Τα πληροφοριακά συστήματα νέας γενιάς για τη διαχείριση πληροφοριών που απευθύνονται στις βιβλιοθήκες: οι νέες τάσεις στην ανάπτυξη σύγχρονων και πλέον λειτουργικών πληροφοριακών συστημάτων που απευθύνονται στις βιβλιοθήκες

Νίκος Παναγιωτάκης

Απ' την ιδέα η ανάπτυξη, Απ' την ανάπτυξη η εξόγκωση, Απ' την εξόγκωση η σκέψη, Απ' τη σκέψη η ανάμνηση, Απ' την ανάμνηση η επιθυμία. Ο λόγος έγινε γόνιμος, Ενώθηκε με το αμυδρό φως Και γέννησε τη νύχτα

Η ΑΥΓΗ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΦΙΛΟΣΟΦΙΑΣ J. BURNET ΚΑΘ.ΠΑΝ/ΜΙΟΥ SAINT ANDREW (ΣΚΟΤΙΑ) ΚΟΣΜΟΓΟΝΙΑ - ΠΟΙΗΜΑ ΤΩΝ ΜΑΟΡΙ

Αγαπητοί φίλοι,

Αποτελεί μεγάλη χαρά τόσο για μένα όσο και για την εταιρεία που εκπροσωπώ, την ITCC A.E, το γεγονός της συμμετοχής μας στο συνέδριο αυτό. Ένας από τους στόχους αυτού του συνεδρίου, εκτός από τις συζητήσεις που αφορούν στο παρόν και το μέλλον των ακαδημαϊκών βιβλιοθηκών στην Ελλάδα, είναι να μοιραστούμε απόψεις που αφορούν στα σύγχρονα μέσα υποστήριξης αυτών, προκειμένου να γίνουν λειτουργικότερες και πλέον προσιτές στους χρήστες τους.

Αν η παρουσίαση γνωστών μέχρι στιγμής εννοιών και πρακτικών βοηθά στην εμπέδωση και κατανόηση τους, η παρουσίαση καινοτόμων ιδεών αποτελεί πόλο έλξης για απόκτηση πρόσθετης γνώσης και εφαρμογών των ιδεών αυτών. Στην παρούσα εισήγηση θα σας παρουσιάσω τις επιστημονικές και τεχνολογικές παραμέτρους που θεμελιώνουν την αρχιτεκτονική και τη λειτουργικότητα των "Πληροφοριακών συστημάτων νέας γενιάς για τις βιβλιοθήκες".

Κατ' αρχάς θα κάνω μια μικρή ιστορική αναδρομή στη χρήση της πληροφορικής από τις βιβλιοθήκες, καθώς επίσης και τις σύγχρονες απαιτήσεις των βιβλιοθηκών από την πληροφορική. Ακολούθως θα παρουσιάσω το θεωρητικό και εννοιολογικό πλαίσιο των τεχνικών και μεθόδων που θα χρησιμοποιηθούν για τη δημιουργία των "Νέας γενιάς πληροφοριακών συστημάτων για τις βιβλιοθήκες".

Βιβλιοθήκες και πληροφορική

Ο όρος βιβλιοθήκη μας είναι γνωστός, καθώς επίσης γνωστό μας είναι και το γεγονός ότι κατά καιρούς (από την αρχαιότητα μέχρι σήμερα) οι υπεύθυ-

νοι για τις βιβλιοθήκες εφάρμοζαν διάφορες μεθόδους για την καλύτερη δυνατή διαχείριση τους, αξιοποιώντας στο έπακρο τα εκάστοτε κρατούντα τεχνολογικά επιτεύγματα.

Η πληροφορική τις τελευταίες δεκαετίες έχει αλλάξει δραστικά επί τα βελτίω το βαθμό αποτελεσματικότητας της όλης διαχειριστικής διαδικασίας στις βιβλιοθήκες. Επειδή κάθε σύστημα στον κόσμο μας ακολουθεί μια εξελικτική πορεία, το αυτό ισχύει και για το στερεο-λογισμικό (υπολογιστές, προγράμματα κλπ. μέσα υποστήριξης) που χρησιμοποιούνται κατά καιρούς από τις βιβλιοθήκες.

Συνηθίζεται να χρησιμοποιείται ο όρος *"γενιά πληροφοριακού συστήματος"*, για να εκφραστεί η εξελικτική αυτή πορεία, που ακολουθεί το πληροφοριακό σύστημα μέσα στα χρόνια, εμφανίζοντας άλμα τέτοιο, ώστε η καινούργια μορφή και λειτουργικότητα του να διαφέρει σημαντικά από την παλαιότερη, σηματοδοτώντας *"σταθμό "* στην όλη εξελικτική του πορεία.

Απαριθμώ τις γενιές των συστημάτων αυτοματισμού βιβλιοθηκών στον παρακάτω πίνακα και παραθέτω τις ιδιότητες εκάστης γενιάς.

