλεχθεί για ανάγνωση από το χρήστη, ως σκορ χρησιμοποιούμε το μέγιστο στην κλίμακα 1-5 που μπορεί ο χρήστης να βαθμολογήσει, δηλαδή 5.

7.3.2 Αποτελέσματα και ανάλυση

Για το πρώτο και δεύτερο πείραμα που ακολουθούν, σταματήσαμε να παρουσιάζουμε άρθρα όταν είχαμε τον απαιτούμενο αριθμό από αξιολογήσεις άρθρων, ο οποίος για την περίπτωσή μας ήταν $R_{min} = 20$.

Στο πρώτο πείραμα, προσπαθήσαμε να εκτιμήσουμε την καλύτερη τιμή για την παράμετρο M όπως αυτή περιγράφεται στην ενότητα 5.4 και στον αλγόριθμο 10. Δηλαδή, το καλύτερο πλήθος άρθρων που θα πρέπει να παρουσιαστούν στο χρήστη και τα οποία ανήκουν σε συστάδα (είτε άρθρων είτε χρηστών) εφόσον μία ή περισσότερες αξιολογήσεις έχουν ανακτηθεί από τον χρήστη. Για την αξιολόγηση των προτάσεων του μηχανισμού χρησιμοποιούμε την μετρική του MAE που περιγράφηκε στην ενότητα 3.7.1.6.2. Πιο συγκεκριμένα, στη σχέση 27, $r(u,i) \in [1,5]$ είναι η πραγματική αξιολόγηση του άρθρου i από τον u (η οποία βρίσκεται όπως περιγράφηκε νωρίτερα) και $r'(u,i) \in [1,5]$ η προβλεπόμενη/εκτιμώμενη προτίμηση του χρήστη u για τα άρθρα που ανήκουν στο χώρο των προτεινόμενων προς αυτόν άρθρων νέων, R'. Για το συγκεκριμένο πείραμα χρησιμοποιήσαμε έναν αυξανόμενο αριθμό από τιμές για το M σε κάθε εκτέλεση, αρχίζοντας με $M = 1$ και τελειώνοντας με $M = 50$. Τα αποτελέσματα παρουσιάζονται στην γραφική παράσταση του σχήματος 33.



Σχήμα 33: Αξιολόγηση των επιλογών του συστήματος για πρόταση προς το χρήστη ώστε να συγκεντρωθούν οι απαραίτητες βαθμολογήσεις άρθρων νέων

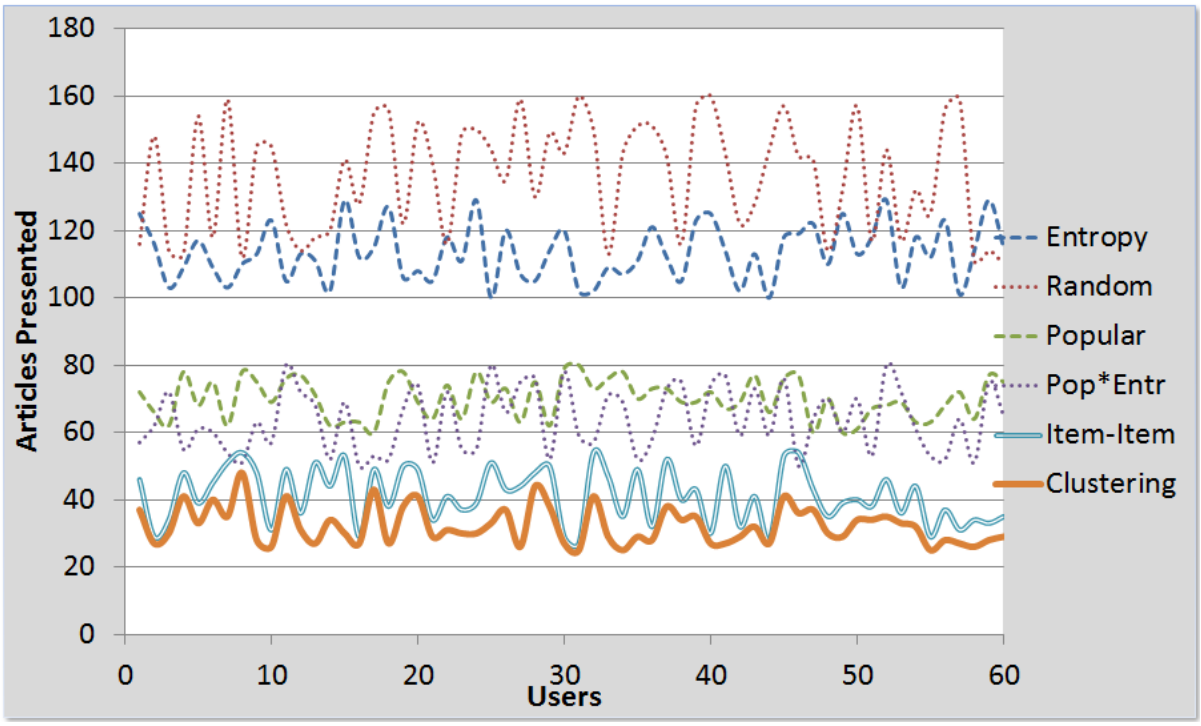
Από την γραφική παράσταση του σχήματος 33, μπορούμε να εντοπίσουμε την καλύτερη τιμή της παραμέτρου M σε σχέση με τις τιμές MAE, δηλαδή: $M = 5$. Η φυσική έννοια αυτού του αποτελέσματος είναι ότι η επιλογή 5 άρθρων από τις συστάδες άρθρων ή άρθρων που διάβασαν χρήστες της συστάδας είναι η καλύτερη επιλογή για την διαμόρφωση των λιστών $M2, M3$ και $M4$ που περιγράφηκαν στον αλγόριθμο 10. Εκτελέσεις της διαδικασίας με χαμηλότερες τιμές M υπέφεραν από λίγες προτάσεις άρθρων από τις εν' λόγω συστάδες, κάτι που με τη σειρά του

οδηγούσε σε χαμηλή απόδοση καθώς και μεγαλύτερη διάρκεια στην διαδικασία προτάσεων άρθρων προς αξιολόγηση. Θεωρούμε επίσης ότι σε ένα τυχών online πείραμα με πραγματικούς χρήστες η απόδοση σε αυτές τις περιπτώσεις θα ήταν ακόμη χειρότερη αν συνυπολογίσουμε τον τυχών εκνευρισμό των χρηστών για την μεγάλη διάρκεια της διαδικασίας. Παρομοίως, τιμές $M > 10$ επίσης υπέφεραν από χαμηλή απόδοση. Αυτό εξηγείται από το γεγονός ότι όταν χρησιμοποιούμε πολλά άρθρα από τις εν' λόγω συστάδες, οι χρήστες μπορούν με δυσκολία να γυρίσουν πίσω και να αποφύγουν τις περαιτέρω προτάσεις από αυτές τις συστάδες. Έτσι για τα πειράματα που ακολουθούν, χρησιμοποιούμε την τιμή $M = 5$.

Για το δεύτερο πείραμα αξιολόγησης της αντιμετώπισης του προβλήματος νέου χρήστη, χρησιμοποιήσαμε τις εξής στρατηγικές προτάσεων άρθρων νέων προς το χρήστη, οι οποίες και περιγράφηκαν στην ενότητα 3.9.1: εντροπίας, τυχαία, δημοφιλίας, ζυγισμένη, προσωποποιημένη στοιχείο προς στοιχείο, καθώς και την προτεινόμενη στρατηγική η οποία αξιοποιεί την πληροφορία συσταδοποίησης. Όταν η διαδικασία συλλογής αξιολογήσεων χρηστών ολοκληρωνόταν για κάθε στρατηγική ($R_{min} = 20$ συλλεγμένες αξιολογήσεις άρθρων), μετρούσαμε το πλήθος των άρθρων που ο χρήστης χρειάστηκε να "δει". Φυσικά, όσο λιγότερα άρθρα χρειαζόταν να δει ο χρήστης, τόσο το καλύτερο δεδομένου ότι γλιτώνουμε χρόνο και προσπάθεια από τη μεριά του. Επίσης, για την personalized item by item στρατηγική, χρησιμοποιήσαμε την προσέγγιση δημοφιλίας για την παρουσίαση των αρχικών άρθρων έως ότου μία αξιολόγηση ληφθεί από τον χρήστη. Η προσέγγιση αυτή είναι παρόμοια και με τη δικιά μας: και εμείς αξιοποιούμε την μέθοδο δημοφιλίας για την παρουσίαση των αρχικών άρθρων έως ότου μία αξιολόγηση ληφθεί από τον χρήστη, όμως φυσικά στη συνέχεια αλλάζουμε την στρατηγική μας και αξιοποιούμε την πληροφορία συσταδοποίησης.

Από τα αποτελέσματα που φαίνονται στο σχήμα 34 μπορούμε να παρατηρήσουμε ότι η προτεινόμενη μεθοδολογία που βασίζεται στην συσταδοποίηση ξεπερνάει όλες τις άλλες, μιας και οι τιμές CI είναι σαφώς χαμηλότερες για κάθε περίπτωση. Βλέπουμε επίσης ότι η τυχαία στρατηγική απαιτούσε να παρουσιαστούν κατά μέσο όρο 135 άρθρα προκειμένου να ανακτηθούν 20 αξιολογήσεις. Το προηγούμενο σε γενικές γραμμές είναι αναμενόμενο δεδομένης της τυχαίας φύσης της εν' λόγω στρατηγικής, αλλά και της μη κανονικότητας των αξιολογήσεων που μπορεί να δίνουν οι χρήστες: κάθε χρήστης ενδιαφέρετε για ένα συγκεκριμένο πεδίο και όχι για όλες τις θεματολογίες που καλύπτονται από τον τεράστιο όγκο άρθρων νέων του συστήματος. Για την περίπτωση της στρατηγικής που βασίζεται στην εντροπία, το πλήθος άρθρων που χρειάστηκε να παρουσιαστούν ήταν κατά μέσο όρων 115, ενώ για την στρατηγική δημοφιλίας ήταν 70. Τα αποτελέσματα για την στρατηγική εντροπίας, παρότι εντυπωσιακά αρνητικά, έχουν μία πιθανή εξήγηση: αυτή η στρατηγική προωθεί λιγότερο δημοφιλή άρθρα, όμως, υπάρχει ευθεία συσχέτιση μεταξύ των δημοφιλών άρθρων και της πιθανότητας ένας νέος χρήστης να ενδιαφέρεται για αυτά. Ως αποτέλεσμα, με το να αγνοεί τα δημοφιλή άρθρα, η στρατηγική αυτή οδηγείται σε χαμηλή απόδοση. Η προσωποποιημένη στοιχείο προς στοιχείο στρατηγική, παρότι πολλά υποσχόμενη με μέσο όρο 41 προτάσεις, ήταν επίσης χειρότερη της προτεινόμενης μεθοδολογίας που αξιοποιεί την συσταδοποίηση όπου χρειάστηκε κατά μέσο όρο 37.5 προτάσεις για να πάρει 20 αξιολογήσεις.

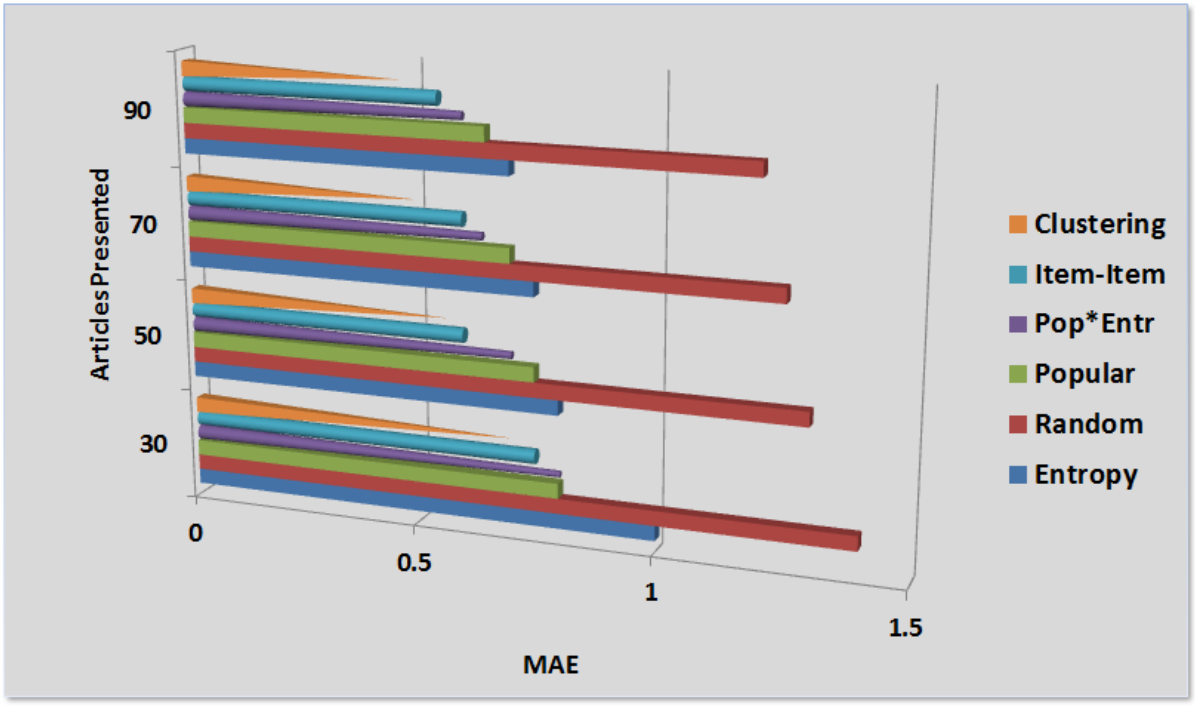
Για το τρίτο μας πείραμα, προσπαθήσαμε να εκτιμήσουμε την ακρίβεια προβλέψεων της προ-



Σχήμα 34: Σύγκριση μεθοδολογιών πρότασης άρθρων σε σχέση με την τεχνική μας που βασίζεται στη συσταδοποίηση

τεινόμενης προσέγγισης σε σχέση με τις προαναφερθείσες στρατηγικές. Ξανά, χρησιμοποιήσαμε την μετρική MAE. Για τον προσδιορισμό των τιμών MAE για κάθε στρατηγική, παρουσιάσαμε για αξιολόγηση στον χρήστη ένα σύνολο από 30, 50, 70 και 90 άρθρα σε 4 ακολουθιακές εκτελέσεις καθεμίας εκ' των 6 στρατηγικών. Τα αποτελέσματα που φαίνονται στο σχήμα 35 δείχνουν τις MAE μεταβολές των στρατηγικών σαν συνάρτηση των άρθρων που παρουσιάστηκαν για αξιολόγηση.

