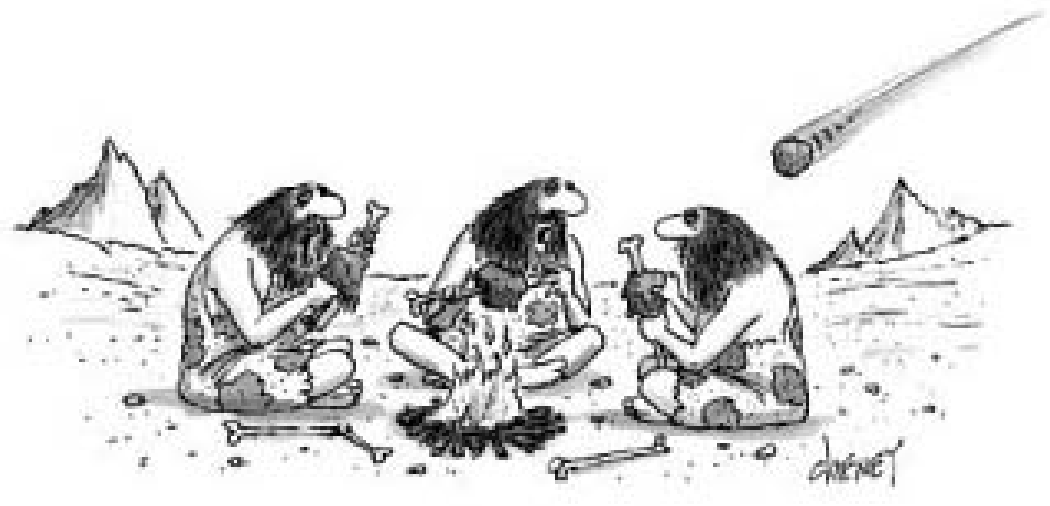
ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ-ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ



"You've got mail."

ΑΥΤΟΜΑΤΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΝΥΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΥΤΟΜΑΤΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΝΥΜΑΤΩΝ

ΕΙΣΗΓΗΤΕΣ: ΚΑΠΙΔΑΚΗΣ ΣΑΡΑΝΤΟΣ
ΠΑΠΑΘΕΟΔΩΡΟΥ ΧΡΗΣΤΟΣ

ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ
ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΥΔΡΑΙΟΥ ΙΩΑΝΝΑ
ΑΜ: Β200082

ΑΥΤΟΜΑΤΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΝΥΜΑΤΩΝ

Πίνακας Περιεχομένων

Περίληψη

Σκοπός της παρούσας εργασίας είναι να αναφέρει και να περιγράψει τα πακέτα λογισμικού που διαχειρίζονται τα ηλεκτρονικά μηνύματα και ειδικότερα αυτά, που επιτελούν μία συγκεκριμένη εργασία, αυτή της κατηγοριοποίησης. Προκειμένου, όμως να γίνει κατανοητή η λειτουργία των σχετικών συστημάτων και ο τρόπος με τον οποίο καθένα από αυτά προσεγγίζουν το ζήτημα της κατηγοριοποίησης, κρίθηκε απαραίτητη η κατανόηση και η περιγραφή της ίδιας της εργασίας. Για τον λόγο αυτό το πρώτο κεφάλαιο αφιερώθηκε στη σχετική διαδικασία.

Η περιγραφή της σχετικής μεθόδου ξεκινά συνοπτικά, από τα πρώτα βήματα, που πραγματοποιήθηκαν σε αυτό τον τομέα, όταν η κατηγοριοποίηση ήταν ακόμα μια από τις εργασίες που εκτελούνταν σε τεκμήρια της παραδοσιακής μορφής και συγκεκριμένα, κυρίως στα πλαίσια των βιβλιοθηκών. Στην συνέχεια, ο τομέας της κατηγοριοποίησης γίνεται αντικείμενο πιο εξελιγμένων επιστημών, όπως αυτή της Τεχνητή Νοημοσύνη, της Εκμάθηση Μηχανής και της Ανάκτηση Πληροφοριών. Σκοπός του ερευνητικού ενδιαφέροντος ήταν να εντάξουν τη σχετική υπηρεσία στο νέο περιβάλλον των ηλεκτρονικών τεκμηρίων και παράλληλα, να επιτύχουν όλο και μεγαλύτερο βαθμό αυτοματοποίησής της. Το μεγάλο ενδιαφέρον της επιστημονικής κοινότητας οφείλεται στα πλεονεκτήματα που παρουσιάζει η κατηγοριοποίηση όσον αφορά την οργάνωση των δεδομένων. Η μείωση του χρόνου και του κόπου οργάνωσης των πληροφοριών, που πλέον επιτελείται υπό νοηματικές κατηγορίες κατανοητές από τον χρήστη και η ευκολία της μετέπειτα πρόσβασης, αναζήτησης και ανάκτησης είναι μόνο κάποια από τα πλεονεκτήματα της διαδικασίας. Τα πλεονεκτήματα αυτά γίνονται ακόμα πιο δυνατά και η χρήση της καθίσταται απαραίτητη μέσα από την αυτοματοποίηση της διαδικασίας.

Δεύτερος σταθμός στην εργασία ήταν η κατανόηση της φύσης των ηλεκτρονικών μηνυμάτων και ο λόγος για τον οποίο αυτή αποτελούσε τροχοπέδη στη διαδικασία της κατηγοριοποίησης. Η περιγραφή της φύσης των μηνυμάτων και η προσέγγιση των διαφόρων συνιστωσών που αποτελούν το πρόβλημα, καταλαμβάνει την πρώτη ενότητα του δεύτερου κεφαλαίου. Οι υπόλοιπες ενότητες του δεύτερου κεφαλαίου αναφέρονται στα διάφορα συστήματα που διαχειρίζονται τα ηλεκτρονικά μηνύματα και που στο πλαίσιο των υπηρεσιών που προσφέρουν στους χρήστες τους περιλαμβάνεται η διαδικασία της κατηγοριοποίησης.

Γενικά, το δεύτερο κεφάλαιο διασπάται σε ενότητες, οι οποίες περιλαμβάνουν συστήματα τα οποία εκτελούν τη διαδικασία της κατηγοριοποίησης σε διαφορετικά χρονικά διαστήματα. Στο τέλος των δύο τελευταίων ενοτήτων πραγματοποιείται σύγκριση των συστημάτων αναλογικά με τις δυνατότητές τους, τη συμβατότητά τους και τις δυνατότητές τους.

ΕΙΣΑΓΩΓΗ

Σε μία γενική θεώρηση, η ηλεκτρονική αλληλογραφία παρουσιάζεται στο προσκήνιο ως νέα μορφή επικοινωνίας με την παγκόσμια διασύνδεση των ηλεκτρονικών υπολογιστών μέσω του Διαδικτύου και την ανάπτυξη των σχετικών τεχνολογιών επικοινωνίας, γνωστών ως *messaging*. Τα *e-mail*, όπως είναι ευρύτερα γνωστά τα ηλεκτρονικά μηνύματα, οφείλουν την ευρεία χρήση τους αλλά και την καθιέρωσή τους, ως μία από τις πιο δημοφιλείς και εύχρηστες υπηρεσίες του Διαδικτύου, χάρη στις ελκυστικές προτάσεις και λύσεις που παρέχουν για την επικοινωνία των χρηστών και την ασφαλή μεταφορά πληροφοριών.

Η χρήση τους ανάγεται στις δεκαετίες 1960-1970, όταν οι εταιρίες που κατείχαν mainframe και mini υπολογιστές, χρησιμοποιούσαν παράλληλα την υπηρεσία ηλεκτρονικών μηνυμάτων. Βάση της σχετικής υπηρεσίας, οι χρήστες των τερματικών που βρίσκονταν συνδεδεμένα με τα υπολογιστικά συστήματα, μπορούσαν να στέλνουν μεταξύ τους μηνύματα. Το πρόσωπο, όμως που χρεώνεται την εφεύρεση της ηλεκτρονικής αλληλογραφίας, με την έννοια που αυτή είναι σήμερα γνωστή, είναι ο Roy Tomlinson, ο οποίος το 1971 έστειλε το πρώτο ηλεκτρονικό μήνυμα με τη μορφή testing 1-2-3 χρησιμοποιώντας το ARPANET [a] .

Έκτοτε, η χρήση των ηλεκτρονικών μηνυμάτων έχει διευρυνθεί τόσο ποσοτικά όσο και λειτουργικά. Τα ηλεκτρονικά μηνύματα κατέληξαν να χρησιμοποιούνται από εκατομμύρια χρήστες παγκοσμίως τόσο για προσωπική τους χρήση όσο και ως τμήμα της εργασίας τους, αντικαθιστώντας ως ένα βαθμό την παραδοσιακή αλληλογραφία. Έρευνες έχουν αποδείξει ότι έχουν συμβάλλει στην ανάπτυξη κατανεμημένων οργανισμών, επιτρέποντας σε υπαλλήλους που βρίσκονται σε

διαφορετικές γεωγραφικές περιοχές να επικοινωνούν σπάζοντας τα όρια χρόνου και χώρου.

Οι e-mail εφαρμογές σχεδιάστηκαν αρχικά για την εκτέλεση «*ασύγχρονης επικοινωνίας*» αλλά πλέον οι νέες λειτουργίες που πρέπει να εκτελέσουν, θα μπορούσαν να ενταχθούν υπό τους όρους : *διαχείριση εργασιών*, *προσωπική αρχειοθέτηση* παράλληλα με την *ασύγχρονη επικοινωνία*. Επειδή, όμως οι σχετικές εφαρμογές δεν είχαν σχεδιαστεί για να εξυπηρετήσουν αυτούς τους τομείς και σε συνδυασμό με τον όγκο των πληροφοριών που μεταφέρονται με αυτά οδήγησαν στο φαινόμενο που ορίζεται στην έρευνα του Whittaker ως *email overload*, δηλαδή, υπέρ-πληθώρα ηλεκτρονικών μηνυμάτων.

Η ύπαρξη του φαινομένου, οφείλεται τόσο στην ευρεία χρήση της ηλεκτρονικής αλληλογραφίας αλλά και λόγω της σπουδαιότητας των πληροφοριών, που μεταφέρονται μέσω αυτής. Οι χρήστες τείνουν να διατηρούν τα μηνύματα που φέρουν σημαντικές πληροφορίες και συνεπώς να παραχωρούν μεγάλο τμήμα του αποθηκευτικού χώρου για την καταχώρισή τους. Πέρα, όμως από την αποθήκευση των μηνυμάτων, απαιτούνται και διαδικασίες, που θα διευκολύνουν την οργάνωση και τη διαχείριση των μηνυμάτων, με τέτοιο τρόπο ώστε να επιτυγχάνεται η απρόσκοπτη πρόσβαση των χρηστών σε αυτά καθώς και η αποτελεσματική και ακριβής ανάκτησή τους.

Η επίλυση του προβλήματος υποθάλπει την τεχνική της αυτόματης κατηγοριοποίησης, δηλαδή, της αυτόματης ταξινόμησης των μηνυμάτων υπό θεματικές κατηγορίες, κατανοητές από τον χρήστη, Με αυτό τον τρόπο, διευκολύνεται τόσο η παρουσίαση των μηνυμάτων στο χρήστη, όσο και η φυλλομέτρηση, η αναζήτηση και η ανάκτηση αποθηκευμένων μηνυμάτων και πληροφοριών, διευκολύνοντας με αυτό τον τρόπο, την γενική εργασία του χρήστη. Η κατηγοριοποίηση του περιεχομένου, είναι μία διαδικασία που εντάχθηκε στις υπηρεσίες των βιβλιοθηκών για την οργάνωση του υλικού τους και τη διευκόλυνση των αναζητήσεων των χρηστών. Με την ίδια λογική χρησιμοποιείται και στα ηλεκτρονικά μηνύματα, υιοθετώντας τις τεχνικές αυτόματης κατηγοριοποίησης ηλεκτρονικών τεκμηρίων, που θα περιγραφτούν στην συνέχεια.

1. Κατηγοριοποίηση Κειμένου

1.1 Προσέγγιση στην Κατηγοριοποίηση

Κατηγοριοποίηση Κειμένου είναι η διαδικασία αυτόματου προσδιορισμού κειμένων φυσικής γλώσσας σε προκαθορισμένες κατηγορίες βάση του περιεχομένου τους [] ή διατυπωμένο διαφορετικά, η διαδικασία περιγραφής κειμένου φυσικής γλώσσας με θεματικές κατηγορίες από ένα προκαθορισμένο σύνολο κατηγοριών. Επιπρόσθετα, ο όρος χρησιμοποιείται για να δηλώσει και τη διαδικασία δημιουργίας εργαλείων λογισμικού, τα οποία είναι ικανά να ταξινομήσουν τεκμήρια γραπτού λόγου ή υπερκείμενα σε προκαθορισμένες κατηγορίες ή κωδικούς θεμάτων [17]. Στη διαδικασία κατηγοριοποίησης ηλεκτρονικών τεκμηρίων, οι κατηγορίες τυπικά χρησιμοποιούνται ως μέσα οργάνωσης των πληροφοριών αλλά και ως γενικής επισκόπησης, στα πλαίσια μίας συλλογής τεκμηρίων [18]

Η μέθοδος οργάνωσης τεκμηρίων ή γενικότερα πληροφοριών σε κατηγορίες έχει τις ρίζες της στον χώρο των Βιβλιοθηκών. Στον χώρο αυτό δημιουργήθηκαν και υιοθετήθηκαν διάφορα συστήματα ταξινόμησης προκειμένου να οργανωθεί το υλικό υπό θεματικές κατηγορίες, ανεξαιρέτως της φυσικής του υπόστασης, προκειμένου να είναι δυνατή η εύκολη αναζήτηση και ανάκτηση του υλικού. Ένα από τα πλέον εύχρηστα σύστημα ταξινόμησης είναι το δεκαδικό σύστημα του Dewey (DDC), όπου σε αριθμητικά οργανωμένες θεματικές κατηγορίες γίνεται η ταξινόμηση του υλικού βάση του περιεχομένου του.

Στη δεκαετία του 90 δόθηκε, όμως μία νέα προσέγγιση στο ζήτημα της ταξινόμησης με τη δημιουργία της πρώτης Web πύλη από το Yahoo. Το Yahoo εισήγαγε ένα δικτυακό ιστότοπο, όπου βρισκόταν οργανωμένη ανά θεματικές κατηγορίες μία πληθώρα πληροφοριών. Για την ένταξη των πληροφοριών σε κατηγορίες, δραστηριοποιείται μία ομάδα ανθρώπινου δυναμικού, που τις

ευρετηριάζει και τις καταχωρεί ανάλογα. Παράλληλα, η κατηγοριοποίηση κειμένου επικέντρωσε το επιστημονικό ενδιαφέρον με την ευρεία εμφάνιση και τη συνεχώς αυξανόμενη διαθεσιμότητα τεκμηρίων σε ψηφιακή μορφή αλλά και χάρη στην επακόλουθη ανάγκη των χρηστών για εύκολη πρόσβαση σε αυτά [b] . Ακρογωνιαίος λίθος της υποστήριξης της πρόσβασης των χρηστών στο υλικό τους θεωρείται η καλή οργάνωσή του.

Η αυτόματη διαχείριση ηλεκτρονικών τεκμηρίων, web σελίδων, ηλεκτρονικών μηνυμάτων και συζητήσεων καθώς και οποιωνδήποτε άλλων πληροφοριών σε ηλεκτρονική μορφή, αποτελεί ενεργό ερευνητικό πεδίο για διάφορες επιστημονικές προσεγγίσεις, όπως είναι η **Τεχνητή Νοημοσύνη,** η **Εκμάθηση Μηχανής** και η **Ανάκτηση Πληροφοριών.** Οι πληροφορίες όμως, που βρίσκονται αποτυπωμένες σε μορφή κειμένου αποτελούν τροχοπέδη για την αυτόματη διαχείριση λόγω διαφόρων δυσκολιών, όπως είναι η έλλειψη δόμησης και περιορισμών του κειμένου, η μεγάλη του έκταση, το ποικίλο περιεχόμενο καθώς και η αναπαράσταση του περιεχομένου στο οποίο εμφανίζονται οι πληροφορίες .

Οι παραπάνω προκλήσεις, οδήγησαν στην εφαρμογή περιορισμών και δόμησης της πληροφορίας σε συστήματα που βασίζονται στην Τεχνητή Νοημοσύνη. Η δομημένη αυτή προσέγγιση είναι λιγότερο λειτουργική σε συστήματα διαχείρισης ηλεκτρονικών μηνυμάτων, τα οποία χαρακτηρίζονται από την έλλειψη δόμησης, κυρίως στο *«σώμα»* των μηνυμάτων.

Στην συγκεκριμένη εργασία, γίνεται αναφορά σε μεθόδους και συστήματα που εφαρμόζουν κυρίως τη *Εκμάθηση Μηχανής* τεχνική ή άλλες και λιγότερο σε συστήματα που υιοθετούν τη μέθοδο *Ανάκτηση Πληροφοριών* ή *Τεχνητή Νοημοσύνη*. Στην συνέχεια επιχειρείται μία προσέγγιση στο ζήτημα της αναπαράστασης του περιεχομένου των ηλεκτρονικών μηνυμάτων αλλά και περιγραφή μεθόδων κατηγοριοποίησης. Η αναφορά των μεθόδων κατηγοριοποίησης γίνεται βάση ενός συγκριτικού στοιχείου, του βαθμού αυτοματοποίησης που παρέχουν στους χρήστες.

1.2 Αναπαράσταση Κειμένου

Η αναπαράσταση του τεκμηρίου ανάγεται σε διεργασία μείζονος σημασίας καθώς αποτελεί την βάση της κατηγοριοποίησης και γενικότερα της διαχείρισης των ηλεκτρονικών μηνυμάτων και τεκμηρίων. Πριν πραγματοποιηθεί οποιαδήποτε από τις προαναφερθείσες διαδικασίες, το περιεχόμενου του κειμένου θα πρέπει να αναπαρασταθεί με τέτοια μορφή, ώστε να γίνει κατανοητό από τον ταξινομητή. Υπάρχουν δύο βασικές προσεγγίσεις στο συγκεκριμένο θέμα όσον αφορά την επιλογή των συστατικών που θα το αντιπροσωπεύσουν.

Η κυρίαρχη μέθοδος, γνωστή ως *σύνολο λέξεων,* κάνει χρήση των λέξεων του κειμένου, τις οποίες μετατρέπει σε μεταβλητές ή χαρακτηριστικά ενός διανύσματος, με το οποίο αντιπροσωπεύεται το εκάστοτε κείμενο [12] . Η χρήση, όμως, όλων των λέξεων του κειμένου δεν είναι δυνατή, καθώς στην αντίθετη περίπτωση θα δημιουργούνταν τεράστια διανύσματα, φαινόμενο, το οποίο χαρακτηρίζεται ως «*high dimensionality*» [3] . Η χρήση πολλών λέξεων και συνεπώς η δημιουργία μεγάλων διανυσμάτων μειονεκτεί καθώς, μεταξύ άλλων επιβραδύνει την ταχύτητα του ταξινομητή καθώς θα πρέπει να επεξεργαστεί μεγάλο αριθμό χαρακτηριστικών. Υπάρχουν διάφοροι μέθοδοι επιλογής χαρακτηριστικών, όπως είναι η *tf\idf*[1] (*term frequency\inverse document frequency*), *mutual information*[2] και *information gain*[3]*,* που χρησιμοποιούνται για να μειωθεί ο αριθμός των χαρακτηριστικών [2]. Ο υπερβολικός, βέβαια, περιορισμός των χαρακτηριστικών που θα χρησιμοποιηθούν για την κατηγοριοποίηση, από την άλλη μεριά, οδηγεί στην πιθανότητα απώλειας βασικών πληροφοριών, που απαιτούνται για μία ακριβή κατηγοριοποίηση []. Συχνά, στις λέξεις που χρησιμοποιούνται, προσδίδεται ένα ειδικό βάρος, διαχωρίζοντας με αυτό τον τρόπο το σύνολο των λέξεων σε περισσότερο ή

[1] Y. Yang and J. Pederson. A comparative study on feature selection in text categorization. In *Proc. of the FourteenthInternational Conference on Machine Learnin*g, 1997.
[2] T. Joachims. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proc. of the Fourteenth International Conference on Machine Learnin*g, 1997.
[3] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.

λιγότερο σημαντικές (*term-weighting*). Βάση της σχετικής τεχνικής, *σύνολο λέξεων,*υποστηρίζουν τη λειτουργία τους διάφοροι αλγόριθμοι ταξινόμησης, που θα αναφερθούν στην συνέχεια .