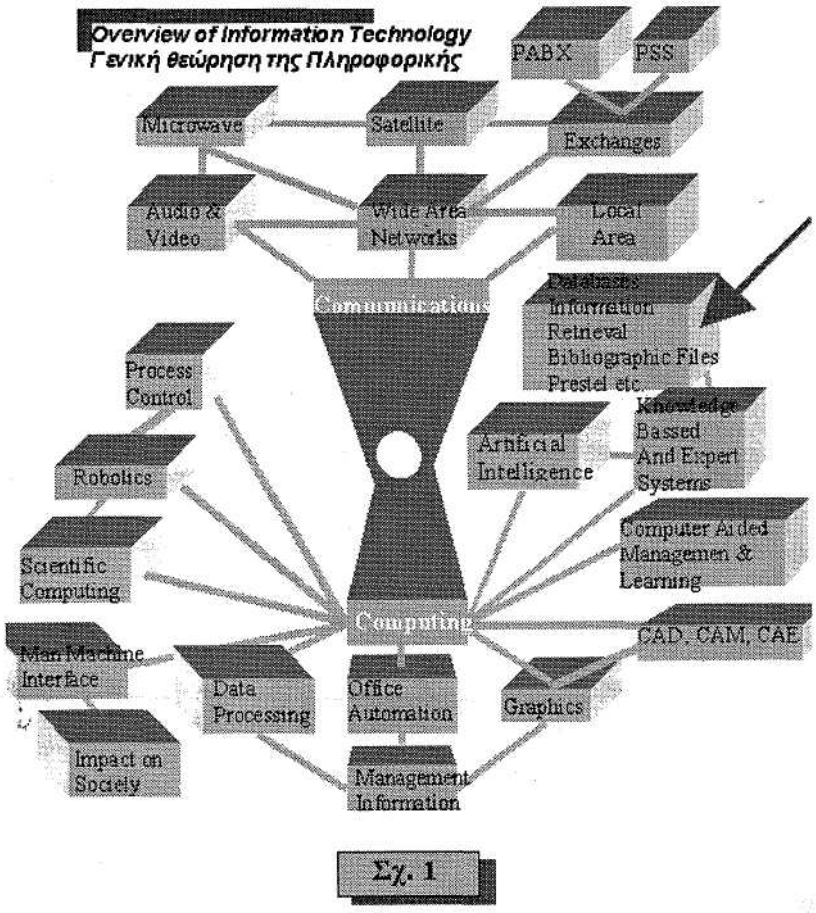
LIBRARY AUTOMATION OVERVIEW – BY GENERATION			
	1ST GEN	2ND GEN	3RD GEN
S/W LANG	Proprietary	C. Assembler	4GL
O/S	Proprietary	Vendor Specific	UNIX, DOS
DBMS	Proprietary	Proprietary	Entity-Relational Object-Oriented
COMMUNICATIONS	Limited	Some interfaces	All standards
IMPORT/EXPORT	None	Some	On-board
PLATFORMS	Locked	Vendor family	Multi-vendor
3RD PARTY S/W	None	Some	Unlimited
REPORTS	Fixed Format	Fixed format	User defined
COLOUR	No	No	Yes
CAPACITY	Limited	Improved	Unlimited
MODULE INTEGRATION	None	Bridges	Seamless
ARCHITECTURE	Shared	Shared	Distributed

Πίνακας 1

Είναι γνωστό σε όλους μας ότι οι βιβλιοθήκες αποτελούν μφνάδα ζωτικής σημασίας όχι μόνο για τα σχολεία, τα πανεπιστήμια, το δημόσιο, δημοτικό τομέα αλλά και για τις επιχειρήσεις και τους οργανισμούς παγκοσμίως.

Τα παραπάνω συστήματα χρησιμοποιούν σήμερα την πληροφορική για την καλύτερη, αποδοτικότερη και ποιοτικότερη προσέγγιση στον τρόπο εργασίας στοχεύοντας συνάμα στην καλύτερη δυνατή και αρτιότερη παροχή υπηρεσιών προς τον πολίτη.

Ακολούθως παραθέτω σχήμα, στο οποίο φαίνεται η θέση των βιβλιογραφικών αρχείων στο γενικότερο πλαίσιο ενός ολοκληρωμένου πληροφοριακού συστήματος.



Σχ. 1

Μια από τις μεγαλύτερες επενδύσεις που θα μπορούσε να κάνει μια βιβλιοθήκη αλλά και ένας οργανισμός ή μια επιχείρηση, είναι στις *"πληροφορίες"*.

Είναι προφανές ότι το πληροφοριακό σύστημα που υποστηρίζει τη βιβλιοθήκη, θα πρέπει να ολοκληρώνεται με διάφανο τρόπο με το υπόλοιπο πλη-

ροφοριακό σύστημα του οργανισμού *(ανοιχτότητα),* η δε *"πληροφορία"* θα πρέπει να είναι *διαχειρίσιμη,* και *εύκολα προσβάσιμη* από τους χρήστες του πληροφοριακού συστήματος.

Η καθημερινή μας εργασία στον ηλεκτρονικό κόσμο περιλαμβάνει χρήση πόρων όπως το Internet, Intranets, Extranets, Network Computing, προσωπικούς υπολογιστές που μπορεί να ευρίσκονται στο σπίτι ή γενικά μακριά από τη βιβλιοθήκη, καθώς επίσης και τα κατάλληλα μέσα υποστήριξης για πρόσβαση σε ηλεκτρονικές πηγές πληροφόρησης και δημοσιεύσεις.

Παρακολουθώντας την εξελικτική πορεία των βιβλιοθηκών σε διεθνές επίπεδο και καταγράφοντας τις ανάγκες, τους προβληματισμούς και τις επιδιώξεις των βιβλιοθηκονόμων, συνάγονται τα ακόλουθα:

- Σήμερα οι βιβλιοθηκονόμοι χρειάζονται περισσότερο από ποτέ άλλοτε να χρησιμοποιήσουν τεχνολογικά επιτεύγματα, πλέον σύγχρονα και έξυπνα εργαλεία, θέτοντας σαν αντικειμενικό σκοπό να καταστήσουν τις βιβλιοθήκες στις οποίες εργάζονται κεντρικούς κόμβους παροχής πληροφοριών στους χρήστες αυτών.
- Η πληροφόρηση των χρηστών δεν περιλαμβάνει μόνο παροχή πληροφοριών από τα στοιχεία του καταλόγου της βιβλιοθήκης ή από τις συλλογές των κλάδων αυτής, αλλά και από συλλογές προερχόμενες από εξωτερικές πηγές, πολυμέσα, έντυπα που έχουν εκδοθεί σε ηλεκτρονική μορφή κλπ,
- Οι χρήστες των βιβλιοθηκών (καθηγητές, φοιτητές, πλατύ κοινό) χρειάζονται να έχουν πρόσβαση σε πηγές πληροφόρησης άσχετα από τη θέση τους ή τη μορφή του υλικού που είναι καταχωρημένες.
- Τα πληροφοριακά συστήματα που χρησιμοποιούν οι βιβλιοθήκες θα πρέπει λοιπόν να γίνουν αρκετά πιο έξυπνα από τα υφιστάμενα, προκειμένου να ικανοποιήσουν τις ανάγκες και τις απαιτήσεις των χρηστών από το χώρο της πληροφόρησης,
- Οι MARC καταλογογραφήσεις, το EDI και η WWW πρόσβαση στα δεδομένα των καταλόγων δεν είναι πλέον αρκετά για να στηρίξουν τις ανάγκες των βιβλιοθηκών στην απόκτηση και διαχείριση της πληροφορίας.
- Οι παραδοσιακές τεχνικές και τα εργαλεία έχουν αγγίξει το ανώτερο όριο σε ό,τι αφορά την απόδοση τους για την επίτευξη των στόχων, που τέθηκαν την εποχή που αυτά σχεδιάστηκαν και υλοποιήθηκαν.

