4

Το Διανυσματικό Μοντέλο

Περιεχόμενα Κεφαλαίου

4.1 Εισαγωγή

Στο κεφάλαιο αυτό μελετούμε το *Διανυσματικό μοντέλο ανάκτησης*, που χρησιμοποιείται εκτενώς στα σύγχρονα συστήματα ανάκτησης. Το Διανυσματικό μοντέλο στηρίζεται στη διανυσματική αναπαράσταση εγγράφων και ερωτημάτων ενώ η ομοιότητα ενός κειμένου και ενός ερωτήματος προσδιορίζεται με τη χρήση ειδικών μετρικών ομοιότητας. Στο Κεφάλαιο 3 μελετήσαμε το εκτεταμένο Boolean μοντέλο, που επίσης χρησιμοποιεί διανύσματα για την αναπαράσταση των εγγράφων. Ωστόσο, το Διανυσματικό μοντέλο είναι γενικότερο, πιο απλό στη χρήση του και χαρακτηρίζεται από πολύ καλή αποτελεσματικότητα.

Αρχικά δίνουμε τις βασικές έννοιες που χαρακτηρίζουν το Διανυσματικό μοντέλο, μελετώντας τους τρόπους αναπαράστασης εγγράφων και ερωτημάτων. Στη συνέχεια εξετάζονται οι μετρικές ομοιότητας που χρησιμοποιούνται και αναλύεται ο τρόπος επεξεργασίας ερωτημάτων. Τέλος αναφέρονται επεκτάσεις του απλού διανυσματικού μοντέλου που έχουν προταθεί στη βιβλιογραφία με στόχο τη βελτίωση των επιδόσεων. Τονίζεται ότι το Διανυσματικό μοντέλο προτάθηκε πριν το εκτεταμένο Boolean μοντέλο. Αυτός είναι και ο λόγος του ότι μερικές από τις τεχνικές που χρησιμοποιεί το εκτεταμένο Boolean μοντέλο αρχικά είχαν εφαρμοστεί στο Διανυσματικό μοντέλο ανάκτησης.

4.2 Βασικές Έννοιες

Το Διανυσματικό μοντέλο ανάκτησης (vector space model) προτάθηκε από τον Salton (και τους συνεργάτες του) [51, 56], έναν από τους σημαντικότερους και πρωτοπόρους ερευνητές στην επιστημονική περιοχή της ανάκτησης πληροφορίας. Κάθε έγγραφο d_j της συλλογής αναπαριστάται με ένα διάνυσμα $\vec{d_j} = (w_{t_1,d_j}, w_{t_2,d_j}, ..., w_{t_M,d_j})$, όπου M είναι το πλήθος των όρων της συλλογής και w_{t_i,d_j} είναι το βάρος του όρου t_i στο έγγραφο d_j. Τονίζεται ότι η τιμή του M εξαρτάται από την προεπεξεργασία που έχουν υποστεί τα έγγραφα. Εάν έχουμε αναπαράσταση πλήρους εγγράφου, η τιμή του M θα είναι ο αριθμός όλων των μοναδικών λέξεων που εμφανίζονται σε όλα τα έγγραφα της συλλογής, ενώ εάν έχει προηγηθεί απαλοιφή άρθρων, ρημάτων και άλλων τύπων τότε η τιμή του M θα είναι σαφώς μικρότερη.

Στην πιο απλή του μορφή, το Διανυσματικό μοντέλο θεωρεί ότι τα βάρη w_{t_i,d_j} είναι είτε 0 είτε 1. Σε περίπτωση που ο όρος t_i περιέχεται στο έγγραφο d_j έχουμε $w_{t_i,d_j} = 1$, ενώ σε διαφορετική περίπτωση έχουμε $w_{t_i,d_j} = 0$. Η διανυσματική αναπαράσταση των εγγράφων γίνεται περισσότερο κατανοητή εάν κατασκευάσουμε τον πίνακα όρων-εγγράφων, τον οποίο καλούμε πίνακα D. Ο πίνακας αυτός έχει

d_1: Ο κομήτης του Χάλλεϋ μας επισκέπτεται περίπου κάθε εβδομήντα έξι χρόνια.
d_2: Ο κομήτης του Χάλλεϋ ανακαλύφθηκε από τον αστρονόμο Έντμοντ Χάλλεϋ.
d_3: Ένας κομήτης διαγράφει ελλειπτική τροχιά.
d_4: Ο πλανήτης Άρης έχει δύο φυσικούς δορυφόρους, το Δείμο και το Φόβο.
d_5: Ο πλανήτης Δίας έχει εξήντα τρεις γνωστούς φυσικούς δορυφόρους.
d_6: Ο Ήλιος είναι ένας αστέρας.
d_7: Ο Άρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

Σχήμα 4.1: *Συλλογή εγγράφων.*

όρος	d_1	d_2	d_3	d_4	d_5	d_6	d_7
κομήτης	1	1	1	0	0	0	0
πλανήτης	0	0	0	1	1	0	1
Χάλλεϋ	1	1	0	0	0	0	0
Άρης	0	0	0	1	0	0	1
Δίας	0	0	0	0	1	0	0
τροχιά	0	0	1	0	0	0	0

Πίνακας 4.1: *Πίνακας όρων-εγγράφων με δυαδικά βάρη.*

M γραμμές και N στήλες, όπου N το πλήθος των εγγράφων της συλλογής και M το πλήθος των όρων της συλλογής. Το κελί του πίνακα στη γραμμή i και τη στήλη j είναι το βάρος w_{t_i,d_j} του όρου t_i στο έγγραφο d_j [1].

Ο πίνακας όρων-εγγράφων για τη μικρή συλλογή εγγράφων του Σχήματος 4.1 δίνεται στον Πίνακα 4.1. Έχουμε θεωρήσει, για λόγους απλότητας, ότι οι όροι που μας ενδιαφέρουν και που χαρακτηρίζουν τη συλλογή εγγράφων είναι οι: κομήτης, πλανήτης, Χάλλεϋ, Άρης, Δίας, τροχιά. Ωστόσο, χωρίς πρόβλημα θα μπορούσαμε να χρησιμοποιήσουμε όλες τις λέξεις που εμφανίζονται στα έγγραφα. Κάθε στήλη του πίνακα αποτελεί το διάνυσμα για το αντίστοιχο έγγραφο. Για παράδειγμα, το διάνυσμα $\vec{d_1}$ του κειμένου d_1 είναι:

$$\vec{d_1} = (1, 0, 1, 0, 0, 0)$$

[1] Στη βιβλιογραφία πολλές φορές χρησιμοποιείται ο πίνακας εγγράφων-όρων που είναι ο ανάστροφος του πίνακα όρων-εγγράφων.

Στο παράδειγμα που μελετούμε, τα διανύσματα των εγγράφων ορίζονται στο χώρο των έξι (6) διαστάσεων. Ο αριθμός των διαστάσεων ισούται με τον αριθμό των όρων που χρησιμοποιούνται για την περιγραφή του περιεχομένου των εγγράφων. Συνήθως, ο αριθμός των όρων είναι πολύ μεγάλος, με αποτέλεσμα τα διανύσματα να ορίζονται σε χώρους πολλών διαστάσεων. Επειδή ο μεγάλος αριθμός διαστάσεων δημιουργεί προβλήματα στην αποδοτική επεξεργασία των ερωτημάτων, έχουν προταθεί τεχνικές μείωσης της διαστασιμότητας. Μία από τις τεχνικές αυτές μελετάται σε ξεχωριστό κεφάλαιο.

