

*Θοδωρής Πιτικάρης, Ιωάννης Τσαγκατάκης,  
Μιχάλης Νικητάκης, Γιώργος Παπαδουράκης*  
Α. Τ. Ε. Ι ΚΡΗΤΗΣ

*Theodoris Pitikaris, John Tsagatakis,  
Michael Nikitakis, George Papadourakis*  
Technological Educational Institute of Crete

*Ψηφιακές βιβλιοθήκες: ανάκτηση πληροφοριών με neural network  
Digital libraries: information retrieval with neural network*

#### ΠΕΡΙΛΗΨΗ

Μια από τις μεγαλύτερες προκλήσεις για την επιστήμη της βιβλιοθηκονομίας, είναι η διαχείριση του συνεχούς αυξανόμενου όγκου πληροφορίας που συγκεντρώνεται στις ψηφιακές βιβλιοθήκες. Οι νέες τεχνολογίες ψηφιοποίησης και παροχής πληροφορίας επηρεάζουν την λειτουργία και την οργάνωση των βιβλιοθηκών. Βασικός στόχος των βιβλιοθηκών δεν είναι μόνο η ανάπτυξη και η διατήρηση έντυπων και ηλεκτρονικών συλλογών αλλά η δυνατότητα παροχής και ανάκτηση πληροφοριών από αυτές. Πέραν από τις κλασικές μεθόδους αναζήτησης και ανάκτησης έχουν αρχίσει να κάνουν την εμφάνιση τους εξελιγμένες εφαρμογές βασισμένες στα τεχνητά νευρωνικά δίκτυα.

Τα νευρωνικά δίκτυα είναι μια μορφή τεχνητής νοημοσύνης, μια προσομοίωση της ανθρώπινης νόησης. Η εφαρμογή τους στο χώρο της βιβλιοθηκονομίας έχει ως στόχο την αναζήτηση ανάκτηση και παρουσίαση πληροφοριών, με τρόπο τέτοιο που να παρουσιάζει με ακρίβεια την σχετικότητα τους ως προς το ζητούμενο θέμα και μάλιστα να συσχετίζει φορείς πληροφορίας που πιθανά να μην αναφέρονται απευθείας στο θέμα αλλά να παρουσιάζουν νοηματική ενότητα με αυτό.

**Λέξεις κλειδιά: ( ψηφιακές βιβλιοθήκες, νευρωνικά δίκτυα, ανάκτηση πληροφοριών)**

#### ABSTRACT

*With the enormous increase in recent years in the number of text databases available on-line, and the consequent need for better techniques to access this information, there has been a strong resurgence of interest in the research done in the area of information retrieval (IR). Evolving from neuro-biological insights, neural network technology gives a computer system an amazing capacity to actually learn from input data. Artificial neural*

*networks have provided solutions to problems normally requiring human observation and thought processes. Relevance feedback is a process where users identify relevant documents in an initial list of retrieved documents, and the system then creates a new query based on those sample relevant documents.*

**Keywords: (digital libraries, neural network, information retrieval)**

## Εισαγωγή

---

Η εξέλιξη της επιστήμης των υπολογιστών και της πληροφόρησης προσέφερε στις βιβλιοθήκες νέα εργαλεία στην παραγωγή και διανομή πληροφορίας και γνώσης.

Οι βιβλιοθήκες κάτω από αυτές τις συνθήκες βρίσκονται σε μια διαρκή αλλαγή τόσο σε διοικητικό όσο και λειτουργικό πεδίο.

Η ψηφιοποίηση των συλλογών(έγγραφα, χάρτες, φωτογραφίες κτλ)τους και η διάθεσης τους μέσα από τον παγκόσμιο ιστό δημιουργεί νέα δεδομένα στην διαχείριση του κύκλου ζωής της πληροφορίας (παραγωγή, περιγραφή, πρόσβαση, ανάκτηση).

Η ψηφιακή βιβλιοθήκη είναι ένα σύνολο υπηρεσιών που περιλαμβάνει την πρόσκτηση, την αποθήκευση, την πρόσβαση και διακίνηση πληροφορίας υλικού που είναι σε ψηφιακή μορφή. Το πλεονέκτημα μιας ψηφιακής βιβλιοθήκης δεν είναι μόνο η οργάνωση των πληροφοριών αλλά η εξασφάλιση εύκολης και ακριβής πρόσβασης σε αυτές.

Η ανάγκη πληροφοριών (τι πραγματικά θέλει ο χρήστης) ,η εξαγωγή πληροφορίας, το φιλτράρισμα πληροφοριών(ακρίβεια δεδομένων μέσα από μεγάλο όγκο πληροφοριών ο χρήστης ανακτά αυτά που πραγματικά των ενδιαφέρουν) η διαλειτουργικότητα, η ταξινόμηση πληροφοριών, η αναπαράσταση πληροφορίας, η αξιολόγηση, είναι προβλήματα τα οποία ζητούν άμεσες λύσεις [2].

Την τελευταία δεκαετία με την αύξηση της παραγωγής πληροφόρησης η ακριβή και ταχεία πρόσβαση στις πληροφορίες γίνεται δυσκολότερη ,τα γνωστά εργαλεία ανάκτησης πληροφοριών καθίστανται αδύναμα και αναποτελεσματικά με αποτέλεσμα ένα μεγάλος αριθμός πληροφοριών να μένει αναξιοποίητος ενώ ταυτόχρονα απαιτείται περισσότερος χρόνος για την αξιολόγηση τους

## Η ανάκτηση πληροφοριών στο κέντρο του ενδιαφέροντος

---

Η ανάκτηση Πληροφοριών (information retrieval) εξετάζει την αντιπροσώπευση , την αποθήκευση, την οργάνωση και την πρόσβαση στα στοιχεία πληροφοριών. Η αντιπροσώπευση και η οργάνωση των στοιχείων πληροφοριών πρέπει να παρέχουν στο χρήστη την εύκολη πρόσβαση στις πληροφορίες για τις οποίες ενδιαφέρεται [13].