Από το σχήμα 35 μπορούμε να παρατηρήσουμε την βελτίωση των MAE τιμών καθώς το πλήθος των άρθρων αυξάνει, κάτι που έχει σημαντικό αντίκτυπο ειδικά στην προτεινόμενη μεθοδολογία με αξιοποίηση συσταδοποίησης: καθώς όλο και περισσότερα άρθρα αξιολογούνται από τον χρήστη, η προσέγγισή μας μπορεί να επιλέγει καλύτερα υποψήφια άρθρα για αξιολόγηση από τον χρήστη χρησιμοποιώντας τα δεδομένα συσταδοποίησης που υπάρχουν στη ΒΔ. Πράγματι, η προτεινόμενη μεθοδολογία δίνει τα χαμηλότερα MAE σκορ για κάθε πλήθος άρθρων που αφορούν τις εκτελέσεις του πειράματος. Ένα ακόμη χρήσιμο αποτέλεσμα που μπορούμε να εξάγουμε από το σχήμα 35 είναι ότι η τυχαία στρατηγική έχει την χειρότερη ακρίβεια πρόβλεψης, επικυρώνοντας ουσιαστικά τις παρατηρήσεις μας στα προηγούμενα πειράματα. Μπορούμε τέλος να δούμε ότι η προσωποποιημένη στοιχείο προς στοιχείοι στρατηγική είναι ξανά σχετικά κοντά στην προτεινόμενη μας μεθοδολογία.



Σχήμα 35: Σύγκριση μεθοδολογιών πρότασης άρθρων σε σχέση με την τεχνική μας που βασίζεται στη συσταδοποίηση

7.4 Προσωποποίηση στο χρήστη / παραγωγή προτάσεων

Προκειμένου να αξιολογήσουμε την απόδοση και ακρίβεια του συστήματός μας όσον αφορά στις παραγόμενες προτάσεις νέων που αξιοποιούν την προσωποποίηση στο χρήστη, εκτελέσαμε ορισμένα πειράματα των οποίων η λογική ακολουθεί την εξής σειρά:

- αξιολόγηση της τρέχουσας αποτελεσματικότητας των παραγόμενων προτάσεων
- εφαρμογή νέας τεχνικής
- επαν-αξιολόγηση και σύγκριση αποτελεσμάτων

Από την παραπάνω διαδικασία προέκυψε ένα σύνολο από δεδομένα τα οποία έδειξαν την συνολική τάση όσον αφορά στα συγκεκριμένα κριτήρια/μετρικές αξιολόγησης που χρησιμοποιήθηκαν.

Στα επόμενα λοιπόν αξιολογείται το κομμάτι προσωποποίησης στο χρήστη, και πιο συγκεκριμένα τα βήματα που περιγράφονται στον αλγόριθμο 9.

7.4.1 Σύνολο δεδομένων

Το σύνολο δεδομένων για τα παρακάτω πειράματα αποτελείται από καταγραφές που περιλαμβάνουν τα μοτίβα πλοήγησης 30 χρηστών του συστήματος. Οι χρήστες αυτοί χρησιμοποιούσαν το σύστημα καθώς οι μεθοδολογίες εφαρμόζονταν μία προς μία, χωρίς να έχουν κάποια γνώση για

τις αλλαγές στο σύστημα. Οι επιλογές των χρηστών, καθώς και οι προτάσεις του συστήματος κατεγράφησαν σε όλη την διαδικασία.

Δεδομένης της φύσης των άρθρων νέων, τα οποία θα πρέπει εν' γένει να είναι "νέα" ώστε να έχουν αξία για τον χρήστη, το σύστημά μας αγνόησε εκείνα με ημερομηνία δημοσίευσης πέραν των 3 μηνών. Ως αποτέλεσμα αυτού, παρότι τα συνολικά δεικτοδοτημένα άρθρα στη ΒΔ του συστήματος ξεπερνούσαν τα 750.000, μόνο 3.000 από αυτά χρησιμοποιήθηκαν για την εν' λόγω πειραματική διαδικασία. Όπως και πριν, τα άρθρα αυτά ανήκουν ομοιόμορφα στις 8 βασικές κατηγορίες του συστήματός μας.

Ως μετρικές αξιολόγησης, χρησιμοποιήθηκαν το MAE (ενότητα 3.7.1.6.2) και το F-measure (ενότητα 3.2.2.3).

7.4.2 Αποτελέσματα και ανάλυση

Για το πρώτο μας πείραμα, καταγράψαμε τις MAE τιμές της διαφοράς μεταξύ των πραγματικών επιλογών των χρηστών, δηλαδή των άρθρων που επιλέχθηκαν για ανάγνωση, και των προτάσεων άρθρων που έδωσε το σύστημα για τις διάφορες φάσεις λειτουργίας:

1. χωρίς καθόλου ευρετικά, απλά προτείνοντας τα πιο πρόσφατα άρθρα που προστέθηκαν στη ΒΔ

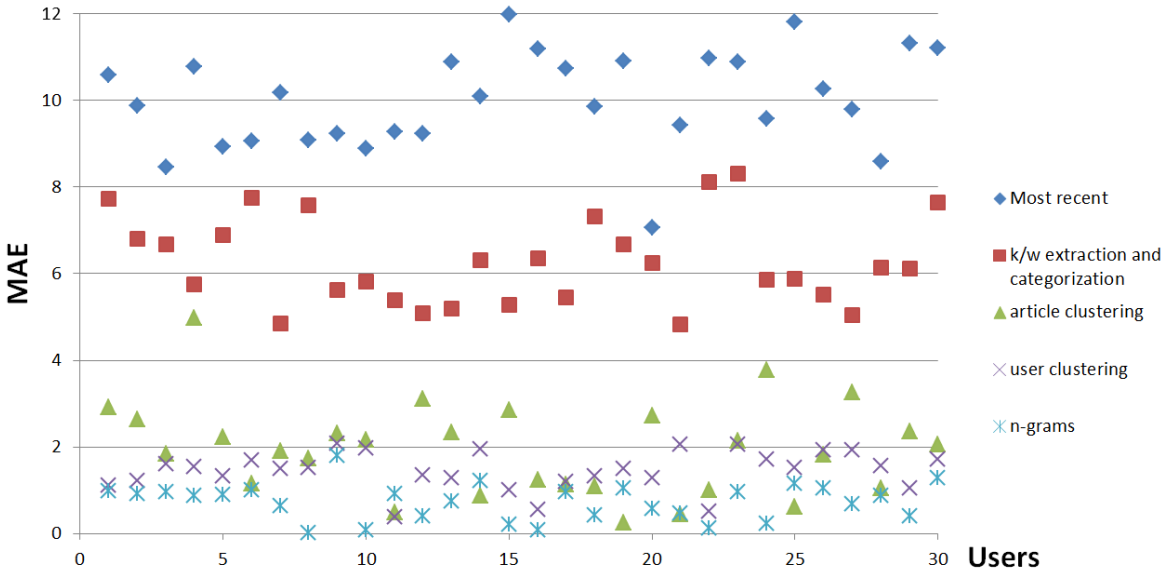
2. όταν η εξαγωγή keywords και η κατηγοριοποίηση εφαρμοζόταν για την παραγωγή προτάσεων, όπως περιγράφεται στη σχέση 32 (περίπτωση λειτουργίας μηχανισμού μεταπτυχιακής εργασίας)

3. όταν εκτός από τα ευρετικά του 2), η πληροφορία συσταδοποίησης άρθρων νέων επίσης αξιοποιούνταν για την παραγωγή προτάσεων

4. όταν εκτός από τα προηγούμενα ευρετικά, η πληροφορία συσταδοποίησης χρηστών επίσης αξιοποιούνταν για την παραγωγή προτάσεων

5. όταν εκτός από τα προηγούμενα ευρετικά, η πληροφορία για n-grams επίσης αξιοποιούνταν για την παραγωγή προτάσεων (σχέση 40)

Στο σχήμα 36 απεικονίζονται τα αποτελέσματα των τιμών MAE που λάβαμε από την πειραματική μας διαδικασία. Από αυτά, μπορούμε να παρατηρήσουμε για γενική τάση για σημαντική μείωση των τιμών MAE τόσο ξεχωριστά με την εφαρμογή κάθε μίας από τις προαναφερθείσες τεχνικές, όσο και συνολικά όταν αξιοποιούνται όλες. Πιο συγκεκριμένα, η μέση τιμή MAE μειώθηκε από 10.01 όταν κανένα ευρετικό/τεχνική δεν εφαρμοζόταν (επιλογή από τα πιο πρόσφατα άρθρα), σε 6.28 όταν η εξαγωγή λέξεων κλειδιών μαζί με την κατηγοριοποίηση εφαρμόστηκαν. Από φυσική άποψη αυτό σημαίνει ότι στην περίπτωση 2), υπήρχαν κατά μέσο όρο 6.28 λάθος προτάσεις προς τον χρήστη. Η αξιοποίηση της πληροφορίας συσταδοποίησης άρθρων μείωσε επίσης σημαντικά την μέση MAE τιμή στο 1.95. Η βελτίωση αυτή συμβαδίζει με τις παρατηρήσεις μας σε προηγούμενα



Σχήμα 36: Τιμές MAE των προτάσεων με χρήση των διαφόρων ευρετικών

πειράματα όσον αφορά στην επίπτωση του αλγορίθμου W-kmeans στην διαδικασία πρωσοποποίησης/προτάσεων προς τον χρήστη. Ακολούθως, η αξιοποίηση και της πληροφορίας συσταδοποίησης χρηστών μείωσε ακόμη περισσότερο την μέση τιμή MAE στο 1.44. Τέλος, η αξιοποίηση της πληροφορίας των εξαγόμενων n-grams οδήγησε την μέση MAE τιμή στο 0.73, υπονοώντας ότι κατά μέσο όρο υπήρχε λιγότερο από 1 λάθος πρόταση προς τον χρήστη ανά συνεδρία.

