Απόστολος Γεωργάκης Κωνσταντίνος Κοτρόπουλος Ιωάννης Πήτας

ΟΡΓΑΝΩΣΗ ΚΑΙ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ ΤΟΥΡΙΣΤΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΓΛΩΣΣΑΣ

Abstract

The efficient information retrieval in the world wide web often requires the formulation of a rather complicated query. Restating this query in the case of poor or inaccurate results might not be always a solution. Current search techniques exhibit some inherent drawbacks which originate from the fact that they rely almost entirely on keywords and disregard the contextual information. The paper presents a method for document organization of domain-specific documents that enables the exploitation of contextual and semantic information in information retrieval tasks. This is done by representing both word themes and documents as vectors and by clustering them according to a specific metric. The clustering algorithm is partially based on the self-organizing map. Self-organizing maps are artificial neural networks that have the property of creating a spatially organised representation of the input vectors. The documents used to train the neural network as well as to test its performance have been derived from the tourism domain.

Εισαγωγή

Μια από τις πιο δύσκολες και επίπονες εργασίες είναι η ανάκτηση πληροφοριών που βρίσκονται αποθηκευμένες στον παγκόσμιο ιστό. Τα σύγχρονα συστήματα ανάκτησης πληροφοριών (ΣΑΠ) βασίζονται σε ερωτήσεις (queries) που υποβάλλονται από τους χρήστες. Η ερώτηση πρέπει να έχει συγκεκριμένη δομή και έχει τη μορφή λέξεων-κλειδιών που συνδέονται μεταξύ τους με τελεστές της άλγεβρας Boole. Συχνά δεν μπορούν να διαμορφωθούν τέτοιες ερωτήσεις, γιατί υπάρχει ελλιπής γνώση σχετικά με το αντικείμενο αναζήτησης. Ωστόσο αναζήτηση θα μπορούσε να εκτελεστεί και με χρήση ενός γνωστού κειμένου ως υπόδειγμα - πρότυπο. Είναι προφανές ότι τεχνικές αναζήτησης που βασίζονται σε λέξεις-κλειδιά δεν μπορούν να χρησιμοποιηθούν στην τελευταία περίπτωση.

Ας εξετάσουμε, τον τρόπο με τον οποίο λειτουργούν ορισμένα από τα συνήθη ΣΑΠ που προσφέρονται στο διαδίκτυο, όπως εκείνα στις πύλες Altavista ή Yahoo [**Ber99**]. Σε όλα, ανεξαιρέτως, χρησιμοποιούνται προγράμματα-ρομπότ (web crawlers) για τη συλλογή ιστοσελίδων προς ταξινόμηση. Οι σελίδες ταξινομούνται με ημι-αυτόματο τρόπο όπου συχνή είναι η επέμβαση ειδικών, όπως στην περίπτωση του Yahoo, όπου την ταξινόμηση των σελίδων αναλαμβάνουν εξειδικευμένοι συντάκτες. Μετά την ταξινόμηση των δεδομένων, ο χρήστης είναι σε θέση να υποβάλει μια ερώτηση στο ΣΑΠ. Το σύστημα επιστρέφει στο χρήστη έναν κατάλογο με ταξινομημένα από πριν κείμενα, τα οποία στην ιδανική περίπτωση είναι συναφή με το ερώτημα του, από τα οποία τα πιο σχετικά εντοπίζονται στην αρχή της λίστας. Το προηγούμενο παράδειγμα περιγράφει ένα ιδανικό σύστημα. Στην πραγματικότητα η απόκριση της μηχανή αναζήτησης, περιέχει άσχετη και περιττή πληροφορία, που την ονομάζουμε συνοπτικά θόρυβο.

Ένα άλλο μεγάλο πρόβλημα που αντιμετωπίζουν τα ΣΑΠ είναι ότι παρόλο που ορισμένα από αυτά μπορούν αρκετά καλά να εντοπίσουν μια ιστοσελίδα που πιθανώς να σχετίζεται με την ερώτηση του χρήστη, αδυνατούν να προτείνουν ιστοσελίδες συναφείς με το περιεχόμενο δοσμένου κειμένου. Επίσης, μπορούν πολύ εύκολα να εξαπατηθούν με σκοπό να ταξινομήσουν σε υψηλότερη θέση κάποιες άσχετες με το θέμα ιστοσελίδες. Τα περισσότερα από τα προβλήματα, που προαναφέρθηκαν, έχουν τις ρίζες τους στο γεγονός ότι οι τεχνικές αναζήτησης βασίζονται σε λέξεις-κλειδιά.