Η ανάκτηση στοιχείων πληροφοριών στα πλαίσια ενός συστήματος IR αποτελείται κυρίως από τον καθορισμό των εγγράφων μιας συλλογής τα όποια περιέχουν τις λέξεις κλει-

διά στην ερώτηση του χρήστη τα οποία δεν είναι αρκετά για την κάλυψη των πληροφοριακών τους αναγκών. Στην πραγματικότητα ο χρήστης ενός συστήματος ανάκτηση πληροφοριών ανησυχεί περισσότερο με την ανάκτηση των πληροφοριών για ένα συγκεκριμένο θέμα από ότι με την ανάκτηση του στοιχείου που ικανοποιεί μια δεδομένη ερώτηση. Μια γλώσσα ανάκτησης στοιχείων στοχεύει στην ανάκτηση όλων των αντικειμένων που ικανοποιούν τους σαφώς καθορισμένους όρους όπως εκείνοι εμφανίζονται σε μια αλγοριθμική έκφραση [14].

Ο χρήστης ενός συστήματος ανάκτησης πληροφοριών πρέπει να μεταφράσει την πληροφοριακή του ανάγκη σε μια ερώτηση στην γλώσσα που παρέχεται από το σύστημα. Αυτό μεταφράζεται ως ένα σύνολο λέξεων που μεταβιβάζουν την σημασιολογία της πληροφοριακής ερώτησης.

Ένα σύστημα (IR) θα πρέπει να ερμηνεύει το περιεχόμενο ενός στοιχείου πληροφορίας που υπάρχει σε συλλογή κειμένων με βάση την σχετικότητα στην ερώτηση του χρήστη. Αυτή η ερμηνεία περιλαμβάνει την εξαγωγή των συντακτικών και σημασιολογικών πληροφοριών από ένα σύνολο εγγράφων και την χρησιμοποίηση αυτών των πληροφοριών που ταιριάζουν στην πληροφοριακή ανάγκη. Κατά συνέπεια η έννοια σχετική είναι στο κέντρο της ανάκτησης πληροφοριών. Στην πραγματικότητα στόχος ενός συστήματος ανάκτησης πληροφοριών είναι να ανακτηθούν όλα τα έγγραφα που είναι σχετικά με μία ερώτηση χρηστών ανακτώντας όσο το δυνατόν λιγότερα άσχετα έγγραφα [13].

### Βιβλιοθήκες και συστήματα ανάκτησης πληροφοριών

Οι βιβλιοθήκες ήταν μεταξύ των πρώτων που υιοθέτησαν τα συστήματα IR για τις πληροφορίες. Η πρώτη γενιά τέτοιων συστημάτων αφορούσε την αυτοματοποίηση του δελτίοκαταλόγου και επέτρεψαν την αναζήτηση με βάση τον συγγραφέα, τίτλο. Στην δεύτερη γενιά η ανάγκη για αποτελεσματικότερη αναζήτηση επέτρεψε την χρήση και την εξαγωγή αποτελεσμάτων με λέξεις κλειδιά. Σήμερα η εστίαση των ερευνών στρέφεται στην βελτίωση των γραφικών διεπαφών, στα χαρακτηριστικά γνωρίσματα υπερκειμένων που βασίζονται σε αρχιτεκτονικές ανοικτών συστημάτων.

Εξετάζοντας τις μηχανές αναζήτησης διαπιστώνεται ότι συνεχίζουν να χρησιμοποιούν τους δείκτες των βιβλιοθηκονόμων. Πρέπει όμως να σημειωθεί ότι πολλά πράγματα έχουν αλλάξει σήμερα. Καταρχάς η πρόσβαση στις πληροφορίες έγινε φτηνότερη ενώ υπάρχει καθολική πρόσβαση σε αυτές ανά πάσα στιγμή και από οποιοδήποτε σημείο κυρίως μέσω του διαδικτύου [1].

Τα προβλήματα πάντως δεν παύουν να υπάρχουν καταρχάς οι άνθρωποι βρίσκουν ακόμα δύσκολο το πώς να ψάχνουν και να βρίσκουν τις πληροφορίες που τους ενδιαφέρουν επομένως η ανάπτυξη τεχνικών που θα επιτρέπουν ανάκτηση υψηλής ποιότητας είναι το

ενα βήμα και δεύτερο η ποιότητα του στόχου ανάκτηση επηρεάζεται πολύ από την αλληλεπίδραση χρηστών με το σύστημα. Ειδικότερα στο πρόβλημα της πολυγλωσσία των ψηφιακών στοιχείων που εμφανίζονται θα πρέπει να υιοθετηθούν αρχιτεκτονικές που σέβονται τη γλωσσική και πολιτιστική ποικιλομορφία. Η πολυγλωσσία είναι μια πτυχή του ευρύτερου ζητήματος της πολυπολιτισμικότητας, που περιλαμβάνει, παραδείγματος χάριν: [5]

- τρόπος με τον οποίο οι ημερομηνίες αντιπροσωπεύονται στα διαφορετικά ημερολόγια,
- Η κατεύθυνση στην οποία το κείμενο παρουσιάζεται και διαβάζεται,
- Πολιτιστικές υποδηλώσεις ορισμένων εικονιδίων και εικονογραμμάτων.

Το interface των πληροφοριακών συστημάτων θα πρέπει να είναι σε θέση να χειριστεί αιτήματα σε φυσική γλώσσα χωρίς αυτό να μειώνει την συνολική απόδοση του συστήματος. Ένα άλλο θέμα που προκύπτει είναι ότι θα πρέπει το οπτικό αποτέλεσμα της αναζήτησης να όνεται με τέτοιο τρόπο ώστε να υπάρχει ξεκάθαρη αναπαράσταση της σχετικότητας μεταξύ του θέματος αναζήτησης και των αποτελεσμάτων που εμφανίζονται

οστα συστήματα IR πρέπει να μπορούν να παρέχουν την δυνατότητα να χειριστούν κείμενα των οποίων η σημασιολογία δεν προέρχεται ή δεν συμφωνεί με τις προβλέψεις τυποποιημένων και ελεγχόμενων λεξιλογίων, ώστε να μπορεί να διευρύνει την εξαγωγή αποτελεσμάτων από έγγραφα τα οποία δεν περιέχουν άμεσες αναφορές προς το ζητούμενο θέμα (link Compaq) Κατά συνέπεια μια καλύτερη κατανόηση της συμπεριφοράς των χρηστών έχει επιπτώσεις στο σχεδιασμό και την επέκταση νέων τεχνικών ανάπτυξης πληροφοριών.

Απο τα παραπάνω γίνεται προφανές ότι η αύξηση του όγκου πληροφοριών δεν προκάλεσε μόνο προβλήματα αποθήκευσης, άλλα αφήνει ακόμα άλυτο το θέμα της βέλτιστης ανάκτησης.