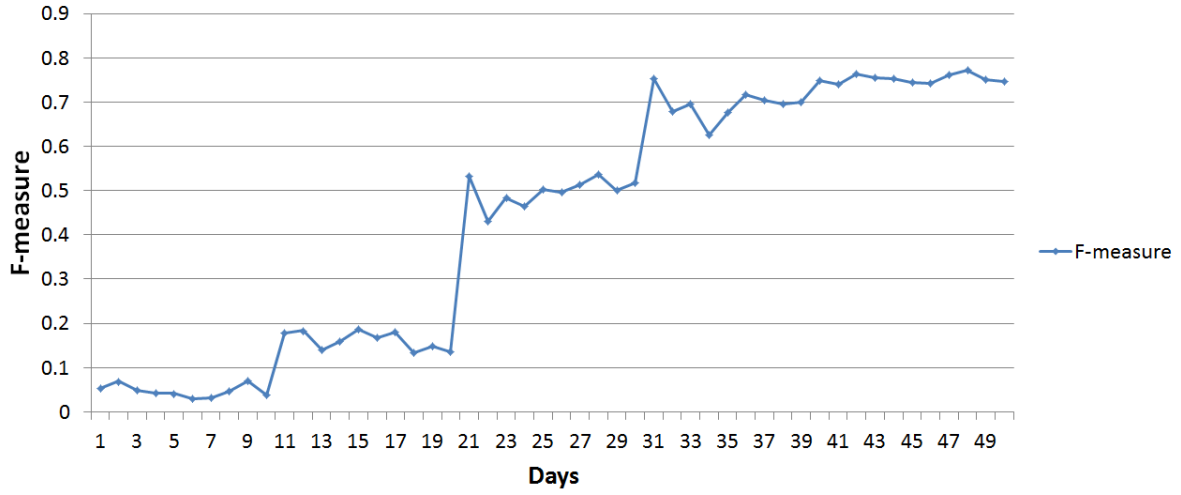
Για το επόμενο πείραμα, προσπαθήσαμε να αξιολογήσουμε την συνολική βελτίωση της αποδοτικότητας του συστήματος όταν κάθε μία από τις προαναφερθείσες μεθοδολογίες εφαρμοζόταν. Ως μετρική αξιολόγησης χρησιμοποιήσαμε το F-measure. Αξιοποιώντας τα ίδια δεδομένα πλοήγησης με πριν, εξάγαμε τις τιμές F-measure για όλους τους χρήστες για μία περίοδο χρήσης του συστήματος 50 ημερών. Στην περίπτωση αυτή, κάθε 10 ημέρες μία νέα μεθοδολογία εφαρμοζόταν με τις προτάσεις να είναι όπως φαίνονται στον πίνακα 12

Ημέρα	**Προσέγγιση παραγωγής προτάσεων**
1-10	μόνο τα πιο πρόσφατα άρθρα
11-20	εξαγωγή keywords και κατηγοριοποίηση
21-30	επίσης συσταδοποίηση άρθρων
31-40	επίσης συσταδοποίηση χρηστών
41-50	επίσης εξαγωγή n-grams

Πίνακας 12: Αλλάζοντας την μεθοδολογία παραγωγής προτάσεων με βάση το χρόνο

Τα παραγόμενα αποτελέσματα F-measure των προτάσεων του συστήματος, εξάγοντας τον μέσο όρο ως προς όλους τους χρήστες, φαίνονται στο σχήμα 37.

Από τα αποτελέσματα του σχήματος 37 μπορούμε να παρατηρήσουμε ότι οι προτάσεις που κάνουν χρήση όλων των προτεινόμενων ευρετικών, υπερτερούν σημαντικά των περιπτώσεων που



Σχήμα 37: Μέσες τιμές F-measure προτάσεων προς τον χρήστη με χρήση των διαφόρων ευρετικών

καθένα από αυτά εφαρμόζεται αυτοτελώς. Πιο συγκεκριμένα, ενώ η μέση τιμή F-measure ξεκινά από περίπου 0.05 για την περίπτωση που μόνο πρόσφατα άρθρα προτείνονται (τιμή πολύ χαμηλή για ένα σύστημα προτάσεων), φτάνει το 0.8 όταν και η αξιοποίηση της πληροφορίας των n-grams εφαρμόζεται. Ξανά, η πληροφορία συσταδοποίησης άρθρων έδωσε μία σημαντική ώθηση στην απόδοση του συστήματος: από 0.15 όταν αξιοποιούνταν η εξαγωγή keywords και η κατηγοριοποίηση, στο 0.47.

Μία ακόμη παρατήρηση είναι ότι η βελτίωση γενικά αυξάνεται μετά από ορισμένες ημέρες χρήσης του συστήματος. Αυτό έχει δύο εξηγήσεις που μπορούμε να δώσουμε: το σύστημα έχει περισσότερα δεδομένα σχετικά με τις επιλογές/προτιμήσεις του χρήστη, και επίσης, το σύστημα έχει περισσότερο χρόνο να παράγει πιο συνεκτικές και γενικά καλύτερες συστάδες. Αρχικά οι τιμές F-measure είναι χαμηλές δεδομένου ότι το σύστημα προτάσεων δεν έχει ακόμη καθορίσει το προφίλ χρήστη σε αποτελεσματικό βαθμό.

Τα παραπάνω αποτελέσματα έχουν επίσης άμεση συνέπεια και από φυσική άποψη σε σχέση με την ποιότητα του παραγόμενου περιεχομένου προτάσεων: τα άρθρα νέων είναι γενικά ενδιαφέροντα για τους χρήστες, ταιριάζοντας στο προφίλ τους και τα περισσότερα από αυτά επιλέγονται για ανάγνωση (έστω και σε μεταγενέστερο στάδιο). Μία ακόμη σημαντική παρατήρηση είναι ότι οι προτάσεις γενικά σταθεροποιούνται γρήγορα στη μέση τιμή τους μέσα στο χρονικό των 10 ημερών, χωρίς πολύ διακύμανση. Αυτό συνεπάγεται και το γεγονός ότι ο μηχανισμός παραγωγής προτάσεων συγκλίνει σχετικά γρήγορα στα ενδιαφέροντα των χρηστών, κάτι που παρατηρήθηκε και σε προηγούμενο πείραμα.

ΚΕΦΑΛΑΙΟ 8

ΣΥΜΠΕΡΑΣΜΑΤΑ

> It is no measure of health to be well adjusted to a profoundly sick society.
>
> *Jiddu Krishnamurti, Indian Philosopher, 1895*

Στο παρόν κεφάλαιο παρουσιάζονται τα συμπεράσματα που προέκυψαν κατά τη διάρκεια εκπόνησης της διδακτορικής διατριβής. Δεδομένου ότι η εργασία που επιτελέστηκε αφορούσε μία πληθώρα συστημάτων που σε σχέση τον μηχανισμό προτάσεων άρθρων νέων, τα αποτελέσματα χωρίζονται σε ενότητες που ανταποκρίνονται στις αναπτυγμένες μεθοδολογίες και στα συμπεράσματα της αντίστοιχης πειραματικής διαδικασίας.

8.1 Το πρόβλημα...

Το υπερμέγεθες διαδίκτυο με την υπέρογκη πληροφορία που διακινείται σε αυτό κάνει την καθημερινή του χρήση δύσκολη για τον χρήστη. Στην εποχή μας και με τα μέσα που διαθέτει ακόμα και ο απλός χρήστης, η προσθήκη περιεχομένου στο διαδίκτυο από τον καθένα, είναι μια διαδικασία το ίδιο εύκολη με την απλή περιαγωγή στο χώρο του παγκόσμιου ιστού. Το πρόβλημα που δημιουργεί αυτή η ανεξέλεγκτη κατάσταση είναι ότι ακόμα οι πιο έμπειροι χρήστες καταναλώνουν πολύ χρόνο στην προσπάθεια εύρεσης πληροφορίας και συγκεκριμένα πηγών ενημέρωσης για τα θέματα που τους ενδιαφέρουν αφού κατακλύζονται από την πληροφορία.

Εστιάζοντας στο πρόβλημα του "κατακλυσμού της πληροφορίας", επικεντρωνόμαστε σε αυτή που διακινείται στο διαδίκτυο και αφορά νέα και γεγονότα. Αυτό που θέλουμε ουσιαστικά, είναι να δημιουργήσουμε μεθοδολογίες με τις οποίες ένα σύστημα προτάσεων θα είναι σε θέση να παρουσιάζει ειδήσεις που δημοσιεύονται στο διαδίκτυο, με τρόπο απλό και έχοντας στο νου μας τον παράγοντα άνθρωπο. Για να το επιτύχουμε αυτό πρέπει να παρέχουμε στο χρήστη πληροφορία η οποία θα μπορεί να προσαρμόζεται σε αυτόν και να του παρέχει ποιοτικό και πλήρες περιεχόμενο για τις εξελίξεις που τον ενδιαφέρουν. Δεν στοχεύουμε στην ανάπτυξη ενός ακόμα portal νέων αφού κάτι τέτοιο δεν θα αντιμετώπιζε το πρόβλημα. Στοχεύουμε αντίθετα στην εξεύρεση τεχνικών που θα επιτελούν αποτελεσματικό φιλτράρισμα της πληροφορίας.

8.2 ...και η αντιμετώπισή του

Στα πλαίσια της μεταπτυχιακής μου εργασίας είχε δημιουργηθεί η πλατφόρμα PeRSSonal, ένα σύστημα αποδελτίωσης άρθρων νέων του παγκοσμίου ιστού. Σκοπός της διδακτορικής διατριβής ήταν η προσθήκη διεργασιών πυρήνα που θα συνέβαλαν στον μετασχηματισμό του σε ένα υβριδικό σύστημα προτάσεων άρθρων νέων. Το σύστημα αυτό δημιουργήθηκε ώστε να περνάει την πληροφορία μέσα από διάφορα στάδια επεξεργασίας. Αρχικά γίνεται το διαπέρασμα των ιστοσελίδων γνωστών news portals με χρήση των RSS Feeds που διαθέτουν και αποθηκεύεται ο html κώδικάς τους μέσω της διαδικασίας του crawling. Στη συνέχεια, από τον html κώδικα αναγνωρίζεται η χρήσιμη πληροφορία που αφορά το συγκεκριμένο άρθρο. Ακολουθώντας το φιλτράρισμα του κειμένου, ο μηχανισμός προχωρά με τη διαδικασία της προεπεξεργασίας κειμένου και εξαγωγής των keywords καθώς και των n-grams. Πρόκειται για τη θεμελιώδη διεργασία σχεδόν όλων των μηχανισμών ανάκτησης πληροφορίας και επομένως επιθυμούμε τα καλύτερα δυνατά αποτελέσματα. Ο μηχανισμός προεπεξεργασίας που κατασκευάστηκε δοκιμάστηκε διεξοδικά ώστε να παράγει σωστές εξόδους και να κρατά τον πλεονασμό σε χαμηλά επίπεδα. Περιλαμβάνει τις διαδικασίες αφαίρεσης σημείων στίξης και αριθμών, ορθογραφικού ελέγχου και διόρθωσης λαθών, εύρεσης των ουσιαστικών του κειμένου, αφαίρεσης των λέξεων που ανήκουν στη λίστα των stopwords, το stemming των λέξεων και φυσικά την αντιστοίχηση των keywords και n-grams που προκύπτουν με τις προτάσεις όπου αυτά εμφανίζονται. Η όλη διαδικασία κάνει εκτεταμένη χρήση regular expressions της βιβλιοθήκης

boost-regex της C++ και είναι υλοποιημένη ώστε να αποφεύγονται περιττοί έλεγχοι ή επανάληψης με στόχο τη βελτίωση της απόδοσης.

Ο μηχανισμός που αναπτύχθηκε συνεχίζει με την κατηγοριοποίηση και περίληψη των άρθρων κάνοντας χρήση πολλών παραμέτρων οι οποίες μπορούν να προσαρμόζονται από το σύστημα δυναμικά ώστε να ανταποκρίνονται στο κείμενο.

Ακολουθεί το υποσύστημα συσταδοποίησης, κεντρικό τμήμα της διδακτορικής διατριβής. Το αφορά τόσο την συσταδοποίηση άρθρων νέων όσο και χρηστών με χρήση του καινοτόμου αλγορίθμου W-kmeans. Η πληροφορία συσταδοποίησης αποθηκεύεται στην ΒΔ για περαιτέρω αξιοποίηση από το υποσύστημα προσωποποίησης που ακολουθεί. Το υποσύστημα προσωποποίησης είναι και αυτό που επιλέγει ποια άρθρα νέων θα πρέπει να παρουσιαστούν στο χρήστη βάσει μίας πληθώρας παραγόντων και ευρετικών. Θεωρούμε το συγκεκριμένο σύστημα ως το τελικό στάδιο της ροής πληροφορίας πριν αυτή εν' τέλει παρουσιαστή στον τελικό χρήστη (με όποιον τρόπο και αν γίνει αυτό).

Τέλος, προτείναμε μία μεθοδολογία αντιμετώπισης του προβλήματος νέου χρήστη για τα συστήματα προτάσεων, κάτι που πιστεύουμε ότι ολοκληρώνει ουσιαστικά τις διεργασίες πυρήνα στις οποίες επαφίεται ένα τέτοιο σύστημα.

Στην συνέχεια παρατίθενται μία πιο αναλυτική παρουσίαση των συμπερασμάτων στα οποία καταλήγει η παρούσα διατριβή ανά τμήμα με το οποίο ασχοληθήκαμε.

8.3 Αξιοποίηση n-grams

Η εξαγωγή και αξιοποίηση των n-grams από τα άρθρα νέων του διαδικτύου, στα πλαίσια του αλγορίθμου συσταδοποίησης W-kmeans που προτείνεται στην διδακτορική διατριβή, όπως περιγράφεται στην ενότητα 5.1.1 και αξιολογείται στην ενότητα 7.1.1, μας οδήγησε σε ορισμένα ορισμένα αξιόλογα συμπεράσματα.

Η πειραματική διαδικασία που τρέξαμε, προκειμένου να βρούμε την καταλληλότερη τιμή για την παράμετρο n στον τομέα των άρθρων νέων, έδειξε ότι η τιμή $n = 3$, δηλαδή η 3-grams, βελτιώνει σημαντικά την απόδοση της συσταδοποίησης. Το παραπάνω αποτελεί επίσης επιβεβαίωση της υπάρχουσας βιβλιογραφίας (δεν υπήρχε όμως κάτι σχετικό για τα άρθρα νέων).