Η νέα μέθοδος που χρησιμοποιείται και υιοθετείται από τα περισσότερα νέα συστήματα της κατηγοριοποίησης, είναι αυτή που κάνει χρήση των εννοιών, που υπόκεινται σε ένα κείμενο, αντί για την καθεαυτού χρήση των λέξεων. Οι έννοιες, που αναφέρονται και ως «*υψηλού επιπέδου χαρακτηριστικά*» στην σχετική βιβλιογραφία, αποτελούν ουσιαστικά προτάσεις ρημάτων ή ουσιαστικών. Ο εντοπισμός τους πραγματοποιείται βάση γλωσσολογικών ή στατιστικών τεχνικών, ενώ παράλληλα, γίνεται χρήση ελεγχόμενου λεξιλογίου ή θησαυρού και διαγραφή όμοιων λέξεων .Οι προτάσεις ουσιαστικών βάση του θησαυρού μειώνονται προτάσεις τριών διαφορετικών νοημάτων, που στην συνέχεια συσχετίζονται με το κείμενο [b] . Η μέθοδος, αυτή αντιμετωπίζει αποτελεσματικά το φαινόμενο της «*high dimensionality»,* ενώ, παράλληλα, η κατηγοριοποίηση αποδίδει μεγαλύτερη ακρίβεια.

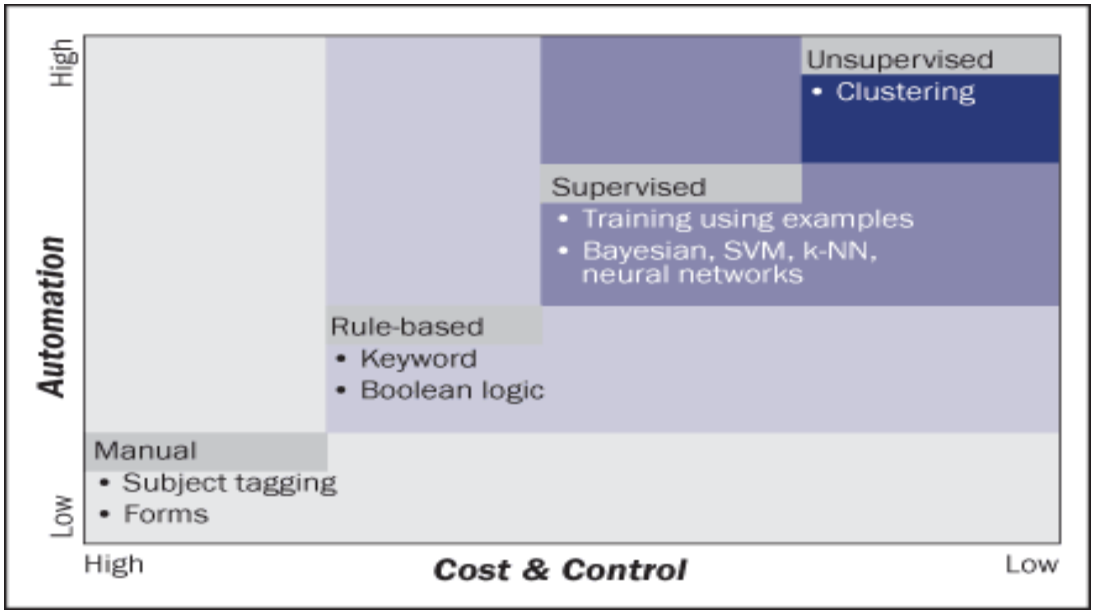
Κοινό χαρακτηριστικό και των δύο μεθόδων είναι ότι το κείμενο υπόκειται σε μία γενική διαδικασία επεξεργασίας, πριν την επιλογή των αντιπροσωπευτικών του δομών είτε αυτές είναι λέξεις ή έννοιες. Η διαδικασία αυτή περιλαμβάνει την απομάκρυνση λέξεων με υψηλό ποσοστό εμφάνισης αλλά μικρής νοηματικής αξίας. Οι λέξεις αυτές είναι γνωστές ως *stop words* και περιλαμβάνουν, άρθρα, συνδέσμους, προθέσεις κλπ. Στην συνέχεια, οι λέξεις υποβάλλονται στη διαδικασία του *stemming,* της αφαίρεσης, δηλαδή επιθεμάτων και προθεμάτων. Τέλος, αφού παραμείνουν μόνο τα βασικά χαρακτηριστικά, το κείμενο αναπαριστάται ανάλογα με τη μέθοδο που υιοθετείται στο σύστημα.

1.3 Τεχνικές Κατηγοριοποίησης[4]

[4] Sebastiani, Fabrizio. Machine Learning in Automated Text Classification

Η νέα ώθηση που δόθηκε στον τομέα της κατηγοριοποίησης με την εμφάνιση και την ευρεία διάθεση των ηλεκτρονικών τεκμηρίων αποδίδεται πρακτικά με την ανάπτυξη διαφόρων ερευνητικών προσεγγίσεων. Κάθε προσέγγιση στηρίζεται σε διαφορετικές τεχνικές για τη δημιουργία ταξινομητών κειμένου, για την αναπαράσταση και διαχείριση του ηλεκτρονικού εγγράφου και πέρα από τα διάφορα πλεονεκτήματα και μειονεκτήματα που παρουσιάζει, προσφέρει στον χρήστη μεγαλύτερο ή μικρότερο βαθμό αυτοματοποίησης, ακρίβειας και αποτελεσματικότητας.

Όπως φαίνεται και στον παρακάτω πίνακα, οι βασικές προσεγγίσεις στην κατηγοριοποίηση ηλεκτρονικού εγγράφου, παρουσιάζουν μία κλιμακωτή αύξηση όσον αφορά τον άξονα της αυτοματοποίησης και αντίστοιχα μείωση στο ζεύγος τιμών χρόνου και κόστους ξεκινώντας από την παραδοσιακή χειρονακτική κατηγοριοποίηση προς την μη-επιβλεπόμενη.



Πίνακας 1- Αναπαράσταση των Διαφόρων Μορφών Κατηγοριοποίησης

- **<u>Χειρωνακτική Κατηγοριοποίηση:</u>**

Αναφέρεται στη βασική μέθοδο κατηγοριοποίησης τεκμηρίων αποτυπωμένων στην παραδοσιακή μορφή που χρησιμοποιείται αυτούσια για την κατηγοριοποίηση

ηλεκτρονικών τεκμηρίων. Η χειρονακτική κατηγοριοποίηση είναι η μέθοδος που χρησιμοποιείται κατά κόρον στις βιβλιοθήκες και στα κέντρα τεκμηρίωσης ή σε οποιοδήποτε άλλο ίδρυμα που στηρίζεται στην οργάνωση του υλικού του. Προϋποθέτει την εργασία του ανθρώπινου παράγοντα, συνήθως ενός ειδικού στον τομέα ή ενός βιβλιοθηκονόμου που αρχικά θα ευρετηριάσει τα τεκμήρια και στη συνέχεια θα στηριχτεί σε ένα ταξινομικό σύστημα ή σε μία οντολογία, προκειμένου να τα θέσει στις κατάλληλες κατηγορίες. Αν και υπάρχουν περιπτώσεις που οι ειδικοί διαφωνούν πάνω σε μία συγκεκριμένη πράξη ταξινόμησης, η μέθοδος απολαμβάνει το μεγάλο ακρίβειας και αποτελεσματικότητας που παρουσιάζει. Το βασικό μειονέκτημα της μεθόδου είναι ο χρόνος και το κόστος που χρειάζεται για να πραγματοποιηθεί μία τέτοιου είδους κατηγοριοποίηση, γεγονός που αμβλύνεται από το μεγάλο αριθμό ηλεκτρονικών τεκμηρίων που μεταφέρονται στις μέρες μας μέσω των διαφόρων διαύλων. Μία εταιρία, δηλαδή, ή ένας οργανισμός μέσου μεγέθους, δεδομένων των διαφόρων τύπων και του αριθμού των τεκμηρίων που εισέρχονται και εξέρχονται μέσω του τοπικού Intranet ή του Internet, θα χρειαζόταν μία ομάδα βιβλιοθηκονόμων που να εργάζεται σε καθημερινή βάση προκειμένου η κατηγοριοποίηση να είναι ακριβής και η βάση δεδομένων τους να είναι ενημερωμένη.

- **Κατηγοριοποίηση βάση Κανόνων**[5]

Η συγκεκριμένη μέθοδος είναι από τις πρώτες «έξυπνες» προσεγγίσεις που επιτεύχθηκαν στο ζήτημα της κατηγοριοποίησης ηλεκτρονικού τεκμηρίου [5]. Η μέθοδος στηρίζεται στην εισαγωγή κανόνων για τον εντοπισμό λέξεων-κλειδιών[6], στις Boolean εκφράσεις (and, or) [b] καθώς και στην γλωσσολογική επεξεργασία του κειμένου.

Κατά την προσέγγιση αυτή, από το κείμενο αφαιρούνται λέξεις με μικρή σημασιολογική αξία[7] , ενώ οι υπόλοιπες χρησιμοποιούνται ως ευρετήριο όρων που περιγράφει το κείμενο.. Οι κανόνες για τον εντοπισμό των keywords είναι της μορφής **If…Then**, δηλαδή στην περίπτωση που εντοπιστεί αυτός ο όρος Χ,

[5] Rule-Based Classification
[6] keyword spotting
[7] stop-words

τότε το κείμενο θα ταξινομηθεί υπό την εξής κατηγορία Ψ. Αρχικά, η προσέγγιση αυτή στηρίχτηκε στην τεχνική της knowledge engineering, που κυριαρχούσε κατά τη δεκαετία του 80 [16] . Βάση της δεδομένης τεχνικής, που στηρίζεται στην κωδικοποίηση της γνώσης του ειδικού, οι κανόνες αυτοί εισάγονταν με το χέρι. Η μέθοδος, όμως παρουσίαζε σοβαρά μειονεκτήματα καθώς η διαδικασία σύνθεσης ενός κανόνα είναι γνωστικά ακριβής και στην περίπτωση λάθους στο συντακτικό του κανόνα, η κατηγοριοποίηση είτε δεν πραγματοποιείται ή χειρότερα πραγματοποιείται σε λάθος κατηγορία [5] . Παράλληλα, όπως αναφέρεται στο άρθρο του W.E. Mackay «*οι χρήστες γενικά αποφεύγουν να παραμετροποιούν το σύστημά τους*».

Στη δεκαετία του 90, εισήχθησαν τα αυτόματα εκπαιδευόμενα συστήματα εξαγωγής κανόνων, που αυτοματοποιούν τη διαδικασία. Ο εντοπισμός κανόνων στα συστήματα αυτά στηρίζεται στην χρήση προκαθορισμένων φορμών βάση της ILP[8] μεθόδου. Την ίδια περίοδο ο Cohen εισήγαγε τον Ripper, έναν αλγόριθμο που εισάγει αυτόματα κανόνες για τον εντοπισμό λέξεων-κλειδιών. Σύμφωνα, με τον Cohen « *η χρήση κανόνων εντοπισμού λέξεων-κλειδιών μπορεί να γίνει εύκολα κατανοητή και διαχειρίσιμη από τον χρήστη, ενώ παράλληλα, ο εντοπισμός των λέξεων κλειδιών είναι πιο χρήσιμος καθώς αποδίδει μία πιο κατανοητή περιγραφή του φίλτρο του ηλεκτρονικού μηνύματος».* Οι κανόνες ήταν της μορφής

cfp←”cfp”∈ subject, “95” ∈ subject
cfp←”cfp”∈ subject, “1995”∈ subject
cfp← “call”∈ body”, papers”∈ body

Το συγκεκριμένο ηλεκτρονικό μήνυμα κατηγοριοποιείται υπό την κλάση cfp στην περίπτωση που περιλαμβάνει το δείγμα ‘cfp’ και το ‘95’ στο πεδίο subject ή αν περιέχει τα δείγματα ‘cfp’ και ‘1995’ στο πεδίο subject ή στην περίπτωση που περιλαμβάνει τα δείγματα ‘call’ και ‘paper’ στο σώμα του μηνύματος.

Μία τεχνική που ακολουθεί τη μέθοδο κατηγοριοποίησης βάση κανόνων είναι τα fuzzy sets. Ένα σύστημα που βασίζεται στα fuzzy set επιτρέπει στους χρήστες τον

[8] ILP-Inductive Learning Process

προσδιορισμό θεμάτων ενδιαφέροντος με τη μορφή ιεραρχίας υπό-εννοιών. Η διάταξη των υπό-εννοιών στη συνέχεια χρησιμοποιείται ως σύνολο κανόνων. Στο μικρότερο επίπεδο, οι υπό-έννοιες ορίζονται ως συνδυασμοί των λέξεων που βρίσκονται στο κείμενο. Οι κανόνες είναι της μορφής «**IF** Χ **THEN** Ψ» ή «**IF** Χ **THEN** Ψ1 **BUT IF ALSO** Z **THEN** Ψ2». Στην πρώτη περίπτωση, αν το Χ είναι τμήμα του κειμένου, τότε και το Ψ είναι τμήμα του κειμένου. Στη δεύτερη περίπτωση αν το Χ είναι τμήμα και το Z όχι, τότε το Ψ1 είναι τμήμα του κειμένου αλλά αν τόσο το Χ όσο και το Z είναι τμήμα του κειμένου, τότε το Ψ2 δεν ανήκει στο κείμενο.

Τα fuzzy sets επιτρέπουν την αναπαράσταση πληροφοριών σχετικά με τη συμμετοχή ενός στοιχείου στο σύνολο. Βασίζονται στη μέτρηση του σχετικού βαθμού, όπου η τιμή που μπορεί να πάρει αυτό το μέτρο ποικίλλει από το μηδέν (0) στο ένα (1). Όσο μεγαλύτερη είναι η τιμή, τόσο περισσότερο το στοιχείο θεωρείται να είναι μέλος του συνόλου. Συνεπώς, η τιμή 0 σημαίνει ότι το συγκεκριμένο στοιχείο δεν αποτελεί τμήμα, ενώ αντίστοιχα η τιμή 1 υποδηλώνει ότι το στοιχείο αποτελεί βασικό μέλος του συνόλου. Παράλληλα, επιτρέπουν τη γενίκευση σημαντικών ιδεών με σκοπό τη διαχείριση συνόλων, όπως είναι η τομή, η ένωση, η γενίκευση [c].

Η κατηγοριοποίηση βάση κανόνων όπως φαίνεται και από το διάγραμμα, θα μπορούσε να χαρακτηριστεί ως η πρώτη μέθοδος που αυτοματοποιεί μέχρι κάποιο βαθμό τη διαδικασία της κατηγοριοποίησης, παραμένει όμως ακόμα σε χαμηλά επίπεδα. Παράλληλα, μειώνει το κόστος της διαδικασίας καθώς και τον έλεγχο που απαιτείται για την εκτέλεσή της. Η συγκεκριμένη μέθοδος είναι κατάλληλη στην περίπτωση που λίγες λέξεις μπορούν να χρησιμοποιηθούν για να περιγράψουν μία κατηγορία, ενώ είναι αποτελεσματική όταν ο αριθμός των κατηγοριών δεν είναι πολύ μεγάλος. Η δαπάνη προσδιορισμού και διατήρησης κατηγοριών γενικά αποφεύγεται για μεγάλης κλίμακας συστήματα κατηγοριοποίησης.

- **Με επίβλεψηΚατηγοριοποίηση**

Η Με επίβλεψηκατηγοριοποίηση είναι η μέθοδος κατηγοριοποίησης που παρουσιάζει το μεγαλύτερο ενδιαφέρον καθώς σε αυτή βασίζονται τα περισσότερα συστήματα κατηγοριοποίησης ηλεκτρονικών μηνυμάτων, ενώ παράλληλα, βάση αυτής της τεχνικής έχει αναπτυχθεί μεγάλος αριθμός ταξινομητών ηλεκτρονικού τεκμηρίου. Η με επίβλεψη κατηγοριοποίηση βασίζεται στην τεχνική της Εκμάθηση Μηχανής, που θα περιγραφτεί στην συνέχεια και χαρακτηρίζεται ως με επίβλεψη γιατί απαιτεί ένα σώμα προκατηγοριοποιημένων τεκμηρίων προκειμένου να εκπαιδευτούν οι ταξινομητές. Το σχετικό σύνολο προκατηγοριοποιημένων τεκμηρίων, χαρακτηρίζεται ως *training set*.

Η Εκμάθηση Μηχανής τεχνική βασίζεται σε μία γενικά εισαγωγική διαδικασία αυτόματης δημιουργίας ταξινομητών κειμένου, οι οποίοι μαθαίνουν από το σύνολο προκατηγοριοποιημένων τεκμηρίων, τα χαρακτηριστικά κάθε κατηγορίας ενδιαφέροντος. Γενικά, για την σχετική τεχνική το πρόβλημα της κατηγοριοποίησης κειμένου αποτελεί μία διαδικασία επιβλεπόμενης εκμάθησης, καθώς η διαδικασία εκμάθησης οδηγείται ή ελέγχεται από την γνώση των κατηγοριών στις οποίες ανήκει το σύνολο των προκατηγοριοποιημένων τεκμηρίων.

Σύμφωνα με τη Εκμάθηση Μηχανής τεχνική, η κατηγοριοποίηση κειμένου μπορεί να οριστεί ως η εργασία προσδιορισμού της εκχώρησης μίας τιμής {0,1} σε μία εγγραφή a ενός δυαδικού πίνακα[9] αποφάσεων. Ο πίνακας είναι της μορφής:

	d_1	...	...	d_j	...	...	d_n
c_1	a_{11}	...	...	a_{1j}	...	...	a_{1n}
...	....	...	...	...	...	...	...
c_i	a_{i1}	...	...	a_{ij}	...	...	a_{in}
...	...	...	...	...	...	...	...
c_m	a_{m1}	...	...	a_{mj}	...	...	a_{mn}

όπου $C = \{ c_1, ..., c_m \}$ είναι το σύνολο των προκαθορισμένων κατηγοριών και $D = \{d_1, ..., d_n\}$ είναι το σύνολο των τεκμηρίων που πρέπει να ταξινομηθούν. Η τιμή 1 για την a_{ij} ορίζει την απόφαση να κατηγοριοποιηθεί το τεκμήριο d_j στην

[9] matrix

κατηγορία c_i, ενώ η τιμή 0 ορίζει την απόφαση να μην κατηγοριοποιηθεί το τεκμήριο d_j υπό την κατηγορία c_i.

Θα πρέπει παράλληλα να σημειωθεί ότι οι κατηγορίες είναι μόνο συμβολικές ετικέτες, χωρίς κάποια επιπλέον γνώση ή νόημα σε αυτές. Επίσης, ο Sebastiani υποστηρίζει *«ότι η ένταξη του κειμένου σε μία κατηγορία θα πρέπει να βασίζεται σε ενδογενή γνώση, δηλαδή, στα σημασιολογικά χαρακτηριστικά του τεκμηρίου και όχι βάση εξωγενών παραγόντων, όπως είναι τα μεταδεδομένα»*. Ο ίδιος βασίζει την άποψη του στο γεγονός ότι *«η ίδια η σημασιολογία του περιεχομένου ενός τεκμηρίου είναι μία υποκειμενική έννοια, συνεπώς η ταξινόμησή του υπό μία κατηγορία δεν θα μπορούσε να αποφασισθεί ντετερμινιστικά»*. Παράλληλα, στο άρθρο του διακρίνει διαφορετικών ειδών κατηγοριοποιήσεις όπως είναι η κατηγοριοποίηση βάση μίας ή πολλαπλών ετικετών και η κατηγοριοστρεφής και η τεκμηριοστρεφής κατηγοριοποίηση.