Η διανυσματική αναπαράσταση των εγγράφων που εξατάσαμε προηγουμένως χρησιμοποιεί δυαδικά βάρη (0 ή 1). Η χρήση δυαδικών βαρών δε λαμβάνει υπόψη τη συχνότητα εμφάνισης του όρου στο έγγραφο, ούτε τον αριθμό των εγγράφων στα οποία εμφανίζεται ο συγκεκριμένος όρος. Εάν ένας όρος εμφανίζεται πολύ συχνά σε ένα έγγραφο τότε η σημαντικότητά του για το έγγραφο αυτό θα πρέπει να είναι μεγαλύτερη από αυτήν ενός όρου που εμφανίζεται μόνο μία φορά. Επίσης, αν ένας όρος εμφανίζεται σε πολλά έγγραφα, τότε δεν αποτελεί χαρακτηριστικό ενός κειμένου. Οι δύο αυτές παρατηρήσεις οδήγησαν τους ερευνητές στη μελέτη εναλλακτικών μεθόδων προσδιορισμού των βαρών w_{t_i,d_j} όπου t_i είναι κάποιος όρος και d_j ένα έγγραφο της συλλογής. Στο Κεφάλαιο 3 μελετήσαμε έναν τρόπο προσδιορισμού βαρών για το εκτεταμένο Boolean μοντέλο ο οποίος στηρίζεται στο σχήμα tf-idf (term frequency - inverse document frequency). Η μέθοδος αυτή εφαρμόστηκε αρχικά για το Διανυσματικό μοντέλο ανάκτησης και για λόγους πληρότητας εξετάζεται συνοπτικά στη συνέχεια. Ο Πίνακας 4.2 περιέχει τα σύμβολα που χρησιμοποιούνται στη συνέχεια.

4.2.1 Υπολογισμός Σημαντικότητας Όρων

Έστω t ένας όρος και d ένα έγγραφο της συλλογής. Η *συχνότητα* (frequency) εμφάνισης του όρου t στο d συμβολίζεται με $f_{t,d}$ και προσδιορίζει τον αριθμό των εμφανίσεων του όρου στο συγκεκριμένο έγγραφο. Για παράδειγμα, με βάση τη συλλογή των εγγράφων που χρησιμοποιούμε, η συχνότητα εμφάνισης του όρου Χάλλεϋ στο έγγραφο d_2 είναι 2, καθώς έχουμε δύο εμφανίσεις του όρου στο έγγραφο.

Η συχνότητα εμφάνισης του όρου στο έγγραφο μπορεί να χρησιμοποιηθεί για να δηλώσει τη σημαντικότητα (βάρος) του όρου για το έγγραφο. Επομένως, μία πρώτη προσέγγιση για τον προσδιορισμό του βάρους $w_{t,d}$ είναι να χρησιμοποιήσουμε τον τύπο:

$$w_{t,d} = f_{t,d} \tag{4.1}$$

σύμβολο	περιγραφή
$\mathcal{D}$	συλλογή εγγράφων
N	πλήθος εγγράφων της συλλογής ($N = \|\mathcal{D}\|$)
$\mathcal{T}$	σύνολο μοναδικών όρων της συλλογής
M	πλήθος όρων ($M = \|\mathcal{T}\|$)
t, t_i	ο όρος t, ο i-οστός όρος (t_i)
d, d_j	το έγγραφο d, το j-οστό έγγραφο της συλλογής (d_j)
q	έγγραφο ερωτήματος
$\mathcal{T}_d$	σύνολο μοναδικών όρων στο έγγραφο d της συλλογής
$\mathcal{T}_q$	σύνολο μοναδικών όρων στο έγγραφο ερωτήματος q
$\mathcal{T}_{q,d}$	σύνολο μοναδικών όρων στο q και d ($\mathcal{T}_{q,d} = \mathcal{T}_q \cap \mathcal{T}_d$)
$f_{t,d}$	αριθμός εμφανίσεων του όρου t στο έγγραφο d
$f_{t,q}$	αριθμός εμφανίσεων του όρου t στο ερώτημα q
f_d	αριθμός εμφανίσεων όλων των όρων στο έγγραφο d ($\sum f_{t,d}$)
$nf_{t,d}$	κανονικοποιημένη συχνότητα εμφάνισης του όρου t στο έγγραφο d
idf_t	αντίστροφη συχνότητα εγγράφων για τον όρο t
n_t	πλήθος εγγράφων που περιέχουν τον όρο t
$nidf_t$	κανονικοποιημένη αντίστροφη συχνότητα εγγράφων για τον όρο t
$w_{t,d}$	σημαντικότητα (βάρος) του όρου t στο έγγραφο d της συλλογής
$w_{t,q}$	σημαντικότητα (βάρος) του όρου t στο έγγραφο ερωτήματος q
$\vec{d_j}$	το διάνυσμα που αντιστοιχεί στο έγγραφο d_j
L_j, L_q	μήκος εγγράφου συλλογής και ερωτήματος
$\|\vec{d_j}\|$	μέτρο του διανύσματος $\vec{d_j}$
θ	γωνία που σχηματίζεται μεταξύ δύο διανυσμάτων
$\cos(\theta)$	το συνημίτονο της γωνίας θ
$\vec{d_j} \bullet \vec{d_k}$	εσωτερικό γινόμενο διανυσμάτων $\vec{d_j}$ και $\vec{d_k}$

Πίνακας 4.2: *Σύμβολα και περιγραφές.*

Αν χρησιμοποιήσουμε τον παραπάνω τύπο για τον υπολογισμό του βάρους $w_{t,d}$, τότε όροι που εμφανίζονται σε μεγάλα έγγραφα ενδεχομένως να έχουν και μεγαλύτερο βάρος, διότι αυξάνεται η πιθανότητα ύπαρξής τους στο έγγραφο. Για το λόγο αυτό, και για να μη γίνεται διάκριση μεταξύ μικρών και μεγάλων εγγράφων, χρησιμοποιείται η *κανονικοποιημένη συχνότητα εμφάνισης* (normalized frequency) που συμβολίζεται με $nf_{t,d}$ η οποία ορίζεται ως εξής:

$$nf_{t,d} = \frac{f_{t,d}}{max_x\{f_{x,d}\}}$$

Το πλήθος των εμφανίσεων ενός όρου σε ένα έγγραφο δηλώνει τη σημαντικότητα του όρου για το έγγραφο αυτό. Ωστόσο, θα πρέπει να παρατηρήσουμε ότι όροι που εμφανίζονται σε πολλά έγγραφα έχουν μικρή διακριτική ικανότητα. Αυτό σημαίνει, ότι αν και οι όροι αυτοί μπορεί να εμφανίζονται πολλές φορές σε κάποια έγγραφα, το γεγονός ότι εμφανίζονται σε πολλά από αυτά μειώνει τη σημαντικότητά τους. Για παράδειγμα, σε μία συλλογή εγγράφων που περιλαμβάνει άρθρα από την επιστημονική περιοχή της ανάκτησης πληροφορίας, είναι λογικό κάποιο έγγραφο να περιέχει πολλές φορές τον όρο **ανάκτηση**. Όμως, είναι επίσης λογικό ο όρος **ανάκτηση** να εμφανίζεται στα περισσότερα έγγραφα της συλλογής. Επομένως, τελικά το βάρος του όρου θα πρέπει να είναι μικρό, καθώς δεν αποτελεί αντιπροσωπευτική λέξη για κανένα έγγραφο της συλλογής. Η παρατήρηση αυτή οδήγησε τους ερευνητές στη χρήση ενός νέου παράγοντα στον υπολογισμό των βαρών $w_{t,d}$. Ο νέος παράγοντας καλείται *αντίστροφη συχνότητα εγγράφων* (inverse document frequency) και συμβολίζεται με idf_t. Αν συμβολίσουμε με N το πλήθος των εγγράφων της συλλογής και με n_t το πλήθος των εγγράφων που περιέχουν τον όρο t, τότε ο παράγοντας αυτός υπολογίζεται για κάθε όρο ξεχωριστά ως εξής:

$$idf_t = \ln\left(\frac{N}{n_t}\right) \tag{4.2}$$

Χρησιμοποιώντας την κανονικοποιημένη συχνότητα εμφάνισης και την αντίστροφη συχνότητα εγγράφων, προκύπτει ένας νέος τρόπος υπολογισμού των βαρών $w_{t,d}$ που είναι:

$$w_{t,d} = nf_{t,d} \cdot idf_t = \frac{f_{t,d}}{max_x\{f_{x,d}\}} \cdot \ln\left(\frac{N}{n_t}\right) \tag{4.3}$$

Παρατηρήστε ότι ο παράγοντας idf_t δεν είναι κανονικοποιημένος. Η κανονικοποίηση του παράγοντα αυτού μπορεί να πραγματοποιηθεί διαιρώντας με το λογάριθμο του πλήθους των εγγράφων, σύμφωνα με τον Turtle [69]. Με τον τρόπο αυτό προκύπτει η *κανονικοποιημένη αντίστροφη συχνότητα εγγράφων* (normalized inverse document frequency) η οποία υπολογίζεται ως εξής:

$$nidf_t = \frac{idf_t}{\ln(N)} = \frac{\ln(N/n_t)}{\ln(N)} \tag{4.4}$$

Χρησιμοποιώντας τους ορισμούς για τους παράγοντες $nf_{t,d}$ και $nidf_t$ προκύπτει ο ακόλουθος τρόπος υπολογισμού των βαρών:

$$w_{t,d} = nf_{t,d} \cdot nidf_t = \frac{f_{t,d}}{max_x\{f_{x,d}\}} \cdot \frac{\ln(N/n_t)}{\ln(N)} \tag{4.5}$$

Στη βιβλιογραφία έχουν προταθεί διάφορες παραλλαγές του τρόπου προσδιορισμού των βαρών $w_{t,d}$ χρησιμοποιώντας ως βάση το σχήμα tf-idf. Για παράδειγμα, οι Salton και Buckley [53] προτείνουν τον ακόλουθο τύπο για τον υπολογισμό των βαρών σε περιπτώσεις όπου η συλλογή εγγράφων αποτελείται από συμβατικά έγγραφα ή από περιλήψεις:

$$\boxed{w_{t,d} = \frac{f_{t,d} \cdot \ln(\frac{N}{n_t})}{\sqrt{\sum_{x \in \mathcal{T}_d} \left(f_{x,d} \cdot \ln(\frac{N}{n_x})\right)^2}}} \tag{4.6}$$

Ένα έγγραφο ερωτήματος q μπορεί να θεωρηθεί και αυτό ως ένα τυπικό έγγραφο και επομένως για τον προσδιορισμό των βαρών $w_{t,q}$ μπορεί να χρησιμοποιηθεί ένας από τους τύπους που αναφέρθηκαν προηγουμένως (π.χ. ο τύπος 4.6). Ωστόσο, η μελέτη των Salton και Buckley [53] έδειξε ότι είναι καλύτερα να χρησιμοποιηθεί ο ακόλουθος τύπος ο οποίος δίνει καλύτερα αποτελέσματα ως προς την ακρίβεια για πολλές γνωστές συλλογές εγγράφων:

$$\boxed{w_{t,q} = \left(0.5 \cdot \frac{f_{t,q}}{max_x\{f_{x,q}\}} + 0.5\right) \cdot \ln\left(\frac{N}{n_t}\right)} \tag{4.7}$$

Οι Zobel και Moffat [74] έχουν μελετήσει την αποτελεσματικότητα πολλών διαφορετικών σχημάτων tf-idf χρησιμοποιώντας συλλογές εγγράφων από το TREC. Η διαφοροποίηση μεταξύ των σχημάτων αυτών οφείλεται στον τρόπο ορισμού της σχετικής συχνότητας εμφάνισης και της αντίστροφης συχνότητας εγγράφων. Από τη μελέτη αυτή προέκυψε το συμπέρασμα ότι δεν υπάρχει κάποιος συνδυασμός που να έχει τα καλύτερα αποτελέσματα για όλα τα ερωτήματα και όλες τις συλλογές εγγράφων. Μερικές από τις εναλλακτικές μεθόδους που παρουσιάζονται στην εργασία [74] θα μελετηθούν παρακάτω, αφού πρώτα εξετάσουμε τον τρόπο υπολογισμού της ομοιότητας μεταξύ εγγράφων.

4.2.2 Υπολογισμός Ομοιότητας Εγγράφων

Λαμβάνοντας υπόψη την προηγούμενη περιγραφή, το ερώτημα που προκύπτει είναι το εξής: με ποιον τρόπο θα ποσοτικοποιήσουμε την ομοιότητα μεταξύ ενός ερωτήματος q και ενός κειμένου d; Θυμίζουμε ότι στην περίπτωση του απλού Boolean μοντέλου η μετρική της ομοιότητας $S_{vector}(q, d)$ μπορεί να λάβει μόνο τις τιμές 0 και 1, ενώ στην περίπτωση του εκτεταμένου Boolean μοντέλου, η ομοιότητα εκφράζεται με μία τιμή στο διάστημα [0,1]. Στην περίπτωση του διανυσματικού μοντέλου, η ομοιότητα ερωτήματος-κειμένου είναι πάλι μία τιμή από το διάστημα [0,1], η οποία όμως υπολογίζεται με εντελώς διαφορετικό τρόπο από αυτόν που χρησιμοποιείται στο εκτεταμένο Boolean μοντέλο. Σημειώνεται ότι ο τρόπος υπολογισμού της ομοιότητας στο Διανυσματικό μοντέλο είναι ανεξάρτητος του τρόπου προσδιορισμού των βαρών.