Επιπλέον, είδαμε ότι δεν είναι αρκετό απλά να συμπεριλάβουμε την πληροφορία των n-grams τυχαία στην εξίσωση ζύγισης. Αντιθέτως, παρουσιάσαμε έναν τρόπο αξιοποίησης τόσο της κλασικής BOW αναπαράστασης των keywords, όσο και της αντίστοιχης αναπαράστασης των n-grams με έναν δυναμικό και ζυγισμένο τρόπο μέσω των παραμέτρων A και B (της συνάρτησης 33). Συμπεράναμε ότι τα καλύτερα αποτελέσματα, όσον αφορά στη συσταδοποίηση άρθρων νέων, παρουσιάζονται όταν τα n-grams ζυγίζονται κατά 30% και τα keywords κατά 70% στο τελικό σύστημα ζύγισης. Θεωρούμε δε, ότι το παραπάνω αποτέλεσμα είναι εξειδικευμένο για τον τομέα τον άρθρων νέων και πιθανά σε άλλο τομέα κειμένων (π.χ. δημοσιεύσεις σε επιστημονικά περιοδικά) να είναι διαφορετικό.

Παράπλευρη συνέπεια της παραπάνω πειραματικής διαδικασίας ήταν η ενίσχυση του συμπεράσματος ότι ο αλγόριθμος W-kmeans ξεπερνά σε αποδοτικότητα τον κλασικό k-means αλγόριθμο,

ακόμα και για την περίπτωση που η εξαγωγή n-grams είναι σε ισχύ.

Από τα παραπάνω είναι προφανές ότι η αξιοποίηση της πληροφορίας των εξαγόμενων n-grams έχει σημαντικά οφέλη σε ένα σύστημα συνεργατικού φιλτραρίσματος που χρησιμοποιεί τεχνικές συσταδοποίησης.

8.4 Συσταδοποίηση

8.4.1 Αξιολόγηση αλγορίθμων βιβλιογραφίας

Όσον αφορά στην συσταδοποίηση άρθρων νέων, ένα από τα πρώτα στάδια της διδακτορικής διατριβής ήταν η αξιολόγηση ορισμένων (των βασικότερων) αλγορίθμων συσταδοποίησης για τον τομέα των άρθρων νέων. Τα αποτελέσματα σε αυτό το κομμάτι που παρουσιάστηκαν στην ενότητα 7.2.1.1 ήταν κομβική σημασίας για την εξέλιξη της διατριβής, αφού μας έγινε σαφές ποιοι αλγόριθμοι και ποιες μετρικές απόστασης άξιζαν της προσοχή μας για τη συνέχεια.

Πιο συγκεκριμένα, είδαμε ότι η χρήση της ομοιότητας συνημιτόνου σε συνδυασμό με τον αλγόριθμο k-means έδιναν σταθερά ικανοποιητικά αποτελέσματα για την πειραματική μας βάση. Επίσης, παρατηρήσαμε ότι οι ιεραρχικοί αλγόριθμοι είχαν γενικά χειρότερα σκορ CI σε σχέση με τους διαμερισματικούς, κάτι που πιθανότητα είχε να κάνει με το γενικά μεγάλο πλήθος μοναδιαίων συστάδων που παρήγαγαν.

Ένα ακόμη σημαντικό συμπέρασμα ήταν ότι η χρήση τεχνικών εύρεσης ρίζας λέξεων (stemming) και εξαγωγής των ουσιαστικών του κειμένου μπορεί να βελτιώσει σημαντικά τα αποτελέσματα συσταδοποίηση, περίπου κατά έναν παράγοντα 5-15% ανάλογα τον εφαρμοζόμενο αλγόριθμο. Δεν εντοπίσαμε κάποια αντίστοιχη έρευνα που να ποσοτικοποιεί κατά παρόμοιο τρόπο ένα τέτοιο γενικά διαισθητικό αποτέλεσμα.

8.4.2 W-kmeans για συσταδοποίηση άρθρων νέων

Έχοντας το παραπάνω αποτέλεσμα υπόψιν μας, προχωρήσαμε στον σχεδιασμό και την υλοποίηση του αλγορίθμου W-kmeans, με βασικό στόχο πάντα, την ενίσχυση των αποτελεσμάτων του τυπικού k-means αλγορίθμου. Το αποτέλεσμα ήταν μία καινοτόμα μεθοδολογία συσταδοποίησης η οποία αξιοποιεί την εξωτερική γνώση του WordNet με διττό τρόπο. Πρώτον, ενισχύοντας την ίδια τη διαδικασία συσταδοποίησης με χρήση των υπερωνύμων του κειμένου - έμμεση πληροφορία γενικά μη εκμεταλλεύσιμη από άλλους αλγορίθμους. Και δεύτερον, παράγοντας χρήσιμες ετικέτες για κάθε παραγόμενη συστάδα. Μια διαδικασία που γρήγορα μπορεί να δώσει μία εύληπτη αποτύπωση του νοηματικού περιεχομένου της κάθε συστάδας άρθρων νέων.

Είδαμε λοιπόν μία μεγάλη βελτίωση στις CI τιμές του αποτελέσματος συσταδοποίησης (περίπου δεκαπλάσια), κάτι που σημαίνει πιο συνεκτικές συστάδες και καλύτερα χωρισμένες μεταξύ τους. Επιπλέον, υπολογίσαμε ότι οι ετικέτες που προκύπτουν από την προτεινόμενη μεθοδολογία ταιριάζουν με υψηλή ακρίβεια σε εκείνες που θα έδινε ένας πραγματικός χρήστης του συστήματος.

8.4.3 Συσταδοποίηση χρηστών συστήματος

Πέρα από τη συσταδοποίηση άρθρων νέων, όπως έχει αναφερθεί αρκετές φορές μέχρι τώρα, το σύστημα έχει τη δυνατότητα να παράγει συστάδες από συνεδρίες χρηστών, και μάλιστα, αξιοποιώντας τον ίδιο αλγοριθμικό πυρήνα, W-kmeans. Η εν' λόγω δυνατότητα παρουσιάστηκε αλγοριθμικά στην ενότητα 5.3.2 και αξιολογήθηκε στην ενότητα 7.2.2.

Τα πειραματικά αποτελέσματα έδειξαν ότι η προτεινόμενη μέθοδος συσταδοποίησης συνεδριών χρηστών με χρήση του W-kmeans μπορεί να παράγει σημαντικά καλύτερες συστάδες, πάλι περίπου 10 φορές καλύτερες σε σχέση με τα CI σκορ, σε σχέση με τον τυπικό k-means αλγόριθμο. Επιπλέον εντοπίσαμε μία μέση βελτίωση της τάξης του 10% όσον αφορά τις τιμές F-measure που έχουν να κάνουν με τις προτάσεις νέων που παράγει το σύστημά μας.

Τα παραπάνω αποτελούν σαφή απόδειξη για την σημαντικότητα της αξιοποίησης της συσταδοποίησης χρηστών σε ένα συνεργατικό σύστημα προτάσεων όπως το προτεινόμενο.

8.4.4 Πρόβλημα νέου χρήστη

Ένα ακόμη θέμα με το οποίο καταπιαστήκαμε στην διδακτορική διατριβή ήταν το πρόβλημα του νέου χρήστη - μία κατάσταση την οποία όλα τα συστήματα προτάσεων πρέπει να αντιμετωπίσουν. Παρουσιάσαμε λοιπόν μία προσωποποιημένη στρατηγική για την αντιμετώπιση αυτού του προβλήματος, μέσω της διαδικασίας ερωτήσεων προς τον χρήστη για αξιολόγηση άρθρων νέων. Υλοποιήσαμε και αξιολογήσαμε μία αλγοριθμική προσέγγιση (αλγόριθμος 10) η οποία χρησιμοποιεί την πληροφορία συσταδοποίησης χρηστών και άρθρων νέων με έναν ευρετικό τρόπο, προκειμένου να αποφασίσει τα επόμενα προς παρουσίαση για αξιολόγηση από τον χρήστη αντικείμενα κατά την διαδικασία εγγραφής του. Η προσέγγισή μας έχει ομοιότητα με την προσωποποιημένη στρατηγική ένα προς ένα, αλλά τα αποτελέσματα έδειξαν ότι αποδίδει καλύτερα συγκριτικά με τις πιο κοινά προτεινόμενες μεθοδολογίες αντιμετώπισης του προβλήματος.

Η πειραματική μας διαδικασία, βασιζόμενοι σε offline αξιολογήσεις χρηστών που προυπήρχαν στη ΒΔ, έδειξε ότι χρησιμοποιώντας 5 άρθρα από κάθε συστάδα (είτε άρθρων είτε χρηστών) στον αλγόριθμο 10, παίρνουμε τις χαμηλότερες τιμές MAE. Αξιοποιώντας αυτό το αποτέλεσμα, βρήκαμε ότι η προσέγγισή μας απαιτεί κατά μέσο όρο την παρουσίαση 32,3 άρθρων προκειμένου να επιτευχθούν 20 αξιολογήσεις χρήστη. Ένα νούμερο που ήταν σημαντικά μικρότερο σε σχέση με οποιαδήποτε άλλη τεχνική που χρησιμοποιείται από παρόμοια συστήματα.

Τέλος, συγκρίναμε τις MAE τιμές της προτεινόμενης τεχνικής με κάθε μία από τις entropy, random, popularity, balanced και personalized item by item στρατηγικές. Κάθε πείραμα, εκτελεσμένο πάνω σε διάφορα πλήθη άρθρων, έδειξε επίσης ότι η μεθοδολογία μας ξεπερνά σε απόδοση κάθε μία από τις προαναφερόμενες στρατηγικές.

8.5 Προσωποποίηση στο χρήστη και σύστημα προτάσεων

Στην παρούσα διατριβή παρουσιάσαμε τις βασικές λειτουργίες πυρήνα ενός πλήρους σύστηματος προτάσεων άρθρων νέων που πηγάζουν από το διαδίκτυο. Αποτυπώθηκε επίσης μία γενική

εικόνα της ροής πληροφορίας για ένα τέτοιου είδους σύστημα, κάτι ανεξάρτητο από τα συγκεκριμένα υποσυστήματα που υλοποιήθηκαν.

Βασικός στόχος του αναπτυσσόμενου συστήματος προτάσεων είναι προφανώς η ανάκτηση προσωποποιημένου περιεχομένου που ταιριάζει στις προτιμήσεις του κάθε χρήστη. Όπως έχει αναφερθεί, η προσέγγιση προτάσεων που ακολουθεί το σύστημά μας είναι υβριδική, έχοντας τόσο χαρακτηριστικά περιεχομένου, όσο και συνεργατικού φιλτραρίσματος. Το παραπάνω αποτελεί σημαντικό προτέρημα του συστήματος που περιγράφηκε, μιας και μπορεί γρήγορα να προσαρμόζεται στα διαρκώς μεταβαλλόμενα ενδιαφέροντα των χρηστών.

Ύστερα λοιπόν από την υλοποίηση και αξιολόγηση του αλγορίθμου W-kmeans στο πλαίσιο της συσταδοποίησης τόσο άρθρων νέων όσο και χρηστών, εκτιμήσαμε τη δυνατότητα συνένωσης των παραπάνω μηχανισμών κάτω από το πρίσμα του συνολικού συστήματος προτάσεων. Πιο συγκεκριμένα, κάνοντας μία απλή ανάλυση των μοτίβων των χρηστών στη ΒΔ που περιγράφηκε στην ενότητα 5.3, ορίσαμε κάποια αλγοριθμικά βήματα προσωποποίησης και παραγωγής προτάσεων προς τον χρήστη. Πιο συγκεκριμένα, καταλήξαμε στον αλγόριθμο προσωποποίησης 9, ο οποίος, αξιοποιώντας την πληροφορία τόσο των ίδιων των άρθρων (στατιστικά στοιχεία των περιεχομένου τους), όσο και από τα υπόλοιπα συστήματα ανάκτησης πληροφορίας του μηχανισμού (περίληψη, κατηγοριοποίηση και συσταδοποίηση), αλλά και τα χρονικά μοτίβα του χρήστη όσον αφορά την ανάγνωση άρθρων, μπορεί και παράγει προτάσεις προσωποποιημένες για τον χρήστη. Οι προτάσεις αυτές εν' συνεχεία αξιολογήθηκαν ως προς την αποτελεσματικότητά τους.