Οι διάφοροι ταξινομητές κειμένου που δημιουργήθηκαν στα πλαίσια της επιβλεπόμενης κατηγοριοποίησης βασίζονται σε διάφορους αλγόριθμους που αναπτύχθηκαν υπό την Εκμάθηση Μηχανής τεχνική. Αναφορικά, μερικοί από αυτούς είναι:

- **Probabilistic Ταξινομητές (Naïve Bayes)**
- **Decision Trees Ταξινομητές**
- **Decision Rule Ταξινομητές**
- **Regression Models Ταξινομητές**
- **On-line Linear Ταξινομητές**
- **Rocchio Ταξινομητής**
- **Neural Networks Ταξινομητές**
- **Example-Based Ταξινομητές**
- **Ταξινομητές βάση των Support Vector Machine**
- **Ταξινομητές Βάση Καθορισμού Threshold**

Στην συνέχεια πραγματοποιείται μία προσπάθεια περιγραφής της λογικής υπό την οποία δρουν οι διάφοροι ταξινομητές. Η περιγραφή περιορίζεται στους ταξινομητές, που χρησιμοποιούνται κατά κύριο λόγο από τα συστήματα διαχείρισης ηλεκτρονικών μηνυμάτων, όπως είναι ο **Naïve Bayes,** και οι **Support Vector Machines.** Κοινό χαρακτηριστικό των ταξινομητών της σχετικής μεθόδου είναι ότι απαιτούν προεκπαίδευση προκειμένου να είναι αποτελεσματικοί αλλά παράλληλα, παρουσιάζουν διάφορα διαφορετικά χαρακτηριστικά, όσον αφορά τις προσεγγίσεις που υιοθετούν για την επεξεργασία του κειμένου.

Probabilistic Ταξινομητές

Ιστορικά, ο πρώτος ταξινομητής κειμένου που εμφανίστηκε στην επιστημονική λογοτεχνία σύμφωνα με τον Sebastiani[10] ανήκε στην κατηγορία των **probabilistic classifiers**. Η ονομασία τους αποδίδεται στο γεγονός ότι κάνουν χρήση των πιθανοτήτων, προκειμένου να εντάξουν τα τεκμήρια υπό τις δέουσες κατηγορίες. Ο υπολογισμός της πιθανότητας εκφράζεται από την εξίσωση :

$$P(c_i|\vec{d}_j) = \frac{P(c_i)P(\vec{d}_j|c_i)}{P(\vec{d}_j)}$$

Εξίσωση 1-Υπολογισμός πιθανότητας

Όπου, P(dj) αντιπροσωπεύει την πιθανότητα ένα αυθαίρετα επιλεγμένο τεκμήριο να έχει διάνυσμα dj ως την αναπαράστασή του και P(ci) η πιθανότητα ότι ένα αυθαίρετα επιλεγμένο τεκμήριο εντάσσεται στην κατηγορία ci. Ο υπολογισμός, όμως της πιθανότητας P(ci/dj) όπως αποδίδεται στην εξίσωση (1) δεν είναι λειτουργικός για τη διαδικασία της κατηγοριοποίησης καθώς φέρει ως αποτέλεσμα ένα μεγάλο αριθμό πιθανών διανυσμάτων dj. Για αυτό το λόγο, υιοθετείται η υπόθεση ότι οι δύο συνιστώσες του διανύσματος, ως αυθαίρετες μεταβλητές είναι στατιστικά ανεξάρτητες η μία από την άλλη. Η υπόθεση ονομάζεται «υπόθεση ανεξαρτησίας» και αποτυπώνεται με την εξής εξίσωση:

[10][Maron,M.1961.Automatic indexing:an experimental inquiry.Journal of the Association for Computing Machinery]

$$P(\vec{d}_j|c_i) = \prod_{k=1}^{|\mathcal{T}|} P(w_{kj}|c_i)$$

Εξίσωση 2-Υπόθεση ανεξαρτησίας

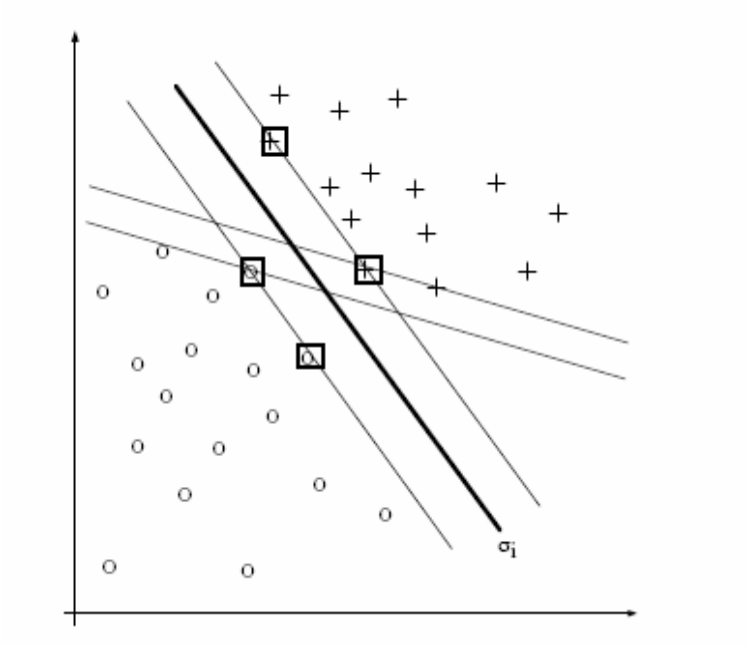
Οι *probabilistic* ταξινομητές που κάνουν χρήση αυτής της εξίσωσης είναι γνωστοί ως **Naïve Bayes**, από τους οποίους προκύπτει μία δυαδική διανυσματική αναπαράσταση του τεκμηρίου. Ο χαρακτηρισμός *naïve*, δηλαδή, αφελής προέρχεται από την χρήση της «υπόθεσης ανεξαρτησίας» των μεταβλητών, καθώς είναι μία κατάσταση που δεν υφίσταται στην πραγματικότητα.

Ουσιαστικά, ο *Naïve Bayes* ακολουθεί την *σύνολο λέξεων* αναπαράσταση του κειμένου και εισάγει δύο διαφορετικούς τρόπους για τη δημιουργία διανύσματος. Ο απλούστερος τρόπος βασίζεται στην καταμέτρηση της παρουσίας ή της απουσίας των λέξεων από το κείμενο. Οι τιμές 0 ή 1 που δηλώνουν την απουσία ή την παρουσία των λέξεων αντίστοιχα, χρησιμοποιούνται για τη δημιουργία του διανύσματος. Ο δεύτερος τρόπος συμπεριλαμβάνει και την καταμέτρηση της συχνότητας της παρουσίας των λέξεων. Παράλληλα, γίνεται η υπόθεση ότι η παρουσία κάθε λέξης είναι ανεξάρτητη από την παρουσία άλλων λέξεων. Στις διάφορες προσεγγίσεις που έχουν πραγματοποιηθεί στη μέθοδο Naïve Bayes χρησιμοποιούνται διαφορετικά σύνολα λέξεων άλλοτε συμπεριλαμβάνοντας όλες τις λέξεις που εμφανίζονται στο κείμενο, άλλοτε μόνο τις λέξεις μέσης συχνότητας ή μικρής συχνότητας [14]. Το σύνολο των λέξεων που θα χρησιμοποιηθεί για την αναπαράσταση του κειμένου έχει επίπτωση στο ποσοστό αποτελεσματικότητας του ταξινομητή, όπως έχουν αποδείξει διάφορες έρευνες[11].

Support Vector Machine

[11] Li, Y. H .and Jain, A.K.1998. Classification of text documents. The Computer Journal 41, 8,537 .546.
Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98,10th European Conference on Machine Learning (Chemnitz, DE, 1998), pp.137 .142.

Η λογική των **Support Vector Machine**, βρίσκεται στην προσπάθεια εύρεσης μίας διαχωριστική γραμμής που να διακρίνει τα θετικά από τα αρνητικά παραδείγματα τεκμηρίων [17]. Σύμφωνα με τον Sebastiani και με πιο μαθηματικούς όρους, *η μέθοδος των SVM, έγκειται στην προσπάθεια ανεύρεσης μεταξύ των επιφανειών ε1, ε2 στο ν-διαστασιακό χώρο που διαχωρίζουν τα θετικά από τα αρνητικά παραδείγματα τεκμηρίων, την επιφάνεια ει, που πραγματοποιεί τον διαχωρισμό με τον καλύτερο δυνατό τρόπο*. Ο καλύτερος δυνατός τρόπος είναι η επιφάνεια ει, που διαχωρίζει τις θετικές από τις αρνητικές τιμές με το μεγαλύτερο δυνατό περιθώριο.



Εικόνα 1-μοντέλο Support Vector Machines

Στην παραπάνω διαγραμματική αναπαράσταση του μοντέλου, οι θετικές και οι αρνητικές τιμές είναι γραμμικά διαχωρισμένες και ορίζονται ως υπερεπίπεδα[12], δηλαδή ως τάξεις δεδομένων. Οι *SVM* επιλέγουν το σύνολο των παράλληλων γραμμών που παρουσιάζουν τη μεγαλύτερη απόσταση ανάμεσα σε δύο στοιχεία του συνόλου.

Σύμφωνα, με τον *Joachim*[13], το μοντέλο των *Support Vector Machines* παρουσιάζουν δύο βασικά πλεονεκτήματα για την κατηγοριοποίηση κειμένου.

[12] hyperplane
[13] Joachims,T.1998.Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98,10th European Conference on Machine Learning (Chemnitz,DE,1998),pp.137 .142.

Αρχικά, δεν απαιτείται επιλογή όρων προκειμένου να κατηγοριοποιηθεί το τεκμήριο βάση αυτών. Επιπλέον, δεν απαιτείται ανθρώπινη ή μηχανική προσπάθεια για την παραμετροποίηση του μοντέλου. Το μοντέλο θεωρητικά εισάγεται με προεπιλεγμένες ρυθμίσεις, οι οποίες όμως έχει αποδειχτεί ότι αποδίδουν βέλτιστα σε οποιοδήποτε σύνολο δεδομένων.

Αξιολόγηση

Στην έρευνα που εξέδωσε ο Sebastiani, η οποία χρησιμοποιήθηκε ως βάση για την παρουσίαση των διαφόρων ταξινομητών στην συγκεκριμένη εργασία, πραγματοποιεί διεξοδική αναφορά και περιγραφή της εισαγωγής της *Εκμάθηση Μηχανής* τεχνικής στο πεδίο της αυτόματης κατηγοριοποίησης κειμένου. Περαιτέρω, ο ίδιος βάση πειραμάτων που διεξήγαγε, αξιολογεί τους διάφορους ταξινομητές συγκριτικά με τρεις συνιστώσες: την «ακρίβεια», την «ανάκληση» και την «αποτελεσματικότητα». Για διευκρίνιση των όρων, *ακρίβεια* θεωρείται το ποσοστό των τεκμηρίων που τέθηκαν στην σωστή κατηγορία από τον αλγόριθμο και ως *ανάκληση*, το ποσοστό των τεκμηρίων που τέθηκαν σε λάθος κατηγορία.

Από τα παραδείγματα που περιγράφηκαν, σύμφωνα με τον Sebastiani, την καλύτερη απόδοση επιτυγχάνουν οι *Support Vector Machines* και ακολούθως τα *Neural Networks*, ενώ οι *Probabilistic Classifiers*, όπως ο *Naïve Bayes* παρουσιάζουν την χειρότερη επίδοση. Δεν δίνεται κάποια άποψη για τα *Decision Trees* λόγω έλλειψης επαρκών δεδομένων. Σύμφωνα με άλλες έρευνες[14] τα *Decision Trees* αποδίδουν τόσο καλά όσο και οι *Support Vector Machines*.

Αντίθετα, όμως με τα ερευνητικά αποτελέσματα, πολλά από τα συστήματα κατηγοριοποίησης ηλεκτρονικών μηνυμάτων φαίνεται να προτιμούν τη μέθοδο *Naïve Bayes* ως βάση για τη διαδικασία κατηγοριοποίησης. Παράλληλα, ευρείας χρήσης τυγχάνει και το μοντέλο των *Support Vector Machines.*

- **Χωρίς επίβλεψη κατηγοριοποίηση**

14

Το κοινό μειονέκτημα όλων των ταξινομητών βάση της επιβλεπόμενης κατηγοριοποίησης, δηλαδή, την ανάγκη για προηγούμενη εκπαίδευσής τους σε ένα σώμα εκ των προτέρων ταξινομημένων τεκμηρίων, αποσκοπεί να επιλύσει η νέα μέθοδος για κατηγοριοποίηση χωρίς επίβλεψη. Το πρόβλημα με την ύπαρξη του *training set*, όπως χαρακτηριστικά αναφέρει ο Ko, J. είναι «*ενώ είναι εύκολη η συλλογή μη προσδιορισμένων τεκμηρίων, δεν είναι το ίδιο απλή η χειρωνακτική κατηγοριοποίησή τους, για τη δημιουργία ενός συνόλου εκπαίδευσης των ταξινομητών*» [11].

Η προτεινόμενη μέθοδος στο σχετικό άρθρο για τη χωρίς επίβλεψη κατηγοριοποίηση, διαιρεί τα τεκμήρια σε προτάσεις και κατηγοριοποιεί κάθε πρόταση χρησιμοποιώντας μία σειρά από λίστες λέξεων-κλειδιών για κάθε κατηγορία. Παράλληλα, κάνει χρήση της τεχνικής «μέτρο ομοιότητας», όπως και ο *k-nn* ταξινομητής, για την εύρεση των παραπλήσια όμοιων προτάσεων. Στην συνέχεια, οι κατηγοριοποιημένες προτάσεις παίρνουν τον ρόλο του *training set* για την εκπαίδευση των ταξινομητών.

Μία άλλη προσέγγιση στο πεδίο της μη-επιβλεπόμενης κατηγοριοποίησης, πραγματοποιείται από τη μέθοδο του *clustering.* Το πλεονέκτημα της συγκεκριμένης μεθόδου είναι ότι εξαλείφει την ανάγκη εκπαίδευσης των ταξινομητών καθώς δεν απαιτεί την ύπαρξη προηγούμενης ταξινόμησης ή δομής των κατηγοριών [b] . Τα συστήματα που χτίζονται βάση αυτής είναι ικανά να αναγνωρίζουν σύνολά ή ομάδες[15] σχετικών τεκμηρίων καθώς και τον συσχετισμό μεταξύ των ομάδων αυτών και με αυτό τον τρόπο να εκτελούν την κατηγοριοποίηση.

Για να πραγματοποιηθεί η ομαδοποίηση-όπως θα μπορούσε να αποδοθεί στα ελληνικά ο όρος clustering- κειμένων υπό κατηγορίες, η συγκεκριμένη μέθοδος εκτελείται σε επίπεδο λέξεων. Έχουν αναπτυχθεί διάφορες τεχνικές για την ομαδοποίηση λέξεων, όπως είναι η αυτόματη[16] ή κατανεμημένη ομαδοποίηση[17], η βάση κανόνων[18] ή βάση της χρήσης ελάχιστων προ-ταξινομημένων

[15] clusters
[16] Automated word Clustering
[17] Bekkerman, Ron. (2002).Distributional Clustering of Words for Text classification. Journal of Machine Learning Research 1 (2002) 1-48.
[18] Han, Hui. Rule-based Word Clustering for TC

παραδειγμάτων[19] καθώς και του co-training [9]. Οι τεχνικές αυτές προκειμένου να αναγνωρίσουν την ομοιότητα των λέξεων κάνουν χρήση τεχνικών, όπως είναι η γλωσσολογική ανάλυση ή η pattern recognition, δηλαδή, αναγνώριση μοτίβων. Στην τελευταία τεχνική, της αναγνώρισης, δηλαδή, μοτίβων, στηρίζουν την χρήση τους πολλά από τα νέα συστήματα κατηγοριοποίησης ηλεκτρονικών μηνυμάτων και τεκμηρίων.

Η χωρίς επίβλεψη κατηγοριοποίηση των τεκμηρίων όπως φαίνεται και στο σχεδιάγραμμα που παρατίθεται στην αρχή το κεφαλαίου, είναι η μέθοδος που βρίσκεται στην κορυφή του άξονα «αυτοματοποίηση», ενώ παράλληλα βρίσκεται στη βάση του άξονα «κόστος και χρόνος». Η κατάταξη αυτή την καθιστά ως τη βασική μέθοδο αυτοματοποίησης της κατηγοριοποίησης και για αυτό τον λόγο τα περισσότερα νέα συστήματα την υιοθετούν ως διακριτή τεχνική ή σε συνδυασμό με κάποιον από τους ταξινομητές που αναφέρθηκαν προηγουμένως.

Με εξαίρεση την χειρονακτική κατηγοριοποίηση που αναφέρθηκε κυρίως ως η απαρχή της σχετικής διαδικασίας, όλες οι άλλες μέθοδοι χρησιμοποιούνται καθεαυτό ή με μικρές τροποποιήσεις για την αυτοματοποίηση της κατηγοριοποίησης ή γενικά της διαχείρισης των ηλεκτρονικών μηνυμάτων. Βέβαια, πέρα από τις κύριες αυτές μεθόδους, στα σχετικά συστήματα υιοθετούνται και άλλες εναλλακτικές όπως είναι βάση μεταδεδομένων ή έχουν μόνο προταθεί όπως είναι η κατηγοριοποίηση ηλεκτρονικών μηνυμάτων βάση οντολογιών.

- **<u>Εναλλακτικοί μέθοδοι κατηγοριοποίησης</u>**

Στο τμήμα αυτό θα αναφερθούν τρεις επιπλέον μέθοδοι κατηγοριοποίησης ηλεκτρονικών μηνυμάτων: βάση οντολογίας, μέσω εκμάθησης πράξεων λόγου και βάση μεταδεδομένων. Οι δύο πρώτες προσεγγίσεις δεδομένης της έρευνας

[19] Zeng, Hua. CBC: Clustering Based Text Classification Requiring Minimal Labeled Data

που πραγματοποιήθηκε στον τομέα των συστημάτων διαχείρισης ηλεκτρονικών μηνυμάτων δεν είναι γνωστό αν υιοθετούνται από κάποιο σύστημα ως βασική ή ως υποστηρικτική λύση. Αντίθετα, η τρίτη μέθοδος γνωρίζει μεγάλη ανάπτυξη και πιθανά θα είναι αυτή που θα επικρατήσει των άλλων τεχνικών στα συστήματα του μέλλοντος.

Στο πανεπιστήμιο της Nevada αναπτύχθηκε η μέθοδος κατηγοριοποίησης ηλεκτρονικών μηνυμάτων βάση οντολογίας[20]. Έναυσμα υπήρξε η ανάγκη διατήρησης των ηλεκτρονικών μηνυμάτων και η διαχείρισή τους ως ηλεκτρονικά αρχεία. Στην πρακτική της φάση, η μέθοδος υποστηρίζεται από ένα περιβάλλον εισαγωγής και διαχείρισης γνώσης, το Protege 2000, το οποίο διατηρεί την οντολογία. Παράλληλα, ως αρχικό πρόγραμμα διαχείρισης των ηλεκτρονικών μηνυμάτων έχει χρησιμοποιηθεί το Lotus Notes.

Η οντολογία είναι μία γραφική ιεραρχική αναπαράσταση της γνώσης που κάνει χρήση κλάσεων (υπερκλάσεις, κλάσεις, υποκλάσεις), στιγμιότυπων και ιδιοτήτων για την εισαγωγή δεδομένων που αφορούν ένα νοηματικό τομέα. Συγκεκριμένα για την κατηγοριοποίηση ηλεκτρονικών μηνυμάτων η χρήση της έγκειται στην εφαρμογή κανόνων για τον προσδιορισμό των χαρακτηριστικών που θα χρησιμοποιηθούν για την κατηγοριοποίηση.

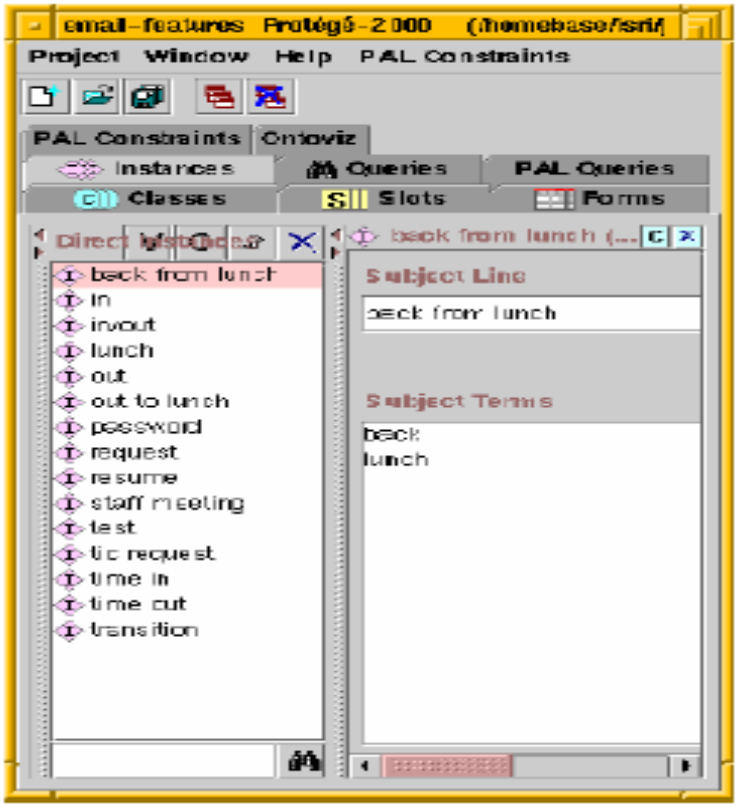
Οι κανόνες που έχουν τη μορφή **«if..then»,** γράφονται σε γλώσσα κατανοητή από το σύστημα Protege. Στην σχετική προσέγγιση υπάρχει άρρηκτη σχέση μεταξύ των κανόνων που εφαρμόζονται και στις κλάσεις της οντολογίας. Ουσιαστικά, οι κανόνες αντανακλούν το περιεχόμενο των κλάσεων. Οι κλάσεις ορίζονται από έννοιες όπως αυτές εξάγονται από δύο πηγές: από τις αρχές διατήρησης και διάθεσης ηλεκτρονικών αρχείων και από το σώμα προ-κατηγοριοποιημένων μηνυμάτων που χρησιμοποιείται για την εκπαίδευση του ταξινομητή. Αναφορικά, ο ταξινομητής που χρησιμοποιείται αποτελεί τροποποίηση του μοντέλου Naive Bayes.