Οι δομές δεδομένων που αξιοποιούν οι σύγχρονες μηχανές αναζήτησης βασίζονται σε μια από τις επόμενες τεχνικές: αντεστραμμένα αρχεία (inverted files), πίνακες καταλήξεων (suffix arrays) ή αρχεία υπογραφών (signature files) [**Bae99**]. Την τελευταία δεκαετία, δίνεται έμφαση στη χρήση των αντεστραμμένων αρχείων. Η δομή αυτή αποτελείται από δύο αρχεία: το αρχείο *λεξιλογίου* και το αρχείο *εμφανίσεων*, τα οποία κατασκευάζονται κατά τη διάρκεια εκπαίδευσης του συστήματος. Στο *λεξιλόγιο* αποθηκεύονται όλες οι λέξεις που εμφανίζονται στα κείμενα προς ταξινόμηση. Στο αρχείο *εμφανίσεων* αποθηκεύονται οι σχετικές θέσεις των λέξεων του λεξιλογίου μέσα στα αρχεία κειμένων. Λέξεις που δεν περιέχουν σημαντική πληροφορία αφαιρούνται αυτόματα από τα κείμενα με μια διαδικασία που λέγεται αποκοπή (stopping). Τα δύο αρχεία μαζί με την ερώτηση του χρήστη, η οποία έχει υποστεί κατάλληλη επεξεργασία, εισάγονται στο ΣΑΠ. Κατά τη φάση της ανάκλησης οι λέξεις της ερώτησης ανιχνεύονται στο αρχείο του *λεξιλογίου*, και αν βρεθούν εκεί, τότε εντοπίζονται με τη βοήθεια του αρχείου *εμφανίσεων* τα αντίστοιχα κείμενα στα οποία απαντώνται οι λέξεις. Από αυτά τα αρχεία το ΣΑΠ διαμορφώνει την απόκριση στο ερώτημα του χρήστη.

Στην εργασία αυτή το ενδιαφέρον μας εστιάζεται στην οργάνωση και ανάκτηση πληροφοριών με τουριστικό περιεχόμενο. Οι πληροφορίες αντλούνται από ιστοσελίδες στον παγκόσμιο ιστό. Αν και μας απασχολεί, σε πρώτη φάση, η επεξεργασία κειμένων στην Αγγλική γλώσσα, περιγράφουμε στο άρθρο πώς θα μπορούσαν οι τεχνικές που συζητούνται να επεκταθούν, ώστε να εφαρμοστούν σε κείμενα της Ελληνικής γλώσσας.

Η εργασία που παρουσιάζεται στη συνέχεια είναι μέρος του ευρωπαϊκού προγράμματος HYPERGEO. Το πρόγραμμα απευθύνεται στους λεγόμενους μετακινούμενους χρήστες (mobile users), οι οποίοι αναζητούν στο διαδίκτυο, πληροφορίες τουριστικού περιεχομένου. Οποιοσδήποτε χρησιμοποιεί προσωπικούς ψηφιακούς βοηθούς (Personal Digital Assistants, PDAs) ή κινητά τηλέφωνα με πρωτόκολλο ασύρματων εφαρμογών (Wireless Application Protocol, WAP) για πρόσβαση στο διαδίκτυο μπορεί να θεωρηθεί μετακινούμενος χρήστης. Με τη βοήθεια τέτοιων συσκευών ο χρήστης μπορεί να έχει πρόσβαση σε οποιοδήποτε ΣΑΠ. Τα προβλήματα που έχουν να αντιμετωπίσουν οι μετακινούμενοι χρήστες είναι:

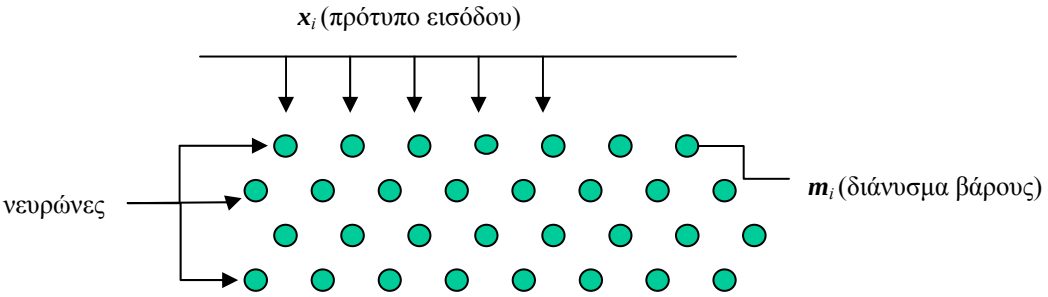
- Οι περιορισμένες δυνατότητες στην απεικόνιση των αποτελεσμάτων.
- Η μειωμένη ακρίβεια των αποκρίσεων των σύγχρονων τεχνικών αναζήτησης.
- Οι περιορισμένες επιδόσεις των συσκευών που προαναφέρθηκαν σε σχέση με τους συνήθεις υπολογιστές.

Η μέθοδος που θα περιγραφεί αποβλέπει στο να βελτιώσει την απόδοση των ΣΑΠ.

Αυτοοργανωνόμενοι χάρτες

Οι αυτοοργανωνόμενοι χάρτες (self-organizing maps, SOMs) είναι νευρωνικά δίκτυα (ΝΔ) ενός επιπέδου στα οποίο δεν συμβαίνει ανατροφοδότηση της εξόδου. Διαμορφώνουν μια μη-γραμμική προβολή -απεικόνιση- από ένα χώρο εισόδου υψηλής διάστασης σε έναν χώρο μικρότερης διάστασης (δισδιάστατο ή τρισδιάστατο πλέγμα). Χρησιμοποιούνται για να ομαδοποιήσουν διανύσματα σύμφωνα με κάποιο συγκεκριμένο χαρακτηριστικό. Ο αλγόριθμος λαμβάνει υπόψη του τις συσχετίσεις των διανυσμάτων εισόδου και υπολογίζει τη βέλτιστη απεικόνιση τους σε πλέγμα του οποίου οι κόμβοι προσεγγίζουν τα διανύσματα εισόδου ελαττώνοντας ένα κριτήριο λάθους, π.χ. το *μέσο τετραγωνικό σφάλμα*.