Η πειραματική αξιολόγηση του συστήματος προτάσεων έδειξε μία βελτίωση της τάξης του 15% όσον αφορά την ποιότητα των παραγόμενων προτάσεων όταν η συσταδοποίηση εφαρμοζόταν (αλγόριθμος 9) σε σύγκριση με την μη εφαρμογή της. Παράλληλα, παρατηρήσαμε μία βελτίωση της τάξης του 0.1 στα F-measure σκορ του συστήματος προτάσεων. Το παραπάνω απέδειξε ότι με την προσθήκη keywords από τις συστάδες χρηστών στις θετικές λίστες των χρηστών, έχουμε μία σαφή βελτίωση στην απόδοση των παραγόμενων προτάσεων, κάτι που δεν είχε ερευνηθεί προηγουμένως στην βιβλιογραφία. Επιπλέον είδαμε ότι η αποδοτικότητα των προτάσεων αυξάνει συνεχώς, καθώς όλο και περισσότεροι χρήστες λαμβάνονται υπόψιν από το σύστημα. Στατιστικά δε, μετρήσαμε ότι οι παραγόμενες προτάσεις ταιριάζουν στις επιλογές χρήστη 7 στις 10 φορές.

Σε επόμενο πείραμα, και αξιοποιώντας την λογική της add-on προσθήκης κάθε μεθοδολογίας που προτείνεται στην διδακτορική διατριβή, μπορέσαμε να υπολογίσουμε την επιμέρους σε κάθε φάση, αλλά και συνολική απόδοση του συστήματος. Σε σχέση με τα MAE σκορ, είδαμε την μέση τιμή να πέφτει από 10.1 όταν τα αποτελέσματα παρουσιαζόταν χωρίς κανένα κριτήριο (τα τελευταία άρθρα), σε 0.73 όταν όλες οι προτεινόμενες μεθοδολογίες είχαν εφαρμοστεί. Σε παρόμοιο αποτέλεσμα καταλήξαμε και όταν για μία περίοδο 50 ημερών μετρήθηκε το F-measure, όπου την ίδια σειρά εφαρμογής των μεθοδολογιών από 0.05, καταλήξαμε στην τιμή 0.8 όταν όλες οι μεθοδολογίες είχαν εφαρμοστεί.

Θεωρούμε ότι τα παραπάνω αποτελέσματα δικαιολογούν πλήρως την αξιοποίηση ενός συνόλου μεθοδολογιών από τα συστήματα προτάσεων. Οι τεχνικές αυτές, δουλεύοντας με ευρετικό τρόπο μπορούν να δώσουν πολύ καλύτερα αποτελέσματα σε σύγκριση με την περίπτωση που καμία ή μόνο

κάποια(ες) από αυτές εφαρμόζεται. Κατά συνέπεια, οι συνεργατικές προσεγγίσεις θα λέγαμε ότι σίγουρα αποδίδουν πολύ καλύτερα όταν εφαρμόζονται σε συνδυασμό με εκείνες που βασίζονται στο ίδιο το περιεχόμενο.

ΚΕΦΑΛΑΙΟ 9

ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

> When a father gives to his son, both laugh; when a son gives to his father, both cry.
>
> *William Shakespeare, English Dramatist, 1564*

Στο τελευταίο αυτό κεφάλαιο, δίνονται ορισμένες μελλοντικές κατευθύνσεις για έρευνα στα πεδία ενδιαφέροντος καθώς και επεκτάσεις που δέχεται το σύστημα που αναπτύχθηκε.

9.1 Γενικές περιοχές μελλοντικής έρευνας

Θα μπορούσαμε να σκεφτούμε και να απαριθμήσουμε πληθώρα προτάσεων για μελλοντική εργασία σε ότι έχει να κάνει με το σύστημα προτάσεων άρθρων νέων που περιγράφηκε στην παρούσα διατριβή. Όπως κάθε σύγχρονο σύστημα προτάσεων/συνεργατικού φιλτραρίσματος αποτελείται από ξεχωριστά τμήματα καθένα από τα οποία έχει σημαντικά περιθώρια βελτίωσης. Πρώτος στόχος μιας μελλοντικής εργασίας θα ήταν λοιπόν η βελτίωση των διαφόρων υποσυστημάτων σε δύο άξονες:

- ταχύτητα λειτουργίας
- αποτελεσματικότητα των τεχνικών

Παρότι όπως ειπώθηκε το σύστημα μπορεί και ανταπεξέρχεται εύκολα στην παρούσα φάση του στο πλήθος των εισερχόμενων κειμένων από τον ιστό, οι πηγές που έχουμε εν ενεργεία όμως δεν είναι πάρα πολλές. Επί της ουσίας δεικτοδοτούμε τα άρθρα νέων από τα περίπου 20 πιο σημαντικά news portals του διαδικτύου, κάτι που φέρνει το σύστημα σε ποσοστό χρησιμοποίησης πόρων περίπου στο 70%. Όπως είναι κατανοητό, όταν το σύστημα κλιμακωθεί ώστε να δεικτοδοτεί εκατοντάδες ή και χιλιάδες portals, θα υπάρξει σημαντικό θέμα κλιμάκωσης που θα αγγίξει όλα τα υποσυστήματα (αν δεν υπάρξει ταυτόχρονη μείωση στη χρήση των πόρων του συστήματος). Κατά συνέπεια, η βελτιστοποίηση τους τόσο σε επίπεδο κώδικα με τις ίδιες τεχνικές, όσο και με χρήση μεθοδολογιών μικρότερης πολυπλοκότητας είναι κάτι που μας ενδιαφέρει για το μέλλον.

Συνολικά για όλα τα υποσυστήματα, η πειραματική διαδικασία εκτελέστηκε χρησιμοποιώντας δεδομένα που είτε υπήρχαν ήδη στη ΒΔ του συστήματος, είτε βασιζόμενοι σε ανάδραση ορισμένων, σχετικά λίγων χρηστών. Αυτό που μας θα μας ενδιέφερε θα ήταν μία πειραματική διαδικασία η οποία θα αξιοποιούσε datasets από διάσημα CF συστήματα, όπως το NetFlix and MovieLens και που αποτελούν τη νόρμα για σύγκριση σε μεγαλύτερη κλίμακα.

9.2 Προεπεξεργασία

Το τμήμα της προεπεξεργασίας κειμένου θα μπορούσε να ενισχυθεί με την χρήση online θησαυρών και όχι μόνο αξιοποιώντας την offline έκδοση του WordNet. Για παράδειγμα η χρήση της Wikipedia για επίκαιρη νοηματική ενίσχυση των keywords του κειμένου θα ήταν μία ενδιαφέρουσα προοπτική μελλοντικής εργασίας.

Όσον αφορά την διαδικασία εξαγωγής n-grams, είναι επιθυμητή η εκτέλεση μεγαλύτερης πειραματικής διαδικασίας ώστε να επιβεβαιωθούν τα αποτελέσματα που παρατηρήσαμε. Με μία μεγαλύτερη και τυχαία επιλογή άρθρων νέων, θεωρούμε ότι θα σιγουρευτούμε για τις καλύτερες τιμές των παραμέτρων n (π.χ. 2-grams, 3-grams) και B (συμμετοχή των εξαγόμενων n-grams στη διαδικασία ζύγισης) όσον αφορά τον τομέα της συσταδοποίησης άρθρων νέων από τον ιστό. Εκτός από το

παραπάνω, στην παρούσα διατριβή δεν αγγίξαμε το ζήτημα της αξιοποίησης των n-grams από τα υπόλοιπα υποσυστήματα του μηχανισμού, όπως αυτό της κατηγοριοποίησης και περίληψης. Αυτό επίσης θα αποτελούσε ένα ενδιαφέρον θέμα για μελλοντική έρευνα.

9.3 Συσταδοποίηση

Σε ότι έχει να κάνει με το τμήμα συσταδοποίησης, υπάρχουν αρκετά περιθώρια βελτίωσης και μελλοντικής εργασίας. Πιο συγκεκριμένα, η βελτιστοποίηση του αυτοματοποιημένου εντοπισμού του πλήθους των συστάδων έχει αρκετό χώρο για μελέτη όσον αφορά τον προτεινόμενο μηχανισμό. Για την ακρίβεια, το ευρετικό που προτάθηκε στην ενότητα 4.3.2.4 δεν έχει αξιολογηθεί αρκετά και ίσως να μην είναι η βέλτιστη επιλογή. Παράλληλα, ο ίδιος ο πυρήνας συσταδοποίησης (αλγόριθμος W-kmeans) έχει αρκετά περιθώρια βελτίωσης μίας και λόγω ότι πηγάζει από τον τυπικό k-means αλγόριθμο, κληρονομεί και τα προβλήματά του όπως αυτά παρουσιάστηκαν στην ενότητα 3.7.1.3.4. Για παράδειγμα η ευαισθησία στις αρχικές συνθήκες και η μη διαχείριση των outliers είναι θέματα που φανερά επηρεάζουν και τον αλγόριθμο W-kmeans. Τέτοια ζητήματα θα συνιστούσαν ένα ενδιαφέρον πεδίο έρευνας για μελλοντική εργασία. Επίσης όσον αφορά στο τμήμα της ονοματοδοσίας συστάδων, παρότι φαίνεται να παράγει ικανοποιητικές ετικέτες, θεωρούμε ότι θα πρέπει να αξιολογηθεί πιο διεξοδικά με ένα μεγαλύτερο σύνολο δεδομένων και αυτοματοποιημένο (χωρίς τις γνώμες χρηστών) τρόπο.

Επιπλέον, ο αλγόριθμος W-kmeans θα μπορούσε εύκολα να ενισχυθεί ώστε να εξάγει από το WordNet επιπλέον πληροφορία που έχει να κάνει με μερώνυμα και συνώνυμα του κειμένου. Μία τέτοια δυνατότητα πιστεύουμε ότι θα βοηθούσε σημαντικά με το πρόβλημα της συνωνυμίας, αν και αυτό εν' μέρει αντιμετωπίζεται ήδη με χρήση των υπερωνύμων. Η επιπλέον αυτή πληροφορία θα μπορούσε να ζυγίζεται με διαφορετικό τρόπο, κάτι που φυσικά θα μπορούσε να αποτελέσει τμήμα μιας μελλοντικής πειραματικής διαδικασίας.

9.4 Προσωποποίηση και παραγωγή προτάσεων

Όσον αφορά το κομμάτι της παραγωγής προτάσεων προς τον χρήστη, στην διδακτορική διατριβή βασιστήκαμε σε απόλυτα κριτήρια αξιολόγησης που έχουν να κάνουν με το MAE ή το F-measure. Παρόλα αυτά, τέτοιου είδους μετρικές, είναι δυνατό να μην αποτυπώνουν την πραγματική εικόνα σε ότι έχει να κάνει με την ικανοποίηση του τελικού χρήστη από τις προτάσεις που του γίνονται (κάτι που προφανώς είναι και το ζητούμενο σε ένα σύστημα προτάσεων). Έτσι μία πιο πλήρης αξιολόγηση θα μπορούσε να συμπεριλαμβάνει και άλλα κριτήρια, όπως: η αποφυγή μεγάλων λαθών όπως π.χ. άρθρα που προσβάλουν ή κάνουν τον χρήστη να νιώθει άβολα, η ποικιλία των προτάσεων σε πολλές θεματολογίες ενδιαφέροντος, κ.α. Εκτός αυτού, οι παραπάνω μετρικές (MAE και F-measure) μπορεί να κάνουν καλή δουλειά στην εξέταση του αν οι προτάσεις μπορούν να ανακτήσουν πληροφορία που ταιριάζει στον χρήστη, δεν κάνουν τόσο καλή δουλειά όμως την αξιολόγηση του κατά πόσο η πληροφορία αυτή είναι "νέα" για αυτόν.

9.5 Παρουσίαση πληροφορίας

Πέρα από τα παραπάνω, το κομμάτι της παρουσίασης της πληροφορίας θα μπορούσε να δεχθεί τεράστιες βελτιώσεις. Το υποσύστημα αυτό δεν ήταν κομμάτι της διδακτορικής διατριβής και ως εκ τούτου δεν ασχοληθήκαμε παραπάνω. Η εύρεση όμως των αποτελεσματικότερων καναλιών επικοινωνίας με τον τελικό χρήστη είναι κάτι εξαιρετικά σημαντικό για την επιτυχία ενός συστήματος προτάσεων. Η πληροφορία θα πρέπει να είναι άμεσα διαθέσιμη, όχι μόνο μπροστά στον υπολογιστή του, αλλά και στο κινητό του, την τηλεόρασή, ή και και γενικότερα σε οποιονδήποτε τρόπο πρόσβασης αυτός επιθυμεί. Το παραπάνω εισάγει μεγάλες προκλήσεις οι οποίες από μόνες τους θα μπορούσαν να αποτελέσουν μία ξεχωριστή εργασία.

9.6 Πρόβλημα νέου χρήστη

Σε σχέση με το πρόβλημα νέου χρήστη, παρότι τα αποτελέσματα της πειραματικής μας διαδικασίας ήταν ενθαρρυντικά, δεν μπορούν να θεωρηθούν οριστικά μιας και έγιναν με χρήση ενός offline στιγμιότυπου της ΒΔ του συστήματος πάνω σε ήδη καταγεγραμμένες βαθμολογήσεις χρηστών. Επομένως, μία μεγαλύτερης κλίμακας πειραματική διαδικασία με online δεδομένα από πραγματικούς χρήστες θα ήταν κάτι ενδιαφέρον για το μέλλον.