Αφού οριστούν τα διάφορα επίπεδα της οντολογίας, ακολουθεί μία διαδικασία μετατροπής των ηλεκτρονικών μηνυμάτων σε διάταξη κατανοητή από το σύστημα

[20] Ontology-based Classification of email

Protege, και συνεπώς να μπορούν να εφαρμοσθούν σε αυτά οι κανόνες κατηγοριοποίησης. Από το αρχικό σύστημα διαχείρισης των μηνυμάτων, το Lotus Notes, τα μηνύματα μετατρέπονται σε XML διάταξη χρησιμοποιώντας μία διάταξη, γνωστή ως DXL (Domino eXtended markup language). Στην συνέχεια, χρησιμοποιείται ένα σύστημα, το οποίο δημιουργήθηκε για αυτό τον σκοπό και το οποίο εξάγει πληροφορίες από τα πεδία της DXL διάταξης των μηνυμάτων και τις οποίες εισάγει ως στιγμιότυπα των κλάσεων. Με το τέλος της σχετικής διαδικασίας, οι κλάσεις της οντολογίας, οι κανόνες και τα ηλεκτρονικά μηνύματα είναι γραμμένα στην ίδια γλώσσα και μπορούν να γίνουν κατανοητά και επεξεργάσιμα από το σύστημα. Με την μετατροπή του μηνύματος και την εισαγωγή του στο σύστημα, γίνεται σύγκριση των στιγμιότυπών του με τα στιγμιότυπα των κλάσεων. Στην περίπτωση που πραγματοποιηθεί ταύτιση των στιγμιότυπων, το μήνυμα κατατάσσεται υπό την εκάστοτε κλάση.

Στην παρακάτω εικόνα δίνεται παράδειγμα της οντολογίας και ενός ηλεκτρονικού μηνύματος, όπως αυτό μετατρέπεται προκειμένου να γίνει κατανοητό.



Εικόνα 2 παράδειγμα οντολογίας

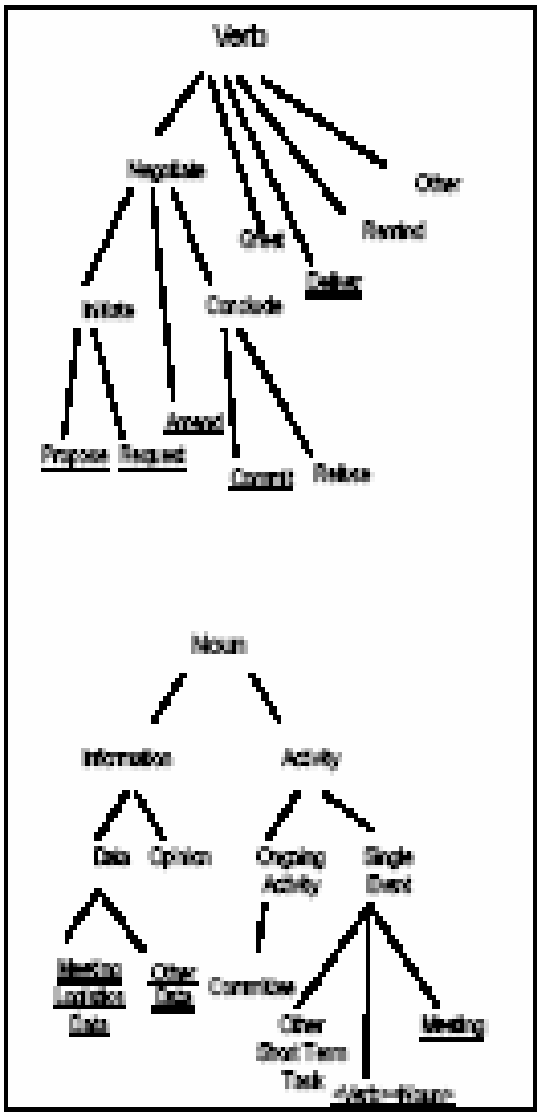
```
([S000046947-email_author-1] of email_author
 (agent_name           "Kathryn Cooper")
 (top_domain_name      "gov")
 (individual_username  "kathryn_cooper")
 (org_domain_name      "ymp")
 (dept_domain_name     "ym"))

([S000046947-email_send_to-1] of email_send_to
 (agent_name           "Nick Connerley")
 (top_domain_name      "gov")
 (individual_username  "nick_connerley")
 (org_domain_name      "ymp")
 (dept_domain_name     "ym"))

([S000046947-email] of email
 (id                  S000046947)
 (body_lengths        9)
 (attachment_counts   0)
 (subject_line        "Lunch")
 (email_author        [S000046947-email_author-1])
 (email_send_to       [S000046947-email_send_to-1]))
```

Εικόνα 3 – παράδειγμα τροποποίησης της διάταξης ηλεκτρονικού μηνύματος

Η κατηγοριοποίηση ηλεκτρονικών μηνυμάτων σε "Speech Acts"[21], δηλαδή, σε «πράξεις λόγου» είναι η προσέγγιση που ερευνάται από τον Cohen προκειμένου να επιτευχθεί η ταξινόμηση των μηνυμάτων βάση της πρόθεσης του δημιουργού. Η πρόθεση αυτή, υποστηρίζεται ότι μπορεί να υποδηλωθεί από τα ζεύγη «ρήμα-ουσιαστικό», τα οποία εξάγονται από το σώμα του μηνύματος και τα οποία μπορούν να αποδώσουν ικανοποιητικά το νόημά του. Η μέθοδος αυτή διαφοροποιείται από την παραδοσιακή θεματική κατηγοριοποίηση, παρόλα αυτά μπορεί να αποδώσει επαρκώς όσον αφορά τον άξονα ακρίβεια και ανάκληση για συγκεκριμένες κατηγορίες. Η κατηγοριοποίηση των μηνυμάτων βάση της πρόθεσης του δημιουργού μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη κυρίως σε περιπτώσεις που ο χρήστης θέλει να εντοπίσει την κατάσταση των δραστηριοτήτων στις οποίες συμμετέχει, φαινόμενο το οποίο έχει οριστεί ως email threading.

[21] Learning to Classify Email into "Speech Acts"

Κοινό της συγκεκριμένης μεθόδου με αυτή που παρουσιάστηκε παραπάνω είναι ότι και οι δύο κάνουν χρήση οντολογίας με τη βοήθεια της οποίας πραγματοποιείται η κατηγοριοποίηση. Στο πλαίσιο της σχετικής μεθόδου, η οντολογία περιλαμβάνει ρήματα και ουσιαστικά., τα οποία σε συνδυασμό περιγράφουν την «πράξη λόγου» που περικλείεται στο μήνυμα. Στην παρακάτω εικόνα δίνεται παράδειγμα μίας συντηγμένης οντολογίας που χρησιμοποιείται.



Εικόνα 4 Παράδειγμα οντολογίας ρημάτων-ουσιαστικών

Από την παραπάνω οντολογία που παρουσιάζεται στην εικόνα, μπορεί να γίνει κατανοητός ο τύπος των ρημάτων και των ουσιαστικών, που επιλέγονται και τα οποία υποδηλώνουν σαφώς την πρόθεση του δημιουργού του μηνύματος όσον αφορά την πρόταση που αποστέλλει προς τον παραλήπτη.

Για την επιλογή των ζευγών ρημάτων- ουσιαστικών και επακόλουθα για τη δημιουργία της οντολογίας πραγματοποιήθηκε έρευνα και ανάλυση σε διάφορα σώματα ηλεκτρονικών μηνυμάτων προκειμένου να βρεθούν καθιερωμένες εκφράσεις. Παράλληλα, γίνεται χρήση ταξινομητή για την κατανόηση του περιεχομένου του μηνύματος, ο οποίος ακολουθεί το μοντέλο των Support Vector Machines. Η δημιουργία της οντολογίας περιλαμβάνει περαιτέρω βήματα, κατά τα οποία πραγματοποιείται εξειδίκευση του περιεχομένου της ώστε να περιλαμβάνει λιγότερο γενικά γλωσσολογικά στοιχεία.

Μία από τις νεώτερες και πλέον υποσχόμενες μεθόδους κατηγοριοποίησης ηλεκτρονικών μηνυμάτων είναι βάση εξαγωγής μεταδεδομένων[22]. Στο σημείο αυτό θα παρουσιαστεί μία από τις προσεγγίσεις που πραγματοποιούνται βάση αυτής της μεθόδου, παρουσιάζοντας τη βασική λειτουργία του συστήματος <!metaMarker>.

Το σύστημα δημιουργήθηκε για να καλύψει τις ανάγκες της κοινότητας των οικονομικών αναλυτών και των πελατών τους. Ουσιαστικά, είναι ένα εργαλείο αυτόματης παραγωγής μεταδεδομένων σε XML διάταξη. Η λειτουργία του, παράλληλα, βασίζεται σε τεχνικές επεξεργασίας φυσικής γλώσσας (NLP). Βάση των σχετικών τεχνικών πραγματοποιείται εξαγωγή μεταδεδομένων που αντιστοιχούν σε προσωπικές πληροφορίες που περικλείονται στην ηλεκτρονική αλληλογραφία των οικονομικών αναλυτών και των πελατών τους. Οι πληροφορίες που εξάγονται προσδιορίζονται με τον όρο προσωπικές καθώς χρησιμοποιούνται για τη δημιουργία διασυνδέσεων μεταξύ των χρηστών και του περιεχομένου του ηλεκτρονικού μηνύματος, γεγονός που επακόλουθα συντελεί στη δημιουργία προφίλ χρηστών.

Ο <!metaMarker> εξάγει στοιχεία μεταδεδομένων που περιλαμβάνουν κύρια ονόματα, αριθμητικές έννοιες καθώς και πληροφορίες που σχετίζονται με το θέμα του μηνύματος. Επίσης, οι δημιουργοί του επηρεαζόμενοι από την ιδέα της «**Speech Acts**» -όπως αυτή περιγράφηκε παραπάνω- συντέλεσαν ώστε το σύστημα να είναι ικανό να εξάγει μεταδεδομένα που αναφέρονται στη διάθεση

[22] Applying Natural language Processing (NLP) Based Metadata Extraction to Automatically Acquire User Preferences.

του δημιουργού, στο σκοπό και στην προτεραιότητα τους. Τα μεταδεδομένα, με αυτό τον τρόπο θεωρητικά διαχωρίζονται σε ρητά και σε υπονοούμενα. Για τη δημιουργία των μεταδεδομένων ακολουθείται μία διαδικασία που περιλαμβάνει επτά βήματα. Στο αρχικό βήμα, το σώμα του ηλεκτρονικού μηνύματος διασπάται στις βασικές προτάσεις του, κάνοντας χρήση των ετικετών <s#1>, </s#1>που υποδηλώνει τα όρια της πρότασης. Στο επόμενο βήμα γίνεται διαχωρισμός των συστατικών που αποτελούν την πρόταση, ενώ στο τρίτο βήμα αφαιρούνται προθέματα και επιθέματα από τις λέξεις. Στο επόμενο βήμα, αρχίζει η επεξεργασία των νοημάτων και στα πλαίσια κάθε πρότασης αναγνωρίζονται και προσδιορίζονται ρητά , με την ετικέτα <pn>, </pn> τα όρια των νοημάτων. Στο πέμπτο βήμα, σε κύρια ονόματα και αριθμητικά δεδομένα προστίθονται προσδιορισμοί που αφορούν τον τύπο των πληροφοριών που περιέχουν, ώστε να μπορεί να πραγματοποιηθεί η κατηγοριοποίηση. Στο έκτο βήμα, εισάγονται στο περιεχόμενο του μηνύματος υπονοούμενα μεταδεδομένα, που αναφέρονται στον σκοπό, τη διάθεση ή την προτεραιότητα του μηνύματος. Το τελευταίο βήμα της διαδικασίας αναφέρεται στην εξαγωγή προτιμήσεων του χρηστή, που είναι ένας συνδυασμός ρητών και υπονοουμένων μεταδεδομένων. Για τη διαδικασία προσδιορισμού μεταδεδομένων χρησιμοποιούνται ταξινομητές όπως ο Naïve Bayes ή ο k-nn.

Αφού ολοκληρωθεί η διαδικασία προσδιορισμού μεταδεδομένων, ακολουθεί η κάθε αυτού διαδικασία της κατηγοριοποίησης, η οποία έγκειται σε δύο φάσεις. Αρχικά, στην εκπαίδευση του συστήματος σε ένα σώμα προ-κατηγοριοποιημένων μηνυμάτων και στην συνέχεια, στην κατηγοριοποίηση εισερχόμενων μηνυμάτων. Το σύστημα προκειμένου να κατηγοριοποιήσει το εκάστοτε μήνυμα εξετάζει το «βαθμό συμμετοχής» του για κάθε κατηγορία, αποδίδοντας τιμές από το μηδέν στο ένα. Στην περίπτωση που το μήνυμα λάβει την τιμή μηδέν αυτόματα η συμμετοχή του στην συγκεκριμένη κατηγορία αποκλείεται, ενώ αντίστοιχα εντάσσεται στην εκάστοτε κατηγορία, αν λάβει τιμή ένα.

2. Ηλεκτρονικά Μηνύματα

Σκοπός αυτού του τμήματος είναι να δώσει μία πιο εξειδικευμένη εικόνα στο ζήτημα της αυτόματης κατηγοριοποίησης ηλεκτρονικών τεκμηρίων, προσανατολίζοντας την έρευνα στον τομέα των ηλεκτρονικών μηνυμάτων. Στην εισαγωγή της εργασίας έγινε αναφορά στα ηλεκτρονικά μηνύματα κυρίως από ιστορική και λειτουργική σκοπιά. Στο τμήμα αυτό επιχειρείται μία εμβάθυνση στην φύση των ηλεκτρονικών μηνυμάτων προκειμένου να κατανοηθούν οι δυσκολίες που δημιουργούνται στη διαχείρισή τους και ειδικότερα στην κατηγοριοποίησή τους. Στην συνέχεια, γίνεται αναφορά και συγκριτική περιγραφή διαφόρων πακέτων λογισμικού, που προσανατολίζονται στην διευκόλυνση του χρήστη αυτοματοποιώντας διαδικασίες που σχετίζονται με τη διαχείριση των ηλεκτρονικών μηνυμάτων.

2.1 Ηλεκτρονικά Μηνύματα – Η Φύση τους

Τα ηλεκτρονικά μηνύματα ως στοιχείο της τάξης των ηλεκτρονικών τεκμηρίων εντάσσονται στην κατηγορία των ήμι-δομημένων δεδομένων [b] . Αν το ηλεκτρονικό μήνυμα θεωρηθεί ως σύνολο πεδίων, η κατάταξη του στα ήμι-δομημένα δεδομένα οφείλεται στο γεγονός ότι περιλαμβάνει πεδία εισαγωγής προκαθορισμένων δεδομένων αλλά και πεδία, των οποίων ο τύπος των δεδομένων που εισάγονται δεν είναι γνωστός ή προσδιορισμένος.

Ένα ηλεκτρονικό μήνυμα περιλαμβάνει τα εξής συστατικά στοιχεία: τον *header*, το *κύριο σώμα* και πιθανά ένα ή περισσότερα *συνημμένα έγγραφα*. Σύμφωνα με τον Klimt, Β. κάθε ηλεκτρονικό μήνυμα *περιλαμβάνει μη-δομημένο κείμενο, κατηγορηματικό και αριθμητικά δεδομένα* [10]. Ο *header* περιλαμβάνει πεδία, όπως η «*ημερομηνία**date*», ο «*αποστολέας**from*», ο «*παραλήπτης**to*» ή «*λίστα κοινοποίησης**cc*» και το «*θέμα**subject*». Σύμφωνα, με την κατηγοριοποίηση του Klimt, το πεδίο «*ημερομηνία*», εντάσσεται στα αριθμητικά δεδομένα, το πεδίο «*θέμα*» και το «*σώμα*» του μηνύματος εντάσσονται στο μη-δομημένο κείμενο, ενώ τα πεδία «*αποστολέας*» και «*παραλήπτης*» ανήκουν στο κατηγορηματικό κείμενο. Οι τύποι των τιμών των πεδίων του κατηγορηματικού κειμένου και των αριθμητικών δεδομένων είναι προκαθορισμένοι και μπορούν να γίνουν εύκολα κατανοητά από τον ταξινομητή. Ο ταξινομητής, δηλαδή, μπορεί να καταλάβει εκ των προτέρων ότι τα δεδομένα που περιλαμβάνονται στο πεδίο «από» αναφέρονται στον αποστολέα του μηνύματος.

Δεν συμβαίνει, όμως το ίδιο και με τα δεδομένα που περιλαμβάνονται στο *σώμα* του ηλεκτρονικού μηνύματος καθώς και στο πεδίο «θέμα». *Το πρόβλημα με τα μη-δομημένα κείμενα είναι ότι οι πιθανές νοηματικές αλληλουχίες*[23] *μπορεί να είναι απεριόριστες και η σημασία των λέξεων μπορεί να αλλάζει ανάλογα με το γενικό περιεχόμενο ή με την σειρά των λέξεων* [3]. Δεδομένου ότι το *σώμα* του μηνύματος γράφεται με τη μορφή plain text, δηλαδή, απλού κειμένου και εφόσον η θεματολογία των ηλεκτρονικών μηνυμάτων είναι απεριόριστη, ο ταξινομητής δεν μπορεί να κατανοήσει το κείμενο ώσπου να εισαχθούν διαδικασίες διαχείρισης του περιεχομένου. Τέλος, τα *συνημμένα έγγραφα*, εμφανίζονται, παράλληλα, σε πληθώρα διατάξεων καθώς μπορεί να περιλαμβάνουν δομημένα έγγραφα, όπως τα PDF ή tax forms, ήμι-δομημένα, όπως τα HTML έγγραφα ή αδόμητα, όπως τα έγγραφα Word ή Notepad. Μπορεί όμως να περιλαμβάνουν και αρχεία, τα οποία δεν σχετίζονται με κείμενο αλλά με φωτογραφίες ή πολυμέσα.

Η δυσκολία της κατηγοριοποίησης έγκειται στο γεγονός ότι θα πρέπει να υπολογιστούν τα δεδομένα ή καλύτερα το περιεχόμενο των δεδομένων όλων των πεδίων, προκειμένου να επιτευχθεί η καλύτερη δυνατή κατηγοριοποίηση. Μία

[23] string

λύση, βέβαια θα ήταν να χρησιμοποιηθούν τα πεδία της επικεφαλίδας και κυρίως ο «*αποστολέας*» ή το «*θέμα*», προκειμένου να ενταχθούν τα ηλεκτρονικά τεκμήρια σε κατηγορίες. Όπως, όμως αποδεικνύει η καθημερινή πρακτική το περιεχόμενο των μηνυμάτων ενός συγκεκριμένου αποστολέα προς έναν συγκεκριμένο παραλήπτη μπορεί να ποικίλλει σε μεγάλο βαθμό. Παράλληλα, παρατηρείται συχνά το φαινόμενο, το περιεχόμενο του μηνύματος να μην έχει καμία σχέση με το θέμα που αναφέρεται στο πεδίο «*θέμα*» [2]. Συνεπώς, κανένα από τα πεδία αυτά δεν είναι αξιόπιστο αρκετά ή λειτουργικό προκειμένου να στηριχθεί σε αυτό η διαδικασία της κατηγοριοποίησης. Θα πρέπει, επίσης να σημειωθεί ότι το πεδίο «*ημερομηνία*» χρησιμοποιείται ελάχιστα κατά την κατηγοριοποίηση, καθώς η πληροφοριακή αξία των (αριθμητικών) δεδομένων του είναι μικρή [] .