Είναι προφανές ότι τα συστήματα νέας γενιάς για διαχείριση των πληροφοριών της βιβλιοθήκης πρέπει να προσφέρουν μια πλέον σύγχρονη προσέγγιση στην αναζήτηση πληροφοριών προχωρώντας αρκετά βήματα μπροστά από τα συμβατικά συστήματα αναζήτησης και ανάκτησης πληροφοριών, προσφέροντας καλυτέρους χρόνους απόκρισης και ακρίβειας στις αναζητήσεις.

Θα πρέπει βεβαίως να κατανοήσουμε ότι για να ικανοποιηθούν οι παραπάνω απαιτήσεις, το λογισμικό της βιβλιοθήκης πρέπει να σχεδιαστεί και να αναπτυχθεί ευθύς εξ αρχής με σύγχρονες μεθόδους, να χρησιμοποιήσει σύγχρονα εργαλεία και τεχνικές για τη διαχείριση και αναζήτηση της πληροφορίας, και φυσικά ότι θα απαιτηθεί κάποιο χρονικό διάστημα από τη στιγμή αυτή (όχι μεγάλο) για να ανακοινωθούν τα πρώτα αποτελέσματα και οι κριτικές.

Τα συστήματα 4ης γενιάς για τις βιβλιοθήκες προβλέπουν τη χρήση των παρακάτω λειτουργικών συστημάτων, τα οποία αποτελούν τον κύριο παράγοντα που διαφοροποιεί τα συστήματα 4ης γενιάς από τα συστήματα 3ης γενιάς.

- Συστήματα αναγνώρισης μορφών με δυνατότητα αυτομάθησης
- Σημασιολογικά δίκτυα
- Συστήματα επεξεργασίας φυσικής γλώσσας

Τα πληροφοριακά συστήματα διαχείρισης βιβλιοθηκών με βάση την τεχνολογία αυτή παίρνουν μια επιπλέον διάσταση. Θα μπορούσαν κάλλιστα να ονομαστούν *πληροφοριακά συστήματα διαχείρισης πληροφοριών* για βιβλιοθήκες. Οι στόχοι που πρέπει να επιτευχθούν μέσα από τη χρήση τέτοιων συστημάτων είναι οι ακόλουθοι:

- Αποδοτικότερος τρόπος διαχείρισης των πόρων της βιβλιοθήκης σε σχέση με τα υπάρχοντα συστήματα, συμβάλλοντας στη μείωση του κόστους λειτουργίας της
- Ευελιξία στον τρόπο αναζήτησης (χρήση συστημάτων πλοήγησης)
- Να επιτευχθούν στην πραγματικότητα συστήματα ΑΝΟΙΧΤΑ και ΓΡΑΜΜΙΚΑ ΕΠΕΚΤΑΣΙΜΑ
- Να προβλέπουν δυνατότητα εγκατάστασης σε πολλαπλές hardware πλατφόρμες και περιβάλλοντα δικτύων
- Να εγκαθίστανται και να έχουν τη δυνατότητα να αξιοποιούν τις δυνατότητες που παρέχουν τα διαφορετικά συστήματα διαχείρισης βάσεων δεδομένων
- Κατά τη σχεδίαση και υλοποίηση αυτών οι κατασκευάστριες εταιρίες να μην έχουν λάβει μονάχα υπόψη την εμπειρία των βιβλιοθηκονόμων, των αναλυτών συστημάτων και των προγραμματιστών, αλλά να έχουν αξιοποιήσει στο έπακρο την εμπειρία των ειδικοτήτων που αναφέρονται ακολούθως:

1. Σχεδιαστές πληροφοριακών συστημάτων
2. Τεχνικοί και Τεχνολόγοι δικτύων
3. Έμπειροι στον τομέα σχεδιασμού συστημάτων διεπαφής ανθρώπου-μηχανής (Man Machine Interfaces)
4. Ειδικοί σε θέματα τεχνητής νοημοσύνης

5. Ειδικοί στην ανάπτυξη έμπειρων συστημάτων και βάσεων γνώσης
6. Επιστήμονες - ειδικοί στη Γνωσιολογία κλπ.

Στις παρακάτω σελίδες παρατίθεται το σύνολο των εννοιών και το θεωρητικό πλαίσιο, στο οποίο βασίζεται η αρχιτεκτονική θεμελίωση των πληροφοριακών συστημάτων $4^{η?}$ γενιάς για τις βιβλιοθήκες.

Η παρουσίαση θα γίνει πολύ συνοπτικά λόγω των χρονικών περιορισμών που υπάρχουν στις εισηγήσεις.

ΣΥΣΤΗΜΑΤΑ ΠΡΟΣΑΝΑΤΟΛΙΣΜΕΝΑ ΣΤΟ ΠΑΡΑΔΟΣΙΑΚΟ TEXT RETRIEVAL ΣΕ ΑΝΤΙΠΑΡΑΘΕΣΗ ΜΕ ΤΑ ΣΥΣΤΗΜΑΤΑ ΑΝΑΓΝΩΡΙΣΗΣ ΜΟΡΦΩΝ APRPs: ΘΕΩΡΗΤΙΚΟ ΚΑΙ ΕΝΝΟΙΟΛΟΓΙΚΟ ΠΛΑΙΣΙΟ

Είναι γεγονός ότι στα υπάρχοντα πληροφοριακά συστήματα κάθε πληροφορία η οποία εισάγεται στο σύστημα θα πρέπει να δεικτοδοτείται προκειμένου να είναι ανακτήσιμη.