ΕΥΡΕΤΗΡΙΟ

C-means, 84
City-block, 88
Clustering Index, 91
CLUTO, 149
Corpus, 167, 170, 181, 185, 186, 191, 195

DBSCAN, 87
Deep Web, 42

Hypernym, 47
Hypernyms, 124, 128

k-means, 79
Kendall's τ, 90

MAE, 91
Manhattan, 88
MATLAB, 150
Meronym, 47
Meta Portals, 44
MySQL, 152

n-grams, 48, 73, 119, 148, 160, 167, 203
news articles, 43
News Portals, 44
news portals, 43, 44

portlet, 43

S-kmeans, 82
SenseClusters, 150
Spearman-rank, 89
Stemming, 46

Vector Space model, 60

W-kmeans, 96, 109, 110, 112, 123, 128, 151, 157, 181, 185, 186, 190, 197, 203–206, 212
Web Portals, 43
Wordnet, 37, 46, 47, 91, 92, 95, 106, 110, 115, 124, 126, 130, 133, 148, 169, 182, 204, 212

Άρθρα Νέων, 43
Ανάκτηση Πληροφορίας, 59
Αξιολόγηση, 165
Απαιτήσεις, 163
Αρχιτεκτονική, 101
Βάση δεδομένων, 153
Διασύνδεση, 160
Εξαγωγή κειμένου, 161
Εξαγωγή λέξεων κλειδιών, 73
Ευκλείδεια απόσταση, 88
Εύρεση συνεδρίων, 131
Ιεραρχικοί αλγόριθμοι, 76
Κατηγοριοποίηση, 75, 162
Μελλοντική εργασία, 209
Μερισματικοί αλγόριθμοι, 78
Μετρικές αξιολόγησης, 63, 88, 90
Ομοιότητα συνημιτόνου, 89
Ονοματοδοσία συστάδων, 94, 129

Περίληψη, 162
Πλήθος συστάδων, 51, 92, 112
Προδιαγραφές, 161
Προεπεξεργασία, 45, 72, 107, 119, 148, 161, 167, 211
Προσωποποίηση, 52, 95, 115, 130, 135, 163, 212
Ροή Πληροφορίας, 104
Στόχοι, 103
Συμπεράσματα, 200
Συνεργατικό Φιλτράρισμα, 45, 67, 68, 71, 115, 141, 204, 211
Συνεργατικό φιλτράρισμα, 67
Συστήματα προτάσεων, 45, 49, 51, 71, 91, 95, 104, 109, 113, 123, 130, 134, 151, 153, 161, 202, 211
Συσταδοποίηση, 48, 51, 75, 109–111, 122–124, 133, 135, 149, 154, 158, 170, 186, 204, 205, 212
Φιλτράρισμα, 45, 65, 67, 70, 202
Φυσική Επεξεργασία Γλώσσας, 48, 49, 57, 58, 73, 74, 149

ΓΛΩΣΣΑΡΙ

agglomerative hierarchical ιεραρχικοί αλγόριθμοι συσταδοποίησης. 76

cluster συστάδα. 75

collaborative filtering συνεργατικό φιλτράρισμα. 45, 67

dendogram δενδρόγραμμα. 77

distance matrix πίνακας ομοιότητας. 88

eigenvalues ιδιοτιμές. 86

eigenvectors ιδιοπίνακες. 86

outliers ακραία αντικείμενα συστάδας. 79, 84, 85, 87, 89, 212

partitional μερισματικοί αλγόριθμοι συσταδοποίησης. 76

recommendation system σύστημα προτάσεων. 45

singleton μοναδιαία συστάδα. 76, 179

synset WordNet Synonym Set. 46

transactional databases βάσεις δεδομένων που βασίζονται σε συναλλαγές. 76

W-kmeans WordNet k-means αλγόριθμος συσταδοποίησης. 37, 38

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

AI Artificial Intelligence. 46, 49, 75

BOW Bag of Words. 48, 63, 73, 111, 119, 168, 203

CI Clustering Index. 154, 158, 167, 171, 181, 186

DBMS DataBase Management System. 65

DBSCAN Density-based spatial clustering of applications with noise. 51, 87

EM Expectation Maximization. 51, 80, 81

HCA Hierarchical Clustering Analysis. 76

IE Information Extraction. 59

IF Information Filtering. 66

IoT Internet of Things. 42

IR Information Retrieval. 49, 50, 59, 63, 65, 73–75, 83, 90, 92

k-NN k Nearest Neighbors. 69, 71

KWs Keywords. 45

LSA Latent Semantic Analysis. 61, 69

LSI Latent Semantic Indexing. 61

MAE Mean Absolute Error. 91, 192, 194, 196, 212

MaxL Maximum Likelihood. 80

MDP Markov decision process. 70

ML Machine Learning. 49, 75

NER Named entity recognition. 58

NLP Natural Language Processing. 49, 57, 73

OCR Optical character recognition. 58

OPTICS Ordering points to identify the clustering structure. 51, 87

PCA Principal Component Analysis. 86, 152

PD Personality Diagnosis. 70

POS Part of Speech. 46, 58

RSoS Residual Sum of Squares. 79

SAHN Sequential Agglomerative Hierarchical Non-overlapping. 78

SVD Singular Value Decompossition. 69, 70, 80

SVM Support Vector Machine. 75

Synsets Synonym sets. 92

TMG Text to Matrix Generator. 151

VSM Vector Space Model. 61, 62

ΑΠ Ανάκτηση Πληροφορίας. 59–62

ΒΔ Βάση Δεδομένων. 44

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] *A Java Example for N-Gram Generation*. A Java Example for N-Gram Generation. http://www.text-analytics101.com/2014/11/what-are-n-grams.html.

[2] *A Simple Ruby N-Gram Generator*. A Simple Ruby N-Gram Generator. http://jasonheppler.org/2012/04/24/a-simple-ruby-ngram-generator/.

[3] Tony Abou-Assaleh et al. "Detection of New Malicious Code Using N-grams Signatures." In: *PST*. 2004, pp. 193–196.

[4] George Adam, Christos Bouras, and Vassilis Poulopoulos. "CUTER: An efficient useful text extraction mechanism". In: *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*. IEEE. 2009, pp. 703–708.

[5] George Adam, Christos Bouras, and Vassilis Poulopoulos. "Efficient extraction of news articles based on RSS crawling". In: *Machine and Web Intelligence (ICMWI), 2010 International Conference on*. IEEE. 2010, pp. 1–7.

[6] George Adam, Christos Bouras, and Vassilis Poulopoulos. *Monitoring rss feeds*. na, 2009.

[7] George Adam, Christos Bouras, and Vassilis Poulopoulos. "Utilizing RSS feeds for crawling the Web". In: *Internet and Web Applications and Services, 2009. ICIW'09. Fourth International Conference on*. IEEE. 2009, pp. 211–216.

[8] Giorgos Adam, Christos Bouras, and Vassilis Poulopoulos. "Image Extraction from Online Text Streams: A Straightforward Template Independent Approach without Training". In: *The 2010 IEEE International Symposium on Mining and Web (MAW 2010), Perth, Australia Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 609–613.

[9] Daniel Aloise et al. "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine Learning* 75.2 (2009), pp. 245–248.

[10] *Amazon*. Online shopping. http://www.amazon.com.

[11] Abdelmalek Amine, Zakaria Elberrichi, and Michel Simonet. *Evaluation of Text Clustering Methods Using WordNet*. 2009.

[12] Nicholas O Andrews and Edward A Fox. *Recent developments in document clustering*. Tech. rep. Tech. rept. TR-07-35. Deparment of Computer Science, Virginia Tech, 2007.

[13] Asim Ansari, Skander Essegaier, and Rajeev Kohli. "Internet recommendation systems". In: *Journal of Marketing research* 37.3 (2000), pp. 363–375.

[14] Ioannis Antonellis, Christos Bouras, and Vassilis Poulopoulos. "Scalable text classification as a tool for personalization". In: *Computer Systems Science and Engineering* 24.6 (2009), p. 399.

[15] *AOL*. AOL. http://www.aol.com/.

[16] Chidanand Apté, Fred Damerau, and Sholom M Weiss. "Towards language independent automated learning of text categorization models". In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc. 1994, pp. 23–30.

[17] David Arthur and Sergei Vassilvitskii. "How Slow is the k-means Method?" In: *Proceedings of the twenty-second annual symposium on Computational geometry*. ACM. 2006, pp. 144–153.

[18] David Arthur and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[19] David Arthur and Sergei Vassilvitskii. "On the worst case complexity of the k-means method". In: (2005).

[20] Javed A Aslam, Emine Yilmaz, and Virgiliu Pavlu. "A geometric interpretation of r-precision and its correlation with average precision". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 573–574.

[21] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*. Vol. 463. ACM press New York, 1999.

[22] Marko Balabanović and Yoav Shoham. "Fab: content-based, collaborative recommendation". In: *Communications of the ACM* 40.3 (1997), pp. 66–72.

[23] *Barnes and Noble*. Barnes and Noble books. http://www.barnesandnoble.com/.

[24] Alberto Barrón-Cedeño and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison". In: *Advances in Information Retrieval*. Springer, 2009, pp. 696–700.

[25] *BBC NEWS*. News portal. http://news.bbc.co.uk/.

[26] Jörg Becker and Dominik Kuropka. "Topic-based vector space model". In: *Proceedings of the 6th International Conference on Business Information Systems*. 2003, pp. 7–12.

[27] N.J. Belkin and W.B. Croft. "Information filtering and information retrieval: two sides of the same coin?" In: *Communications of the ACM* 35.12 (1992), pp. 29–38.

[28] Timothy C Bell, John G Cleary, and Ian H Witten. *Text compression*. Prentice-Hall, Inc., 1990.

[29] Andreas Berg, Renate Meyer, and Jun Yu. "Deviance information criterion for comparing stochastic volatility models". In: *Journal of Business & Economic Statistics* 22.1 (2004), pp. 107–120.

[30] Jeff A Bilmes et al. "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". In: *International Computer Science Institute* 4.510 (1998), p. 126.

[31] *Bing*. Search engine. http://www.bing.com/.

[32] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[33] *Boost Libraries for C++*. Website. http://www.boost.org/.

[34] C. Bouras, V. Poulopoulos, and V. Tsogkas. "PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries". In: *Data and Knowledge Engineering Journal, Elsevier Science Vol. 64, Issue 1* (2008).

[35] Christos Bouras and Vassilis Poulopoulos. "Dynamic user context web personalization in meta-portals". In: *Computers and Communications (ISCC), 2010 IEEE Symposium on*. IEEE. 2010, pp. 925–930.

[36] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. "Adaptation of RSS feeds based on the user profile and on the end device". In: *Journal of Network and Computer Applications* 33.4 (2010), pp. 410–421.

[37] Christos Bouras, Vassilis Poulopoulos, and Vassilis Tsogkas. "Efficient Summarization Based On Categorized Keywords." In: *Dmin*. Citeseer. 2007, pp. 285–291.

[38] Christos Bouras et al. "Trash article detection using categorization techniques." In: *IADIS AC (1)*. Citeseer. 2009, pp. 51–58.

[39] Paul S Bradley and Usama M Fayyad. "Refining Initial Points for K-Means Clustering." In: Citeseer. 1998.

[40] *C Clustering Library*. The C Clustering Library. http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm.

[41] Igor Cadez et al. "Visualization of navigation patterns on a web site using model-based clustering". In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2000, pp. 280–284.

[42] Fazli Can and Esen A Ozkarahan. "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". In: *ACM Transactions on Database Systems (TODS)* 15.4 (1990), pp. 483–517.

[43] G. Cao. "Support Vector Machine Active Learning with Applications to Text Classification". In: ().

[44] *Cgicc. A C++ class library for writing CGI applications*. Website. http://www.cgicc.org/.

[45] S. Chakrabarti et al. "Using taxonomy, discriminants, and signatures for navigating in text databases". In: *Proceedings of the 23rd VLDB Conference* (1997), pp. 446–455.

[46] Sonny Han Seng Chee, Jiawei Han, and Ke Wang. "Rectree: An efficient collaborative filtering method". In: *Data Warehousing and Knowledge Discovery*. Springer, 2001, pp. 141–151.

[47] Chun-Ling Chen, Frank SC Tseng, and Tyne Liang. "An integration of fuzzy association rules and WordNet for document clustering". In: *Knowledge and information systems* 28.3 (2011), pp. 687–708.

[48] Li Chen and Pearl Pu. "Preference-based organization interfaces: aiding user critiques in recommender systems". In: *User Modeling 2007*. Springer, 2007, pp. 77–86.

[49] Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir. "Preventing shilling attacks in online recommender systems". In: *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM. 2005, pp. 67–74.

[50] *Clustering with Matlab*. Clustering with Matlab. http://www.dcorney.com/ClusteringMatlab.html.

[51] *CLUTO*. CLUTO - Software for Clustering High-Dimensional Datasets. http://glaros.dtc.umn.edu/gkhome/views/cluto.

[52] *CNN*. News portal. http://www.cnn.com/.