Το ΝΔ αποτελείται από ένα σύνολο νευρώνων, οι οποίοι είναι τοποθετημένοι στους κόμβους ενός τετραγωνικού ή εξαγωνικού πλέγματος και είναι εφοδιασμένοι με ένα διάνυσμα βαρών (*weight vector*). Η δομή του αυτοοργανωνόμενου χάρτη δείχνεται παραστατικά στο Σχήμα 1. Σε κάθε νευρώνα υπάρχει επίσης ένας μετρητής που μετράει το πλήθος των διανυσμάτων που έχουν εκχωρηθεί στο συγκεκριμένο νευρώνα. Επομένως, αν μπορούσαμε να αναπαραστήσουμε τα κείμενα ή τις λέξεις με διανύσματα, τότε θα



Σχήμα 1: Η δομή ενός αυτοοργανωνόμενου χάρτη με νευρώνες σε εξαγωνική τοπολογία.

μπορούσαμε να χρησιμοποιήσουμε την τεχνική αυτή για να οργανώσουμε τα κείμενα ή τις λέξεις σε ομοειδείς ομάδες. Μια τέτοια αναπαράσταση συζητείται στην επόμενη ενότητα. Ο τρόπος λειτουργίας του νευρωνικού δικτύου είναι ο ακόλουθος: Αρχικά όλα τα διανύσματα βαρών στους νευρώνες έχουν τυχαίες τιμές. Από το σύνολο τον διανυσμάτων εισόδου επιλέγεται τυχαίως ένα διάνυσμα το οποίο τροφοδοτείται στο ΝΔ. Κάθε νευρώνας υπολογίζει την Ευκλείδεια απόσταση μεταξύ του διανύσματος εισόδου και του διανύσματος βαρών του. Εναλλακτικά, θα μπορούσε να υπολογιστεί το εσωτερικό γινόμενο το οποίο είναι συσχετισμένο με την Ευκλείδεια απόσταση ως εξής: όσο μικρότερη γίνεται η Ευκλείδεια απόσταση, τόσο μεγαλύτερο γίνεται το εσωτερικό γινόμενο. Ο πλησιέστερος νευρώνας στο διάνυσμα εισόδου ονομάζεται νικητής. Το

διάνυσμα εισόδου εκχωρείται στην ομάδα διανυσμάτων που αντιπροσωπεύεται από το νικητή. Κατά συνέπεια ο μετρητής του νικητή αυξάνεται κατά μια μονάδα. Ταυτόχρονα, τα διανύσματα βαρών των νευρώνων προσαρμόζονται σύμφωνα με την εξίσωση:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t)\left[\boldsymbol{x}(t) - \boldsymbol{m}_i(t)\right] \quad (1)$$

όπου το t υποδηλώνει το χρόνο, $\boldsymbol{m}_i(t)$ είναι το διάνυσμα βάρους του i-οστού νευρώνα τη χρονική στιγμή t, $\boldsymbol{x}(t)$ είναι το διάνυσμα εισόδου τη χρονική στιγμή t και $h_{ci}(t)$ είναι η συνάρτηση γειτνίασης. Η συνάρτηση γειτνίασης ορίζει την επιρροή που ασκεί ο νικητής νευρώνας στους γειτονικούς του νευρώνες. Συνήθως η συνάρτηση γειτνίασης έχει τη μορφή πυρήνα εξομάλυνσης.

Μετά από αρκετές επαναλήψεις της διαδικασίας ορισμένες από τις τοπολογικές σχέσεις του χώρου εισόδου διατηρούνται και τα διανύσματα βαρών του δικτύου μετατρέπονται σε μια οργανωμένη απεικόνιση τους. Κείμενα τα οποία από άποψη περιεχομένου είναι σχετικώς συναφή θα παρουσιαστούν (με μεγάλη πιθανότητα) σε περιοχές του χάρτη οι οποίες είναι γειτονικές. Ένα μεγάλο πλεονέκτημα των SOM είναι ότι έχουν τη δυνατότητα να ανακαλύπτουν σχέσεις μεταξύ κειμένων οι οποίες είναι κρυμμένες σε μεγάλες διαστάσεις και να τις αναπαριστούν σε ένα δισδιάστατο πλέγμα.

Η εφαρμογή των SOM στην οργάνωση και ανάκτηση πληροφοριών που θα παρουσιαστεί στην επόμενη ενότητα εντάσσεται στη στατιστική επεξεργασία φυσικής γλώσσας [**Man99**]. Το έναυσμα γι' αυτή τη στροφή αποτέλεσε η στενή συνεργασία μεταξύ των κοινοτήτων της επεξεργασίας φυσικής γλώσσας και της αναγνώρισης φωνής, στην οποία στατιστικές τεχνικές, όπως η θεωρία αποφάσεων του Bayes, οι τεχνικές ελάττωσης εντροπίας, τα Μαρκοβιανά μοντέλα, η διανυσματική κβάντιση κ.α. είναι ισχυρώς θεμελιωμένες και χρησιμοποιούνται με εξαιρετική επιτυχία.