Από την προηγούμενη συζήτηση προκύπτει ότι ένα έγγραφο μπορεί να θεωρηθεί ως ένα διάνυσμα σε έναν πολυδιάστατο χώρο. Για παράδειγμα, η κάθε στήλη του Πίνακα 4.1 αντιστοιχεί σε ένα έγγραφο της συλλογής. Επομένως, το κάθε έγγραφο μπορεί να θεωρηθεί ως διάνυσμα στο χώρο των 6 διαστάσεων. Ο αριθμός των διαστάσεων καθορίζεται από το πλήθος των όρων που χρησιμοποιούνται για την περιγραφή των εγγράφων. Αν συμβολίσουμε με $\vec{d_j}$ το διάνυσμα του εγγράφου d_j, τότε:

$$\vec{d_j} = (w_{t_1,d_j}, w_{t_2,d_j}, ..., w_{t_M,d_j})$$

όπου M είναι ο συνολικός αριθμός των όρων που χρησιμοποιείται για την αναπαράσταση των εγγράφων.

Ακολουθώντας την ίδια τακτική, μπορούμε να εκφράσουμε το διάνυσμα ενός εγγράφου ερωτήματος q το οποίο εκφράζει την ανάγκη πληροφορίας κάποιου χρήστη. Το βάρος του όρου t στο έγγραφο ερωτήματος q συμβολίζεται με $w_{t,q}$. Η βασική διαφορά του εγγράφου ερωτήματος από ένα έγγραφο της συλλογής είναι ότι το πρώτο είναι συνήθως πολύ μικρότερο από το δεύτερο.

Από τη στιγμή που έχουμε στη διάθεσή μας τις διανυσματικές αναπαραστάσεις των εγγράφων της συλλογής και του εγγράφου του ερωτήματος το εύλογο ερώτημα που προκύπτει είναι πως μπορεί να προσδιοριστεί ο βαθμός ομοιότητας μεταξύ ενός ερωτήματος και ενός εγγράφου της συλλογής. Μία απλή και προφανής μέθοδος υπολογισμού της ομοιότητας είναι με τη χρήση της Ευκλείδιας απόστασης μεταξύ των αντίστοιχων διανυσματικών αναπαραστάσεων. Αν συμβολίσουμε με $\vec{q}$ και $\vec{d}$ το διάνυσμα του εγγράφου του ερωτήματος q και του εγγράφου d της συλλογής, τότε ορίζουμε ως $D_e(q, d)$ την Ευκλείδια απόστασή τους:

$$\boxed{D_e(q,d) = \sqrt{\sum_{i=1}^{M} |w_{t_i,q} - w_{t_i,d}|^2}} \tag{4.8}$$

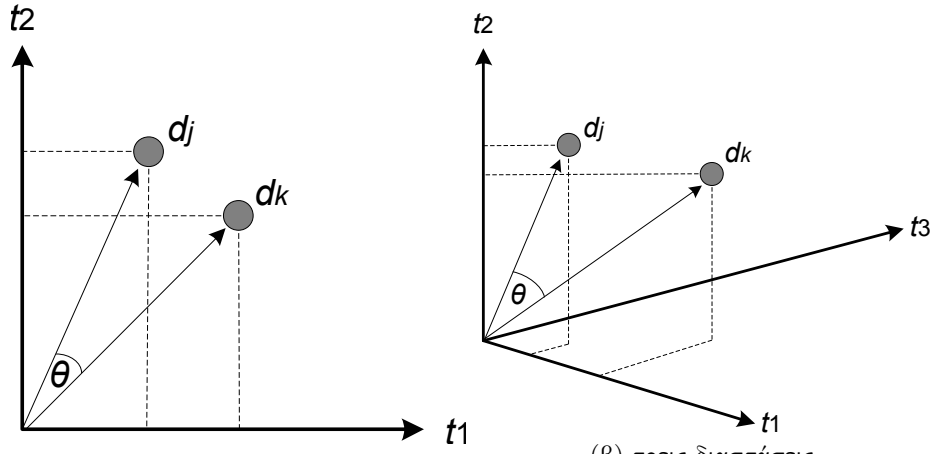
Όσο αυξάνει ή τιμή $D_e(q,d)$ τόσο θεωρούμε ότι μειώνεται η ομοιότητα μεταξύ q και d. Με μια πρώτη ματιά, φαίνεται ότι αυτό το μέτρο ανομοιότητας καλύπτει τις ανάγκες μας. Ωστόσο, υπάρχει ένα σοβαρό πρόβλημα που αξίζει να σχολιαστεί. Συνήθως, το έγγραφο του ερωτήματος είναι αρκετά μικρότερο σε σχέση με τα έγγραφα της συλλογής. Αυτό σημαίνει ότι οι περισσότερες συνιστώσες του διανύσματος $\vec{q}$ θα είναι μηδενικές. Επίσης, όσο μεγαλύτερο είναι ένα έγγραφο, τόσο αυξάνει ο αριθμός των μη-μηδενικών συνιστωσών. Αυτό σημαίνει ότι στην ουσία τιμωρούνται τα μεγαλύτερα έγγραφα της συλλογής τα οποία ακόμη και αν σχετίζονται με το ερώτημα, λόγω της Ευκλείδιας απόστασης, η απόσταση τους από το ερώτημα θα είναι μεγάλη.

Μία άλλη προσέγγιση για τον υπολογισμό της ομοιότητας μεταξύ q και d είναι να χρησιμοποιηθεί το *εσωτερικό γινόμενο* (inner product) των διανυσμάτων, το οποίο συμβολίζεται με $\vec{q} \bullet \vec{d}$. Αν συμβολίσουμε με $S_{inner}(q,d)$ τη συνάρτηση που επιστρέφει την ομοιότητα, τότε έχουμε:

$$\boxed{S_{inner}(q,d) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{M} w_{t_i,q} \cdot w_{t_i,d}} \tag{4.9}$$

Όσο πιο όμοια είναι τα διανύσματα $\vec{q}$ και $\vec{d}$ τόσο η συνάρτηση $S_{inner}(q,d)$ λαμβάνει μεγαλύτερες τιμές. Το αντίστροφο συμβαίνει όταν τα διανύσματα είναι ανόμοια. Η εφαρμογή του εσωτερικού γινομένου για τον υπολογισμό της ομοιότητας έχει το μειονέκτημα ότι τιμωρούνται τα μικρότερα έγγραφα, σε αντίθεση με την Ευκλείδια απόσταση η οποία τιμωρεί τα μεγαλύτερα. Για να ξεπεραστεί αυτό το πρόβλημα, προτάθηκε η χρήση μίας συνάρτησης ομοιότητας που βασίζεται στο συνημίτονο της γωνίας που σχηματίσουν τα δύο διανύσματα στο χώρο. Ας εξετάσουμε τον τρόπο κατασκευής της συνάρτησης αυτής. Θα χρειαστούμε τον ορισμό του μέτρου ενός διανύσματος. Το *μέτρο* του διανύσματος $\vec{d}$, συμβολίζεται με $|\vec{d}|$ και δίνεται από τον ακόλουθο τύπο με άμεση εφαρμογή του Πυθαγόρειου θεωρήματος:

$$|\vec{d}| = \sqrt{\sum_{i=1}^{M} w_{t_i,d}^2}$$



(α) δύο διαστάσεις (β) τρεις διαστάσεις

Σχήμα 4.2: *Γωνία μεταξύ διανυσμάτων σε δύο και τρεις διαστάσεις.*

Είναι γνωστό ότι το εσωτερικό γινόμενο δύο διανυσμάτων ισούται με το γινόμενο των μέτρων τους επί το συνημίτονο της μεταξύ τους γωνίας. Χρησιμοποιώντας την ιδιότητα αυτή για τα διανύσματα $\vec{q}$ και $\vec{d}$ έχουμε:

$$\vec{q} \bullet \vec{d} = |\vec{q}| \cdot |\vec{d}| \cdot \cos(\theta)$$

Η επίλυση της παραπάνω εξίσωσης ως προς $\cos(\theta)$ δίνει έναν τρόπο υπολογισμού του συνημιτόνου της γωνίας που σχηματίζεται μεταξύ των διανυσμάτων. Με τον τρόπο αυτό έχουμε έναν εναλλακτικό τρόπο υπολογισμού της ομοιότητας. Όταν μικραίνει η γωνία θ, μεγαλώνει η ποσότητα $\cos(\theta)$ και αντιστρόφως. Όταν τα δύο διανύσματα ταυτίζονται, τότε έχουμε $\theta = 0$ μοίρες επομένως $\cos(\theta) = 1$. Από την άλλη πλευρά, όταν τα διανύσματα είναι κάθετα μεταξύ τους, τότε $\theta = 90$ μοίρες, επομένως $\cos(\theta) = 0$. Σημειώνεται ότι η γωνία μεταξύ των διανυσμάτων δεν μπορεί να ξεπερνά τις 90 μοίρες εφόσον οι συντεταγμένες είναι θετικοί πραγματικοί αριθμοί και επομένως εστιάζουμε στο άνω δεξί τεταρτημόριο του συστήματος συντεταγμένων. Αν συμβολίσουμε με S_{cosine} τη συνάρτηση ομοιότητας συνημιτόνου, τότε έχουμε:

$$S_{cosine}(q,d) = \cos(\theta) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}| \cdot |\vec{d}|} = \frac{\sum_{i=1}^{M} w_{t_i,q} \cdot w_{t_i,d}}{\sqrt{\sum_{i=1}^{M} w_{t_i,q}^2} \cdot \sqrt{\sum_{i=1}^{M} w_{t_i,d}^2}} \tag{4.10}$$

Στο Σχήμα 4.2 παρουσιάζεται ένα απλό παράδειγμα προσδιορισμού της γωνίας μεταξύ δύο διανυσμάτων για το χώρο των δύο και των τριών διαστάσεων. Στην πραγματικότητα η διαστασιμότητα του χώρου είναι πολύ μεγαλύτερη και καθορίζεται από το πλήθος των όρων που χρησιμοποιούνται για την αναπαράσταση των εγγράφων.

Η εφαρμογή του τύπου του συνημιτόνου είναι ανεξάρτητη από τον τρόπο υπολογισμού των βαρών $w_{t,d}$. Ο τύπος εφαρμόζεται τόσο στην περίπτωση δυαδικών βαρών όσο και στην περίπτωση που τα βάρη είναι πραγματικοί αριθμοί.

4.3 Εναλλακτικές Μέθοδοι

Στις προηγούμενες παραγράφους μελετήσαμε τη βασική μεθοδολογία που ακολουθείται από το Διανυσματικό μοντέλο ανάκτησης, που περιλαμβάνει δύο άξονες: (α) τον προσδιορισμό της σημαντικότητας των όρων στα έγγραφα και (β) τον υπολογισμό του βαθμού ομοιότητας μεταξύ εγγράφων. Και για τους δύο άξονες έχουν προταθεί διάφορες εναλλακτικές λύσεις, μερικές από τις οποίες εξετάζονται στη συνέχεια.

Υπενθυμίζεται ότι στη γενική περίπτωση, οι ποσότητες $w_{t,d}$ (το βάρος του όρου t στο έγγραφο d) και $w_{t,q}$ (το βάρος του όρου t στο έγγραφο ερωτήματος q) υπολογίζονται με βάση το σχήμα tf-idf:

$$w_{t,d} = tf_{t,d} \cdot idf_t \quad \text{και} \quad w_{t,q} = tf_{t,q} \cdot idf_t$$

Χρησιμοποιώντας διαφορετικούς τρόπους υπολογισμού των ποσοτήτων $tf_{t,d}$ και idf_t προκύπτει ένα πλήθος διαφορετικών εκφράσεων για τις ποσότητες $w_{t,d}$ και $w_{t,q}$. Στους Πίνακες 4.3 και 4.4 δίνονται μερικές από τις εκφράσεις υπολογισμού των ποσοστήτων tf και idf αντίστοιχα που έχουν μελετηθεί στη βιβλιογραφία. Επίσης, ο Πίνακας 4.5 παρουσιάζει διαφορετικούς τρόπους υπολογισμού του *μήκους* ενός εγγράφου, ενώ στον Πίνακα 4.6 δίδονται οι βασικότερες μετρικές υπολογισμού της ομοιότητας μεταξύ ενός εγγράφου της συλλογής και ενός εγγράφου ερωτήματος. Τέλος, στον Πίνακα 4.7 δίνονται οι δύο εναλλακτικές λύσεις που έχουν προταθεί για τον υπολογισμό των βαρών $w_{t,q}$ και $w_{t,d}$. Είναι προφανές, ότι υπάρχουν πολλοί διαφορετικοί συνδυασμοί που προκύπτουν επιλέγοντας διαφορετικούς σχηματισμούς για την κάθε ποσότητα. Έτσι, ανάλογα με την έκφραση που θέλουμε να χρησιμοποιήσουμε, επιλέγεται η αντίστοιχη γραμμή από τους πίνακες. Η αποτελεσματικότητα μίας πληθώρας συνδυασμών έχει μελετηθεί πειραματικά στην εργασία [74]. Ένα από τα σημαντικά αποτελέσματα της πειραματικής αυτής μελέτης είναι ότι καμία μέθοδος δεν παρουσιάζει την καλύτερη

περιγραφή	$tf_{t,d}$
δυαδικός σχηματισμός	1 ή 0
συνήθης σχηματισμός	$f_{t,d}$
λογαριθμικός σχηματισμός	$1 + \ln(f_{t,d})$
κανονικοποιημένος σχηματισμός	$\frac{f_{t,d}}{\max_x\{f_{x,d}\}}$
εναλλακτικός κανονικοποιημένος σχηματισμός Το C είναι μία σταθερά η οποία αν λάβει τιμές μεταξύ 0.3 και 0.5 έχει τα καλύτερα αποτελέσματα	$C + (1 - C) \cdot \frac{f_{t,d}}{\max_x\{f_{x,d}\}}$

Πίνακας 4.3: *Εναλλακτικές εκφράσεις υπολογισμού της ποσότητας $tf_{t,d}$.*

περιγραφή	idf_t
δυαδικός σχηματισμός	1
1ος λογαριθμικός σχηματισμός	$\ln\left(\frac{N}{n_t}\right)$
2ος λογαριθμικός σχηματισμός	$\ln\left(1 + \frac{N}{n_t}\right)$
3ος λογαριθμικός σχηματισμός	$\frac{\ln(N/n_t)}{\ln(N)}$
υπερβολικός σχηματισμός	$\frac{1}{n_t}$
1ος κανονικοποιημένος σχηματισμός	$\ln\left(1 + \frac{\max_x\{n_x\}}{n_t}\right)$
2ος κανονικοποιημένος σχηματισμός	$\ln\left(\frac{N-n_t}{n_t}\right)$

Πίνακας 4.4: *Εναλλακτικές εκφράσεις υπολογισμού της ποσότητας idf_t.*

αποτελεσματικότητα σε όλες τις περιπτώσεις.