[53] N. Collier, C. Nobata, and J. Tsujii. "Extracting the names of genes and gene products with a hidden Markov model". In: *Proceedings of the 18th conference on Computational linguistics-Volume 1* (2000), pp. 201–207.

[54] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. "Data preparation for mining world wide web browsing patterns". In: *Knowledge and information systems* 1.1 (1999), pp. 5–32.

[55] Dan Cosley et al. "SuggestBot: using intelligent task routing to help people find work in wikipedia". In: *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM. 2007, pp. 32–41.

[56] Matt Crane. "The new user problem in collaborative filtering". In: *University of Otago* (2011).

[57] Sanjoy Dasgupta and Yoav Freund. "Random projection trees for vector quantization." In: *IEEE Transactions on Information Theory* 55.7 (2009), pp. 3229–3242.

[58] William HE Day and Herbert Edelsbrunner. "Efficient algorithms for agglomerative hierarchical clustering methods". In: *Journal of classification* 1.1 (1984), pp. 7–24.

[59] Inderjit S Dhillon, James Fan, and Yuqiang Guan. "Efficient clustering of very large document collections". In: *Data mining for scientific and engineering applications*. Springer, 2001, pp. 357–381.

[60] Inderjit S Dhillon and Dharmendra S Modha. "Concept decompositions for large sparse text data using clustering". In: *Machine learning* 42.1-2 (2001), pp. 143–175.

[61] S. Dumais and H. Chen. "Hierarchical classification of Web content". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (2000), pp. 256–263.

[62] Susan T Dumais. "Latent semantic analysis". In: *Annual review of information science and technology* 38.1 (2004), pp. 188–230.

[63] Magdalini Eirinaki and Michalis Vazirgiannis. "Web mining for web personalization". In: *ACM Transactions on Internet Technology (TOIT)* 3.1 (2003), pp. 1–27.

[64] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. "Collaborative filtering recommender systems". In: *Foundations and Trends in Human-Computer Interaction* 4.2 (2011), pp. 81–173.

[65] Michael D Ekstrand et al. "Automatically building research reading lists". In: *Proceedings of the fourth ACM conference on Recommender systems*. ACM. 2010, pp. 159–166.

[66] *Excite*. Excite. http://www.excite.com/.

[67] *Expat*. XML parsing library. http://expat.sourceforge.net/.

[68] Martin Farach. "Optimal suffix tree construction with large alphabets". In: *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*. IEEE. 1997, pp. 137–143.

[69] *Finding the Right Number of Clusters in k-Means and EM Clustering: v-Fold Cross-Validation*. http://www.ccs.neu.edu/home/futrelle/teaching/isu535sp2004/finalpapers/clusteringIntro.html.

[70] *foxnews.com*. News portal. http://www.foxnews.com/.

[71] Yongjian Fu, Kanwalpreet Sandhu, and Ming-Yi Shih. "Clustering of web users based on access patterns". In: *Proceedings of the 1999 KDD Workshop on Web Mining*. Citeseer. 1999.

[72] J. Furnkranz, T. Mitchell, and E. Riloff. "A case study in using linguistic phrases for text categorization on the WWW". In: *Learning for Text Categorization: Proceedings of the 1998 AAAI/ICML Workshop* (1998), pp. 98–05.

[73] Johannes Fürnkranz. "A study using n-gram features for text categorization". In: *Austrian Research Institute for Artifical Intelligence* 3.1998 (1998), pp. 1–10.

[74] *gd*. A graphics library for fast image creation. http://libgd.org/.

[75] *Gentoo Linux*. Website. http://www.gentoo.org/.

[76] Tarek F Gharib, Mohammed M Fouad, and Mostafa M Aref. "Fuzzy document clustering approach using WordNet lexical categories". In: *Advanced Techniques in Computing Sciences and Software Engineering*. Springer, 2010, pp. 181–186.

[77] J. Gimenez and L. Marquez. "Svmtool: A general pos tagger generator based on support vector machines". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 2004, pp. 43–46.

[78] Nadav Golbandi, Yehuda Koren, and Ronny Lempel. "Adaptive bootstrapping of recommender systems using decision trees". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 595–604.

[79] Nadav Golbandi, Yehuda Koren, and Ronny Lempel. "On bootstrapping recommender systems". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM. 2010, pp. 1805–1808.

[80] David Goldberg et al. "Using collaborative filtering to weave an information tapestry". In: *Communications of the ACM* 35.12 (1992), pp. 61–70.

[81] Moises Goldszmidt and Mehran Sahami. "A probabilistic approach to full-text document clustering". In: (1998).

[82] Nathaniel Good et al. "Combining collaborative filtering with personal agents for better recommendations". In: *AAAI/IAAI*. 1999, pp. 439–446.

[83] *Google*. Search engine. http://www.google.com/.

[84] *Google News*. News Portal. http://news.google.com/.

[85] *GUN Aspell*. spell checker. http://aspell.net/.

[86] Abdelmoula El-Hamdouchi and Peter Willett. "Comparison of hierarchic agglomerative clustering methods for document retrieval". In: *The Computer Journal* 32.3 (1989), pp. 220–227.

[87] John Hannon, Mike Bennett, and Barry Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches". In: *Proceedings of the fourth ACM conference on Recommender systems*. ACM. 2010, pp. 199–206.

[88] Stephen Paul Harter. "A probabilistic approach to automatic keyword indexing". PhD thesis. University of Chicago, 1974.

[89] John A Hartigan and Manchek A Wong. "Algorithm AS 136: A k-means clustering algorithm". In: *Applied statistics* (1979), pp. 100–108.

[90] Birgit Hay, Geert Wets, and Koen Vanhoof. "Clustering navigation patterns on a website using a sequence alignment method". In: *Intelligent Techniques for Web Personalization: IJCAI* (2001), pp. 1–6.

[91] Jon Herlocker, Joseph A Konstan, and John Riedl. "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms". In: *Information retrieval* 5.4 (2002), pp. 287–310.

[92] Jonathan L Herlocker et al. "An algorithmic framework for performing collaborative filtering". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1999, pp. 230–237.

[93] Will Hill et al. "Recommending and evaluating choices in a virtual community of use". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co. 1995, pp. 194–201.

[94] Thomas Hofmann. "Latent semantic models for collaborative filtering". In: *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), pp. 89–115.

[95] Timo Honkela. "Self-organizing maps in natural language processing". PhD thesis. Helsinki University of Technology Espoo, Finland, 1997.

[96] *htmltidy*. Tidy the layout and correct errors in HTML and XML documents. http://tidy.sourceforge.net/.

[97] Anette Hulth. "Improved automatic keyword extraction given more linguistic knowledge". In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics. 2003, pp. 216–223.

[98] *icu*. International Components for Unicode. http://www.icu-project.org/.

[99] *iGoogle*. iGoogle. http://www.igoogleportal.com.

[100] *Indiatimes*. Indiatimes. http://www.indiatimes.com.

[101] *internetlivestats.com*. Internet Statistics. http://www.internetlivestats.com.

[102] Hosein Jafarkarimi, Alex Tze Hiang Sim, and Robab Saadatdoost. "A naive recommendation model for large databases". In: *International Journal of Information and Education Technology* 2.2 (2012), pp. 216–219.

[103] Taeho Jo. "Evaluation function of document clustering based on term entropy". In: *Proc. of 2nd International Symposium on Advanced Intelligent System*. 2001, pp. 95–100.

[104] Taeho Jo and Malrey Lee. "The evaluation measure of text clustering for the variable number of clusters". In: *Advances in Neural Networks–ISNN 2007*. Springer, 2007, pp. 871–879.

[105] T. Joachims. *Text Categorization with Suport Vector Machines: Learning with Many Relevant Features*. Springer-Verlag London, UK, 1998.

[106] Thorsten Joachims. "Optimizing search engines using clickthrough data". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 133–142.

[107] Simmi John and R Boobathiraj. "High Dimensional Hierarchical Data Clustering using SVM with Kernel Region Approximation Indexing". In: *International Journal of Computer Science & Communication Networks* 2.4 (2012).

[108] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[109] K.S. Jones. "Exhaustivity and specificity". In: *Journal of Documentation* 28.1 (1972), pp. 11–21.

[110] Kyung-Yong Jung, Dong-Hyun Park, and Jung-Hyun Lee. "Hybrid collaborative filtering and content-based filtering for improved recommender system". In: *Computational Science-ICCS 2004*. Springer, 2004, pp. 295–302.

[111] Nishikant Kapoor et al. *A study of citations in users' online personal collections*. Springer, 2007.

[112] George Karypis. *CLUTO-a clustering toolkit*. Tech. rep. DTIC Document, 2002.

[113] George Karypis. "Evaluation of item-based top-n recommendation algorithms". In: *Proceedings of the tenth international conference on Information and knowledge management*. ACM. 2001, pp. 247–254.

[114] Samuel Kaski et al. "WEBSOM–self-organizing maps of document collections". In: *Neurocomputing* 21.1 (1998), pp. 101–117.

[115] J. Kazama et al. "Tuning Support Vector Machines for Biomedical Named Entity Recognition". In: *Proc. of the Workshop on Natural Language Processing in the Biomedical Domain (at ACL'2002)* (2002), pp. 1–8.

[116] David J Ketchen and Christopher L Shook. "The application of cluster analysis in strategic management research: an analysis and critique". In: *Strategic management journal* 17.6 (1996), pp. 441–458.

[117] Shehroz S Khan and Amir Ahmad. "Cluster center initialization algorithm for< i> K</i>-means clustering". In: *Pattern recognition letters* 25.11 (2004), pp. 1293–1302.

[118] Teuvo Kohonen et al. "Self organization of a massive document collection". In: *Neural Networks, IEEE Transactions on* 11.3 (2000), pp. 574–585.

[119] D. Koller and M. Sahami. "Hierarchically classifying documents using very few words". In: *Proceedings of the Fourteenth International Conference on Machine Learning* (1997), pp. 170–178.

[120] Joseph A Konstan and John Riedl. "Recommender systems: from algorithms to user experience". In: *User Modeling and User-Adapted Interaction* 22.1-2 (2012), pp. 101–123.

[121] Yehuda Koren and Robert Bell. "Advances in collaborative filtering". In: *Recommender Systems Handbook*. Springer, 2011, pp. 145–186.

[122] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems". In: *Computer* 42.8 (2009), pp. 30–37.

[123] Jim ZC Lai, Tsung-Jen Huang, and Yi-Ching Liaw. "A fast k-means clustering algorithm using cluster center displacement". In: *Pattern Recognition* 42.11 (2009), pp. 2551–2556.

[124] Frederick Wilfrid Lancaster and Emily Gallup. *Information retrieval on-line*. Tech. rep. 1973.

[125] Danial Lashkari and Polina Golland. "Convex clustering with exemplar-based models". In: *Advances in neural information processing systems*. 2007, pp. 825–832.

[126] Yanjun Li and Soon M Chung. "Parallel bisecting k-means with prediction clustering algorithm". In: *The Journal of Supercomputing* 39.1 (2007), pp. 19–37.

[127] *libcurl*. Curl grabber. http://curl.haxx.se/.

[128] *libpng*. Portable Network Graphics library. http://www.libpng.org/.

[129] *libstemmer*. Multi-language stemmers. http://snowball.tartarus.org/download.php.

[130] Tie-Yan Liu. "Learning to rank for information retrieval". In: *Foundations and Trends in Information Retrieval* 3.3 (2009), pp. 225–331.

[131] Xin Liu et al. "Document clustering with cluster refinement and model selection capabilities". In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, pp. 191–198.

[132] R Lletı et al. "Selecting variables for< i> k</i>-means cluster analysis by using a genetic algorithm that optimises the silhouettes". In: *Analytica Chimica Acta* 515.1 (2004), pp. 87–100.

[133] Pasquale Lops, Marco Degemmis, and Giovanni Semeraro. "Improving social filtering techniques through WordNet-Based user profiles". In: *User Modeling 2007*. Springer, 2007, pp. 268–277.

[134] *Lycos*. Lycos. http://www.lycos.com.

[135] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. "The planar k-means problem is NP-hard". In: *WALCOM: Algorithms and Computation*. Springer, 2009, pp. 274–285.

[136] Markus Maier, Matthias Hein, and Ulrike von Luxburg. "Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters". In: *Theoretical Computer Science* 410.19 (2009), pp. 1749–1764.

[137] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[138] James H Martin and D Jurafsky. "Speech and language processing". In: *International Edition* (2000).

[139] A. McCallum and K. Nigam. "A comparison of event models for naive bayes text classification". In: *AAAI-98 Workshop on Learning for Text Categorization* 752 (1998).

[140] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[141] Prem Melville, Raymond J Mooney, and Ramadass Nagarajan. "Content-boosted collaborative filtering for improved recommendations". In: *AAAI/IAAI*. 2002, pp. 187–192.

[142] Prem Melville and Vikas Sindhwani. "Recommender systems". In: *Encyclopedia of machine learning*. Springer, 2010, pp. 829–838.

[143] Arnd Kohrs–Bernard Merialdo. "Improving Collaborative Filtering For New-Users By Smart Object Selection". In: ().