Εφαρμογή στην οργάνωση και ανάκτηση πληροφορίας από κείμενα

- **Εκπαίδευση**

Σκοπός μας είναι να αναπτύξουμε και να εφαρμόσουμε τεχνικές που αξιοποιούν το νευρωνικό δίκτυο που περιγράψαμε στην οργάνωση δεδομένων και την ανάκτηση πληροφοριών. Τομέας ενδιαφέροντος είναι τα κείμενα τουριστικού περιεχομένου. Για το σκοπό αυτό, συλλέχθηκαν 633 ιστοσελίδες που περιέχουν 118.868 λέξεις. Οι ιστοσελίδες αυτές συγκροτούν το σώμα κειμένων (*corpus*) που θα χρησιμοποιηθεί για την εκπαίδευση και έλεγχο των επιδόσεων του συστήματος ανάκτησης πληροφοριών. Η επιλογή των κειμένων ακολούθησε τους ακόλουθους κανόνες:

- Αντιπροσωπευτικότητα (βασικότερο)
- Ορθή κατά το δυνατό χρήση της αγγλική γλώσσα
- Πεπερασμένο μέγεθος

Πρέπει να τονίσουμε ότι σώμα κειμένων μεροληπτεί κατά το ότι πληροφορίες σχετικές με την Ελλάδα και την Ισπανία αναπαρίστανται πληρέστερα σ' αυτό.

Ας περιγράψουμε τα στάδια προ-επεξεργασίας στα οποία υποβλήθηκε το σώμα κειμένων για να απαλλαγεί από το θόρυβο. Κατ' αρχήν το σώμα κειμένων σχολιάστηκε, με ανθρώπινη παρέμβαση, ώστε να κωδικοποιηθούν ονόματα ανθρώπων και εθνολογικών ομάδων, τοπωνύμια, νομίσματα και ονόματα φαγητών.

Κατόπιν αφαιρέθηκε κάθε κώδικας HTML που υπήρχε στα κείμενα, όπως επίσης και στοιχεία των κειμένων που δεν περιέχουν χρήσιμη πληροφορία (π.χ. αριθμοί, σύμβολα, άρθρα, προθέσεις, αντωνυμίες καθώς και τα σημεία στίξης). Οι μη-Αγγλικοί όροι δεν αφαιρέθηκαν αυτόματα, αλλά με ανθρώπινη παρέμβαση. Τούτο έγινε γιατί σε πολλές περιπτώσεις τέτοιοι όροι είναι τοπωνύμια ή ονόματα τοπικών φαγητών. Εδώ χρήζει ιδιαίτερης αναφοράς ο χειρισμός της περίπτωση της τελείας (.) ως σημείου στίξης. Η τελεία δεν αφαιρέθηκε από τα κείμενα για τον εξής λόγο: Όπως θα δούμε στη συνέχεια, τα διανύσματα που θα κατασκευάσουμε για κάθε *θέμα* λέξης (stem) θα βασιστούν σε στατιστικά στοιχεία εμφάνισης δυάδων θεμάτων λέξεων ή δίγραμμων (bigrams). Δηλαδή δεν μας ενδιαφέρει η συχνότητα εμφάνισης του *i*-στου θέματος λέξης από μόνη της, αλλά οι συχνότητες των ζευγών που περιέχουν το *i*-στο θέμα λέξη είτε ως πρώτο όρο τους είτε ως δεύτερο (π.χ. **inform** reserv και addit **inform**). Η διαδικασία αυτή έχει νόημα μέσα στα όρια της κάθε πρότασης. Γι' αυτό το λόγο και δεν αφαιρείται η τελεία, ώστε να μην χαθούν τα όρια των προτάσεων.

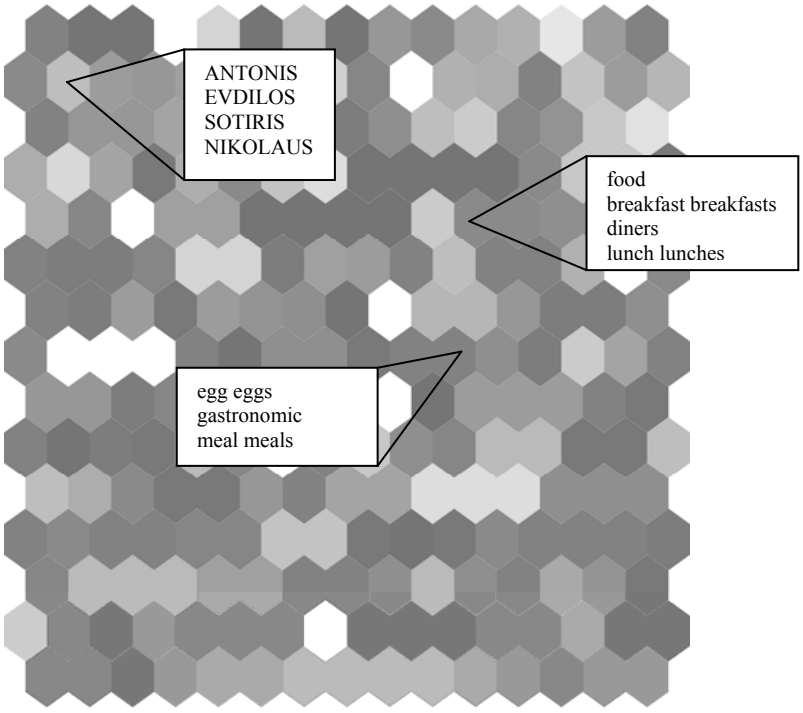
Κατόπιν, αφαιρέθηκαν οι καταλήξεις των λέξεων ώστε να προκύψουν τα θέματα των λέξεων αυτών σε μια διαδικασία που ονομάζεται *stemming* [**Fra92**]. Με τη χρήση του αλγορίθμου του *Porter* [**Por80**] προέκυψαν 40.857 θέματα λέξεων, από τα οποία μοναδικά ήταν τα 6249. Αντίστοιχη διαδικασία είναι επίσης εφικτή για την Ελληνική γλώσσα [**Nik00**], παρά το πλούσιο κλιτικό της σύστημα. Η αφαίρεση των καταλήξεων έγινε για τον εξής σκοπό: Αν χρησιμοποιούσαμε τις λέξεις όπως ήταν με τις καταλήξεις τους, τότε η συχνότητα εμφάνισης της κάθε λέξης θα ήταν αρκετά μικρότερη.