Σκοπός μας είναι να δοθεί μια πλήρη κάλυψη των σημαντικότερων ιδεών που έχουν εμφανιστεί για την αυτοματοποιημένη ανάκτηση πληροφοριών στην ψηφιακή εποχή άλλα συγχρόνως παρουσιάζεται μια μέθοδο ανάκτησης που στηρίζεται σε τεχνολογίες τεχνητής νοημοσύνης και ειδικότερα στα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) Neural network

Η διαδικασία της ανάκτησης

Για την περιγραφή της διαδικασίας της ανάκτησης είναι απαραίτητο να καθοριστεί η βάση δεδομένων κειμένων και συγκεκριμένα α) τα έγγραφα που χρησιμοποιούνται β) οι διαδικασίες που εκτελούνται στο κείμενο γ) το πρότυπο κειμένων δηλ η μορφή κειμένων και ποια στοιχεία μπορούν να ανακτηθούν (2).

Μόλις καθοριστεί η λογική άποψη κειμένων ορίζεται ένας δείκτης κειμένου ο οποίος ορίζει την δομή των πληροφοριών και επιτρέπει την γρήγορη ανάκτηση των δεδομένων. Πολλές φορές η ερωτήσεις των χρηστών είναι ανεπαρκής με αποτέλεσμα οι κακώς διατυπωμένες ερωτήσεις να οδηγούν σε φτωχή ανάκτηση.

Σε ένα συμβατικό σύστημα ανάκτησης πληροφοριών το καταχωρημένο κείμενο προσδιορίζεται από τα σύνολα λέξεων κλειδιών τα οποία ονομάζονται όροι ευρετηρίων (index terms). Τα αιτήματα για τις πληροφορίες εκφράζεται ως συνάρτηση των λογικών πράξεων (and, not, or) και των όρων αναζήτησης. Οι όροι που χαρακτηρίζουν το καταχωρημένο κείμενο μπορούν να σημανθούν χειροκίνητα από το εκπαιδευμένο προσωπικό ή εναλλακτικά, με χρήση αυτόματων μεθόδων δημιουργίας ευρετηρίων. Στα πλήρη συστήματα ανάκτησης κειμένου αποφεύγεται η ανάλυση περιεχομένου (indexing) με την χρησιμοποίηση των λέξεων που βρίσκονται στο σώμα των εγγραφών για τον προσδιορισμό του νοηματικού περιεχομένου [23].

Όλες οι υπάρχουσες προσεγγίσεις στην ανάκτηση κειμένων είναι βασισμένες στους σχετικούς όρους που βρίσκονται στο κείμενο. Σε μια τυπική προσέγγιση προσδιορίζονται οι μεμονωμένες λέξεις που εμφανίζονται στα έγγραφα. Στην πιο απλή περίπτωση το αίτημα αναζήτησης αποτελείται από ένα όρο αναζήτησης, δηλαδή μια ακολουθία αλφαριθμητικών χαρακτήρων. Σε πιο πολύπλοκες περιπτώσεις το αίτημα αναζήτησης αποτελείται από πολλούς όρους αναζήτησης οι οποίοι συνδέονται με λογικούς τελεστές. Σε αυτές τις περιπτώσεις πέραν από την εφαρμογή του αλγόριθμου (substring δοκιμή) για κάθε όρο, πραγματοποιείται και το πρόσθετο βήμα εφαρμογής των λογικών τελεστών στα αποτελέσματα του αρχικού αλγόριθμου [22].

Ο χρόνος αναζήτησης είναι γραμμικός ως προς το μέγεθος των εγγράφων, αλλά εκθετικός ως προς το μέγεθος της συμβολοσειράς αναζήτησης. Ο αλγόριθμος για τη substring δοκιμής είναι ο ακόλουθος: "Συγκρίνονται οι χαρακτήρες της συμβολοσειράς αναζήτησης με τους αντίστοιχους χαρακτήρες του εγγράφου. Εάν δεν υπάρχει το επιθυμητό αποτέλεσμα της εύρεσης του συγκεκριμένου substring, μετατοπίζεται η συμβολοσειρά αναζήτησης κατά μία θέση προς τα δεξιά και συνεχίζεται η ίδια διαδικασία έως ότου είτε βρεθεί η συμβολοσειρά είτε τελειώσει το τέλος έγγραφο".

Ο παραπάνω αλγόριθμος αν και είναι απλός και σχετικά ακριβείς στην εφαρμογή του δεν είναι αρκετά γρήγορος [23].

Μια άλλη προσέγγιση της ανάκτησης κειμένων είναι τα "αρχεία υπογραφών". Σε αυτήν την μέθοδο, κάθε έγγραφο παράγει μια συμβολοσειρά δυαδικών ψηφίων ("υπογραφή"), χρησιμοποιώντας hashing στις λέξεις του κατά την κωδικοποίησή τους. Οι προκύπτουσες υπογραφές εγγράφων καταχωρούνται διαδοχικά σε ένα χωριστό αρχείο (ar-

χείο υπογραφών) το οποίο είναι πολύ μικρότερο από το αρχικό κείμενο, και μπορεί να επεξεργαστεί πολύ γρηγορότερα.

Το βασικό πλεονέκτημα των αρχείων υπογραφών είναι η μεγάλη ταχύτητα με την οποία πραγματοποιείται η επεξεργασία τους. Εντούτοις, το μέγεθος αρχείων υπογραφών είναι ανάλογο προς το μέγεθος των κειμένων και αυτό αποτελεί πρόβλημα για τις ογκώδεις συλλογές κειμένων όπως τις ψηφιακές βιβλιοθήκες. Τα αρχεία υπογραφών περιέχουν τους σχετικούς όρους από τα έγγραφα. Τέτοιοι κομματιασμένοι όροι καλούνται υπογραφές και τα αρχεία που τους περιέχουν είναι αρχεία υπογραφών. Υπάρχουν πολύ τρόποι ώστε να εξαχθούν οι υπογραφές από τα έγγραφα. Οι τρεις βασικές μέθοδοι είναι 1) WS (υπογραφές του Word), 2) Sc (κωδικοποίηση), Π.Χ. (συμπίεση BitBlock) και 3) RL (συμπίεση RunLength).

Εως τώρα η μέθοδος των «αρχείων υπογραφών» έχει εφαρμοστεί σε μικρές βάσεις κειμένων οι οποίες δεν είναι πάντα αντιπροσωπευτικές. Αντίθετα μέχρι σήμερα δεν υπάρχει εφαρμογή της παραπάνω μεθόδου σε μεγάλες βάσης κειμένων όπως είναι μια ψηφιακή βιβλιοθήκη [25].