Δεικτοδότηση

Ορισμός: Με τον όρο "δεικτοδότηση" ορίζεται η εργασία εκείνη που επιτρέπει τον προσδιορισμό και επιλογή των στοιχείων εκείνων ("λέξεις κλειδιά") που είναι αναγκαία για να περιγράψουν το τεκμήριο,

- είναι το εργαλείο προσδιορισμού της θέσης καθενός τεκμηρίου
- είναι "εργαλείο διασύνδεσης" τεκμηρίων
- είναι μια υπηρεσία η οποία προσθέτει αξία στη διαδικασία για αναζήτηση στοιχείων και αυξάνει κατά πολύ την ακρίβεια και την πληρότητα των μεταγενέστερων προσπαθειών ανάκτησης στοιχείων.

Είναι γνωστό ότι η δεικτοδότηση ευρίσκει εφαρμογή όχι μόνο σε βιβλιογραφικές βάσεις δεδομένων αλλά και σε κάθε άλλη βάση δεδομένων με χαρακτήρα εμπορικό, βιομηχανικό κλπ.

Προβλήματα Δεικτοδότησης

Κίνδυνος κακής εφαρμογής όρων, από τον υπεύθυνο να ορίζει τις λέξεις κλειδιά
Απαιτούμενος χρόνος σημαντικός
Υψηλό κόστος εφαρμογής εκσφαλμάτωσης και αναδιάρθρωσης δεικτών

Τρόπος αποφυγής προβλημάτων

Δημιουργία λεξικών, θησαυρών, θεματικών επικεφαλίδων
OGR επεξεργασία

Χρήση συστημάτων αναζήτησης κειμένων

Προβλήματα Δεικτοδότησης και OCR Επεξεργασίας

Ο όρος OCR επεξεργασία είναι το ακρωνύμιο των όρων Optical Character Recognition - Οπτική αναγνώριση χαρακτήρων.

Αναφέρθηκε προηγουμένως σαν ένας τρόπος αποφυγής των προβλημάτων που προκύπτουν από τη δεικτοδότηση. Με την OCR επεξεργασία, καταβάλλεται κυρίως, ακόμα και σήμερα, η προσπάθεια εισαγωγής μεγάλων κειμένων από έντυπα σε βάσεις δεδομένων, με χρήση συσκευών Οπτικής ανάγνωσης (Optical Character Readers), τα οποία μετά από σχετική επεξεργασία δημιουργούν "λέξεις κλειδιά".

Μετά από χρήση της OCR τεχνολογίας επί σειρά ετών απεδείχθησαν τα ακόλουθα:
- Δεν υπάρχει ακρίβεια 100% κατά τη φάση μετατροπής εικόνας σε κείμενο
- Ακρίβεια 96-97%
- Δυσανάγνωστα αποτελέσματα
- Ανάγκη επαναπληκτρολόγησης στοιχείων
- Επανέλεγχος ακεραιότητας κειμένου
- Απαίτηση σημαντικού χρόνου
- Αύξηση κόστους επεξεργασίας
- Αύξηση όγκου βάσης κατά 150-250%

Η Προτεινόμενη Λύση

Χρήση νέων τεχνολογιών για αναζητήσεις με ασαφή προσέγγιση, κάνοντας χρήση των συστημάτων αναγνώρισης μορφών, με δυνατότητα αυτομάθησης κατά τη φάση της επεξεργασίας, σε συνδυασμό με:
- Σημασιολογικά δίκτυα
- Συστήματα επεξεργασίας φυσικής γλώσσας
- Λοιπές άλλες τεχνικές με τις οποίες τα σύγχρονα συστήματα προσπαθούν να μιμηθούν τον τρόπο με τον οποίο σκέπτεται ο άνθρωπος

ΑΝΑΓΝΩΡΙΣΗ ΜΟΡΦΩΝ : PATTERN RECOGNITION
Ορισμός: Η διαδικασία ορισμού και κατηγοριοποίησης της μορφής των α-ντικειμένων σε κάποιο περιβάλλον π.χ. άνθρωποι, μηχανές κλπ.
- Η δυνατότητα αναγνώρισης διαφοράς στη μορφή ανάμεσα σε γράμματα από λέξεις
- Η δυνατότητα διαχωρισμού της μορφής της καρέκλας από το τηλέφωνο
- Του Α από το Ω

- Του χάρτη της Βραζιλίας από αυτόν της Ελλάδος

Στο ερώτημα *"Είναι σημαντική η χρήσί] συστημάτων Pattern Recognition : αναγνώριω] μορφών στις βιβλιοθήκες? "*

Η θέση των ειδικών είναι καταφατική. Ακολούθως συνοψίζονται οι αιτίες για τη θετική τους τοποθέτηση.

Αιτιολόγηση:

Όλη η γνώση βασίζεται στην αναγνώριση μορφών

Στην παρούσα εισήγηση ενδιαφέρει πώς συμβάλλει στην αναγνώριση κειμένου που ευρίσκεται σε μορφή εικόνας, μάλιστα δε με πολλές ασάφειες στη μορφή του κειμένου. Π.χ. η δυνατότητα να διαβάσει αυτό που φαινομενικά εμφανίζεται σαν

TAE CHT σαν **THE CAT**

Το σύστημα αναγνωρίζει ότι το **TAE** δε συναντάται σαν λέξη σε προτάσεις όπως το THE

Τα συστήματα για αναγνώριση μορφών, με δυνατότητα μάθησης κατά τη φάση της επεξεργασίας συμβάλλουν στα ακόλουθα:
- Βοηθούν στην αναζήτηση με τρόπο ασαφούς ταύτισης
- Μπορούν να μαθαίνουν
- Αυτοδιοργανώνονται

ΔΙΑΦΟΡΕΣ APRPS ΑΠΟ ΤΑ ΠΑΡΑΔΟΣΙΑΚΑ TEXT RETRIEVAL ΣΥΣΤΗΜΑΤΑ

Ακολούθως επισημαίνονται οι σημαντικότερες διαφορές μεταξύ των παραδοσιακών συστημάτων Text Retrieval και των συστημάτων που βασίζονται στα συστήματα αναγνώρισης μορφών APRPS.