Κατόπιν υπολογίστηκε για κάθε θέμα ένα διάνυσμα. Έστω **e** διάνυσμα διαστάσεως ίσης με το πλήθος των διακεκριμένων θεμάτων που απαντώνται στο σώμα κειμένων, ας πούμε N_w. Για να υπολογίσουμε ένα διάνυσμα για κάθε θέμα μπορούν να χρησιμοποιηθούν οι εξής σχέσεις [**Bec98**]:

$$\boldsymbol{x}_i = \frac{1}{N_i} \sum_l n_{il} \boldsymbol{e}_l \qquad (2)$$

$$\boldsymbol{x}_i = \frac{1}{N_i}\begin{bmatrix} \sum_{\substack{l=1 \\ l\neq i}}^{N_w} n_{il}^{-}\, \boldsymbol{e}_l \\ \varepsilon \boldsymbol{e}_i \\ \sum_{\substack{m=1 \\ m\neq i}}^{N_w} n_{mi}^{+}\, \boldsymbol{e}_m \end{bmatrix} \tag{3}$$

όπου n_{il}^{-} είναι ο αριθμός των εμφανίσεων της l λέξης πριν από την i λέξη, n_{mi}^{+} είναι ο αριθμός των εμφανίσεων της i λέξης πριν από την m λέξη και n_{il} είναι το πλήθος των εμφανίσεων της l λέξης στο σώμα των κειμένων αμέσως πριν από την λέξη i. Οι εξισώσεις (2) και (3) βασίζονται στην έννοια των δίγραμμων. Η αποτελεσματικότητα των δίγραμμων στην Αγγλική γλώσσα εδράζεται στην αυστηρή σειρά εμφάνισης των λέξεων (word order). Στην εφαρμογή μας χρησιμοποιήσαμε την Εξ. (3).



Σχήμα 2: Χάρτης Κατηγοριών Λέξεων που προκύπτει από την εκπαίδευση ενός ΝΔ 232 νευρώνων με 6249 θέματα λέξεων στο σώμα κειμένων.

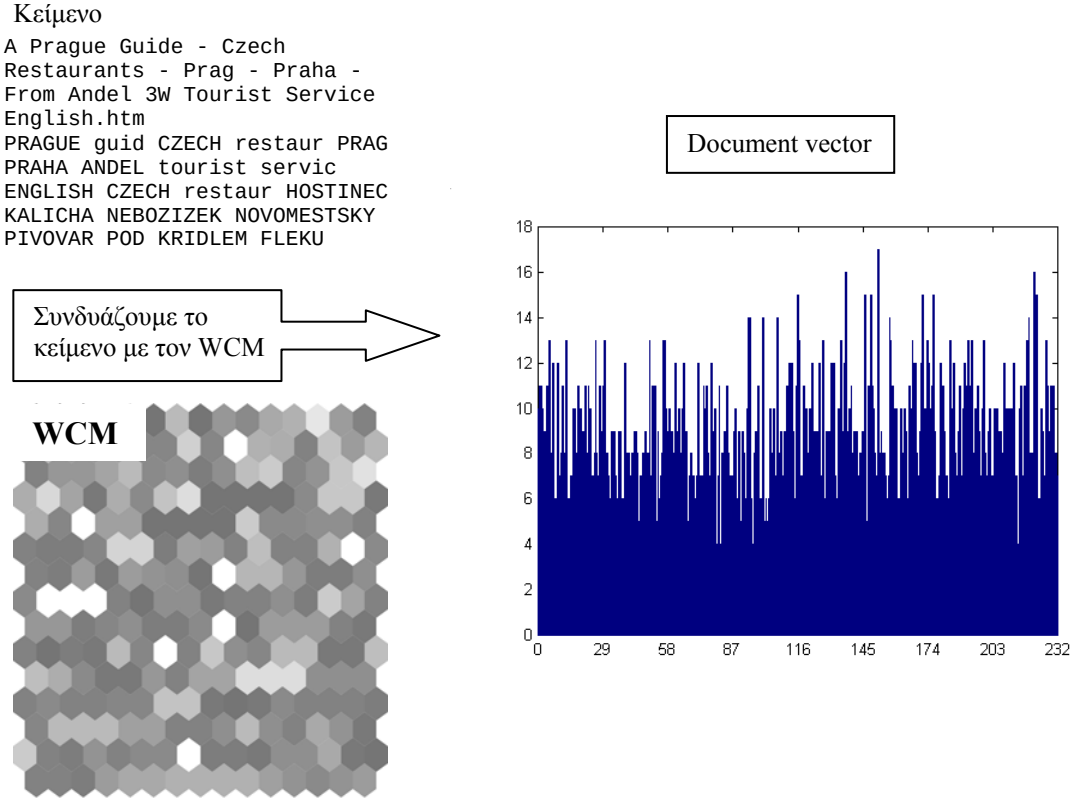
Όλα τα διανύσματα $\boldsymbol{x}_i$ παρουσιάστηκαν, με τυχαία σειρά για ικανοποιητικό πλήθος επαναλήψεων στο νευρωνικό δίκτυο. Για κάθε διάνυσμα εισόδου, το δίκτυο βρίσκει το νικητή και τροποποιεί τα διανύσματα βαρών του νικητή και των γειτονικών του νευρώνων προς την κατεύθυνση του διανύσματος $\boldsymbol{x}_i$. Το αποτέλεσμα της διαδικασίας αυτής είναι ο λεγόμενος *χάρτης κατηγοριών λέξεων* (ΧΚΛ). Κάθε κόμβος του ΧΚΛ χαρακτηρίζεται από τις λέξεις που απεικονίζονται σ' αυτόν (Σχήμα 2). Το ζητούμενο είναι να υπολογιστούν ομάδες λέξεων όμοιων από εννοιολογική σκοπιά (π.χ. food, breakfast, diner, κτλ). Στην ουσία τελικώς ευελπιστούμε ότι θα προκύψει ένας χάρτης που θα περιέχει συλλογές συνωνύμων. Η ομαδοποίηση των θεμάτων των λέξεων και κατ' επέκταση των ίδιων των λέξεων μας επιτρέψει να μειώσουμε ακόμα περισσότερο τη διάσταση του χώρου των κειμένων. Έτσι, αντί να έχουμε N_w διαφορετικά θέματα λέξεων όταν προχωρούμε στην κατασκευή των διανυσμάτων κειμένων έχουμε έναν πολύ μικρότερο αριθμό ομάδων λέξεων.