Μια άλλη μέθοδο ανάκτησης κειμένου είναι αυτή της αντιστροφής στην οποία κάθε έγγραφο μπορεί να αντιπροσωπευθεί από ένα κατάλογο (keywords - λέξεις κλειδιά) τα οποία περιγράφουν το περιεχόμενο του εγγράφου για την ανάκτησης του [27]. Η γρήγορη ανάκτηση μπορεί να επιτευχθεί με την εξέταση των λέξεων κλειδιών στο κείμενο. Οι λέξεις κλειδιά καταχωρούνται, αλφαβητικά, στο "αρχείο ευρετηρίων", για κάθε λέξη κλειδί διατηρούμε έναν κατάλογο δεικτών.

### Αναζήτηση κειμένων στο διαδίκτυο

Ο παγκόσμιος ιστός εξελίσσεται σε μια καθολική αποθήκη της ανθρώπινης γνώσης και του πολιτισμού καθιστώντας αναγκαία την περιγραφή των πληροφοριών (καταλογογράφηση) που υπάρχουν στους πόρους του διαδικτύου.

Τα metadata δεν είναι πλήρης στοιχεία αλλά είναι ένα είδος συντροφικού ταξιδιώτη μια πυξίδα, που περιγράφουν την πληροφορία που υπάρχει σε κάθε ψηφιακό αντικείμενο (έγγραφο, εικόνα, βίντεο, ήχος).

Τα μεταδεδομένα είναι πληροφορίες κατανοητές από τους υπολογιστές του διαδικτύου άλλα αόρατες για τον τελικό χρήστη, υπό την έννοια ότι δεν έχουν καμιά επίδραση στην οπτική εμφάνιση της σελίδας [19].

### Η σημασία ενός συστήματος IR

Η σημασία ενός συστήματος πληροφοριών IR είναι να παρέχει α) Αυξημένες δυνατό-

τητες πρόσβασης β) ανεξαρτησία στην χρήσης των πληροφοριακών πόρων γ) να επιτρέπει επεκτάσεις ώστε οι ιδιαίτερες ανάγκες μιας δεδομένης εφαρμογής να μπορούν να προσαρμοστούν.

Όλα τα παραπάνω αυτοματοποιημένα συστήματα αποτελούν χρήσιμα εργαλεία στην προσπάθεια μας να διαχειριστούμε την πληροφορία, όμως χαρακτηρίζονται από ελλείψεις ακρίβειας και επιδόσεων

Πάνω σε αυτά τα προβλήματα οι τεχνολογίες των ΤΝΔ έχουν να προτείνουν κάποιες λύσεις οι οποίες φαίνονται να ανταποκρίνονται στις προκλήσεις που προκύπτουν στην ψηφιακή εποχή [14].

### Τι είναι τα Τεχνητά Νευρωνικά Δίκτυα ΤΝΔ

Οι άνθρωποι δεν είναι πιο γρήγοροι ούτε πιο ακριβείς από τους υπολογιστές. Οι σημαντικοί υπολογιστές φαίνονται πολύ πιο ικανοί στο να εκτελούν πολύπλοκες αριθμητικές και λογικές πράξεις. Σε πολλούς όμως τομείς η ανθρώπινη ικανότητα υπερτερεί κατά πολύ των δυνατοτήτων των μηχανών. Για παράδειγμα οι άνθρωποι αναγνωρίζουν πολύ πιο εύκολα αντικείμενα και κατανοούν τις σχέσεις μεταξύ τους μέσα στο φυσικό περιβάλλον, έστω κι αν αυτά είναι παραμορφωμένα ή δεν είναι ορατά εξ ολοκλήρου. Η ικανότητα της μάθησης μέσω εμπειρίας είναι ένα από τα κύρια χαρακτηριστικά της ανθρώπινης νοημοσύνης. Επιπλέον η ανθρώπινη μνήμη μπορεί να αποθηκεύσει μεγάλη ποσότητα πληροφοριών και έχει την δυνατότητα να βρίσκει συσχετιζόμενες πληροφορίες μέσα σε αυτή χωρίς μεγάλη προσπάθεια. Αντίθετα οι υπολογιστές έχουν την δυνατότητα να απομνημονεύουν τεράστιες ποσότητες πληροφοριών, αλλά είναι δύσκολο να κάνουν αυτή την γνώση ένα αποτελεσματικό μέρος των δραστηριοτήτων τους.

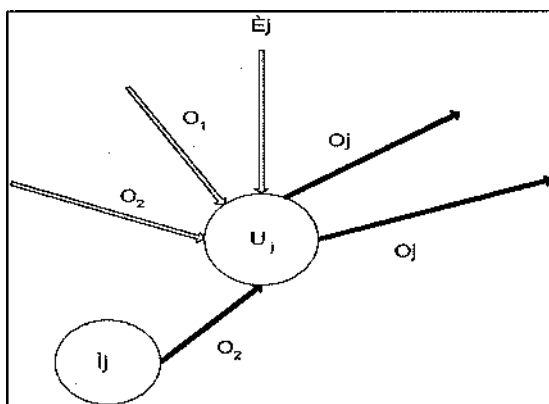
Μια πιθανή εξήγηση είναι ότι ο ανθρώπινος εγκέφαλος και οι υπολογιστές λειτουργούν εντελώς διαφορετικά. Οι άνθρωποι είναι πιο έξυπνοι από τους υπολογιστές επειδή ο εγκέφαλος τους χρησιμοποιεί μια βασική υπολογιστική αρχιτεκτονική η οποία είναι πιο κατάλληλη για την αντιμετώπιση της φυσικής επεξεργασίας πληροφοριών. Ο εγκέφαλος είναι ένας απίστευτα δυνατός υπολογιστής. Μόνο το περίβλημά του περιέχει πάνω από εκατομμύρια νευρώνες. Από κάθε νευρώνα ξεκινούν χιλιάδες ίνες οι οποίες συναντούν ίνες από άλλους νευρώνες σε σημεία που αποκαλούνται συνάψεις. Πιστεύεται ότι η μνήμη και η γνώση αποθηκεύονται στις συνάψεις αυτές και όχι σε συγκεκριμένους νευρώνες. Κάθε νευρώνας στέλνει σήματα διέγερσης ή αποδιέγερσης σε άλλους νευρώνες. Η κατάσταση ενεργοποίησης ενός νευρώνα εξαρτάται από τα σήματα διέγερσης ή αποδιέγερσης που λαμβάνει από τους νευρώνες με τους οποίους συνδέεται. Ο συνδυασμός του τεράστιου αυτού αριθμού απλών υπολογιστικών στοιχείων (νευρώνες) δίνει ένα πολύ ισχυρό επακόλουθο.