Παράλληλα, προκειμένου να γίνει ευκολότερη η κατανόηση του περιεχομένου των δεδομένων που περιλαμβάνονται στο σώμα του μηνύματος, είχε προταθεί η μέθοδος να ενταχθεί ένα βασικό επίπεδο δόμησης των δεδομένων του σχετικού πεδίου[24]. Μία τέτοια προσέγγιση, όμως αντιτίθεται στις βασικές αρχές δημιουργίας των ηλεκτρονικών μηνυμάτων και κυρίως στην ευκολία χρήσης που προσφέρει στους χρήστες του [2]. Επίσης, διάφορα συστήματα διαχείρισης του περιεχομένου των ηλεκτρονικών μηνυμάτων διασπούν κατά την εισαγωγή τους στο σύστημα το κύριο σώμα των μηνυμάτων από τα συνημμένα και τα διαχειρίζονται ως διαφορετικές οντότητες. Αυτό σημαίνει ότι τα αποθηκεύουν και τα κατηγοριοποιούν ανεξάρτητα το ένα από το άλλο, διαταράσσοντας με αυτό τον τρόπο την ενότητα του αρχείου.

Συχνά, επίσης, παρατηρείται το φαινόμενο δημιουργίας αλληλουχιών των ηλεκτρονικών μηνυμάτων. Το φαινόμενο, που απαντάται στην λατινογενή βιβλιογραφία, ως *email threading [10],* συνίσταται από μία σειρά εισερχόμενων μηνυμάτων που αποτελούν απαντήσεις σε παλαιότερα μηνύματα και τα οποία είχαν αποσταλεί από τον ίδιο αποστολέα προς έναν ή περισσότερους παραλήπτες. Δημιουργείται, με αυτό τον τρόπο μία νοηματική και φυσική αλληλουχία των μηνυμάτων, η οποία είναι σημαντικό να διατηρηθεί. Η διατήρησή της είναι μείζονος σημασίας για τον πομπό και τον δέκτη των μηνυμάτων, καθώς πιθανά, ο

[24]Palme J. (1984): *You have 134 Unread Mail! Do you want to read them know?* Proceedings of IFIP Wg 6.5 Working Conference on Computer-Based Message Services.

ενδιαφερόμενος, θα θέλει να ανακτήσει το σύνολο αυτών των μηνυμάτων προκειμένου να ανατρέξει στα διάφορα στάδια μετεξέλιξης του σχετικού θέματος.

Το πρόβλημα, με τη διατήρηση της αλληλουχίας των ηλεκτρονικών μηνυμάτων είναι διττό. Αρχικά, χαρακτηριστικό γνώρισμα των σχετικών μηνυμάτων είναι ότι φέρουν κοινό το πεδίο «*θέμα*» συνήθως με το δεδομένο «***RE:***», που σημαίνει «*in reply to*», δηλαδή «*σε απάντηση του προηγούμενου μηνύματος*». Το γεγονός αυτό θα μπορούσε να οδηγήσει τον ταξινομητή να καταχωρήσει το μήνυμα ως *SPAM.* Παράλληλα, αν και το θέμα του μηνύματος παραμένει το ίδιο, το περιεχόμενο του σώματος σταδιακά εξελίσσεται και μεταβάλλεται, που σημαίνει ότι τα χαρακτηριστικά του διαφοροποιούνται. Για ένα ταξινομητή, που δεν κατανοεί την αλληλουχία, αυτό σημαίνει ότι το θέμα αλλάζει και συνεπώς θα πρέπει να ενταχθεί σε διαφορετικές κατηγορίες. Αν και έχει πραγματοποιηθεί έρευνα[25] στη δημιουργία ταξινομητών ικανών να κατανοήσουν τις αλληλουχίες, παρόλα αυτά, δεν είναι πολλά τα συστήματα, που τους εισάγουν.

Ένα άλλο σημαντικό πρόβλημα είναι ότι οι κατηγορίες υπό τις οποίες εντάσσονται τα ηλεκτρονικά τεκμήρια δεν είναι στατικές αλλά και μεταξύ τους παρουσιάζουν διαφορετικά χαρακτηριστικά. Με τον όρο στατικές εννοείται ότι δεν παραμένουν αμετάβλητες ή αναλλοίωτες. Με την πάροδο του χρόνου τα ενδιαφέροντα του χρήστη αλλάζουν και συνεπώς τροποποιούνται και τα περιεχόμενα των κατηγοριών. Κάποιες κατηγορίες μπορεί να σταματήσουν να χρησιμοποιούνται, ενώ σε άλλες να τροποποιηθεί το θεματικό τους αντικείμενο. Αυτό σημαίνει ότι, ενώ ένας ταξινομητής που ήταν αποτελεσματικός για κάποια χρονική περίοδο μετά τις μεταβολές στις κατηγορίες, είναι πιθανό να μειωθεί η αποδοτικότητά του. Το πρόβλημα αυτό αντιμετωπίζεται με τη μέθοδο «*incremental learning*», δηλαδή *σταδιακή εκπαίδευση στις αλλαγές* [16].

Επίσης, οι φάκελοι του χρήστη που χρησιμοποιούνται ως εικονικές κατηγορίες μπορεί να παρουσιάζουν διαφορετικά χαρακτηριστικά ως προς το μέγεθός τους, την ομοιογένεια των περιλαμβανομένων τεκμηρίων αλλά και τον αριθμό των μηνυμάτων που περιλαμβάνουν. Οι παράγοντες αυτοί επηρεάζουν την κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων καθώς θα πρέπει να ληφθούν

[25] D. D. Lewis, K. A. Knowles: Threading Electronic Mail: A Preliminary Study. In Information Processing and Management, 33(2): 209-217, 1997

υπόψη για την επιλογή των αντιπροσωπευτικών δομών δημιουργίας κατηγοριών. Αυτό συνεπάγεται την εξαγωγή κάποιων χαρακτηριστικών με την εκ των προτέρων επεξεργασία των φακέλων και χρήση αυτών ως παραμέτρους για την επιλογή του συνόλου τεκμηρίων που θα χρησιμοποιηθεί για την εκπαίδευση του ταξινομητή αλλά και για να υπολογιστεί η ιεραρχία τους [18].

Λόγω της αυξημένης χρήσης που παρουσιάζουν τα ηλεκτρονικά μηνύματα- έρευνες έχουν αποδείξει ότι ένας μέσος χρήστης μπορεί να λαμβάνει μέχρι και τριάντα ή σαράντα μηνύματα ημερεσίως- αλλά και λόγω της σπουδαιότητας των πληροφοριών που μεταφέρουν-εταιρικά δεδομένα- έχουν συνασπιστεί ειδικοί κανόνες που αφορούν τον τρόπο διαχείρισης και αποθήκευσης των ηλεκτρονικών μηνυμάτων. Πλέον τα ηλεκτρονικά μηνύματα που μεταδίδονται μέσω του εταιρικού intranet ή μέσω διαδικτύου, θεωρούνται ηλεκτρονικά αρχεία και συνεπώς θα πρέπει να υποβάλλονται σε ειδικές διαδικασίες διατήρησης. Το σύνολο αυτό των οδηγιών, τουλάχιστον για τα δεδομένα των Ηνωμένων Πολιτειών υπόκεινται στις αρχές του DoD 5015.2.

2.2 Λογισμικό για την Κατηγοριοποίηση Ηλεκτρονικών Μηνυμάτων

2.2.1 Agents[26]

Οι **Intelligent Agents** είναι ειδικά προγράμματα υπολογιστών που χρησιμοποιούν τεχνικές *Τεχνητής Νοημοσύνης* προκειμένου να βοηθήσουν τους χρήστες στην εκτέλεση εργασιών που στηρίζονται στην λειτουργία υπολογιστή [20].

Οι **agents** αλλάζουν ριζικά τον τρόπο εκτέλεσης των εργασιών καθώς λειτουργούν σαν ένα είδος προσωπικών βοηθών, η αποστολή των οποίων είναι

[26] Για τη γενική περιγραφή της λειτουργίας των agent, χρησιμοποιήθηκε το άρθρο του Φλωρινά, Ν. «Intelligent Agents». URL:www.dide.flo.sch.gr/Plinet/Tutorials/ Tutorials-IntelligentAgents.html.

να μειώσουν τον φόρτο εργασίας και την υπερφόρτωση πληροφοριών από τον desktop του χρήστη. Ουσιαστικά, ο *agent* λειτουργεί ως ένας μεσολαβητής μεταξύ του χρήστη και του υπολογιστή, παρεμβαλλόμενος στην μεταξύ τους επικοινωνία. Με τη βοήθεια των *agents*, αντί να υπάρχει αλληλεπίδραση που να προκαλείται από τον χρήστη με τον υπολογιστή, μέσω των δικών του εντολών ή ενεργειών, ο χρήστης συνεργάζεται με τους *agents* που έχει δημιουργήσει και στους οποίους έχει αναθέσει διάφορες εργασίες. Κάποιες από τις εργασίες στις οποίες μπορεί να βοηθήσει τον χρήστη χωρίς να περιορίζεται σε αυτές, είναι: *το φιλτράρισμα και η ανάκτηση πληροφοριών, ο χειρισμός αλληλογραφίας, η οργάνωση συναντήσεων ακόμα και η επιλογή βιβλίων, ταινιών, μουσικής.*

Για τη δημιουργία ενός *agent*, απαιτείται η μετατροπή του προγράμματος του τελικού χρήστη σε *interface agent* (χωρίς όμως αυτός να είναι ο μοναδικός τρόπος). Για παράδειγμα, ένας χρήστης μπορεί να δημιουργήσει έναν *electronic mail sorting agent*, δημιουργώντας ένα σύνολο από κανονισμούς που επεξεργάζονται τα εισερχόμενα μηνύματα και τα ταξινομούν σε διαφορετικές κατηγορίες-φακέλους.

Ο **electronic mail agent** μπορεί να μάθει να ιεραρχεί τα μηνύματα βάση σπουδαιότητας (prioritize), να διαγράφει, να προωθεί, να ταξινομεί και να αρχειοθετεί την ηλεκτρονική αλληλογραφία κατά εντολή του χρήστη. Οι εργασίες αυτές εκτελούνται βάση της διαδικασίας παρακολούθησης από τον *agent* των εργασιών του χρήστη. Στο αρχικό επίπεδο, ο *agent* συνεχώς «*κοιτάει πάνω από τον ώμο*» του χρήστη, δηλαδή τον παρακολουθεί καθώς αυτός διαχειρίζεται την ηλεκτρονική του αλληλογραφία. Καθώς, ο χρήστης εκτελεί τις εργασίες, ο *agent* απομνημονεύει όλα τα ζεύγη κατάστασης-ενέργειας που δημιουργούνται.

Για παράδειγμα, αν ο χρήστης αποθηκεύσει ένα συγκεκριμένο μήνυμα ή το διαγράψει, ο *agent* προσθέτει στη μνήμη του μία περιγραφή της κατάστασης και των ενεργειών που εκτέλεσε ο χρήστης. Παράλληλα, καταγράφει τον αποστολέα, τον παραλήπτη, τις λέξεις-κλειδιά του θέματος και άλλες σχετικές ενέργειες. Όταν, πραγματοποιείται μία νέα κατάσταση π.χ. εισαγωγή νέου μηνύματος, ο *agent* προσπαθεί να προβλέψει τις ενέργειες του χρήστη, συγκρίνοντας τη με τα παραδείγματα που έχει καταχωρημένα στη μνήμη του. Μετά την σύγκριση των καταστάσεων και την εύρεση της πλέον κοντινής, ο *agent* αποφασίζει στο ποια

ενέργεια θα πρέπει να ληφθεί. Η τεχνική που χρησιμοποιείται από τους *agent*s για την σύγκριση των καταστάσεων, ονομάζεται *μετρικό απόστασης* και ουσιαστικά είναι ένα «*ζυγισμένο άθροισμα των διαφορών για τα χαρακτηριστικά που αποτελούν μία κατάσταση-μερικά χαρακτηριστικά έχουν μεγαλύτερη βαρύτητα από τα άλλα*[27]»

Η βαρύτητα του κάθε χαρακτηριστικού εξαρτάται από τον *agent*. Κατά την απομνημόνευση των ζευγών κατάσταση-ενέργεια, ο *agent* καθορίζει τους συσχετισμούς[28] ανάμεσα στα χαρακτηριστικά και στις ενέργειες που αναλαμβάνονται. Για παράδειγμα, ο *agent* μπορεί να εντοπίσει ότι το πεδίο «*από*» ενός μηνύματος έχει υψηλή συσχέτιση με το αν ο χρήστης διαβάζει το μήνυμα, ενώ το πεδίο «*ημερομηνία*» δεν σχετίζεται. Οι συσχετίσεις που εντοπίζονται αναφέρονται ως «*βαρύτητες*» (weights) στην τεχνική *μετρικό απόστασης*.

Σε ένα δεύτερο επίπεδο, ο *agent* παράλληλα με την πρόβλεψη των ενεργειών, μετράει και το επίπεδο εμπιστοσύνης σε κάθε πρόβλεψη. Το επίπεδο εμπιστοσύνης καθορίζεται από διάφορες παραμέτρους, όπως είναι το: αν όλες οι γειτονικές εκτιμήσεις έχουν προτείνει την ίδια ενέργεια ή όχι, πόσο κοντά βρίσκονται οι πιο κοντινές γειτονικές εκτιμήσεις και από τα πόσα παραδείγματα έχει απομνημονεύσει. Παράλληλα, ορίζονται δύο *threshold*- τα οποία μπορούν να οριστούν ως όρια- που καθορίζουν το πώς χρησιμοποιεί ο *agent* κάθε πρόβλεψή του. Στην περίπτωση που το επίπεδο εμπιστοσύνης[29] βρίσκεται πάνω από το όριο, που τον παρακινεί σε εκτέλεση ενέργειας, τότε, ο *agent* αυτόνομα εκτελεί την ενέργεια. Ταυτόχρονα, στέλνει μία αναφορά προς τον χρήστη προκειμένου να τον ενημερώσει για την ενέργεια που πραγματοποίησε.

Στην αντίθετη περίπτωση, όταν το επίπεδο εμπιστοσύνης βρίσκεται κάτω του σχετικού ορίου, ο *agent* στέλνει πρόταση προς τον χρήστη αλλά αναμένει την επιβεβαίωσή του ή την άρνησή του, προτού φέρει εις πέρας την εργασία. Ο χρήστης μπορεί από μόνος του να τροποποιήσει τα επίπεδα των ορίων, σε κλίμακα, όπου αυτός αισθάνεται κατάλληλη.

[27] Φλωρινάς, Ν: Intelligent Agents
[28] correlation
[29] confidence level

Για τη διαχείριση των ηλεκτρονικών μηνυμάτων έχουν προταθεί διάφοροι *agent*s, οι οποίοι στηρίζονται σε διαφορετικές τεχνικές για την αναπαράσταση των μηνυμάτων και εξαγωγή των χαρακτηριστικών. Στην συνέχεια, γίνεται προσπάθεια περιγραφής των βασικών ερευνητικών προσεγγίσεων στη δημιουργία *agent*s και στην λειτουργία τους. Παράλληλα, οι περισσότεροι *email clients,* όπως το *Outlook, cMail, Eudora* στηρίζονται στην χρήση *agents* για το φιλτράρισμα της ηλεκτρονικής αλληλογραφίας του χρήστη [2] .

Το σύστημα **Information Lens**[30] αναπτύχθηκε το 1987 ως ένα πρωτοποριακό εργαλείο, στο οποίο οι *agent*s θα βοηθούσαν τους χρήστες να βρουν, να φιλτράρουν και να ταξινομήσουν μεγάλες ποσότητες ηλεκτρονικών πληροφοριών. Το σύστημα είχε έναν κεντρικό *server* – με το όνομα «*anyone*»- που λάμβανε μηνύματα που περιελάμβαναν το «*anyone*» ως διεύθυνση από τον *email server*. Με την αυτόματη ταξινόμηση και περιοδική ανάκτηση μηνυμάτων από το *mailbox*, ο κεντρικός *server* έστελνε το μήνυμα σε διάφορους παραλήπτες, των οποίων οι κανόνες το επέλεγαν. Οι κανόνες αποτελούνταν από μία δοκιμή/test και μία δράση/action. Στην περίπτωση που ένα μήνυμα ικανοποιούσε τη δοκιμή, τότε η δράση που οριζόταν από τον κανόνα, εκτελείτω στο μήνυμα. Οι κανόνες δημιουργούνταν μέσω ενός *editor* με την εισαγωγή τους στα πεδία της σχετικής φόρμας. Η μέθοδος αυτή εισαγωγής κανόνων ήταν πολύ χρηστική για άπειρους χρήστες. Το σύστημα ήταν γραμμένο βάση της *LISP*.

Ένας παρόμοιος *agent* με την τεχνική του *Lens*, είναι ο **Postman**[31], που δημιουργήθηκε ως προσωπικός *agent* για το φιλτράρισμα και την ταξινόμηση των μηνυμάτων βάση κανόνων. Οι κανόνες αυτοί πρέπει να τεθούν χειρονακτικά. Ο Postman αν και παρουσιάζει ομοιότητες με τον Lens όσον αφορά την εισαγωγή κανόνων, παρόλα αυτά είναι πιο εύχρηστος καθώς δημιουργήθηκε πάνω στο Pine σύστημα ηλεκτρονικών μηνυμάτων και παράλληλα, είναι γραμμένος βάση της Java.

[30] Tom W. Malone.; Grant, K. R.; Turbak, F. A.; Brobst, S. A.; and Cohien, M. D. Intelligent information-sharing systems. Communications of the ACM 30:390-402, 1987.

[31] Jiang Chen and Haiwei Ye , **POSTMAN - An EMAIL Filtering Agen**t, University of Montreal, 1999.

Ο **Maxim** [13] είναι ένας από τους πρώτους *agents* που δημιουργήθηκαν για την αυτοματοποίηση της διαχείρισης των ηλεκτρονικών μηνυμάτων. Το σύστημα βασιζόταν στην εισαγωγή κανόνων βάση της αποθήκευσης ζευγών κατάστασης-δράσης. Τα ζευγάρια αυτά δημιουργούνταν με την παρακολούθηση των ενεργειών του χρήστη. Ο *Maxim* όταν παρατηρούσε μία κατάσταση παρόμοια μία που είχε ξανασυναντήσει, συνιστούσε την κατάλληλη δράση. Τα χαρακτηριστικά που χρησιμοποιούνταν για την κωδικοποίηση του νοήματος μίας δράσης, περιελάμβαναν τον παραλήπτη και τον αποστολέα του μηνύματος, τις λέξεις-κλειδιά από το «θέμα», την κατάσταση του μηνύματος και τέλος αν υπήρξε απάντηση προς τον αποστολέα. Το σύστημα, παράλληλα, χρησιμοποιούσε δύο όρια: το πρώτο όριο εμπιστοσύνης χρησιμοποιούταν προκειμένου να καθορίσει αν έπρεπε να κάνει πρόταση προς τον χρήστη ή όχι, ενώ το δεύτερο όριο επέτρεπε στον χρήστη να υποδείξει στον *agent* αν θα έπρεπε να αυτοματοποιήσει τη διαδικασία εκτέλεσης της πράξης χωρίς προηγούμενη επιβεβαίωση.

Το σύστημα λειτουργούσε στον client ηλεκτρονικών μηνυμάτων, **Eudora**, παρέχοντας τη δυνατότητα οργάνωσης και κατηγοριοποίησης των μηνυμάτων. Η μέθοδος για την πρόβλεψη των κινήσεων του χρήστη, βασιζόταν στην **MBR**[32]. Η λειτουργία της *MBR* βασίζεται στην αναζήτηση ταυτίσεων μεταξύ των πρόσφατων και των περασμένων καταστάσεων. Στην περίπτωση που μία τέτοια ταύτιση πραγματοποιηθεί, στην συνέχεια προβλέπει την δράση του χρήστη. Ο *Maxims* εκτελεί απευθείας την πράξη ή παρέχει στον χρήστη συντομεύσεις πλήκτρων για την διευκόλυνση του χρήστη.