Τα διαφορετικά επίπεδα γκρι στον χάρτη παραπέμπουν σε διαφορετικές πυκνότητες, δηλαδή αριθμό λέξεων στους νευρώνες. Αποχρώσεις του γκρι κοντά στην τιμή

255 (άσπρο χρώμα) αντιστοιχούν σε ολιγομελείς ομάδες συνωνύμων, ενώ επίπεδα του γκρι κοντά στο 0 (μαύρο χρώμα) σε μεγαλύτερες πυκνότητες λέξεων.

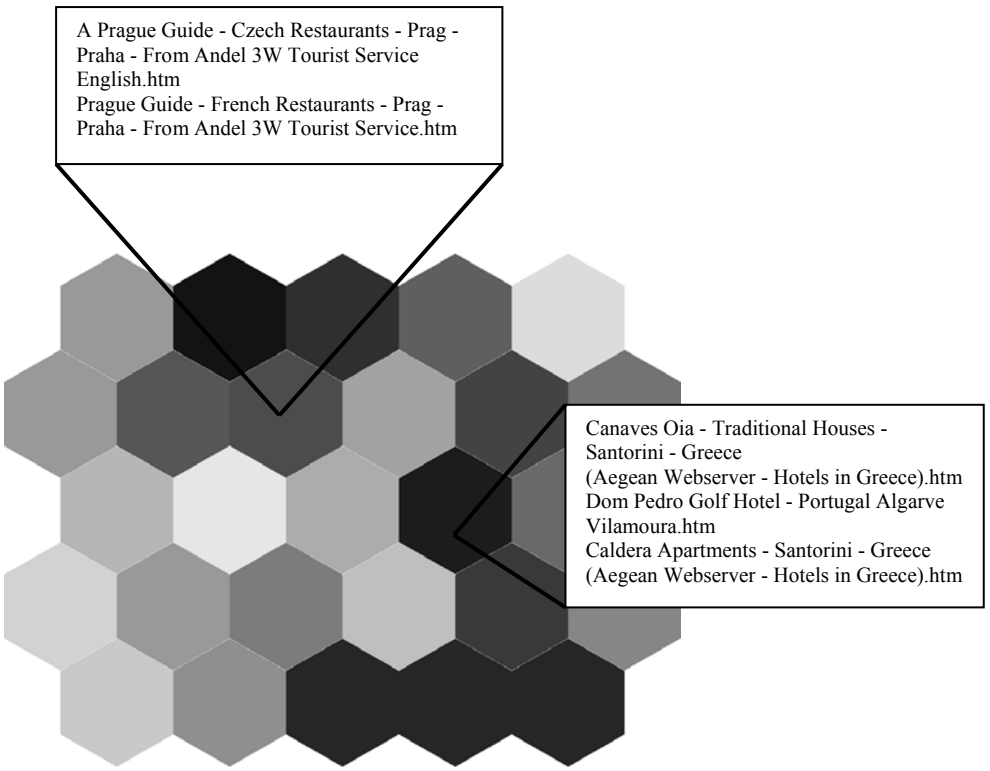
Κατόπιν για κάθε κείμενο του σώματος κειμένων κατασκευάζουμε ένα ιστόγραμμα ομάδων λέξεων, το διάνυσμα κειμένου $\boldsymbol{a}_k$. Η διάσταση του διανύσματος $\boldsymbol{a}_k$. είναι ίση με τον αριθμό των ομάδων λέξεων που εντοπίζονται στο ΧΚΛ. Το ιστόγραμμα αυτό κατασκευάστηκε ως εξής: για κάθε θέμα λέξης σε ένα κείμενο εντοπίζεται στο ΧΚΛ η ομάδα λέξεων, στην οποία εντάχθηκε το υπό μελέτη θέμα. Το πλήθος εμφάνισης της κατηγορίας λέξεων, τότε αυξάνεται κατά ένα. Κατόπιν το ιστόγραμμα μπορεί να εξομαλυνθεί, για να βελτιωθεί η ευρρωστία σε μικρές αλλαγές της εισόδου. Η εξομάλυνση του ιστογράμματος επιτυγχάνεται με συνέλιξη του ιστογράμματος με ένα Gaussian πυρήνα. Η διαδικασία κατασκευής ενός διανύσματος κειμένου φαίνεται στο Σχήμα 3.