Το τεχνητό νευρωνικό δίκτυο σχεδιάστηκε ώστε να αντιγράψει τα χαρακτηριστικά και τις λειτουργίες των βιολογικών νευρώνων του ανθρώπινου εγκεφάλου. Πρόκειται για ένα σύστημα επεξεργασίας πληροφορίας σχεδιασμένο με βάση την λογική συνδεσμολογία των νευρώνων του ανθρώπινου εγκεφάλου. Αποτελείται από ένα μεγάλο αριθμό πολλαπλά διασυνδεδεμένων επεξεργαστών (νευρώνες) που εργάζονται σε πλήρη συμφωνία μεταξύ τους, και εκτελούν διάφορους υπολογισμούς πάνω σε δεδομένα εισόδου, τα τμήματα αυτά συνδέονται μεταξύ τους με αμφίδρομα κανάλια επεξεργασίας που μεταφέρουν κωδικοποιημένες πληροφορίες με την μορφή αριθμών

Τα τεχνητά νευρωνικά δίκτυα διέπονται από κανόνες εκπαίδευσης που προσαρμόζονται με βάση τα πρότυπα των τιμών εισόδου και έχουν την δυνατότητα μάθησης μέσα από τα παραδείγματα (καταχώρηση εμπειρικής γνώσης) και παράσχουν λύσεις στα προβλήματα που απαιτούν την ανθρώπινη παρατήρηση και σκέψη (διαθέσιμη γνώση για χρήση).

Πώς λειτουργεί ένα τεχνητό νευρωνικό δίκτυο

**Ένα γενικό πλαίσιο λειτουργίας των ΤΝΔ**



Σχήμα 1 Γενικό μοντέλο ΤΝΔ

Σε ένα ΤΝΔ υπάρχουν τα εξής τμήματα

- Ένα σύνολο επεξεργαστικών μονάδων

Το πρώτο στάδιο του ορισμού ενός ΤΝΔ είναι ο καθορισμός ενός συνόλου από επεξεργαστικές μονάδες. Σε ένα δίκτυο μια επεξεργαστική μονάδα είναι ένας τεχνητός νευρώνας με περιορισμένη μνήμη και επεξεργαστική ισχύ. Στο σχήμα 1 οι νευρώνες παριστάνονται γραφικά με κύκλους. Κάθε τεχνητός νευρώνας δέχεται εισόδους από νευρώνες με τους οποίους συνδέεται και υπολογίζει μια τιμή εξόδου σαν συνάρτηση των ει-

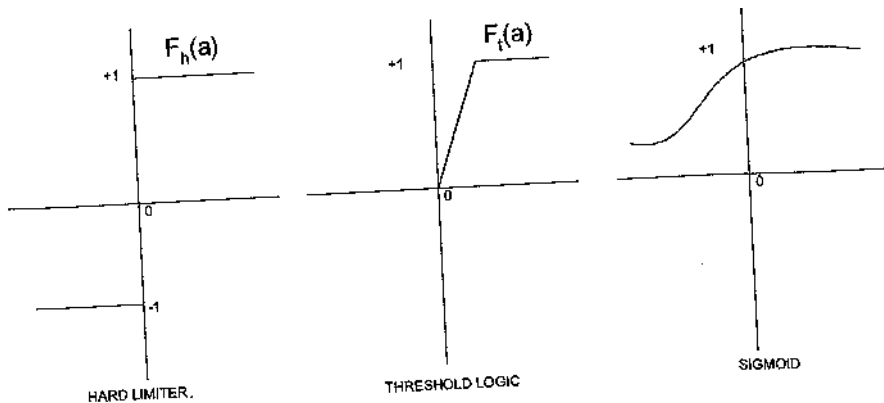


σόδων του, την οποία την διοχετεύει με την σειρά του σε άλλους νευρώνες με τους οποίους επικοινωνεί. Το σύστημα λειτουργεί παράλληλα και πολλές μονάδες έχουν δυνατότητα να πραγματοποιούν ταυτόχρονα τους υπολογισμούς τους. Υπάρχουν τρεις κατηγορίες μονάδων : μονάδες εισόδου οι οποίες λαμβάνουν τα δεδομένα εισόδου από εξωτερικές πηγές, μονάδες εξόδου οι οποίες στέλνουν τα αποτελέσματα εκτός συστήματος και κρυμμένες μονάδες. Οι κρυμμένες μονάδες δεν είναι ορατές στον εξωτερικό κόσμο και οι εισοδοί τους καθώς και οι εξοδοί τους βρίσκονται εντός του Τ.Ν.Δ.

- Κατάσταση ενεργοποίησης

Είναι χρήσιμο να κατανοήσουμε την επεξεργασία στο ΤΝΔ σαν μία εξέλιξη στο χρόνο μιας μορφής ενεργοποίησης του συνόλου των μονάδων. Η ενεργοποίηση αυτή  $a_j(t)$  κάποιος μονάδας ( $u_j$ ) αντιπροσωπεύει την κατάσταση της στην χρονική στιγμή ( $t$ ). Διαφορετικές υποθέσεις για την τιμή ενεργοποίησης που μπορεί να έχει μια μονάδα γίνονται από διάφορα μοντέλα ΤΝΔ μοντέλα που υπάρχουν. Οι τιμές αυτές μπορεί να είναι συνεχείς και διακριτές. Εάν είναι συνεχείς τότε μπορεί να είναι φραγμένες ή μη φραγμένες. Εάν είναι διακριτές τότε μπορεί να είναι δυαδικές ή να παίρνουν τιμές από ένα μικρό σύνολο τιμών. Διαφορετικές υποθέσεις για τις τιμές ενεργοποίησης οδηγούν σε μοντέλα με λίγο διαφορετικά χαρακτηριστικά.