Στη συνέχεια δίνεται ένα συγκεκριμένο παράδειγμα προσδιορισμού των ποσοτήτων. Έστω ότι θέλουμε να εκφράσουμε την ομοιότητα μεταξύ ενός ερωτήματος q και ενός εγγράφου της συλλογής d χρησιμοποιώντας τη μέθοδο του συνημιτόνου. Αυτό σημαίνει ότι πρέπει να επιλεγεί η δεύτερη γραμμή του Πίνακα 4.6. Για τη χρήση του 2ου λογαριθμικού σχηματισμού για τον υπολογισμό της ποσότητας idf_t πρέπει να επιλεγεί η τρίτη γραμμή του Πίνακα 4.4. Η χρήση του σχηματισμού tf για τον υπολογισμό της ποσότητας $w_{t,d}$ προϋποθέτει την επιλογή της πρώτης γραμμής του Πίνακα 4.7, ενώ η χρήση του λογαριθμικού σχηματισμού για την ποσότητα $tf_{t,d}$ προϋποθέτει την επιλογή της τρίτης γραμμής του Πίνακα 4.3. Για τη χρήση του διανυσματικού σχηματισμού για τον υπολογισμό της ποσότητας L_d πρέπει να επιλέξουμε τη δεύτερη γραμμή του Πίνακα 4.5. Παρατηρούμε ότι έως

περιγραφή	L_d
μοναδιαίος σχηματισμός	1
διανυσματικός σχηματισμός	$\sqrt{\sum_{x \in \mathcal{T}_d} w_{x,d}^2}$
1ος προσεγγιστικός σχηματισμός	$\lvert\mathcal{T}_d\rvert$
2ος προσεγγιστικός σχηματισμός	$\sqrt{\lvert\mathcal{T}_d\rvert}$
3ος προσεγγιστικός σχηματισμός	$\log_2(\lvert\mathcal{T}_d\rvert)$
4ος προσεγγιστικός σχηματισμός	f_d
5ος προσεγγιστικός σχηματισμός	$\sqrt{f_d}$

Πίνακας 4.5: *Εναλλακτικές εκφράσεις υπολογισμού του μήκους L_d (L_q) ενός εγγράφου d (ερωτήματος q).*

περιγραφή	$S_{vector}(q,d)$
εσωτερικό γινόμενο	$\sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})$
μέθοδος συνημιτόνου	$\frac{1}{L_q \cdot L_d} \cdot \sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})$
απλή πιθανοτική μετρική	$\sum_{t \in \mathcal{T}_{q,d}} (C + idf_t)$
σύνθετη πιθανοτική μετρική	$\sum_{t \in \mathcal{T}_{q,d}} (C + idf_t) \cdot tf_{t,d}$
εναλλακτικό εσωτερικό γινόμενο	$\sum_{t \in \mathcal{T}_{q,d}} \frac{w_{t,d}}{L_d}$
μέθοδος Dice	$\frac{2}{L_q^2 + L_d^2} \cdot \sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})$
μέθοδος Jaccard	$\frac{\sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})}{L_q^2 + L_d^2 - \sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})}$
μέθοδος επικάλυψης	$\frac{\sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})}{\min(L_q^2, L_d^2)}$

Πίνακας 4.6: *Εναλλακτικές εκφράσεις υπολογισμού ομοιότητας $S_{vector}(q,d)$.*

περιγραφή	$w_{t,d}$
σχηματισμός tf	$tf_{t,d}$
σχηματισμός tf-idf	$tf_{t,d} \cdot idf_t$

Πίνακας 4.7: *Εναλλακτικές εκφράσεις υπολογισμού βαρών $w_{t,d}$ (και $w_{t,q}$).*

τώρα έχουμε προσδιορίσει πλήρως τις ποσότητες που σχετίζονται με το έγγραφο d της συλλογής. Συνεχίζουμε με την επιλογή του τρόπου προσδιορισμού των ποσοτήτων που αφορούν στο ερώτημα q. Επιλέγουμε το σχηματισμό tf-idf για την ποσότητα $w_{t,q}$ (δεύτερη γραμμή του Πίνακα 4.7), επιλέγουμε το λογαριθμικό

περιγραφή	έκφραση
συνάρτηση ομοιότητας	$S_{vector}(q,d) = \frac{1}{L_q \cdot L_d} \cdot \sum_{t \in \mathcal{T}_{q,d}} (w_{t,q} \cdot w_{t,d})$
υπολογισμός idf_t	$idf_t = \ln\left(1 + \frac{N}{n_t}\right)$
υπολογισμός $w_{t,d}$	$w_{t,d} = tf_{t,d}$
υπολογισμός $tf_{t,d}$	$tf_{t,d} = 1 + \ln(f_{t,d})$
υπολογισμός L_d	$L_d = \sqrt{\sum_{x \in \mathcal{T}_d} w_{x,d}^2}$
υπολογισμός $w_{t,q}$	$w_{t,q} = tf_{t,q} \cdot idf_t$
υπολογισμός $tf_{t,q}$	$tf_{t,q} = 1 + \ln(f_{t,q})$
υπολογισμός L_q	$L_q = 1$

Πίνακας 4.8: *Παράδειγμα προσδιορισμού συγκεκριμένου μοντέλου.*

σχηματισμό για την ποσότητα $tf_{t,q}$ (τρίτη γραμμή του Πίνακα 4.3) και τέλος επιλέγουμε το μοναδιαίο σχηματισμό για την ποσότητα L_q (πρώτη γραμμή του Πίνακα 4.5). Με βάση τις προηγούμενες επιλογές, παρατηρούμε ότι έχουν προσδιοριστεί όλες οι ποσότητες. Το μοντέλο που σχηματίζεται συνοψίζεται στον Πίνακα 4.8.

Παράδειγμα 4.1

Ο τρόπος υπολογισμού των παραμέτρων του μοντέλου θα γίνει περισσότερο κατανοητός με ένα παράδειγμα. Έστω ότι ένας χρήστης ενδιαφέρεται για την εύρεση πληροφοριών σχετικά με τον κομήτη του Χάλλεϋ. Αν συμβολίσουμε με q την αναπαράσταση της πληροφοριακής ανάγκης του χρήστη, τότε έχουμε q = {κομήτης, Χάλλεϋ}. Είναι προφανές ότι το ερώτημα αποτελείται από δύο όρους, t_1 = κομήτης και t_2 = Χάλλεϋ. Στόχος μας είναι να υπολογίσουμε το βαθμό ομοιότητας μεταξύ του ερωτήματος q και των εγγράφων της συλλογής του Σχήματος 4.1. Προφανώς, δε θα ασχοληθούμε καθόλου με τα έγγραφα που δεν περιέχουν κανέναν από τους δύο όρους του ερωτήματος. Αυτό σημαίνει ότι θα πρέπει να υπολογίσουμε όλες τις τιμές για τις ποσότητες που εμφανίζονται στον Πίνακα 4.8 για τα έγγραφα που περιέχουν έναν ή και τους δύο όρους του ερωτήματος. Τα έγγραφα που τελικά θα βαθμολογηθούν είναι τα d_1, d_2 και d_3. Αρχικά θα υπολογίσουμε τις τιμές idf_{t_1} και idf_{t_2}. Εφόσον ο όρος κομήτης εμφανίζεται σε τρία έγγραφα και ο όρος Χάλλεϋ εμφανίζεται σε δύο έχουμε $n_{t_1} = 3$ και $n_{t_2} = 2$. Επομένως, προκύπτουν οι ακόλουθες τιμές: $idf_{t_1} = 1.203$ και $idf_{t_2} = 1.504$.