Τα μειονεκτήματα που παρουσιάζονται στην δράση του *Maxims* σχετίζονται με τον αλγόριθμο που χρησιμοποιείται, καθώς απαιτείται χρόνος για την απόκτηση εμπειρίας που θα τον καταστήσει έμπιστο. Στο σύστημα υιοθετείται η συνεργατική προσέγγιση μεταξύ *agent*. Σύμφωνα, με τον Segal [] μία τέτοια προσέγγιση δεν είναι αποτελεσματική καθώς θεωρεί αμφίβολη την εκπαιδευτική διαδικασία μεταξύ *agent*s. Παράλληλα, ο *Maxims* δεν μπορεί να μάθει από μηνύματα, που έχουν ήδη κατηγοριοποιηθεί, δεδομένου ότι η λογική του βασίζεται στην πρόβλεψη της γενικότερης έννοιας της δράσης, αντί στην κατανόηση της πιο απλής έννοιας, του φακέλου προορισμού. Τέλος, ο Segal υποστηρίζει ότι το βασικό μειονέκτημα του

[32] Memory-based Reasoning

Maxims είναι ότι ο αλγόριθμος που χρησιμοποιεί δεν υποστηρίζει το *incremental learning*.

Ο **MailCat** [15] είναι ένας από τους *agent* που έχει αναπτυχθεί ως intelligent assistant για την οργάνωση των ηλεκτρονικών μηνυμάτων. Ο *MailCat* χρησιμοποιώντας έναν ταξινομητή κειμένου που μεταλλάσσεται ανάλογα με τις συνήθειες του χρήστη, προβλέπει τους τρεις πιο πιθανά κατάλληλους φακέλους για την κατηγοριοποίηση του συγκεκριμένου μηνύματος, ενώ παράλληλα, παρέχει συντομεύσεις πλήκτρων προκειμένου να διευκολύνει τον χρήστη στην εισαγωγή του μηνύματος. Στην περίπτωση που ένας από τους φακέλους που προτάθηκαν από το *MailCat* είναι σωστός, ο χρήστης χρειάζεται μόνο να τον επιβεβαιώσει με το ποντίκι, προκειμένου το μήνυμα να μεταφερθεί στον κατάλληλο φάκελο. Παράλληλα, στην περίπτωση λάθους πρόβλεψης, ο *MailCat* δημιουργήθηκε με τέτοια λογική, ώστε να μην υπάρχει αρνητική επίδραση στον χρήστη, ενώ να είναι παράλληλα δυνατή η παράβλεψη των προτάσεών του.

Ουσιαστικά, ο *MailCat* σε αντίθεση με άλλους *agent*s δεν προσπαθεί να προβλέψει τις πράξεις του χρήστη, απλά να τον διευκολύνει στην επιλογή των φακέλων. Σύμφωνα, με τους δημιουργούς του, η πρόβλεψη των πράξεων του χρήστη, δηλαδή η πράξη εκμάθησης είναι πιο δύσκολη *καθώς το σύνολο των πιθανών ενεργειών περιλαμβάνει περισσότερες των μία μετακινήσεων του δεδομένου μηνύματος από τον φάκελο «εισερχόμενα» σε ένα άλλο φάκελο και παράλληλα, τα δεδομένα που απαιτούνται για την εκμάθηση τέτοιων ενεργειών είναι δύσκολά να αποκτηθούν.*

Ο *MailCat* δημιουργήθηκε ως add-on, δηλαδή ως επιπρόσθετο προϊόν στον client ηλεκτρονικών μηνυμάτων του Lotus Notes. Η επιλογή αυτή έγινε λόγω της διευκόλυνσης που παρέχει ο Notes για τη δημιουργία επιπρόσθετων συστατικών μέσω του C++ API που περιέχει. Η διεπιφάνεια χρήστη του *MailCat* δημιουργήθηκε τροποποιώντας τη βασική φόρμα σχεδιασμού του *Notes*, προκειμένου να συμπεριληφθούν τα τρία επιπλέον κουμπιά που χρησιμοποιούνται για τη μετακίνηση των φακέλων.

Για την εισαγωγή του ταξινομητή δημιουργήθηκαν δύο ειδικά προγράμματα, που ονομάζονται *daemons* και οι οποίοι είναι ικανοί να «διαβάζουν» και να «γράφουν» την βάση δεδομένων ηλεκτρονικών μηνυμάτων του *Notes*. Ο πρώτος *daemon* είναι υπεύθυνος για την ταξινόμηση των νέων μηνυμάτων, ελέγχοντας κάθε εξήντα δευτερόλεπτα για την εμφάνιση νέων μηνυμάτων. Όταν, ο *daemon* βρίσκει νέα μηνύματα, τα ταξινομεί χρησιμοποιώντας την προηγούμενη γνώση του *MailCat*, και παράλληλα, εισάγει τις κατάλληλες συντομεύσεις πλήκτρων στο μήνυμα. Σύμφωνα με τους δημιουργούς ο λόγος για τον οποίο το *MailCat* δεν κατηγοριοποιεί τα μηνύματα από μόνο αλλά κάνει χρήση του daemon είναι καθαρά για λόγους ταχύτητας της κατηγοριοποίηση. Η χρήση, όμως του daemon έχει αρνητικές επιπτώσεις στην απόδοση του ταξινομητή, καθώς μειώνει την ακρίβειά του. Τα μηνύματα ταξινομούνται στο *MailCat* μόνο κατά την εισαγωγή τους. Παρόλα αυτά, ο *MailCat* συνεχώς αναβαθμίζει τον ταξινομητή του κατά την ένταξη των μηνυμάτων.

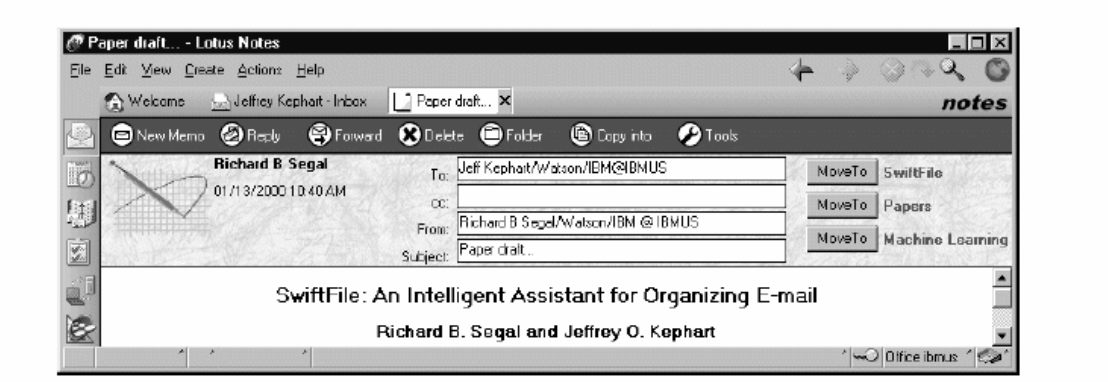
Ο ταξινομητής που χρησιμοποιείται από το MailCat, είναι ο AIM[33] που βασίζεται στο μοντέλο tf/Idf (term frequency/ inverse document frequency). Η επιλογή του σχετικού μοντέλου δημιουργίας ταξινομητή έγινε βάση της επιθυμίας για παροχή υποστήριξης στο incremental learning, δηλαδή στην αέναη εκπαίδευση. Ο ταξινομητής απαιτεί την προηγούμενη εκπαίδευση σε ένα training set, που συνήθως είναι τα εκ των προτέρων ταξινομημένα μηνύματα στο mailbox του χρήστη. Ο δεύτερος *daemon* χρησιμοποιείται για την υποστήριξη του incremental learning. Κάθε φορά που ανιχνεύεται η μετακίνηση ή η διαγραφή ενός μηνύματος, ο daemon ανανεώνει τον ταξινομητή ανάλογα.

Σύμφωνα, με πειράματα που διεξήχθησαν για την διερεύνηση της αποτελεσματικότητας του *MailCat*, αποδείχτηκε ότι προσφέρει στους χρήστες τους, ένα ποσοστό ακρίβειας ίσο με την τάξη του 80-90% για έναν χρήστη που έχει μέχρι και εξήντα φακέλους. Ουσιαστικά, η ακρίβεια του *MailCat* δεν ξεπερνά το 60-80% αλλά δεδομένου ότι παρέχει στον χρήστη τη δυνατότητα επιλογής τριών πιθανών φακέλων κατηγοριοποίησης, το ποσοστό ακρίβειας αυξάνεται. Παράλληλα, το ποσοστό ανάκλησης είναι 20-40% που, όμως έχει μικρή σημασία

[33] Barret & Selker 1995

καθώς η πραγματοποίηση λαθών στην κατηγοριοποίηση, μπορούν εύκολα να υπερκαλυφθούν.

Παρόμοια με τη γενική λειτουργία του *MailCat,* είναι η λειτουργία του **SwiftFile** [16] **.** Ο *SwiftFile* είναι ένας έξυπνος βοηθός για την υποστήριξη της οργάνωσης των ηλεκτρονικών μηνυμάτων του χρήστη και αναπτύχθηκε ως επιπρόσθετο συστατικό για τον *Notes*. Χρησιμοποιεί έναν ταξινομητή κειμένου, προκειμένου να προβλέψει την κατηγορία υπό την οποία θα ενταχθεί το συγκεκριμένο μήνυμα και παράλληλα, παρέχει συντομεύσεις πλήκτρων προς τις αντίστοιχες κατηγορίες/ φακέλους για τη διευκόλυνση του χρήστη. Όπως και το *MailCat*, χρησιμοποιεί ως ταξινομητή τον *AIM*, με σκοπό να εξασφαλίσει την συνεχή ανανέωση και εκπαίδευση του ταξινομητή.



Εικόνα 5- διεπιφάνεια χρήστη του SwiftFile.

Η ακρίβεια του *SwiftFile* ανέρχεται στο 73-90%, δεδομένου ότι παρέχει στον χρήστη την επιλογή από τρεις πιθανούς φακέλους. Σύμφωνα, με την έρευνα που πραγματοποιήθηκε κατά τα πειράματα για την αξιολόγησή του, η αποδοτικότητά του θα έπεφτε στο 52-76% αν το *SwiftFile* πραγματοποιούσε απευθείας κατηγοριοποίηση, Παράλληλα, το επίπεδο λάθους του, ανέρχεται στο 10-27%, το οποίο είναι μικρότερο από το επίπεδο λάθους του *MailCat,* ενώ αντίστροφα το *MailCat* παρουσιάζει υψηλότερο ποσοστό ακρίβειας. Θα πρέπει να αναφερθεί, ότι σύμφωνα με τους δημιουργούς του *SwiftFile* ο ταξινομητής βάση του *tf/idf,* αρχίζει να αποδίδει αφού εκπαιδευτεί πάνω σε εκατό μηνύματα.

Μία άλλη προσέγγιση στους *agent*s, είναι αυτή που ακολουθείται για τη δημιουργία του **Re:*agent*** [3] , ο οποίος είναι ικανός να φιλτράρει, να ιεραρχήσει

τα μηνύματα βάση μείζονος σημασίας, να τα φορτώσει σε palmtop, αλλά και να τα προωθήσει σε voicemail. Η τεχνική που χρησιμοποιείται για την εκτέλεση των παραπάνω λειτουργιών, βασίζεται στην εξαγωγή χαρακτηριστικών από τα ηλεκτρονικά μηνύματα βάση Εκμάθηση Μηχανής τεχνικών.

Τα χαρακτηριστικά αυτά, δεν είναι μεμονωμένες λέξεις από το σώμα των μηνυμάτων αλλά αντίθετα χρησιμοποιούνται ολόκληρα τα ηλεκτρονικά μηνύματα προκειμένου να καθοριστούν υψηλού επιπέδου χαρακτηριστικά. Τα χαρακτηριστικά αναπαριστούν υπάρχουσες έννοιες στο σώμα των ηλεκτρονικών μηνυμάτων και για αυτό ονομάζονται *concept features*. Για να εξαχθούν αυτά τα χαρακτηριστικά, αρχικά πραγματοποιείται ομαδοποίηση παρόμοιων μηνυμάτων και εκ ων υστέρων, συνδυασμός των κοινών πληροφοριών σε διάνυσμα. Με αυτό τον τρόπο, η εξαγωγή των χαρακτηριστικών ανάγεται σε διαδικασία προσδιορισμού της παρουσίας υψηλών χαρακτηριστικών στο πλαίσιο των μηνυμάτων.

Σημαντικό είναι να σημειωθεί ότι ο τύπος των χαρακτηριστικών δεν είναι προκαθορισμένος αλλά ορίζεται αυτόματα από τον τύπο των δεδομένων ή τον τύπο της εργασίας. Παράλληλα, τα στοιχεία των χαρακτηριστικών είναι στατιστικές περιλήψεις των δεδομένων που χρησιμοποιούνται για την εκπαίδευση του ταξινομητή και μπορούν να περιλαμβάνουν π.χ. ποσοστά της παρουσίας λέξεων –κλειδιών. Τέλος, υποστηρίζεται ότι η χρήση χαρακτηριστικών υψηλού επιπέδου επιλύει το πρόβλημα της high-dimensionality, καθώς ο ταξινομητής χρειάζεται να μάθει μόνο λίγα σχετικά χαρακτηριστικά.

Πάνω στην εξαγωγή των χαρακτηριστικών στηρίζεται η όλη λογική της αρχιτεκτονικής του *Re:agent*. Σε αυτά, εφαρμόζονται αλγόριθμοι βάση της *Εκμάθηση Μηχανής* τεχνικής. Για την κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων, ο *Re:agent*, αρχικά, δημιουργεί ένα διάνυσμα που μετράει την παρουσία κάθε χαρακτηριστικού στα εισερχόμενα μηνύματα. Κάθε ένα από αυτά, προσθέτει ένα ή δύο στοιχεία στο διάνυσμα χαρακτηριστικών, το οποίο χρησιμοποιείται από τον αλγόριθμο εκμάθησης πράξης για την κατηγοριοποίηση του μηνύματος.

Στα πειράματα που διεξήχθησαν τα ηλεκτρονικά μηνύματα αναπαραστάθηκαν ως διανύσματα βάση της tf/idf μεθόδου και στο σύνολο των μηνυμάτων που χρησιμοποιήθηκαν, εφαρμόστηκαν αλγόριθμοι βάση των *neural networks* και του *k-nearest neighbor*. Αποδείχτηκε, χωρίς να δίνονται στοιχεία για την συνολική απόδοση του *agent*, ότι ο *κ-nearest neighbor* υπερτερούσε συγκριτικά με τα *neural network* σε ποσοστό 96% προς το 82% που υπέδειξε ο τελευταίος.

Ο **Cool *Agent* Personal Assistant** [2] αναπτύχθηκε στα εργαστήρια λογισμικού της **Hewlett Packard** ως συστατικό του **PIA** *(Personal Information Assistant)*, ενός συστήματος που αποσκοπεί στην ευρετηρίαση και διαχείριση του περιεχομένου από διάφορες πηγές διάχυσης πληροφοριών. Ο *Cool Agent Personal Assistant* αποσκοπεί στην ενεργή διαχείριση προσωπικών, ομαδικών και οργανωτικών πληροφοριών. Βασικό συστατικό του είναι ο **PEA** (*Personal Email Assistant*), ο οποίος παρέχει ένα παραμετροποιήσιμο περιβάλλον βάση Εκμάθηση Μηχανής τεχνικών, για την υποστήριξη δραστηριοτήτων διαχείρισης ηλεκτρονικών μηνυμάτων. Ένα από τα βασικά πλεονεκτήματα του είναι ότι σχεδιάστηκε να λειτουργεί είτε με είτε χωρίς υποδομή για *agent* και ότι είναι συμβατό με διάφορα συστήματα ηλεκτρονικών μηνυμάτων, όπως είναι το *Exchange* και το *Outlook*.

Ο *PEA* υποστηρίζει την ιεράρχηση μηνυμάτων βάση προτεραιότητας, την κατηγοριοποίηση μηνυμάτων και το φιλτράρισμα ασήμαντων πληροφοριών-χρησιμοποιώντας τεχνικές ταξινόμησης και κανόνες-, ενώ, παράλληλα, διατηρεί έναν ανεστραμμένο ευρετήριο όλων των πρόσφατων και αρχειοθετημένων μηνυμάτων, όπου ο χρήστης μπορεί να πραγματοποιήσει αναζήτηση. Τέλος, παρέχει τη δυνατότητα vacation response, όπου ο Vacation *agent* χρησιμοποιεί τους *agent* contact και calendar για να απαντήσει σε εισερχόμενα μηνύματα.

Μελλοντικές εργασίες που προβλέπονται να εισαχθούν στις ήδη υπάρχουσες του PEA περιλαμβάνουν: τη δυνατότητα δημιουργίας περιλήψεων, κυρίως όταν πολλά μηνύματα εισέρχονται πάνω σε ένα ζήτημα, τη δυνατότητα χρήσης του ηλεκτρονικού μηνύματος ως μηχανισμού που θα πυροδοτήσει κάποια άλλη ενέργεια, τη δυνατότητα αυτόματης απάντησης μέσω προ-καθορισμένων φορμών και τέλος τη δυνατότητα παρακολούθησης των εισερχόμενων μηνυμάτων, για

μηνύματα που αναφέρονται στον προγραμματισμό συναντήσεων (meeting minder).

Η αρχιτεκτονική του *PEA* βασίζεται στη δημιουργία διάφορων συστατικών για τη διαχείριση των ηλεκτρονικών μηνυμάτων, τα οποία μπορούν να λειτουργήσουν ξεχωριστά ή στο πλαίσιο του *agent*. Η δομή τους έχει τη μορφή ενός δέντρου-ιεραρχία κλάσεων, το οποίο μπορεί εύκολα να τροποποιηθεί με την εισαγωγή νέων συστατικών, όπως είναι τα φίλτρα, οι κανόνες και οι ταξινομητές.

Τα βασικά συστατικά διαχείρισης για τα ηλεκτρονικά μηνύματα είναι τα φίλτρα, που ονομάζονται *EmailHandlers*/*email filters*. Ουσιαστικά, αποτελούν μία υποκλάση στην δενδροειδή αναπαράσταση. Για κάθε νέο μήνυμα, οι handlers που μπορούν να χρησιμοποιηθούν είναι: *contacts, classify* και *index.* Ο handler *contact* αποσκοπεί στο να επικοινωνήσει μη τη Βάση Δεδομένων και να εισάγει νέες ιδιότητες προκειμένου να διαχωρίσει και να διαχειριστεί επαφές. Ο *classify* χρησιμοποιεί τις ιδιότητες των μηνυμάτων και τις προσδιορισμένες επαφές ως δεδομένα εισόδου στον ταξινομητή, ρυθμίζοντας με αυτό τον τρόπο τις επιπλέον ρυθμίσεις που θα θέσουν σε λειτουργία διαδοχικούς κανόνες ή EmailHandlers. Ο *Index* εξάγει λέξεις-κλειδιά από το μήνυμα, ενώ παράλληλα διατηρεί ένα σύνολο από ανεστραμμένα ευρετήρια. Παράλληλα, υπάρχουν και άλλοι handlers που σχετίζονται με την μετακίνηση, διαγραφεί και ιεράρχηση των ηλεκτρονικών μηνυμάτων ή φακέλων.

Ο ταξινομητής που χρησιμοποιείται από τον *PEA*, είναι ο **WECA**. Ο *WECA*[34], ουσιαστικά είναι ένα open source λογισμικό, που είναι γραμμένο σε Java και «τρέχει» σε οποιαδήποτε πλατφόρμα, ενώ υιοθετεί και αλγόριθμους βάση της Εκμάθηση Μηχανής τεχνικής για τη διαχείριση data mining προβλημάτων. Ένας ταξινομητής που υιοθετείται από το *WECA* είναι ο *LIBSVM,* βάση του μοντέλου Support Vector Machine. Η κατηγοριοποίηση στον *PEA,* πραγματοποιείται μέσα από ένα συνδυασμό τεχνικών Εκμάθηση Μηχανής και εισαγωγής κανόνων (Java/JESS/XML).