Όταν υπολογιστούν τα διανύσματα των κειμένων $\boldsymbol{a}_i$ ξαναεκτελείται ο αλγόριθμος



Σχήμα 3: Η διαδικασία κατασκευής ενός διανύσματος κειμένου.

εκπαίδευσης του νευρωνικού δικτύου για την κατασκευή ενός δεύτερου χάρτη που λέγεται



Σχήμα 4: Χάρτης κειμένων που προκύπτει από την εκπαίδευση ΝΔ 27 νευρώνων σε 172 κείμενα.

χάρτης κειμένων (ΧΚ). Σ' αυτό το βήμα το νευρωνικό δίκτυο αποτελείται από ένα πλέγμα με λιγότερους νευρώνες από ό,τι στην πρώτη φάση κατά την οποία κατασκευάστηκε ο ΧΚΛ. Στα πειράματα μας χρησιμοποιήσαμε ποικίλες διαστάσεις για τα πλέγματα: από 15 έως 25 νευρώνες σε κάθε πλευρά στο ΧΚΛ και μέχρι 7 νευρώνες για το ΧΚ (Σχήμα 4).

Το τελευταίο βήμα της διαδικασίας είναι η εποπτική αναπαράσταση του ΧΚΛ και του ΧΚ. Το αποτέλεσμα αυτού του σταδίου φαίνεται στα Σχήματα 2 & 4. Κάθε νευρώνας αναπαρίσταται ως ένα εξάγωνο. Η επιλογή των εξαγώνων οφείλεται στην τοπολογία του δικτύου.

- **Αναζήτηση**

Το βασικότερο πλεονέκτημα της μεθόδου που περιγράψαμε είναι ότι επιτρέπει αναζήτηση ως προς το περιεχόμενο κάποιου άλλου κειμένου που ορίζει ο χρήστης ή ως προς ερώτημα που διατυπώνεται από το χρήστη σε φυσική γλώσσα. Για την ακρίβεια, οι αναζητήσεις υλοποιούνται ως εξής:

- **Ως προς λέξη-κλειδί**: Αφού προσδιοριστούν τα θέματα των λέξεων ως προς τις οποίες ο χρήστης επιθυμεί να κάνει αναζήτηση, εντοπίζονται στο ΧΚΛ οι κατηγορίες των θεμάτων. Από τις κατηγορίες στις οποίες ανήκουν τα θέματα των λέξεις και με τη βοήθεια δομών αντεστραμμένων αρχείων, όπως στις συνήθεις μηχανές αναζήτησης, αναζητούνται τα αρχεία, στα οποία απαντώνται οι λέξεις αυτές ή οι συνωνυμικές προς αυτές.
- **Ως προς το περιεχόμενο**: Το κείμενο, το οποίο αποτελεί το υπόδειγμα ως προς το οποίο θα γίνει η αναζήτηση των συναφών κειμένων, περνά όλα τα στάδια επεξεργασίας που περιγράφηκαν στη φάση της εκπαίδευσης. Κατόπιν, με τη βοήθεια του ΧΚΛ που έχει ήδη κατασκευαστεί, υπολογίζεται το αντίστοιχο διάνυσμα κειμένου. Κατόπιν βρίσκουμε το νευρώνα με τη δυνατότερη απόκριση στο ΧΚ. Τα κείμενα που έχουν εκχωρηθεί στο συγκεκριμένο νευρώνα κατά τη διάρκεια της εκπαίδευσης αναμένεται να έχουν ανάλογο περιεχόμενο με το κείμενο εισόδου.

Η δεύτερη περίπτωση παρουσιάζει ιδιαίτερο ενδιαφέρον. Στην περίπτωση που ο χρήστης έχει ήδη στα χέρια του ένα κείμενο που περιγράφει, για παράδειγμα, το μουσείο του Λούβρου στο Παρίσι και θέλει να βρει σχετικά κείμενα-ιστοσελίδες δεν είναι υποχρεωμένος να συνθέσει κάποια ερώτηση (query) και κατόπιν να την υποβάλλει σε μια μηχανή αναζήτησης. Αρκεί να δώσει το κείμενο αυτό ως είσοδο στη μηχανή που περιγράψαμε. Μετά το τέλος της διαδικασίας, ο χρήστης θα πάρει ως απόκριση, κείμενα με ανάλογο περιεχόμενο προς αυτό που έδωσε ως υπόδειγμα.