Στη συνέχεια, για κάθε όρο και κάθε έγγραφο υπολογίζονται οι τιμές $tf_{t,d}$. Με απλές μαθηματικές πράξεις παίρνουμε: $tf_{t_1,d_1} = 1$, $tf_{t_1,d_2} = 1$, $tf_{t_1,d_3} = 1$,

$tf_{t_2,d_1} = 1$, $tf_{t_2,d_2} = 1.693$. Ομοίως υπολογίζουμε και τους παράγοντες tf για το ερώτημα q και έχουμε: $tf_{t_1,q} = 1$ και $tf_{t_2,q} = 1$. Εφόσον έχουν προσδιοριστεί οι τιμές tf και idf μπορούμε πλέον να προσδιορίσουμε τις τιμές των παραμέτρων $w_{t,d}$ και $w_{t,q}$ για τους όρους t_1 και t_2: $w_{t_1,d_1} = 1$, $w_{t_1,d_2} = 1$, $w_{t_1,d_3} = 1$, $w_{t_2,d_1} = 1$, $w_{t_2,d_2} = 1.693$, $w_{t_1,q} = 1.203$ και $w_{t_2,q} = 1.504$. Πριν τον υπολογισμό της συνάρτησης ομοιότητας απομένει ο προσδιορισμός των τιμών L_d και L_q. Με βάση τον Πίνακα 4.8 και τις προηγούμενες τιμές έχουμε: $L_{d_1} = \sqrt{11} = 3.316$, $L_{d_2} = 3.296$, $L_{d_3} = 2.23$ και $L_q = 1$.

Τέλος, εφαρμόζουμε τη συνάρτηση ομοιότητας (πρώτη γραμμή του Πίνακα 4.8) και λαμβάνουμε το βαθμό ομοιότητας των εγγράφων d_1, d_2 και d_3 ως προς το ερώτημα q. Για παράδειγμα, ο υπολογισμός της ποσότητας $S_{vector}(q, d_1)$ γίνεται ως εξής:

$$S_{vector}(q, d_1) = \frac{w_{t_1,d_1} \cdot w_{t_1,q} + w_{t_2,d_1} \cdot w_{t_2,q}}{L_{d_1}} = 0.816$$

$$S_{vector}(q, d_2) = \frac{w_{t_1,d_2} \cdot w_{t_1,q} + w_{t_2,d_2} \cdot w_{t_2,q}}{L_{d_2}} = 1.131$$

$$S_{vector}(q, d_3) = \frac{w_{t_1,d_3} \cdot w_{t_1,q}}{L_{d_3}} = 0.539$$

Από τις παραπάνω βαθμολογίες είναι προφανές ότι το πιο σχετικό έγγραφο της συλλογής, ως προς το ερώτημα $q = \{$κομήτης, Χάλλεϋ$\}$, είναι το έγγραφο d_2 με βαθμολογία 1.131. Το δεύτερο σχετικότερο έγγραφο είναι το d_1 με βαθμολογία 0.816 και ακολουθεί το d_3 με βαθμολογία 0.539. Παρατηρήστε ότι η βαθμολογία του d_2 είναι μεγαλύτερη της μονάδας! Αυτό οφείλεται στο γεγονός ότι δε διαιρέσαμε με το μέτρο του διανύσματος του ερωτήματος (ποσότητα L_q), αφού δε θα αλλάξει η σχετική σειρά των εγγράφων στην τελική κατάταξη. □

4.4 Πλεονεκτήματα και Μειονεκτήματα

Το βασικό πλεονέκτημα του Διανυσματικού μοντέλου ανάκτησης είναι η δυνατότητά του να βαθμολογεί τα έγγραφα με βάση την ομοιότητά τους ως προς κάποιο ερώτημα. Όσο περισσότερο σχετικά είναι δύο έγγραφα τόσο μικρότερη θα είναι η γωνία των αντίστοιχων διανυσμάτων τους και τόσο μεγαλύτερη θα είναι η τιμή του συνημιτόνου της μεταξύ τους γωνίας. Χρησιμοποιώντας διαφορετικούς ορισμούς για τις ποσότητες $w_{t,d}$ (το βάρος του όρου t στο έγγραφο d) μπορούμε να έχουμε ένα σύνολο διαφορετικών μοντέλων.

Το δεύτερο σημαντικό πλεονέκτημα του μοντέλου είναι ο σχετικά απλός τρόπος υλοποίησής του, κάθώς στηρίζεται σε απλές μαθηματικές πράξεις. Βέβαια, σε περίπτωση που τα έγγραφα έχουν μεγάλο μήκος και ο αριθμός των όρων είναι μεγάλος (π.χ., μερικές χιλιάδες) τότε ενδεχομένως ο προσδιορισμός της ομοιότητας μεταξύ δύο εγγράφων να απαιτεί σημαντικό χρόνο. Ωστόσο, για τις τυπικές περιπτώσεις όπου το έγγραφο του ερωτήματος αποτελείται από μερικούς όρους, οι υπολογισμοί της ομοιότητας με βάση τον τύπο του συνημιτόνου πραγματοποιούνται γρήγορα.

Το τρίτο σημαντικό πλεονέκτημα του Διανυσματικού μοντέλου είναι η υποστήριξη μερικής ταύτισης. Ένα έγγραφο που περιέχει ένα υποσύνολο των όρων του ερωτήματος δε θα λάβει μηδενικό βαθμό. Αυτό είναι πολύ σημαντικό λαμβάνοντας υπόψη ότι μπορεί να μην υπάρχει κανένα έγγραφο που να περιέχει όλους τους όρους του ερωτήματος.

Ένα από τα μειονεκτήματα του Διανυσματικού μοντέλου είναι η υπόθεση ότι οι όροι των εγγράφων είναι ανεξάρτητοι. Αυτό οδηγεί στη θεώρηση ότι έχουμε ένα ορθοκανονικό σύστημα αξόνων βάσει του οποίου ορίζονται τα διανύσματα των εγγράφων και των ερωτημάτων. Αυτή η υπόθεση δεν είναι απολύτως σωστή καθώς υπάρχουν όροι που δεν είναι ανεξάρτητοι και επομένως η εμφάνιση του ενός εξαρτάται από την εμφάνιση των άλλων. Δύο βασικές αιτίες που βλάπτουν την ανεξαρτησία των όρων είναι η *πολυσημία* και η *συνωνυμία*. Στην περίπτωση της πολυσημίας, ένας όρος μπορεί να έχει διαφορετικό νόημα ανάλογα με το είδος και το περιεχόμενο του εγγράφου ενώ στην περίπτωση της συνωνυμίας δύο όροι που γράφονται εντελώς διαφορετικά, μπορεί να έχουν το ίδιο ακριβώς νόημα (συνώνυμοι όροι). Ωστόσο, υιοθετώντας την ανεξαρτησία των όρων απλοποιείται η διαδικασία του προσδιορισμού της ομοιότητας.