Text Retrieval

Δημιουργία δεικτών με βάση το κείμενο => Προϋποθέτει μετατροπή εικόνας σε κείμενο (OCR) επεξεργασία

APRPS

- Βλέπουν την πληροφορία, όπως αυτή έχει αποθηκευθεί σαν δυαδικά νούμερα

- Τα δυαδικά νούμερα αποτελούν τη βάση δεικτοδότησης|και όχι τα κείμενα

Π.χ.
P – 00010000
1 – 01101001
X – 01111000
T – 00010100
E – 01100101
X – 0111000

- Είναι μοντελοποιημένα όπως τα βιολογικά συστήματα με χρήση νευρωνικών δικτύων για επεξεργασία πληροφοριών
- Διαθέτουν δυνατότητα αυτο-οργάνωσης μιας μάθησης
- Διαθέτουν δυνατότητα ανάμνησης ψηφιακών μορφών
- Παρέχουν αυτόματη δεικτοδότηση ψηφιακών μορφών με δημιουργία αυτό-βελτιστοποιουμένης μνήμης μορφών

Αποτελέσματα
Ταύτιση αντικειμένων με χρήση fuzzy logic (τεχνικές ασάφειας).

Αναζήτηση όρων, οι οποίοι εισάγονται με λάθος πληκτρολόγηση, δε θα προκαλέσει κατ' ανάγκη αρνητική απάντηση από το σύστημα (εκτός εάν δεν υπάρχει ο όρος αυτός).
Βοηθά το χρήστη να κάνει την εργασία αναζήτησης ευρύτερη ή στενότερη χρησιμοποιώντας τη δυνατότητα για ταύτιση μορφών.

Μετρήσιμα μεγέθη
Ελαχιστοποίηση κόστους δεικτοδότησης (ο παραδοσιακός τρόπος δεικτο-δότησης των εγγράφων απαιτεί 1250-25000δρχ./έγγραφο ανάλογα με το βαθμό πολυπλοκότητας της δεικτοδότησης).
Αυτόματη δεικτοδότηση.
Μηδενισμός κόστους διόρθωσης κειμένου από OCR επεξεργασία, ο οποίος είναι περίπου 1000-2 5 ΟΟδρχ./σελίδα.
Ελαχιστοποίηση των απαιτήσεων σε χώρους δίσκου.
Αύξηση ταχύτητας επεξεργασίας κατά την αναζήτηση.
Αριστοποίηση διαχείρισης πόρων συστήματος.

Οφέλη
Αυτόματη ψηφιακή δεικτοδότηση με δυνατότητα αυτοαναδιοργάνωσης.

Ελαχιστοποίηση κόστους δεικτοδότησης.

Αποφυγή σφαλμάτων πληκτρολόγησης ή κατηγοριοποίησης.

Φυσική ανοχή γλωσσικών σφαλμάτων.

Ελαχιστοποίηση αναγκών για διόρθωση σφαλμάτων OCR επεξεργασίας.

Υψηλή ακρίβεια αναζητήσεων.

Παροχή υψηλού αισθήματος ασφάλειας στο χρήστη ότι το σύστημα θα του παράσχει πληροφορίες άσχετα από δικό του κακό πληκτρισμό σε όρους αναζήτησης.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ
Η επεξεργασία προτάσεων που εισάγονται ή διαβάζονται από το σύστημα, το οποίο απαντά επίσης με προτάσεις με τρόπο τέτοιο που να θυμίζει απαντήσεις μορφωμένου ανθρώπου.
Βασικό ρόλο παίζει η γραμματική, το συντακτικό, η ανάλυση των εννοιολογικών στοιχείων και γενικά της γνώσης, για να γίνει κατανοητή η ανθρώπινη γλώσσα από τη μηχανή.

Εργασίες του Επεξεργαστή Φυσικής Γλώσσας

Lexical Analysis	Λεξικογραφική Ανάλυση
Syntactic Analysis	Συντακτική Ανάλυση
Semantic Analysis	Σημασιολογική Ανάλυση
Discourse Analysis	Ανάλυση ομιλίας - Πραγματεία
Pragmatic Analysis	Πραγματική-Ρεαλιστική Ανάλυση

Η Γνώση σε ένα Επεξεργαστή Φυσικής Γλώσσας
Lexical Knowledge
Syntactic Knowledge
Semantic Knowledge
Discourse Knowledge
Pragmatic Knowledge
Application-specific Knowledge
General Knowledge

Η Μηχανή του Επεξεργαστή Φυσικής Γλώσσας - NLP
Ο επεξεργαστής φυσικής γλώσσας εκτελεί τις παρακάτω ΕΡΓΑΣΙΕΣ:

- Διάβασε την εισαχθείσα πρόταση

- Επεξεργάσου την εισαχθείσα πρόταση
  Εκτέλεσε λεξικογραφική ανάλυση: λεξικογραφικός διαχωρισμός των λέξεων της πρότασης

  Αναζήτησε αυτές στο λεξικό

  Εκτέλεσε συντακτική ανάλυση: προσδιορισμός του τύποί) των λέξεων με χρήση συντακτικού

  Εκτέλεσε σημασιολογική ανάλυση: κατανόηση των λέξεων

  Εκτέλεσε πραγματική ανάλυση: προσδιόρισε τον τύπο της απάντησης που θα δημιουργήσεις

- Δημιούργησε την κατάλληλη απάντηση

- *Η γνώση* φυλάσσεται σε "δομές γνώσης", που διαχωρίζουν αυτήν σε οργανωμένους τύπους γνώσης

- *Μηχανισμός ανάστροφης ιχνηλασιμότιμας (Backtracking mechanism):* βοηθά τον επεξεργαστή να εγκαταλείψει μια ανεπιτυχή αναζήτηση και να προσπαθήσει πάνω σε εναλλακτικά μονοπάτια

- *Μηχανισμός εξαγωγής συμπερασμάτων (Inference engine):* Εξάγει νέα γνώση από προϋπάρχουσα. Π.χ. στην εξίσωση Α+Β=Γ, η πρόταση Γ είναι η νέα γνώση που προκύπτει από τις προτάσεις Α και Β

ΑΝΤΙΚΕΙΜΕΝΙΚΟΙ ΣΤΟΧΟΙ ΤΟΥ NLP

Να μιμηθεί τις δυνατότητες διαλόγου που έχουν μορφωμένοι άνθρωποι.
Να κατανοήσει μια εισαγόμενη πρόταση και να δημιουργήσει την κατάλληλη απάντηση.