Συμπέρασμα

Κύριος σκοπός της εργασίας ήταν να παρουσιάσει τα πρώτα αποτελέσματα που έχουν αποκτηθεί από την ερευνητική εργασία που έχει επιτελεστεί μέχρι σήμερα στο Εργαστήριο Τεχνητής Νοημοσύνης και Ανάλυσης Πληροφοριών του ΑΠΘ στο ερευνητικό έργο HYPERGEO για την οργάνωση και την ανάκτηση πληροφοριών τουριστικού ενδιαφέροντος. Μολονότι το διαθέσιμο σώμα κειμένων είναι μικρό, τα πρώτα αποτελέσματα, που παρουσιάστηκαν στην εργασία, κρίνονται ικανοποιητικά. Στο μέλλον όταν θα διαθέτουμε μεγαλύτερα σώματα κειμένων, θα εκτελέσουμε πειράματα μεγαλύτερης κλίμακας. Σε επόμενη φάση, σκοπεύουμε να αυτοματοποιήσουμε όλες τις διαδικασίες που στην παρούσα φάση έγιναν με την ανθρώπινη παρέμβαση, όπως: προγράμματα-ρομπότ (web crawlers) για τη συλλογή ιστοσελίδων, σχολιασμός (annotation) κτλ.

Εξετάζεται, επίσης, η περίπτωση της πλήρους απόρριψης του πρώτου βήματος που οδηγεί στην κατασκευή του χάρτη κατηγοριών λέξεων. Τούτο θα διευκόλυνε την

επέκταση και χρήση της παρούσας τεχνολογίας στην οργάνωση κειμένων που έχουν συνταχθεί στην Ελληνική. Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί το σύστημα προσδιορισμού θεμάτων λέξεων [Nik00] σε συνδυασμό με το σύστημα που περιγράψαμε. Τέλος αξίζει να διερευνηθεί η αντικατάσταση των δίγραμμων από n-γραμμα (n-grams) για να αντιμετωπιστεί η πιο χαλαρή συντακτική δομή της Ελληνικής.

Βιβλιογραφία

[**Bae99**] Baeza - Yates, R. and Ribeiro - Neto, B. 1999. *Modern Information Retrieval.* ACM.

[**Bec98**] Becchetti, C. and Ricotti, L. P. 1998. *Speech Recognition: Theory and C++ Implementation.* New York: J. Wiley.

[**Ber99**] Berry, M. W. and Browne, M. 1999. *Understanding Search Engines: Mathematical Modelling and Text Retrieval*. SIAM.

[**Fra92**] Frakes, W. B. and Baeza - Yates, R. 1992. *Information Retrieval: Data structures and Algorithms*. Englewood Cliffs, N. J.: Prentice-Hall, Inc.

[**Hay99**] Haykin, S. 1999. *Neural Networks, A Comprehensive Foundation.* Englewood Cliffs, N. J.: Prentice-Hall, Inc.

[**Kas98a**] Kaski, S. Lagus, K. Honkela, T. and Kohonen, T. 1998. "Statistical Aspects of the WEBSOM System in Organizing Document Collections," *Computing Science and Statistics*, 29: 281-290.

[**Kas98b**] Kaski, S. 1998. "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proc. 8th of Int. Conf. on Neural Networks*, IEEE, 1: 413-418.

[**Kas99**] Kaski, S. 1999. "Fast winner search for SOM-based monitoring and retrieval of high-dimensional data," in *Proc. of the 9th Int. Conf. on Artificial Neural Networks*, Edinburgh.

[**Koh97**] Kohonen, T. 1997. *Self-Organizing Maps.* Springer Verlag.

[**Koh98**] Kohonen, T. 1998. "Self-organization of very large document collections: State of the art," in *Proc. of the 8th Int. Conf. on Artificial Neural Networks*, 1: 65-74, Springer Verlag.

[**Man99**] Manning, D. and Schutze, H. 1999. *Foundation of Statistical Natural Language processing.* Cambbridge, M. A.: MIT Press.

[**Nik00**] Nikolaidis S. and Kalamboukis T. Z. 2000. "An Evaluation of Stemming Algorithms with Modern Greek," *Advances in Informatics*, 1: 212-222.

[**Oak98**] Oak, M. P. 1998. *Statistics for corpus linguistics.* Cambridge University Press.

[**Por80**] Porter, M. F. 1980. "An algorithm for suffix stripping," *Program* 14: 130-137.

[**Rit89**] Ritter, H. and Kohonen, T. 1989. "Self-Organizing Semantic Map" *Biol. Cybernetics* 61: 241-254.

[**Sal83**] Salton, G. and McGill, M. J. 1983. *Introduction to modern information retrieval.* New York: McGraw Hill.

[**Tuk77**] Tukey, J. W. 1977. *Exploratory Data Analysis.* Addison-Wesley.

Λέξεις-κλειδιά: *Αυτοοργανωνόμενοι χάρτες, ανάκτηση πληροφοριών, στατιστική επεξεργασία γλώσσας, μοντέλο διανυσμάτων κειμένου, μηχανές αναζήτησης στο διαδίκτυο*.