[34] Waikato Environment for Knowledge Analysis

Η κατηγοριοποίηση πραγματοποιείται μέσω της εξής διαδικασίας: Κάθε ταξινομητής θέτει μία ιδιότητα *true* ή *false (*Boolean classifier*),* στην *EmailEvent* (υπό-κλάση στην δενδροειδή αναπαράσταση των συστατικών) ή κάποιο ποσοστό ή επίπεδο εμπιστοσύνης (numeric classifier). Δίνοντας τη δυνατότητα, με αυτό τον τρόπο, σε άλλους *EmailHandlers*να έχουν πρόσβαση στις ταξινομήσεις. Στην συνέχεια, εφαρμόζονται κανόνες στις προσδιορισμένες τιμές και ιδιότητες, συμπεριλαμβανομένων των ιδιοτήτων που δημιουργήθηκαν από τον ταξινομητή. Στην περίπτωση που υπάρχει ταύτιση κανόνα, ταυτόχρονα θα τεθούν σε λειτουργία οι αντίστοιχες *Actions* στον *EmailHandler* και με αυτό τον τρόπο θα πραγματοποιηθεί η κατηγοριοποίηση.

Από τους *agent* που αναφέρθηκαν παραπάνω την καλύτερη απόδοση έχει ο *Re:agent* με χρήστη των *neural network*, καθώς η ακρίβειά του προσεγγίζει το 96,9%, ενώ το μέγιστο της απόδοσης του *MailCat αγγίζει το 80% .* Ο τελευταίος παρουσιάζει το προτέρημα της γρήγορης εκτέλεσης της κατηγοριοποίησης στα 0,3 δευτερόλεπτα. Παράλληλα, ο *PEA* παρουσιάζει το πλεονέκτημα της εύκολης εισαγωγής του σε πολλά συστήματα ηλεκτρονικών μηνυμάτων, ενώ αντίθετα οι περισσότεροι από τους προαναφερθέντες *agent*s δημιουργήθηκαν ως επιπρόσθετα συστατικά στα σχετικά συστήματα.

2.2.2 Εργαλεία Κατηγοριοποίησης

Η κατηγοριοποίηση ηλεκτρονικών μηνυμάτων μπορεί να πραγματοποιηθεί από διαφόρων ειδών εργαλεία. Τα εργαλεία αυτά μπορεί να δρουν μεμονωμένα, ανεξάρτητα, δηλαδή, από κάποιο σύστημα αλλά παράλληλα, μπορεί να αποτελούν υποσύστημα ενός γενικότερου συστήματος, με τη μορφή υπηρεσίας. Τα συστήματα αυτά, δεδομένης της ανάγκης για διαχείριση των πληροφοριών, εμφανίζονται σε μία πληθώρα κατηγοριών. Αναφορικά, οι κατηγορίες στις οποίες υποκύπτουν τα συστήματα αυτά, περιλαμβάνουν: *μηχανές αναζήτησης*, *πύλης*, *knowledge management* (KM), *content management* (CM), *content record management* (CRM), *electronic document record management* (EDRM) κα. Κάποια, από τα συστήματα, που μόλις αναφέρθηκαν διαχειρίζονται τόσο ηλεκτρονικά τεκμήρια- συμπεριλαμβάνοντας τα μηνύματα- όσο και τεκμήρια της

παραδοσιακής μορφής, που παράγονται στα πλαίσια ενός οργανισμού ή μίας εταιρίας και σχετίζονται με τη μεταφορά πληροφοριών (π.χ. Fax).

Επίσης, με την αύξηση της χρήσης των ηλεκτρονικών μηνυμάτων και τις σπουδαιότητας των πληροφοριών που μεταφέρουν αναπτύχθηκαν ειδικά συστήματα αρχειοθέτησης, που παρέχουν στους χρήστες τη δυνατότητα διαχείρισης του κύκλου ζωής των μηνυμάτων, ενώ έκαναν την εμφάνιση τους και συστήματα αυτόματης απάντησης, γνωστά ως responders, τα οποία αυτοματοποιούν τη διαδικασία απάντησης των ηλεκτρονικών μηνυμάτων μέσω φορμών. Η σημασία της κατηγοριοποίησης σε όλα αυτά τα συστήματα είναι εμφανής: *για την καλύτερη εξυπηρέτηση των χρηστών, απαιτείται οργάνωση, η οποία παρέχεται κυρίως με την ταξινόμηση των μηνυμάτων υπό κατηγορίες, είτε είναι εικονικοί φάκελοι είτε ταξινομία ή file plan*.

Παράλληλα, μία άλλη διάκριση που θα μπορούσε να πραγματοποιηθεί στα εργαλεία κατηγοριοποίησης είναι βάση του χρόνου που πραγματοποιείται η διαδικασία. Η κατηγοριοποίηση, δηλαδή, μπορεί να πραγματοποιηθεί πριν την εισαγωγή των μηνυμάτων στο σύστημα του χρήστη ή του οργανισμού ή παράλληλα, μετά την εισαγωγή τους και αποθήκευσή τους σε τοπικούς ή κεντρικούς servers ή σε άλλους αποθηκευτικούς χώρους.

2.2.2.1 Εργαλεία Απευθείας Κατηγοριοποίησης

2.2.2.1.1 PoPfile

Το **PoPfile** [e] είναι ένα εργαλείο κατηγοριοποίησης ηλεκτρονικών μηνυμάτων, που ταξινομεί αυτόματα τα ηλεκτρονικά μηνύματα, ενώ παράλληλα, περιλαμβάνει λειτουργία για το φιλτράρισμα των spam μηνυμάτων. Τα *spam* μηνύματα είναι μία ειδική κατηγορία ηλεκτρονικών μηνυμάτων, τα οποία στέλνονται από έναν αποστολέα σε πολλούς παραλήπτες και συνήθως είναι προσβλητικού, διαφημιστικού αλλά συχνά και επικίνδυνου περιεχομένου, καθώς είναι δυνατό να περιέχουν ιούς, που μπορούν να προσβάλλουν τον υπολογιστή. Λόγω των διαστάσεων που έχει πάρει το φαινόμενο των *spam* μηνυμάτων, έχουν αναπτυχθεί πολλά συστήματα που να κατηγοριοποιούν τα μηνύματα σε spam ή

σε μη-spam. Τα συστήματα, όμως αυτά δεν αποτελούν αντικείμενο της συγκεκριμένης έρευνας.

Το *PoPfile* λειτουργεί ως (PoP 3, SMTP, NNTP, IMAP)[35] *proxy* ανάμεσα στον client και στον server ηλεκτρονικών μηνυμάτων του χρήστη. Όταν, το πρόγραμμα εγκαθίσταται για πρώτη φορά στο σύστημα του χρήστη δεν πραγματοποιεί κανενός είδους κατηγοριοποίηση- για αυτό τον λόγο χαρακτηρίζεται ως dump- περιμένει να μάθει από τις προσωπικές ενέργειες κατηγοριοποίησης του χρήστη και με την πάροδο του χρόνου, γίνεται όλο και πιο λειτουργικό και ανεξάρτητο. Τα μηνύματα καθώς εισάγονται στο σύστημα, σκανάρονται και στην συνέχεια μαρκάρονται ανάλογα με τις ρυθμίσεις του χρήστη. Η κατηγοριοποίηση μπορεί να γίνει βάση *spam* ή *good* αλλά και βάση προσωπικών κριτηρίων, όπως είναι personal, work, important κλπ. Τα κριτήρια βάση των οποίων θα γίνει η κατηγοριοποίηση των μηνυμάτων ονομάζονται *backets* και είναι παρόμοια με τους εικονικούς φακέλους. Τα μηνύματα σε κάθε φάκελο, μαρκάρονται με διαφορετικό τρόπο- βάση του θέματός τους ή της επικεφαλίδας- ώστε ο χρήστης να μπορεί να δώσει εντολές στο πρόγραμμα ηλεκτρονικών μηνυμάτων του, να τα διαχειρίζεται βάση των προτιμήσεών του. Ο ταξινομητής που εφαρμόζεται για την κατανόηση των μηνυμάτων είναι βάση του μοντέλου *Naïve Bayes.* Άλλες λειτουργίες του *PoPfile* περιλαμβάνουν:

- Διαγραφή μηνυμάτων με έγκριση του χρήστη.
- Μπορεί να εκπαιδευτεί να δρα στα spam και να τα διαγράφει ή να τα μεταφέρει σε άλλο φάκελο.
- Παρέχει μία ειδική επιλογή *quarantine*, απομόνωσης, δηλαδή του μηνύματος. Κατά τη διαδικασία αυτή, αποθηκεύει το μήνυμα σε ένα ειδικό χώρο (container), από όπου ο χρήστης μπορεί να πληροφορηθεί για το περιεχόμενο του μηνύματος, χωρίς ουσιαστικά, να το ανοίξει.
- Τέλος, παρέχει ελαστικό και ακριβές φιλτράρισμα των μηνυμάτων αλλά απαιτείται εκ μέρους του χρήστη μία βασική γνώση των ηλεκτρονικών μηνυμάτων και των *POP λ*ογαριασμών, προκειμένου να είναι αποτελεσματικό.

[35] πρωτόκολλα επικοινωνίας και μετάδοσης ηλεκτρονικών μηνυμάτων

Το *PoPfile* αναπτύχθηκε από τον *John Graham Cumming* το 2003 και εμφανίστηκε με την έκδοση *PoPfile 0.19.1*. Διατηρείται ως open source λογισμικό από μία ομάδα χρηστών και ειδικών. Σε γενικές γραμμές, τα μειονεκτήματα του *PoPfile* αφορούν την ανάγκη για χρήση μεγάλου τμήματος της μνήμης και του cpu του υπολογιστή, ενώ λειτουργεί καλύτερα, όταν χρησιμοποιείται μόνο από ένα χρήστη. Η τρέχουσα έκδοσή του είναι *PoPfile v0.22.2,* η δημιουργία της οφείλεται στην εμφάνιση νέων ιών, που το νέο πακέτο λόγω διαφόρων βελτιώσεων είναι ικανό να καταπολεμήσει.

Οι βελτιώσεις αφορούν: τη δημιουργία νέου «packing list» το οποίο θα ελέγχει αυτόματα ότι τα σωστά Perl συστατικά έχουν εγκατασταθεί στο Popfile, προκειμένου αυτό να λειτουργεί σωστά. Περιλαμβάνει, επίσης ανανεώσεις στις πιο πρόσφατες εκδόσεις των Perl (v.5.8.4), PDK (v6.0), SQLite (v2.8.15). Παρέχει, παράλληλα, νέα επιλογή για τη γρήγορη διαγραφή μηνυμάτων από την Βάση Δεδομένων, ενώ θα πραγματοποιείται και back-up στην βάση κάθε μία ώρα. Άλλες βελτιώσεις περιλαμβάνουν, τη δημιουργία νέου bag fix (καταπολέμηση ιού) για την Κορεάτική γλώσσα αλλά και παροχή νέας διεπιφάνειας χρήστη για τα απλοποιημένα και τα παραδοσιακά κινέζικα. Το προφίλ της νέας έκδοσης *v0.22.2,* παρουσιάζεται στον παρακάτω πίνακα:

	PoPfile v0.22.2
supported platforms	Windows 98/ME/2000/XP Mac OS X Linux Unix
Email clients	Microsoft Exchange/Outlook
size	4890 KB

2.2.2.1.2 Nexor Interceptor

Το λογισμικό της ***Nexor,*** ο ***Nexor Interceptor*** αποτελεί παράδειγμα των συστημάτων που παρέχουν αυτόματη δρομολόγηση των ηλεκτρονικών μηνυμάτων στους κατάλληλους χρήστες.

Ο *Nexor Interceptor f[]* βασίζει την λειτουργία του για τη δρομολόγηση και διανομή της ηλεκτρονικής αλληλογραφίας στους χρήστες, στην ικανότητά του να αναλύει αυτόματα το περιεχόμενο των ηλεκτρονικών μηνυμάτων και των συνημμένων αρχείων τους. Για την επίτευξη αυτών των δυνατοτήτων αλλά και για την κατηγοριοποίηση των μηνυμάτων, η *Nexor* εισήγαγε στο σύστημά της, την τεχνολογία *αναγνώρισης μοτίβων (pattern recognition)*, που έχει αναπτυχθεί από την ***Autonomy***[36]***.***

Τα μηνύματα είναι πιθανό να στέλνονται σε μία γενική διεύθυνση π.χ. info@bank.com και η λειτουργία του *Nexor Interceptor* ανάγεται στην ανάλυση του περιεχομένου των ηλεκτρονικών μηνυμάτων, προκειμένου να δρομολογηθούν προς το κατάλληλο τμήμα της σχετικής εταιρίας. Αν, δηλαδή, η εταιρία είναι μία τράπεζα, όπως αναφέρεται και στο παράδειγμα, τότε ο *Nexor Interceptor,* θα πρέπει να το δρομολογήσει κατάλληλα, στο τμήμα δανείων ή υποθηκών, ανάλογα με τις λέξεις, που περιέχονται στο μήνυμα. Παράλληλα, τα κατηγοριοποιεί βάση του θέματός τους.

Γενικά, το σύστημα έχει το ρόλο ενός είδους φρουρού για το περιεχόμενο της εκάστοτε εταιρίας. Αυτό σημαίνει ότι συλλαμβάνει τόσο τα εισερχόμενα όσο και τα εξερχόμενα μηνύματα, ταυτοποιεί τον παραλήπτη και τον αποστολέα τους, ενώ, παράλληλα, αποκρυπτογραφεί και ελέγχει το περιεχόμενο τους. Κατά τον έλεγχο του περιεχομένου, αποτρέπει την είσοδο στο σύστημα μηνυμάτων spam ή γενικά αμφίβολου περιεχομένου ή αποστολέα, ενώ για τα εξερχόμενα μηνύματα, εμποδίζει την αποστολή αυτών που θεωρεί ότι περιλαμβάνουν σημαντικά δεδομένα της εταιρίας.

[36] Περιγραφή της σχετικής τεχνολογίας, πραγματοποιείται σε επόμενο τμήμα της εργασίας με την αναφορά του αντίστοιχου συστήματος της Autonomy.

2.2.2.2 Διαχείριση Ηλεκτρονικών Μηνυμάτων Σε Πραγματικό Χρόνο

Στο σημείο αυτό, αναφέρονται εργαλεία διαχείρισης ηλεκτρονικών μηνυμάτων, που εκτελούν οποιαδήποτε εργασία που σχετίζεται με τη διαχείριση των μηνυμάτων, κατά την αποθήκευσή τους στον κεντρικό ή τοπικό server του χρήστη ή του οργανισμού.

2.2.2.2.1 Interwoven

Η ***Interwoven*** [g] είναι μία από τις εταιρίες που προσανατολίζεται στη διαχείριση των ηλεκτρονικών μηνυμάτων με μία σειρά από εργαλεία για το *Microsoft Outlook* και για τον *Lotus Notes*. Η λύση της *Interwoven* βασίζεται στην παροχή δύο βασικών client, του ***DeskSite*** και του ***Worksite***, που ανάγουν το πρόβλημα της διαχείρισης σε ένα σύστημα διαχείρισης τεκμηρίων (RMS). Η προσέγγιση που πραγματοποιείται βασίζεται στην σύλληψη των μηνυμάτων βάση περιεχομένου και στη συγχώνευσή τους με άλλα σχετικά τεκμήρια σε ένα κοινό σύστημα διαχείρισης. Οι *client* αυτοί δεν διαφοροποιούνται μόνο μεταξύ τους, όσον αφορά, δηλαδή τις υπηρεσίες που προσφέρουν αλλά και ανάλογα με το γενικότερο πρόγραμμα προσφοράς υπηρεσιών για ηλεκτρονικά μηνύματα στο οποίο εισάγονται, δηλαδή, στο Outlook ή στο Lotus Notes.

Σε γενικές γραμμές, οι *client* της *Interwoven* παρέχουν τις βασικές υπηρεσίες ενός προγράμματος διαχείρισης τεκμηρίων, όπως είναι η δημιουργία νέου εγγράφου, το άνοιγμα, η αποθήκευση κ.α. Παράλληλα, μπορούν να εργαστούν με προγράμματα, όπως το *Microsoft Office, Novel GroupWise και WordPerfe*ct. Όσον αφορά, τη λειτουργία τους με το *Outlook* και το *Lotus Notes*, παρέχουν τη δυνατότητα **drag and drop** των ηλεκτρονικών μηνυμάτων από τη διεπιφάνεια των προγραμμάτων στη δική τους διεπιφάνεια χρήστη. Μέσω της διεπιφάνειας των δύο *clients*, ο χρήστης μπορεί να επεξεργαστεί τα ηλεκτρονικά μηνύματα ως απλά τεκμήρια, επιτρέποντας την οργάνωσή τους, την αποθήκευσή τους και την επαναχρησιμοποίηση τους. Ένα βασικό στοιχείο στην αρχιτεκτονική τους είναι ο

MetaTagger Content Intelligence Server [] οποίος λειτουργεί ως η βάση για μία σειρά εργασιών. Στα πλαίσια των εργασιών αυτών περιλαμβάνεται η εισαγωγή μεταδεδομένων στο περιεχόμενο όλων των τεκμηρίων του οργανισμού, συμπεριλαμβανομένων των ηλεκτρονικών μηνυμάτων. Η μηχανή περιεχομένου που δρα στα πλαίσια του server κατηγοριοποιεί αυτόματα το περιεχόμενο των τεκμηρίων σε μία ή περισσότερες ταξινομίες. Παράλληλα, ο *server* παράγει περιλήψεις και λέξεις κλειδιά, εντοπίζει υψηλού επιπέδου νοήματα από το περιεχόμενο των τεκμηρίων, ενώ τέλος εξάγει διάφορους τύπους δεδομένων, όπως ημερομηνίες, ονόματα πελατών, εταιριών και προϊόντων.

Η προσέγγιση της *Interwoven* για τη διαχείριση περιεχομένου στο *MetaTagger* βασίζεται στη μέθοδο *content intelligence*, δηλαδή, *έξυπνο περιεχόμενο.* Προκειμένου το περιεχόμενο να χαρακτηριστεί έξυπνο, θα πρέπει να εισαχθούν σε αυτό μεταδεδομένα. Προκειμένου να υπάρχει συνοχή και ακρίβεια στον προσδιορισμό των μεταδεδομένων, γίνεται χρήση ελεγχόμενων λεξιλογίων. Ένα ελεγχόμενο λεξιλόγιο είναι ένα καθιερωμένο σύνολο λέξεων, φράσεων ή και κωδικών στη μορφή επίπεδης ή ιεραρχικής ταξινομίας. Τα στοιχεία αυτά (λέξεις, φράσεις, κωδικοί) μπορούν να χρησιμοποιηθούν για να προσδιορίσουν κάθε τμήμα του περιεχομένου ενός τεκμηρίου και συνεπώς να χρησιμοποιηθούν ως μεταδεδομένα. Σημαντικό είναι ότι κάθε εφαρμογή μπορεί να χρησιμοποιεί το δικό της ελεγχόμενο λεξιλόγιο και συνεπώς να έχει τα δικά της θεματικά πεδία μεταδεδομένων. Για την εφαρμογή, δηλαδή, που διαχειρίζεται τα ηλεκτρονικά τεκμήρια, μπορούν να οριστούν διαφορετικά πεδία, που να περιλαμβάνουν τα πεδία της επικεφαλίδας, αποστολέας ή ημερομηνία, καθώς, επίσης, να οριστούν στοιχεία και για το σώμα του ηλεκτρονικού μηνύματος.

Δύο βασικές λειτουργίες για τις οποίες γίνεται η εισαγωγή μεταδεδομένων στο περιεχόμενο είναι η *αναγνώριση* και η *κατηγοριοποίηση*. Η αναγνώριση περιλαμβάνει τη διαδικασία σκαναρίσματος του περιεχομένου των τεκμηρίων ή μηνυμάτων προκειμένου να εντοπιστούν σε αυτό σημαντικές οντότητες, όπως είναι εταιρίες, πρόσωπα, προϊόντα. Ο *MetaTagger* παρέχει αλφαριθμητικούς προσδιοριστές (identifiers) για τον εντοπισμό θεμάτων και κυρίων ονομάτων.

Η κατηγοριοποίηση στο *MetaTagger* πραγματοποιείται αυτόματα χρησιμοποιώντας τη μέθοδο pattern recognition και ελεγχόμενα λεξιλόγια. Αρχικά, χρησιμοποιεί

ένα σύνολο προκατηγοριοποιημένων τεκμηρίων (ή μηνυμάτων στην περίπτωση των ηλεκτρονικών μηνυμάτων) για να μάθει τα μοτίβα των λέξεων και φράσεων που αναπαριστούν τις διάφορες κατηγορίες . Μετά, τη φάση της εκπαίδευσης, ο *MetaTagger* μπορεί να αναλύσει οποιαδήποτε κείμενο και να καθορίσει την κατάλληλη κατηγορία ή κατηγορίες για αυτό.