- Συνάρτηση εισόδου εξόδου



Σχήμα 2 Συναρτήσεις Εξόδου

Οι μονάδες αλληλοεπιδρούνται μεταβιβάζοντας και λαμβάνοντας και λαμβάνοντας σήματα από γειτονικές μονάδες. Η ισχύς των σημάτων και κατά συνέπεια ο βαθμός αλληλοεπίδρασης καθορίζεται από τον βαθμό ενεργοποίησης κάθε μονάδας  $u_j$  αντιπροσωπεύει μια συνάρτηση εξόδου  $f_j$  η οποία μετασχηματίζει την παρούσα κατάσταση  $a_j(t)$  της μονάδας σε ένα σήμα εξόδου  $o_j(t) = f_j[a_j(t)]$  Σε μερικά μοντέλα το σήμα εξόδου είναι το ίδιο με την κατά-

στάση ενεργοποίησης της μονάδας δηλ  $F(x)=\chi$ . Στις περισσότερες περιπτώσεις, η  $F(x)=\chi$  είναι κάποια οριακή συνάρτηση (threshold function)

Κατά αυτό τον τρόπο, μία μονάδα επηρεάζει μια γειτονική της, μόνο όταν η τιμή ενεργοποίησης της είναι μεγαλύτερη από μια προκαθορισμένη τιμή.

- Πρότυπο συνδέσεων μεταξύ μονάδων

Οι μονάδες συνδέονται μέσω του πρότυπου σύνδεσης το οποίο αποτελεί την γνώση του συστήματος και καθορίζει πώς θα ανταποκριθεί το σύστημα σε μια τυχαία είσοδο. Γενικά η κατάσταση ενεργοποίησης δίνεται από την εξίσωση:

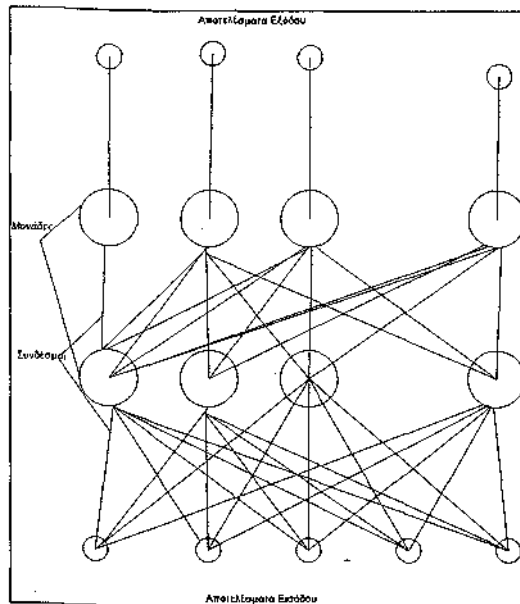
$$\alpha_j(t) = \sum_k^i w_{jk} O_k(t) + \theta_j$$

όπου  $w_{jk}$  είναι οι συντελεστές βαρών των εισόδων της μονάδας  $u_j$  και  $\theta_j$  είναι ένα εσωτερικό όριο που χαρακτηρίζει την ίδια την μονάδα. Σε μια τέτοια περίπτωση, η αναπαράσταση του προτύπου σύνδεσης καθορίζεται από τον προσδιορισμό του συντελεστή βάρους κάθε σύνδεσης του δικτύου. Θετική τιμή ενός συντελεστή βάρους αναπαριστά μια είσοδο διέγερσης και αρνητική τιμή αναπαριστά μια είσοδο αποδιέγερσης. Ο συντελεστής  $w_{jk}$  είναι θετικός εάν η μονάδα  $U_k$  διεγείρει την μονάδα  $u_j$ , αρνητικός αν την αποδιεργεί και μηδέν αν η μονάδα  $U_k$  δεν συνδέεται με την μονάδα  $u_j$ . Η απόλυτη τιμή του συντελεστή βάρους  $w_{jk}$  καθορίζει την ισχύ της σύνδεσης.

- Τοπολογία δικτύου

Συνήθως οι μονάδες του δικτύου διατάσσονται σε ξεχωριστές δομές οι οποίες καλούνται «στρώματα» (layer). Τα πρωταρχικά μοντέλα αποτελούνται από ένα απλό στρώμα όπου κάθε είσοδος του δικτύου συνδέεται με όλες τις μονάδες του στρώματος αυτού. Οι πληροφορίες ρέουν διαμέσω του στρώματος από τις εισόδους και στις εξόδους χωρίς ανάδραση των εξόδων. Τα μοντέλα αυτής της κλάσης ονομάζονται feed-forward μοντέλα.

Στα δίκτυα πολλαπλών στρωμάτων τα μεσαία στρώματα ονομάζονται κρυμμένα (hidden layers) και αποτελούνται από κρυμμένες μονάδες. Έχουν εισαχθεί επίσης μοντέλα τα οποία χρησιμοποιούν την έννοια των feed-backward συνδέσεων. Τα δίκτυα αυτά (ανάδρασης) βελτιώνουν την διαδικασία εκπαίδευσης και μάθησης του δικτύου τροφοδοτώντας προς τα πίσω αποτελέσματα (σφάλματα) από το επόμενο στρώμα σε ένα προηγούμενο.



Σχήμα 3 Μοντέλο Πολλαπλών *feed-forward*

- Κανόνας μάθησης

Ο κανόνας μάθησης είναι ο τρόπος εύρεσης των σωστών συντελεστών βαρών έτσι ώστε οι σωστές μορφές ενεργοποίησης να δημιουργηθούν κάτω από κατάλληλες συνθήκες. Για να αλλάξουμε την δομή επεξεργασίας ή της γνώσης σε ένα ΤΝΔ χρειάζεται να τροποποιήσουμε το πρότυπο συνδέσεων. Οι αλλαγές αυτές μπορούν να γίνουν με τους εξής τρόπους: 1) ανάπτυξη καινούργιων συνδέσεων 2) Αφαίρεση παλιών συνδέσεων.

### Γιατί να χρησιμοποιούμε Νευρωνικά δίκτυα.

Τα νευρωνικά δίκτυα με την δυνατότητα τους να εξάγουν έννοιες-αποτελέσματα από περίπλοκα ή ανακριβή στοιχεία μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να ανιχνεύσουν τάσεις που είναι σύνθετες να παρατηρηθούν από άλλα υπολογιστικά συστήματα.

Τα πλεονεκτήματα ενός νευρωνικού δικτύου είναι

#### **1. Προσαρμοσμένη μάθηση (supervised -associative learning)**

Το ΤΝΔ εκτελεί εργασίες βασισμένο σε δεδομένα (ζεύγη εισόδων-εξόδων) που έχουν δοθεί κατά την εκπαίδευση του.