Τέλος αξίζει να σημειωθεί ότι ο τρόπος ανάθεσης των βαρών στους όρους αν και διαισθητικά φαίνεται να έχει νόημα, δε στηρίζεται σε κάποιο μαθηματικό φορμαλισμό και θα μπορούσε να χαρακτηριστεί ακόμη και αυθαίρετος. Η επιλογή συγκεκριμένων τιμών για τα βάρη έχει επιβεβαιωθεί με πειραματικές μελέτες ότι έχει καλά αποτελέσματα αλλά δεν μπορεί να τεκμηριωθεί με μαθηματική ανάλυση.

4.5 Σύνοψη και Περαιτέρω Μελέτη

Το Διανυσματικό μοντέλο ανάκτησης προτάθηκε επίσημα από τον Salton το 1975 [56] και αποτελεί το πιο διαδεδομένο μοντέλο ανάκτησης. Ενώ το Boolean μοντέλο στηρίζεται στη Θεωρία Συνόλων, το Διανυσματικό μοντέλο βασίζεται κυρίως στη Γραμμική Άλγεβρα. Ο υπολογισμός της ομοιότητας μεταξύ δύο εγγράφων ή μεταξύ ενός εγγράφου και ενός ερωτήματος πραγματοποιείται με τη

χρήση του συνημιτόνου της γωνίας που σχηματίζεται μεταξύ των αντίστοιχων διανυσμάτων στο χώρο των M διαστάσεων, όπου M είναι το πλήθος των μοναδικών όρων που περιέχονται στα έγγραφα και χρησιμοποιούνται για την αναπαράσταση των εγγράφων.

Το Διανυσματικό μοντέλο καλύπτεται επαρκώς σε όλα τα βιβλία του χώρου. Ο ενδιαφερόμενος μπορεί να ανατρέξει στα αντίστοιχα κεφάλαια των βιβλίων [3, 38, 73]. Επίσης, προτείνουμε τη μελέτη της εργασίας [56] που αποτελεί την πρόταση του Διανυσματικού μοντέλου, και των εργασιών [53, 74] στις οποίες παρουσιάζονται διάφορες εκδοχές του μοντέλου. Η εργασία [74] αποτελεί επέκταση της [53], όπου παρουσιάζονται διαφορετικές εκδοχές του Διανυσματικού μοντέλου, ανάλογα με τις επιλογές. Μεγάλο ενδιαφέρον παρουσιάζει επίσης η εργασία [36] στην οποία περιλαμβάνεται μία μελέτη σχετικά με απλοποιήσεις που μπορούν να εφαρμοστούν στο Διανυσματικό μοντέλο με στόχο την ταχύτερη επεξεργασία των ερωτημάτων, αλλά χωρίς να βλάπτεται σημαντικά η αποτελεσματικότητα.

Επίσης, κρίνεται πολύ σημαντική η ενασχόληση με το θρυλικό σύστημα SMART, το οποίο μπορεί ο αναγνώστης να προμηθευτεί από τη διεύθυνση [62]. Στη διεύθυνση αυτή υπάρχουν επίσης και διάφορες συλλογές εγγράφων (μεταξύ των οποίων οι CACM, ISI, MED και CRAN) που μπορούν να χρησιμοποιηθούν σε συνδυασμό με το σύστημα SMART.

4.6 Ασκήσεις

4.1 Ποιές είναι οι σημαντικότερες διαφορές μεταξύ του Διανυσματικού και του Λογικού μοντέλου;

4.2 Να περιγράψετε τη διαδικασία υπολογισμού των βαρών στο Διανυσματικό μοντέλο.

4.3 Ποιά συνάρτηση χρησιμοποιείται για τον προσδιορισμό της ομοιότητας μεταξύ ενός ερωτήματος q και ενός εγγράφου d;

4.4 Για ποιό λόγο η χρήση της Ευκλείδιας απόστασης δεν είναι καλή πρακτική για τον προσδιορισμό της ομοιότητας;

4.5 Ποιά είναι τα βασικά μειονεκτήματα του Διανυσματικού μοντέλου;

4.6 Ο αριθμός των διαστάσεων καθορίζεται από το πλήθος των μοναδικών όρων της συλλογής που είναι συνήθως αρκετά μεγάλος. Να συζητήσετε για τα προβλήματα που ενδεχομένως δημιουργούνται από το μεγάλο αριθμό διαστάσεων.

4.7 Να κατασκευάσετε ένα πρόγραμμα που να διαβάζει τη συλλογή εγγράφων CRAN και για κάθε ερώτημα q της συλλογής να υπολογίζει το βαθμό ομοιότητας μεταξύ του q και κάθε εγγράφου d χρησιμοποιώντας μία από τις δυνατές εκφράσεις του Διανυσματικού μοντέλου.

4.8 Να κατασκευάσετε πρόγραμμα που να διαβάζει τη συλλογή εγγράφων MED και στη συνέχεια να υπολογίζει για κάθε ερώτημα q της συλλογής το βαθμό ομοιότητας με κάθε έγγραφο d. Στη συνέχεια, να υπολογίσετε το πλήθος των σχετικών εγγράφων εάν θεωρήσουμε ότι ενδιαφερόμαστε για τα top-20 έγγραφα της συλλογής. Δοκιμάστε την αποτελεσματικότητα για διαφορετικές εκφράσεις του Διανυσματικού μοντέλου.

4.9 Με βάση τη συλλογή εγγράφων του Σχήματος 4.1 και θεωρώντας ότι τα βάρη $w_{t,d}$ και $w_{t,q}$ υπολογίζονται με τη βοήθεια των σχέσεων 4.6 και 4.7 ενώ η ομοιότητα δύο εγγράφων προσδιορίζεται από τη σχέση 4.10, να προσδιορίσετε τον πίνακα ομοιότητας της συλλογής. Ο πίνακας αυτός είναι ένας συμμετρικός πίνακας $N \times N$ όπου N ο αριθμός των εγγράφων. Το κάθε κελί του πίνακα στη γραμμή i και τη στήλη j περιέχει μία πραγματική τιμή που δηλώνει το βαθμό ομοιότητας μεταξύ των εγγράφων d_i και d_j. Σχολιάστε το αποτέλεσμα.

4.10 Να δώσετε ένα παράδειγμα με το οποίο να φαίνεται ότι αν δε ληφθούν μέτρα, κάποια μεγάλα έγγραφα μπορεί να λάβουν μεγαλύτερο βαθμό από μικρότερα χωρίς να περιέχουν κατ'ανάγκη και περισσότερους όρους του ερωτήματος.