ΣΗΜΑΣΙΟΛΟΓΙΑ
Ορισμός: Η μελέτη του τρόπου με τον οποίο αναπαριστάται η έννοια μιας λέξης στο μυαλό του ανθρώπου.

Σημασιολογική Ανάλυση
Ορισμός: Είναι η εργασία προσδιορισμού της σημασίας λέξεων μέσα σε πρόταση.

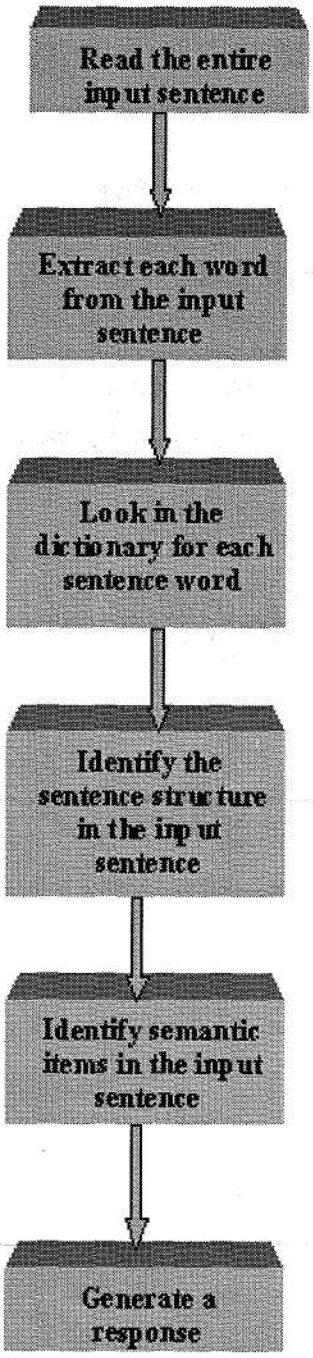
Λειτουργίες
Προσδιορισμός της έννοιας στην πρόταση.

Απεικόνιση-αντιγραφή της έννοιας αυτής στη βάση γνώσης.

Semantic Networks

Ορισμός: Το σημασιολογικό δίκτυο είναι ένας διγράφος με ετικέτες που χρησιμοποιείται για να περιγράψει σχέσεις (συμπεριλαμβανομένων των ιδιοτήτων) αντικειμένων, ιδεών, καταστάσεων ή δράσεων.

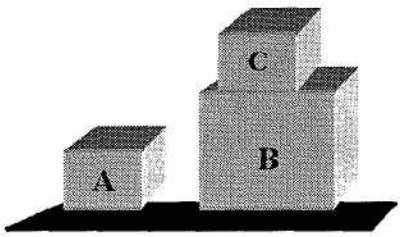
NLP PROCESS DIAGRAM



Οφέλη από τη χρήση των σημασιολογικών δικτύων

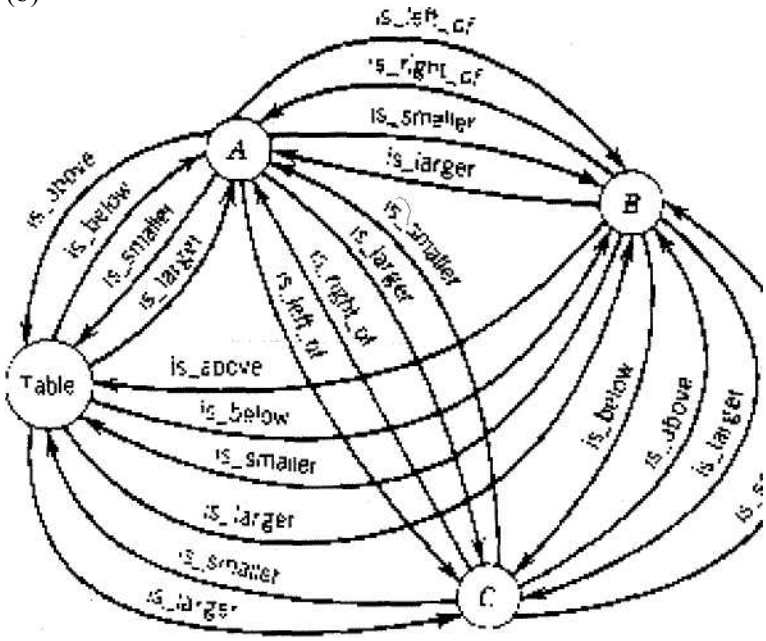
Εύκολη-αυτόματη ανεύρεση της πληροφορίας που απαιτείται.
Χρήση όλων των δυνατοτήτων που παρέχει η επεξεργασία φυσικής γλώσσας.
Ενσωμάτωση συντακτικών κανόνων, μορφολογικών, πραγματικών νοημάτων λέξεων.
(a)

Παραδείγματα Σημασιολογικών δικτύων'



Τραπέζι

(b)



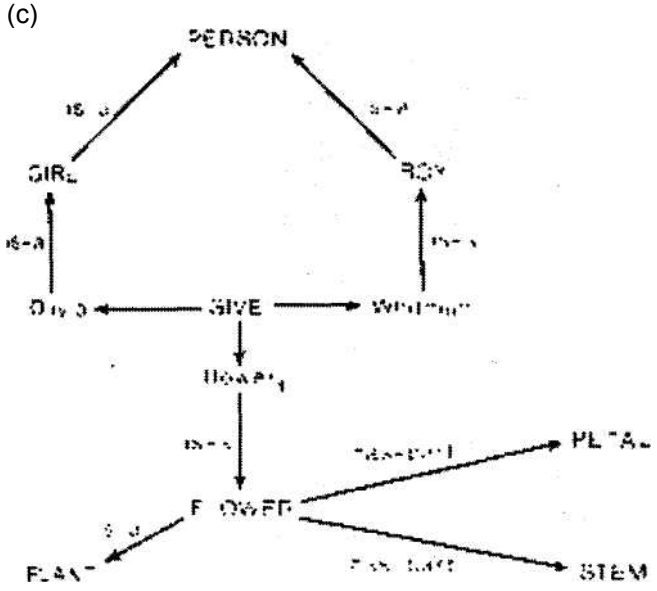
Semantic net representation of "blocks-world"