#### **2. Αυτοοργάνωση (self-organization)**

Όπου μια ομάδα εξόδου εκπαιδεύεται για να ανταποκρίνεται σε ομάδες προτύπων που

υπάρχουν στην είσοδο. Εδώ σε σχέση με την προσαρμοσμένη μάθηση δεν υπάρχουν εκ των προτέρων καθορισμένα σύνολα κατηγοριών στα οποία θα ταξινομηθούν τα πρότυπα άλλα αναπτύσσει την δική του αναπαράσταση στα ερεθίσματα εισόδου.

### **3. Λειτουργία σε πραγματικό χρόνο**

Οι υπολογισμοί ενός ΤΝΔ μπορούν να εκτελεστούν παράλληλα, ενώ συγχρόνως παρουσιάζουν ανοχή στην διαστρέβλωση εικόνων-ηρώτύπων(π.χ αμφισημία λέξεων ) Δυνατότητα γενίκευσης

### **4. Δυνατότητα διόρθωση λαθών**

Μερική καταστροφή του δικτύου έχει ως αποτέλεσμα τη μείωση της ταχύτητας επεξεργασίας. Τα μερικώς κατεστραμμένα νευρωνικά δίκτυα είναι σε θέση να αναδιατάσσουν τις διεργασίες επεξεργασίας του δικτύου διοχετεύοντας όλη τη διαδικασία επεξεργασίας πληροφοριών στο υγιές τμήμα του νευρωνικού δικτύου.

### **Μειονεκτήματα**

- Οι γνώσεις πάνω στο τρόπο που ο ανθρώπινος εγκέφαλος εκπαιδεύεται για να επεξεργαστεί πληροφορίες είναι ακόμα συγκεχυμένες και οι θεωρίες αφθονούν
- Οι μαθηματικές θεωρίες που χρησιμοποιούνται είναι ακόμα υπό ανάπτυξη και δεν μπορούν να εγγυηθούν την απόδοση ενός νευρωνικού δικτύου.

### **LATENT SEMANTIC ANALYSIS**

Ένας αλγόριθμος τεχνητών νευρωνικών δικτύων που παρουσιάζει ιδιαίτερο ενδιαφέρον είναι ο LSA (Latent Semantic Analysis), ανήκει στην κατηγορία των Self organized Networks (δίκτυο αυτοεπίβλεψης) [21] και χρησιμοποιείται για την εξαγωγή του περιεχομένου του κειμένου και την παραγωγή μιας περίληψης για αυτό.

Το κύριο χαρακτηριστικό του είναι ότι δεν κάνει χρήση λεξικών ή γραμματικών κανόνων. Σε αντίθεση με τις κλασικές μεθόδους που βασίζονται στην καταμέτρηση των όμοιων λέξεων του κειμένου. Ο LSA βασίζεται στην μέτρηση της νοηματικής συσχέτισης η οποία αντανακλάται στην νοηματική ομοιότητα μεταξύ συνώνυμων, ουσιαστικών, επιθέτων και άλλων μερών του λόγου τα οποία τείνουν να χρησιμοποιηθούν κατά παρόμοιο τρόπο. Επιπρόσθετα η αυτοματοποίηση της όλης διαδικασίας, επιτρέπει όχι μόνο την επεξεργασία μεγάλου όγκου κειμένων, αλλά δίδεται η δυνατότητα να εξεταστούν περαιτέρω πτυχές της δομής των κειμένων.

Κάθε ενότητα αναπαρίσταται με ένα σταθμισμένο διάνυσμα της συχνότητας εμφάνισης όλων των λέξεων της. Το διάνυσμα που προκύπτει αναπαριστά το νοηματικό περιεχόμενο της ενότητας σε ένα αφηρημένο διανυσματικό χώρο εννοιών. Με τη χρήση της

singular value decomposition (SVD) [18] και λαμβάνοντας υπόψη μόνο της ανώτερες ιδιοτιμές που προκύπτουν επιτυγχάνεται από την μια δραστική μείωση της διαστασιμότητας του χώρου εννοιών, πράγμα που έχει δραστικές επιδράσεις στους απαιτούμενους υπολογιστικούς χώρους ενώ παράλληλα αποκαλύπτεται μία αφαιρετική δομή η οποία συσχετίζει μεμονωμένες λέξεις με το νοηματικό περιεχόμενο στο οποίο εμφανίζονται.

Η συσχέτιση ενισχύεται περαιτέρω μέσω μιας αυτοργανούμενης διαδικασίας ομαδοποίησης με χρήση ενός T.N.Δ Kohonen ή με κάποιο άλλο αλγόριθμο ομαδοποίησης (clustering).

Όταν μια ερώτηση τεθεί στο σύστημα αρκεί να μορφοποιηθεί ένα συμβατό διάνυσμα λέξεων και να βρεθούν οι ομάδες στις οποίες ταξινομείται αυτό το διάνυσμα. Συνήθως αυτό γίνεται με βάση τις αποστάσεις του ερωτήματος από τις ομάδες.

Η αφαιρετική δομή που προκύπτει συσχετίζει μεμονωμένες λέξεις με το νοηματικό περιεχόμενο των κειμένων στις οποίες εμφανίζεται είτε ή ίδια είτε συγγενείς λέξεις. Δηλαδή είναι δυνατό να συσχετιστεί μια λέξη με μία έννοια ακόμα και αν δεν εμφανίζεται εκεί. Τέτοιου είδους συσχετισμοί είναι εξαιρετικά δύσκολο να γίνουν με κλασικές στατιστικές μεθόδους επεξεργασίας φυσικής γλώσσας. Πρέπει να σημειωθεί ότι στη διαδικασία αυτή δεν λαμβάνονται υπόψη πληροφορίες για την συντακτική δομή του κειμένου ή την σειρά εμφάνισης των λέξεων.

Όσο μεγαλύτερο σε πλήθος και νοηματικό εύρος είναι η συλλογή κειμένων που χρησιμοποιήθηκε για την κατασκευή του μητρώου, τόσο πιο ισχυροί εννοιολογικοί συσχετισμοί αναπτύσσονται.

### Συμπεράσματα

Ο Χώρος της βιβλιοθηκονομίας και της επιστήμησ της πληροφόρησης έχει πολλά να κερδίσει από την χρήση των νευρωνικών δικτύων. Η δυνατότητα τους να μαθαίνουν μέσω παραδειγμάτων τα καθιστά ευέλικτα και αποτελεσματικά.