[144] T. Mitchell et al. "Machine Learning". In: *Annual Review of Computer Science* 4.1 (1990), pp. 417–433.

[145] Koji Miyahara and Michael J Pazzani. "Collaborative filtering with the simple Bayesian classifier". In: *PRICAI 2000 Topics in Artificial Intelligence*. Springer, 2000, pp. 679–689.

[146] D. Mladenic and M. Grobelnik. "Word sequences as features in text-learning". In: *Proceedings of ERK-98, the Seventh Electrotechnical and Computer Science Conference* (1998), pp. 145–148.

[147] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. "Automatic personalization based on Web usage mining". In: *Communications of the ACM* 43.8 (2000), pp. 142–151.

[148] *MSN*. MSN. http://www.msn.com.

[149] *Mysq++ C++ API interface to the MySQL database*. Website. http://www.mysql.org/downloads/api-mysql++.html.

[150] *MySQL, Opensource database*. http://www.mysql.com.

[151] Makoto Nagao and Shinsuke Mori. "A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese". In: *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics. 1994, pp. 611–615.

[152] Peter Náther. "N-gram based Text Categorization". In: *Lomonosov Moscow State Univ* (2005).

[153] *Naver*. Naver. http://www.naver.com.

[154] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants". In: *Learning in graphical models*. Springer, 1998, pp. 355–368.

[155] *Netvibes*. Netvibes. http://www.netvibes.com/.

[156] *News Junkies*. News Portal. http://www.newsjunkie.info/.

[157] *News me*. News Portal. http://www.news.me/.

[158] *NGramCounter*. TextToolbox NGramCounter - A Free Web API for N-Gram Generation. https://www.mashape.com/rxnlp/texttoolbox.

[159] Hien Nguyen and Peter Haddawy. "The decision-theoretic video advisor". In: *AAAI-98 Workshop on Recommender Systems*. 1998, pp. 77–80.

[160] C. Nobata, N. Collier, and J. Tsujii. "Automatic term identification and classification in biology texts". In: *Proc. of the 5th NLPRS* (1999), pp. 369–374.

[161] Richard Nock and Frank Nielsen. "On weighting clustering". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.8 (2006), pp. 1223–1235.

[162] Sebastian Nowozin and Gökhan Bakir. "A decoupled approach to exemplar-based unsupervised learning". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 704–711.

[163] Alauddin Yousif Al-Omary and Mohammad Shahid Jamil. "A new approach of clustering based machine-learning algorithm". In: *Knowledge-Based Systems* 19.4 (2006), pp. 248–258.

[164] *Open Directory Project*. Website. http://www.dmoz.org.

[165] *openssl*. Full-strength general purpose cryptography library (including SSL and TLS). http://www.openssl.org/.

[166] Rong Pan, Peter Dolog, and Guandong Xu. "KNN-based clustering for improving social recommender systems". In: *Agents and Data Mining Interaction*. Springer, 2013, pp. 115–125.

[167] Seung-Taek Park et al. "Naïve filterbots for robust cold-start recommendations". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 699–705.

[168] Dmitry Y Pavlov and David M Pennock. "A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains". In: *Advances in Neural Information Processing Systems*. 2002, pp. 1441–1448.

[169] José Manuel Pena, Jose Antonio Lozano, and Pedro Larranaga. "An empirical comparison of four initialization methods for the< i> K</i>-Means algorithm". In: *Pattern recognition letters* 20.10 (1999), pp. 1027–1040.

[170] *personews*. News Portal. http://news.csd.auth.gr/.

[171] *Perssonal*. Perssonal, the most powerfull meta-portal. http://perssonal.cti.gr.

[172] István Pilászy and Domonkos Tikk. "Recommending new movies: even a few ratings are more valuable than metadata". In: *Proceedings of the third ACM conference on Recommender systems*. ACM. 2009, pp. 93–100.

[173] Huseyin Polat and Wenliang Du. "SVD-based collaborative filtering with privacy". In: *Proceedings of the 2005 ACM symposium on Applied computing*. ACM. 2005, pp. 791–795.

[174] Jay M Ponte and W Bruce Croft. "A language modeling approach to information retrieval". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, pp. 275–281.

[175] Amruta Purandare and Ted Pedersen. "SenseClusters: finding clusters that represent word senses". In: *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics. 2004, pp. 26–29.

[176] Mian Qin, Scott Buffett, and Michael Fleming. "Predicting user preferences via similarity-based clustering". In: (2008).

[177] Al Mamunur Rashid, George Karypis, and John Riedl. "Learning preferences of new users in recommender systems: an information theoretic approach". In: *ACM SIGKDD Explorations Newsletter* 10.2 (2008), pp. 90–100.

[178] Al Mamunur Rashid et al. "Getting to know you: learning new user preferences in recommender systems". In: *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM. 2002, pp. 127–134.

[179] Diego Reforgiato Recupero. "A new unsupervised method for document clustering by using WordNet lexical and conceptual relations". In: *Information Retrieval* 10.6 (2007), pp. 563–579.

[180] *Rediff*. Rediff. http://www.rediff.com.

[181] Paul Resnick et al. "GroupLens: an open architecture for collaborative filtering of netnews". In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM. 1994, pp. 175–186.

[182] *Reuters*. News portal. http://www.reuters.com/.

[183] Pravin Revankar and Jyoti Dahiwele. "Web Usage Mining". In: *5th National conference*. 2011, pp. 10–11.

[184] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[185] J. Rissanen. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc. River Edge, NJ, USA, 1989.

[186] Stephen E Robertson and K Sparck Jones. "Relevance weighting of search terms". In: *Journal of the American Society for Information science* 27.3 (1976), pp. 129–146.

[187] Sura Rodpongpun, Vit Niennattrakul, and Chotirat Ann Ratanamahatana. "Selective subsequence time series clustering". In: *Knowledge-Based Systems* 35 (2012), pp. 361–368.

[188] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. "Akaike information criterion statistics". In: *Dordrecht, The Netherlands: D. Reidel* (1986).

[189] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, USA, 1986.

[190] Gerard Salton, Edward A Fox, and Harry Wu. "Extended Boolean information retrieval". In: *Communications of the ACM* 26.11 (1983), pp. 1022–1036.

[191] Gerard Salton, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing". In: *Communications of the ACM* 18.11 (1975), pp. 613–620.

[192] Badrul Sarwar et al. "Item-based collaborative filtering recommendation algorithms". In: *Proceedings of the 10th international conference on World Wide Web*. ACM. 2001, pp. 285–295.

[193] Julian Sedding and Dimitar Kazakov. "WordNet-based text document clustering". In: *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*. Association for Computational Linguistics. 2004, pp. 104–113.

[194] *SenseClusters*. SenseClusters. http://senseclusters.sourceforge.net/.

[195] Guy Shani, Ronen I Brafman, and David Heckerman. "An MDP-based recommender system". In: *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 2002, pp. 453–460.

[196] Upendra Shardanand and Pattie Maes. "Social information filtering: algorithms for automating "word of mouth"". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co. 1995, pp. 210–217.

[197] Luo Si and Rong Jin. "Flexible mixture model for collaborative filtering". In: *ICML*. Vol. 3. 2003, pp. 704–711.

[198] N. Slonim and N. Tishby. "The power of word clusters for text classification". In: *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research* (2001).

[199] Michael Steinbach, George Karypis, Vipin Kumar, et al. "A comparison of document clustering techniques". In: *KDD workshop on text mining*. Vol. 400. 1. Boston. 2000, pp. 525–526.

[200] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. "Impact of similarity measures on web-page clustering". In: *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*. 2000, pp. 58–64.

[201] Xiaoyuan Su and Taghi M Khoshgoftaar. "A survey of collaborative filtering techniques". In: *Advances in artificial intelligence* 2009 (2009), p. 4.

[202] Xiaoyuan Su and Taghi M Khoshgoftaar. "Collaborative filtering for multi-class data using belief nets algorithms". In: *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*. IEEE. 2006, pp. 497–504.

[203] Na Tang and V Rao Vemuri. "User-interest-based document filtering via semi-supervised clustering". In: *Foundations of Intelligent Systems*. Springer, 2005, pp. 573–582.

[204] Loren Terveen and Will Hill. "Beyond recommender systems: Helping people help each other". In: *HCI in the New Millennium* 1 (2001), pp. 487–509.

[205] *Text to Matrix Generator*. Text to Matrix Generator. http://scgroup20.ceid.upatras.gr:8000/tmg/.

[206] *The Apache Web Server*. Website. http://httpd.apache.org/.

[207] *The GNU Compiler Collection*. Website. http://www.netbeans.org/.

[208] *The PHP language runtime engine: CLI, CGI and Apache2 SAPIs*. Website. http://www.php.net/.

[209] Pucktada Treeratpituk and Jamie Callan. "Automatically labeling hierarchical clusters". In: *Proceedings of the 2006 international conference on Digital government research*. Digital Government Society of North America. 2006, pp. 167–176.

[210] Yuen-Hsien Tseng. "Generic title labeling for clustered documents". In: *Expert Systems with Applications* 37.3 (2010), pp. 2247–2254.

[211] Yuen-Hsien Tseng et al. "Toward generic title generation for clustered documents". In: *Information Retrieval Technology*. Springer, 2006, pp. 145–157.

[212] Esko Ukkonen. "On-line construction of suffix trees". In: *Algorithmica* 14.3 (1995), pp. 249–260.

[213] Cornelis J Van Rijsbergen. "A non-classical logic for information retrieval". In: *The computer journal* 29.6 (1986), pp. 481–485.

[214] Giannis Varelas et al. "Semantic similarity methods in wordNet and their application to information retrieval on the web". In: *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM. 2005, pp. 10–16.

[215] Andrea Vattani. "The hardness of k-means clustering in the plane". In: *Manuscript, accessible at http://cseweb. ucsd. edu/avattani/papers/kmeans_ hardness. pdf* 617 (2009).

[216] J. Verbeek. "An information theoretic approach to finding word groups for text classification". In: *Institute for Lang uage, L og ic and Comp utation, University of Am sterdam* (2000).

[217] David L Weakliem. "A critique of the Bayesian information criterion for model selection". In: *Sociological Methods & Research* 27.3 (1999), pp. 359–397.

[218] Ian H Witten et al. "Text mining: A new frontier for lossless compression". In: *Data Compression Conference, 1999. Proceedings. DCC'99*. IEEE. 1999, pp. 198–207.

[219] SK Michael Wong, Wojciech Ziarko, and Patrick CN Wong. "Generalized vector spaces model in information retrieval". In: *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1985, pp. 18–25.

[220] *WordNet*. WordNet, a lexical database for English. http://wordnet.princeton.edu/.

[221] *worldwidewebsize.com*. World Wide Web size. http://www.worldwidewebsize.com/.

[222] *Xerces*. A validating XML parser. http://xerces.apache.org/xerces-c/.

[223] *yahoo.com*. Search engine. http://www.yahoo.com/.

[224] Y. Yang and J.O. Pedersen. "A comparative study on feature selection in text categorization". In: *Proceedings of the Fourteenth International Conference on Machine Learning* 97 (1997).

[225] *YouTube*. Video sharing. http://www.youtube.com/.

[226] Clement T Yu and Gerard Salton. "Precision weighting—an effective automatic indexing method". In: *Journal of the ACM (JACM)* 23.1 (1976), pp. 76–88.

[227] Sławomir Zadrożny and Katarzyna Nowacka. "Fuzzy information retrieval model revisited". In: *Fuzzy Sets and Systems* 160.15 (2009), pp. 2173–2191.

[228] Oren Zamir and Oren Etzioni. "Web document clustering: A feasibility demonstration". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, pp. 46–54.

[229] *Zhang, L., N-Gram Extraction Tools*. N-Gram Extraction Tools. http://homepages.inf.ed.ac.uk/lzhang10/ngram.html.

[230] Ying Zhao and George Karypis. "Empirical and theoretical comparisons of selected criterion functions for document clustering". In: *Machine Learning* 55.3 (2004), pp. 311–331.

[231] Shi Zhong and Joydeep Ghosh. "Scalable, Balanced Model-based Clustering." In: *SDM*. SIAM. 2003, pp. 71–82.

[232] M. Zhu. *Recall, precision and average precision*. Tech. rep. 2004.

[233] Cai-Nicolas Ziegler et al. "Improving recommendation lists through topic diversification". In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 22–32.

[234] Δημήτριος Ζεϊμπέκης. "Φασματικές μέθοδοι ανάκτησης πληροφορίας, εργαλεία λογισμικού και εφαρμογές". PhD thesis. Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, 2009.

[235] Β. Τσόγκας. "Προσωποποιημένη προβολή περιεχομένου του διαδικτύου σε desktop εφαρμογή με τεχνικές ανάκτησης δεδομένων, προεπεξεργασίας κειμένου, αυτόματης κατηγοριοποίησης και εξαγωγής περίληψης". In: *Μεταπτυχιακή Διπλωματική Εργασία, Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Η/Υ και Πληροφορικής* (2008).

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.