(a) System

(b) Semantic net

Pattern recognition: statistical, structural, and neural approaches / Robert J. Schalkoff

(c)

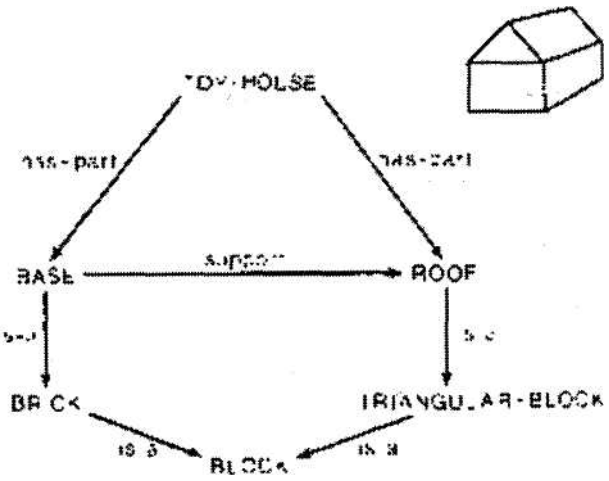


Examples of networks:

(c) is a network showing some of the linkages between a particular proposition and general conceptual knowledge about human beings and flowers

(d) is a semantic network expressing some of the structure of a toy house

(d)



Παράδειγμα

Θεωρείστε τα ακόλουθα αντικείμενα, κατηγορίες, σχέσεις κατηγοριών και δραστηριότητες.

Αντικείμενα	**Κατηγορίες**
Αγόρι	Άνθρωπος
Σκυλί	Θηλαστικό
Γάτα	Θηλαστικό
Τριαντάφυλλο	Φυτό

Κατηγορίες	**Κατηγορίες**
Άνθρωπος	Θηλαστικό
Φυτό	Φυτό

Κατηγορίες	**Ενέργειες**
Θηλαστικό	Ονειρεύεται
Φυτό	Δεν ονειρεύεται

ΕΡΩΤΗΣΗ. Ποιο θα ήταν το αποτέλεσμα, αν η ακόλουθη πρόταση εισαχθεί σε επεξεργαστή φυσικής γλώσσας

- "Το τριαντάφυλλο ονειρεύεται"

σύμφωνα με τους παραπάνω κανόνες?

ΑΠΑΝΤΗΣΗ: Θα είναι αρνητική

ΑΙΤΙΑ: Το τριαντάφυλλο είναι φυτό, και τα φυτά δεν ονειρεύονται.

Πριν κλείσει αυτή η εισήγηση θα ήθελα να σας υποβάλω μια ακόμα ερώτηση:

Υπάρχουν συστήματα διαχείρισης βιβλιοθηκών τα οποία να έχουν σχεδιαστεί χρησιμοποιώντας τις τεχνολογίες που ανέφερα προηγουμένως;

Η απάντηση είναι "ΝΑΙ"

Είναι τα Μέσα Υποστήριξης Q Series και η Information Quest (IQ) Από την Electronic Online Systems International (EOSi)

Η Σειρά Q είναι το:
- Σύστημα νέας γενιάς για τη διαχείριση πληροφοριών στις βιβλιοθήκες
- Ολοκληρωμένο
- Σύγχρονο και
- Πραγματικά ΑΝΟΙΧΤΟ
- Πλήρως παραμετρικό
- Καλύπτει όχι μόνο όλες τις λειτουργικές περιοχές των βιβλιοθηκών, αλλά και πολύ περισσότερα

Η Σειρά Q, με ή χωρίς την Information Quest (IQ), υλοποιεί με ένα σύγχρονο μοναδικό τρόπο το όραμα των βιβλιοθηκονόμων να καταστήσουν τις βιβλιοθήκες τους κόμβους πληροφοριών χάριν της κοινωνίας των χρηστών αυτών.

Τεχνολογικό Υπόβαθρο

Η Σειρά Q χτίστηκε χρησιμοποιώντας τα ακόλουθα εργαλεία:
- PowerBuilder σαν εργαλείο υλοποίησης .
- Το ORACLE RDBMS
- Και για μηχανή αναζήτησης, τον τρόπο δεικτοδότησης του Excalibur (APRPS, Semantics, NLP)

Το λειτουργικό σύστημα είναι τα Windows NT.

Στόχος

Να παρέχει τη Σειρά Q κάτω από οποιοδήποτε γνωστό RDBMS, σε NT ή UNIX πλατφόρμες.

Information Quest (IQ)

Η IQ είναι μια νέα ολοκληρωμένη υπηρεσία ηλεκτρονικής πληροφόρησης της Dawson Holding Company.

Δημιουργήθηκε από την EOSi που είναι κλάδος της Dawson Holding Company.

Παρέχει συνδέσμους στο WWW για επιστημονικά, τεχνικά, ιατρικά, περιοδικά, περιοδικά για τις επιχειρήσεις κλπ. μέσω Internet.

Περιλαμβάνει μια βάση με περισσότερες από 7.000.000 εγγραφές από 12.000 περιοδικά που εκδόθηκαν από το 1990.

Ο WebOPAC της Σειράς Q συνδέεται απευθείας με την IQ, για να παράσχει σε σας τη δυνατότητα πρόσβασης στα στοιχεία της βάσης, στις περιλήψεις, και στα περιεχόμενα πλήρους κειμένου, όπου είσαστε συνδρομητές με τρόπο διάφανο και άσχετα με τη γεωγραφική σας θέση.

Παραπομπές που ευρίσκονται μέσα από την ισχυρή μηχανή αναζητήσεων της IQ, μπορούν να δείχνουν στους χρήστες τα σχετιζόμενα τοπικά έντυπα αποκτήματα στη βιβλιοθήκη τους ή υποβοηθά αυτούς να τα δουν σε σύνδεση online σε μορφή Adobe PDF™ (Portable Document Format) or RealPage™ μορφή. Επίσης μπορούμε να ζητήσουμε να σταλούν τα αντικείμενα αυτά σε ηλεκτρονική ή σε μορφή Τηλε-επιστολής.