Τα ΤΝΔ από την μια παρέχουν γρήγορη και λεπτομερή πρόσβαση στο υλικό και από την άλλη μπορούν να χρησιμοποιηθούν από απλούς χρήστες με ελάχιστες υλικοτεχνικές απαιτήσεις και δεξιότητες.

Με τον τρόπο αυτό ενισχύεται ο ρόλος της βιβλιοθήκης διευκολύνοντας και υποστηρίζοντας τις ερευνητικές και εκπαιδευτικές δραστηριότητες της ακαδημαϊκής κοινότητας.

Το πιο ελπιδοφόρο από την πλευρά των νευρωνικών δικτύων είναι η πιθανότητα ότι κάποια μέρα «συνειδητά» δίκτυα θα κάνουν την εμφάνιση τους, δημιουργώντας συνθή-

κες αλληλοελέγχου σφαλμάτων μεταξύ τεχνίτης και ανθρώπινης εφυήας, άλλωστε πολλοί υποστηρίζουν ότι η συνείδηση είναι μια μηχανική ιδιότητα επομένως η αναπαράστασης της είναι ρεαλιστική .

Ίσως από κάποιους η χρήση της τεχνητής νοημοσύνης σε συστήματα ανάκτησης πληροφοριών να χαρακτηριστεί ως υπερβολικό επιχείρημα με αβέβαια αποτελέσματα, όμως στην πρόσκληση της ψηφιακής εποχής θα πρέπει να απαντήσουμε με ανοικτό μυαλό και διάθεση για πειραματισμό

### Βιβλιογραφία

---

- 1) Blair D. and Maron M (1985) An evaluation of retrieval effectiveness for full-text document-retrieval system. *Communications of ACM* 28(3) -.289-299
- 2) Buckland, M.K(1991) Information as thing *Journal of the American society for information science* 42:586-588
- 3) Buckland, M.K(1997) What is a document. *Journal of the American society for information science* 48(9): 804-809
- 4) Harter, S.P& Hert, C.A (1997) Evaluation of information retrieval systems: approaches, issues and methods. *Annual review of information science and technology* 32:3-
- 5) Croft, B. Lucia T.(1989) Retrieving documents by plausible inference :an experimental study. *Information processing and Management* 25(6) 599-614.
- 6) Γ. Παπαδουράκης και Γ. Μπεμπής "Τεχνητά Νευρωνικά Δίκτυα Ι", Τεχνικά Χρονικά Επιστημονική Έκδοση Τ.Ε.Ε.- Επιστημονική Περιοχή Β, V. 11, N.3, pp.57-73, 1991
- 7) DAMERAU, F.j., 'Automated language processing', *Annual Review of Information Science and Technology*, 11,107-161 (1976).
- 8) SPARCK JONES, K. and KAY, M., *Linguistics and Information Science*, Academic Press. New York and London (1973).
- 9) MONTGOMERY, C.A., "Linguistics and information science", *Journal of the American Society for Information Science*, 23,195-219 (1972).
- 10) FARRADANE, J., RUSSELL, j. M. and YATES-MERCER, A., "Problems in information retrieval. Logical jumps in the expression of information", *Information Storage and Retrieval*, 9, 65-77 (1973).
- 11) COATES, E.J., "Some properties of relationships in the structure of indexing languages", *Journal of Documentation*, 29, 390-404 (1973).

- 12) GHOSE, A. and DHAWLE, A.S., "Problems of thesaurus construction", *journal of the American Society for Information Science*, 28, 211 -217
- 13) LANCASTER, F.W., *Information Retrieval Systems: Characteristics, Testing and Evaluation*, Wiley, New York (1968)
- 14) KOCH EN, M., *Principles of Information Retrieval*, Melville Publishing Co., Los Angeles, California (1974).
- 15) BOOKSTEIN, A. and SWANSON, D.R., "Probabilistic models for automatic indexing", *journal of the American Society for Information Science*, 25, 312-318 (1974).
- 16) BELKIN, N.J., "Information concepts for information science", *journal of Documentation*, 34, 55-85 (1978).
- 17) BARBER, A.S., BARRACLOUGH, E.D. and GRAY, W.A. "On-line information retrieval as a scientist's tool", *Information Storage and Retrieval*, 9, 429-44- (1973).
- 18) W. Berry and Susan T. Dumais and Gavin W. O' Brien. Using Linear Algebra for Intelligent Information Retrieval. <http://www.es.utk.edu/~berry/> (20/9/2002)
- 19) Dublin core metadata. URL: [http://purl.oclc.org/metada/dublin\\_core](http://purl.oclc.org/metada/dublin_core).
- 20) Dushan Badal and Michael Davies. Retrieval of unstructured text. URL: <http://citeseer.nj.nec.com/21436.html>.
- 21) Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis, *journal of the American Society of Information Science*, 41 (6):391 -407, 1990.
- 22) Christos Faloutsos and Douglas W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, 1995.
- 23) Marcos Andre Goncalves, Robert K. France, Edward A. Fox, and Tamas E. Doszkocs. Marian: Searching and querying across heterogeneous federated digital libraries. URL: <http://citeseer.nj.nec.com/455519.html>.
- 24) Robert Hochman. Software reliability engineering: An evolutionary neural network approach. URL: <http://citeseer.nj.nec.com/hochman97software.html>.
- 25) Timo Honkela. On the use of self-organizing map (som) in linguistic visualization. URL: <http://citeseer.nj.nec.com/honkela96use.html>.
- 26) Panayiotis Zaphiris. User-centered evaluation of an on-line modern greek language course. URL: <http://citeseer.nj.nec.com/442742.html>.

- 27) A. Jennings and H. Higuchi. A personal news service based on a user model neural network, 1992. URL: <http://citeseer.nj.nec.com/jennings92personal.html>.
- 28) Dik Lun Lee, Young Man Kim, and Gaurav Patel. Efficient signature file methods for text retrieval. Knowledge and Data Engineering, 7(3):423-435, 1995. URL: <http://citeseer.nj.nec.com/15285.html>.
- 29) Liming Ren The. Fast search algorithms for weight-partitioned signature files. URL: <http://citeseer.nj.nec.com/245790.html>.
- 30) P. Zaphiris and et al. Participatory design of an online greek language course. URL: <http://citeseer.nj.nec.com/451729.html>.
- 31) Vienna University of Technology URL: <http://www.ifs.tuwien.ac.at/~andi/somlib/>