Χρησιμοποιεί την ίδια ισχυρή τεχνολογία με αυτή της **Σειράς Q** σε ό,τι αφορά την αναζήτηση πληροφοριών, παρέχοντας στους χρήστες τη δυνατότητα να εύρουν τη σωστή πληροφορία.

Λίγα λόγια για την EOSi

Η Electronic Online Systems International (EOSi) αποτελεί εταιρεία με παγκόσμια διάσταση στον τομέα λογισμικού και παροχής υπηρεσιών που αφορούν στη διαχείριση πληροφοριών. Παρέχει μια ολοκληρωμένη σειρά προϊόντων και υπηρεσιών προκειμένου να αντιμετωπίσει τις ανάγκες των βιβλιοθηκών όλων των τύπων και διαστάσεων. Διαθέτει 40 περιφερειακά γραφεία και κέντρα υποστήριξης σε όλο τον κόσμο, τα οποία την υποβοηθούν να παράσχει υπηρεσίες σε 6500 πελάτες παγκοσμίως.

Παραδείγματα από τη διεθνή πρακτική.
NSF Center for Intelligent Information Retrieval CUR

Objectives

To develop advanced techniques for text analysis and retrieval.

To integrate those techniques with other data management techniques in distributed environments.

Research Projects

Advanced retrieval techniques: The project aims to Improve the ability of the retrieval engine to accurately locate relevant text objects (such as documents or passages) - Work done mostly using INQUERY retrieval engine, based on the "inference net" probabilistic model of retrieval.

Indexing and Natural language Processing: The project aims to Improve the analysis of natural language queries and texts to produce better representation of content, work is being done in phrase-based representations and recognizing entities in text.

Routing and filtering: The project aims to develop an engine that can effectively deal with the real-time demands of large volume text feeds being processed against thousands of user profiles.

Distributed information retrieval: The project aims to the development of techniques for locating relevant text databases, merging results, and efficient retrieval in both local and wide area network enviroments.

Browsing and query formulation: The project aims to the development of corpus-based techniques to support query formulation and access through browsing rather than by query. One focus is the automatic determination of relationships between phrasal concepts.

Text extraction: The project aims to develop techniques to automatically extract entities, attributes and relationships from text, and to integrate this capability with the retrieval techniques. This work involves more sophisticated Natural Language Processing (NLP) than is typically needed for information retrieval, and is more domain knowledge-intensive.

Integration with database systems: The project aims to develop approaches to integrating information retrieval and database system query languages, query processing, and object management. The whole project also aims to satisfy performance constraints of large text-based applications.

The Library of Congress Digital Library Effort

The Library of Congress was in 1995 in the process of evaluating new technologies to provide on-line access to parts of their huge collections. As part of the "American Memory" project, the Library has worked with the CIIR

(Center for Intelligent Information Retrieval), to provide text-basied access to collections of photographs, speeches, and books. The initial prototype has used a Mosaic interface that incorporates some aspects of an advanced text retrieval interface, such as Natural Language Queries (NLQ), ranked output, and relevance feedback.

Future versions of the system will incorporate more advanced features of INQUERY such as probabilistic field-based retrieval and phrase-based representations.

Some of the research issues that can be addressed in the context of this project include effective query processing for image-based queries (what types of queries do people ask in an image database?) and indexing strategies for MARC records, which have many short fields of varying importance.

The British Library Initiatives for Access Projects,

The British Library holds over 18.000.000 volumes and is one of the world's greatest treasure houses of written information from every age and culture. In 1993 it published its Strategic *Objectives for the year 2000,* which made a commitment to providing access to the collections using digital and networking technologies for onsite and remote users.

Initiative For Access, a program of 20 development projects, was inaugurated in 1993 to investigate hardware and software platforms for the digitization and subsequent networking of a range of library materials. The objectives of the Library are apart from enhancing library services and facilitating access, to establish standards for storage, indexing, retrieval, and transmission of data, together with issues involved with digitization of material and its provision over networks.

Major projects within IFLA include the following.
- The Patent Express Jukebox
- The Electronic Beowulf
- Electronic Photo Viewing System
- The Network OPAC

From British Library's pilot Electronic Photo Viewing System
Image scanned from Communications of the ACM April 1995/vol38, no 4 p.65
Other projects in the Initiatives for Access program include:

- Digitization of ageing microfilm which will provide searchable indexes for popular microfilm collections.
- Testing of Excalibur PixTex/EFS for catalog conversion and other appli cations.

- Multimedia publications program; "Medieval Realms" and "Inventors and Inventions" are two of the Library's first interactive publications.

Through these applications, and others to follow, the library is seeking to further its strategic objectives of becoming a major holder and supplier of digital data by the beginning of the new century.

Βιβλιογραφία

Cognition : Exploring the science of the mind I Daniel Reisberg.

Cognitive science : an introduction I Neil A. Stillings ... [et al.], 2nd ed. A Bradford book.

COMMUNICATIONS of the ACM : Digital Libraries April 1995-Volume 38, Number 4.

Developing Natural Language Interfaces : processing human conversations I Russel Suereth. Published by Me Graw-Hill Publishing Company.

Digital image processing I William K. Pratt. A Wiley-Interscience publication.

Foundations of cognitive science I edited by Michael I. Posner. A Bradford book.

Image Storage and Retrieval Systems : A new approach to records management (J. Ranade IBM Series) / Marc D. Alleyrand. Published by Me Graw-Hill Publishing Company.

Management Information Systems : Briefing 1 I Consultant editor: R.V. Franks. Published by Kogan Page Ltd in association with the Chartered Institute of Management Accountants.

The Management of information systems I Gary W Dickson. (McGraw-Hill series in management information systems).

Pattern Recognition : Statistical, structural and neural approaches I Robert Schalkoff. Published by John Willey & Sons, Inc.

Unified theories of cognition I Allen Newell. (The William James lectures, 1987).