Η τρέχουσα έκδοση του *MetaTagger* είναι η 3.5 του 2003. Τα βασικά νέα χαρακτηριστικά που παρουσιάζει είναι: η ανεξαρτησία από τον αποθηκευτικό χώρο των δεδομένων, ανεπτυγμένη διαχείριση ταξινομιών και υποστήριξη μεγαλύτερου τύπου διατάξεων δεδομένων.

2.2.2.2.2 Mobius

Μία άλλη εταιρία που εισάγει τη διαχείριση των ηλεκτρονικών μηνυμάτων στο γενικότερο προϊόν της για τη διαχείριση του περιεχομένου είναι η ***Mobius*** με το ***ViewDirect TCM*** (**T***otal* **C***ontent* **M***anagement*) [h] **.** Το ***ViewDirect E-mail management*** είναι η υπηρεσία που διαχειρίζεται τα ηλεκτρονικά μηνύματα στο πλαίσιο του γενικότερου συστήματος διαχείρισης περιεχομένου.

Το *ViewDirect E-mail management* παρέχει στον χρήστη ένα πλήρες πακέτο υπηρεσιών διαχείρισης των ηλεκτρονικών μηνυμάτων, συμπεριλαμβάνοντας τις δυνατότητες σύλληψης, οργάνωσης, κατηγοριοποίησης, αποθήκευσης καθώς και διατήρησης και διάθεσης. Η σύλληψη των μηνυμάτων γίνεται από τα ηλεκτρονικά γραμματοκιβώτια (mailboxes) των χρηστών και από τον Microsoft Exchange, ενώ στην συνέχεια οργανώνει τα μηνύματα σύμφωνα με την ταξινομική δομή που έχει οριστεί από το οργανισμό. Η συλλογή των μηνυμάτων μέσω του *ViewDirect* συνεπάγεται τη διαγραφή των μηνυμάτων από τους τοπικούς αποθηκευτικούς χώρους των χρηστών. Το βήμα αυτό πραγματοποιείται προκειμένου να γίνει αποσυμφόρηση του κεντρικού server του οργανισμού από το βάρος της αποθήκευσης των ηλεκτρονικών μηνυμάτων, καθιστώντας τον με αυτό τον τρόπο ικανό να λειτουργεί καλύτερα. Κατά τη μεταφορά τους, τα ηλεκτρονικά μηνύματα αποθηκεύονται σε ένα on-line αποθηκευτικό χώρο, ενώ παράλληλα, διατηρούν την αρχική τους διάταξη. Τα μηνύματα που μεταφέρονται αντικαθιστούνται στη διεπιφάνεια του *Outlook* με ένα σημάδι. Τέλος, στα μηνύματα αυτά, εφαρμόζονται αρχές διατήρησης και διάθεσης που η περίοδος τους ορίζεται από τον οργανισμό, ενώ, παράλληλα, παρέχεται η δυνατότητα ασφαλούς πρόσβασης.

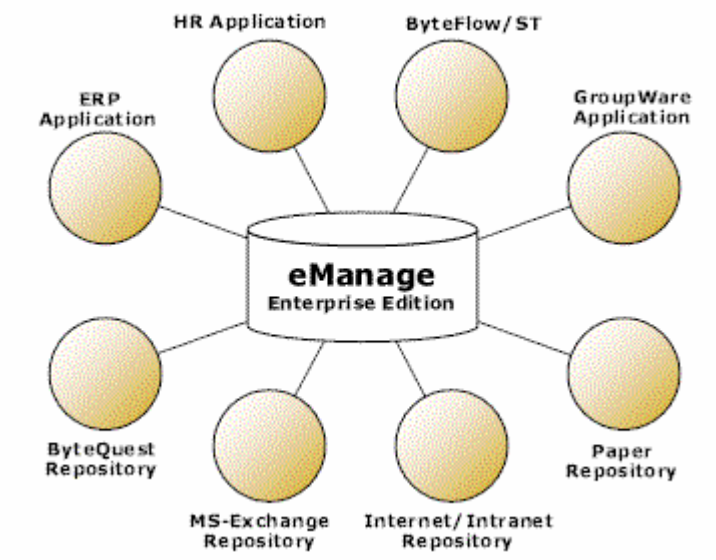
Η οργάνωση των ηλεκτρονικών μηνυμάτων στο *ViewDirect* πραγματοποιείται σε μία ταξινομική δομή, όπως αυτή έχει οριστεί από τους φακέλους των χρηστών ή βάση μίας χρονολογικής δομής ή τέλος βάση μίας υβριδικής προσέγγισης των δύο προαναφερθέντων μεθόδων. Εναλλακτικά, η οργάνωση της δομής ταξινόμησης των ηλεκτρονικών μηνυμάτων μπορεί να αντικατοπτρίζει τις λειτουργίες του οργανισμού (τμήματα, υπηρεσίες, γραφεία κλπ). Η κατηγοριοποίηση των μηνυμάτων βασίζεται στην εξαγωγή μεταδεδομένων. Τα μεταδεδομένα μπορεί να περιλαμβάνουν τα πεδία της επικεφαλίδας, όπως αποστολέας, ημερομηνία και θέμα αλλά μπορούν, παράλληλα να οριστούν και άλλα, ώστε τα μηνύματα να κατηγοριοποιούνται βάση του περιεχομένου. Σημαντικό είναι ότι ο χρήστης μπορεί να ορίσει ποια από τα μηνύματα θέλει να κατηγοριοποιηθούν και να εξαλείψει αυτά τα οποία δεν περιέχουν κάποια πληροφοριακή αξία.

Τέλος, το *ViewDirect TCM* παρέχει διάφορες μεθόδους αναζήτησης, οργάνωσης και επισκόπησης των ηλεκτρονικών μηνυμάτων μαζί με το γενικό περιεχόμενο των δεδομένων του οργανισμού. Το *ViewDirect E-mail Management* είναι συμβατό μόνο με το *Microsoft Outlook* καθώς και με τους servers *Microsoft Windows Server 2000, 2003, XP, 98, Microsoft Exchange 2003. 2000, 5.5,* με τις Βάσεις Δεδομένων SQL*, Sybase, IBM DB2, Oracle 9i, MSDE* καθώς και με τα προγράμματα του *Microsoft Office XP, 2000, 2003*. Τέλος, παρέχει υποστήριξη σε διάφορες γλώσσες.

2.2.2.2.3 ByteQuest

Την ίδια λογική με τα συστήματα που προαναφέρθηκαν εντάσσεται και το σύστημα διαχείρισης ηλεκτρονικών μηνυμάτων της ***ByteQuest***, ***eManage*** [i] ***.*** Το συγκεκριμένο σύστημα παρουσιάζει μεγάλη πορεία μετεξελίξεων γιατί αν και αρχικά εισήχθηκε στην αγορά ως προϊόν της συγκεκριμένης εταιρίας, στη συνέχεια ακολούθησαν διάφορες εκδόσεις του, που πραγματοποιήθηκαν σε συνεργασία με άλλες εταιρίες. Η τελευταία, πραγματοποιήθηκε από τη ***Mobius,*** που μόλις αναφέρθηκε και η οποία το 2004 προσάρτησε στο σύστημα της συστατικά του *eManage,* όπως η τεχνολογία διαχείρισης ηλεκτρονικών τεκμηρίων και μηνυμάτων.

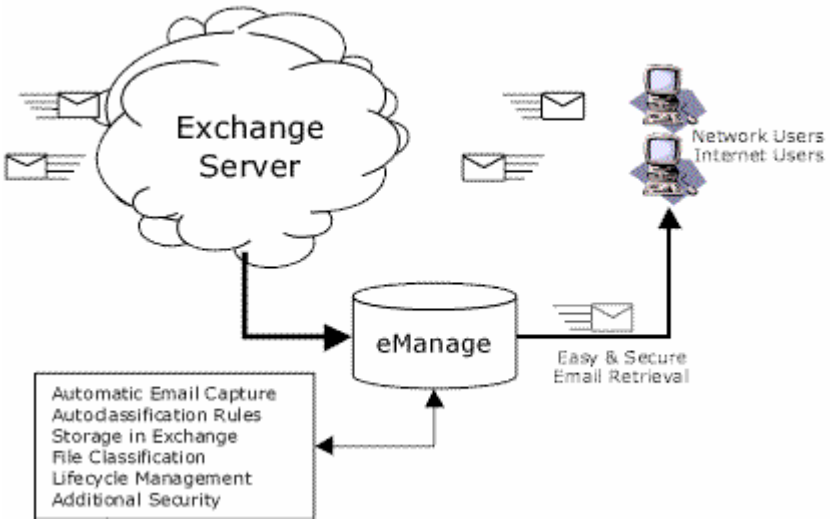
Το *eManage* παρέχει μία σειρά υπηρεσιών για τη διαχείριση ηλεκτρονικών μηνυμάτων για *email clients*, όπως το *Microsoft Outlook*. Στις υπηρεσίες του περιλαμβάνονται η σύλληψη εισερχόμενων και εξερχόμενων ηλεκτρονικών μηνυμάτων, η αποθήκευση, η οργάνωση και διαχείρισή τους, η παροχή ασφαλούς πρόσβασης και τέλος η συγχώνευσή τους σε ένα κοινό αποθηκευτικό χώρο με τα υπόλοιπα τεκμήρια που παράγονται στον εκάστοτε οργανισμό. Η τελευταία υπηρεσία είναι σημαντική καθώς με αυτό τον τρόπο, παρέχεται στους χρήστες ένα κοινό σημείο πρόσβασης σε όλα τα δεδομένα, ενώ παράλληλα, τόσο τα τεκμήρια όσο και τα ηλεκτρονικά μηνύματα υπόκεινται στους ίδιους κανόνες για αποθήκευση, αρχειοθέτηση, διατήρηση και διάθεση, που ορίζονται από τον οργανισμό.



Εικόνα 6-επικοινωνία e-manage με εξωτερικά συστήματα αρχείων

Η σύλληψη των μηνυμάτων, εισερχόμενων και εξερχόμενων πραγματοποιείται σε πραγματικό χρόνο, χωρίς να διακόπτεται η διαδικασία μεταφοράς τους,. Στο σχήμα που ακολουθεί παρουσιάζεται η διαδικασία σύλληψης των τεκμηρίων. Στη συνέχεια, τα μηνύματα αποθηκεύονται σε ειδικούς αποθηκευτικούς χώρους στα πλαίσια του *Exchange,* που ονομάζονται MS Exchange Libraries. Το χαρακτηριστικό των αποθηκευτικών αυτών χώρων έγκειται στο ότι είναι διαφανή για τους χρήστες. Παράλληλα, η σύλληψη των ηλεκτρονικών μηνυμάτων μπορεί

να γίνει και από τον *Exchange*, που στη συνέχεια τα μηνύματα αποθηκεύονται στους *MS Exchange Public Folders.*



Εικόνα 7- διαδικασία σύλληψης ηλεκτρονικών μηνυμάτων από το e-Manage

Η κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων πραγματοποιείται στη μηχανή αυτόματης κατηγοριοποίησης. Η διαδικασία στηρίζεται στην εισαγωγή κανόνων από τους ίδιους τους χρήστες στον βοηθό κατηγοριοποίησης, *autoclassification wizard*. Οι κανόνες εφαρμόζονται στα πεδία της επικεφαλίδας αλλά και στο σώμα του μηνύματος, ενώ, παράλληλα είναι δυνατό να κατηγοριοποιηθούν και βάση σπουδαιότητας ή και ευαισθησίας του περιεχομένου. Παράλληλα, εφαρμόζονται μεταδεδομένα, τα οποία αποθηκεύονται μαζί με τα μηνύματα και τα συνημμένα έγγραφα που μπορεί να περιλαμβάνουν στον κοινό αποθηκευτικό χώρο μαζί με τα άλλα τεκμήρια. Η κατηγοριοποίηση των μηνυμάτων πραγματοποιείται σε ιεραρχικούς *Knowledge Maps.*

Τα *Knowledge Maps* παρέχουν, παράλληλα, παραμετροποιήσιμα προφίλ για τους φακέλους και τα μηνύματα που περιλαμβάνουν. Είναι δυνατό, επίσης, να πραγματοποιηθεί μέσω αυτών η πρόσβαση και η αναζήτηση πλήρους κειμένου στο σύνολο των ηλεκτρονικών μηνυμάτων. Η ανάκτηση μηνυμάτων για το Outlook πραγματοποιείται μέσω των βασικών φορμών που διαθέτει (Outlook 98/200). Τέλος, το *eManage* μπορεί να εισαχθεί σε άλλες εφαρμογές του οργανισμού για διαχείριση- τεκμηρίων.

Το eManage υποστηρίζει *Windows 95/98/NT/2000, Microsoft Exchange, MS Outlook 98/2000* και Βάσεις δεδομένων που υποστηρίζουν το πρότυπο *ODBS*, όπως είναι ο *MS SQL Server, η Oracle, η IBM DB2, η Sybase και η Sybase Anywhere*.

Η πρώτη βελτίωση που πραγματοποιήθηκε στην αρχική έκδοση του *eManage* επιτεύχθηκε με την εισαγωγή στο σύστημα, της μηχανής ερμηνείας[37] περιεχομένου, ***AmikaHighlighter*** της ***Amika***. Το πλεονέκτημα που επέφερε αυτή η εισαγωγή είναι η βελτίωση της αποδοτικότητας της μηχανής αυτόματης κατηγοριοποίησης του *eManage.* Η *AmikaHighlighter* βάση του Τεχνητή Νοημοσύνη λογισμικού που ενσωματώνει στο σύστημά της, επιτυγχάνει τον προσδιορισμό λέξεων κλειδιών, φράσεων και προτάσεων. Με αυτό τον τρόπο βελτιώνεται ο προσδιορισμός του κειμένου και συνεπώς αυξάνεται η αποδοτικότητα της κατηγοριοποίησης. Παράλληλα, ένα άλλο πλεονέκτημα που παρουσιάζει η εισαγωγή της *AmikaHighlighter* είναι η βελτίωση της διαδικασίας αναζήτησης. Η σχετική βελτίωση πραγματοποιείται με τη δημιουργία περιλήψεων των ηλεκτρονικών μηνυμάτων από τα δεδομένα εξόδου της μηχανής.

Το 2000 το *eManage* συγχωνεύθηκε με το προϊόν αρχειοθέτησης ηλεκτρονικών μηνυμάτων, ***Ixos-ExchangeArchive*** της **Ixos.** Το σχετικό προϊόν αρχειοθέτησης της *Ixos* παρέχει ένα αυτοματοποιημένο και διαδραστικό περιβάλλον διαχείρισης των ηλεκτρονικών μηνυμάτων στα πλαίσια του *MS Exchange*. Αναφορικά μόνο θα αναφερθεί μία διακριτή υπηρεσία του, που εισάγει ένα διαφορετικό στοιχείο από την προσέγγιση του *eManage.* Το *Ixos-ExchangeArchive* κατά την αρχειοθέτηση των μηνυμάτων, τα διαγράφει από τον Exchange ή από τους τοπικούς servers και τα αποθηκεύει σε οπτικά ή σε άλλα μέσα. Με αυτό τον τρόπο, μειώνεται ο κατειλημμένος αποθηκευτικός χώρος των servers και βελτιώνεται η αποδοτικότητά τους. Παρά τη μετακίνηση των ηλεκτρονικών μηνυμάτων, οι χρήστες εξακολουθούν να έχουν πρόσβαση σε αυτά και να μπορούν να τα διαχειρίζονται.

[37] engine content interpretation

Με σκοπό την βελτίωση της αρχειοθέτησης των ηλεκτρονικών μηνυμάτων και την παροχή περισσότερων υπηρεσιών στον χρήστη, πραγματοποιήθηκε η συγχώνευση του *eManage v. 4.5* με το ***EAS 3.0*** της ***EAS***. Στη νέα αυτή έκδοση, η διαδικασία της σύλληψης και της κατηγοριοποίησης των ηλεκτρονικών μηνυμάτων ακολουθεί το πρότυπο του *eManage.* Αφού, τα μηνύματα έχουν αποθηκευτεί στους *Exchange Public Folders* βάση μίας προγραμματισμένης εργασίας, αρχίζει η διαδικασία αρχειοθέτησης τους. Ο *EAS 3.0* απομακρύνει τα μηνύματα από τον *Exchange* και τα μεταφέρει σε μία νέα αποθηκευτική περιοχή στο δικτυακό αποθηκευτικό χώρο. Η περιοχή αυτή ονομάζεται *Document Store.* Παρά τη μετακίνησή τους, τα μηνύματα παραμένουν προσβάσιμα στους χρήστες είτε από τη διεπιφάνεια του *Outlook* είτε μέσω του *eManage* ή του *EAS.* Σημαντικό στοιχείο αποτελεί ότι κατά τη μεταφορά των ηλεκτρονικών μηνυμάτων στο νέο αποθηκευτικό χώρο, τα μηνύματα συμπιέζονται κατά 80% του αρχικού τους μεγέθους. Η μετακίνηση και η συμπίεση των μηνυμάτων παρέχει το γνωστό πλεονέκτημα της βελτίωσης της αποδοτικότητας του server.

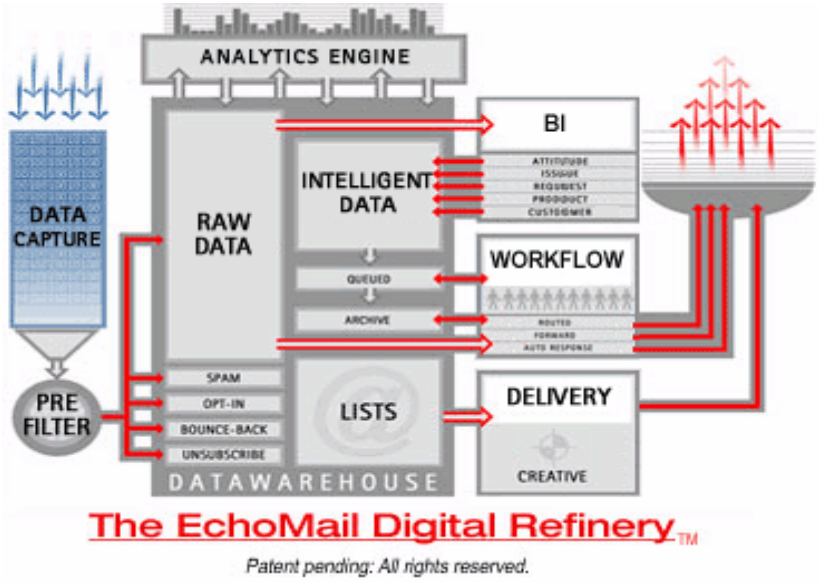
2.2.2.2.4 EchoMail

Μία διαφορετική προσέγγιση στο ζήτημα της κατηγοριοποίησης ηλεκτρονικών μηνυμάτων δίνεται από την ***EchoMail*** [j] με το ομώνυμο προϊόν της. Οι υπηρεσίες διαχείρισης μηνυμάτων εισάγονται σε ένα σύστημα μάρκετινγκ των σχέσεων πελατών (e-Customer Relationship Marketing). Το σύστημα παρέχει μία σειρά υπηρεσιών, όπως αυτές διαφαίνονται στη συνέχεια, με σκοπό την γρήγορη και εύκολη εξυπηρέτηση των πελατών του εκάστοτε οργανισμού. Οι υπηρεσίες που παρέχει είναι:

- Σύλληψη και διαχείριση
- Ανάλυση, δρομολόγηση και παρακολούθηση όλων των μηνυμάτων
- Αποθήκευση, αναζήτηση (mine) και πραγματοποίηση ερωτημάτων
- Δημιουργία mailing list και αποστολή ομαδικών μηνυμάτων

Ο κύριος σκοπός του *EchoMail* είναι η υποστήριξη της υπηρεσίας εξυπηρέτησης των πελατών της εκάστοτε εταιρίας, μέσω της κατανόησης των αναγκών τους. Οι ανάγκες αυτές, συνήθως καταγράφονται σε ηλεκτρονικά μηνύματα, που

αποστέλλονται από τους πελάτες στην εν λόγω εταιρία. Για αυτό τον λόγο το *EchoMail* μέσω της τεχνολογίας **XIVA** κατηγοριοποιεί την ηλεκτρονική αλληλογραφία, βάση των «ιδιοτήτων» των πελατών. Οι «ιδιότητες» εξάγονται από τα ηλεκτρονικά μηνύματα, βάση της *pattern recognition* μεθόδου.



Εικόνα 8-Απεικόνιση της λειτουργίας του EchoMail

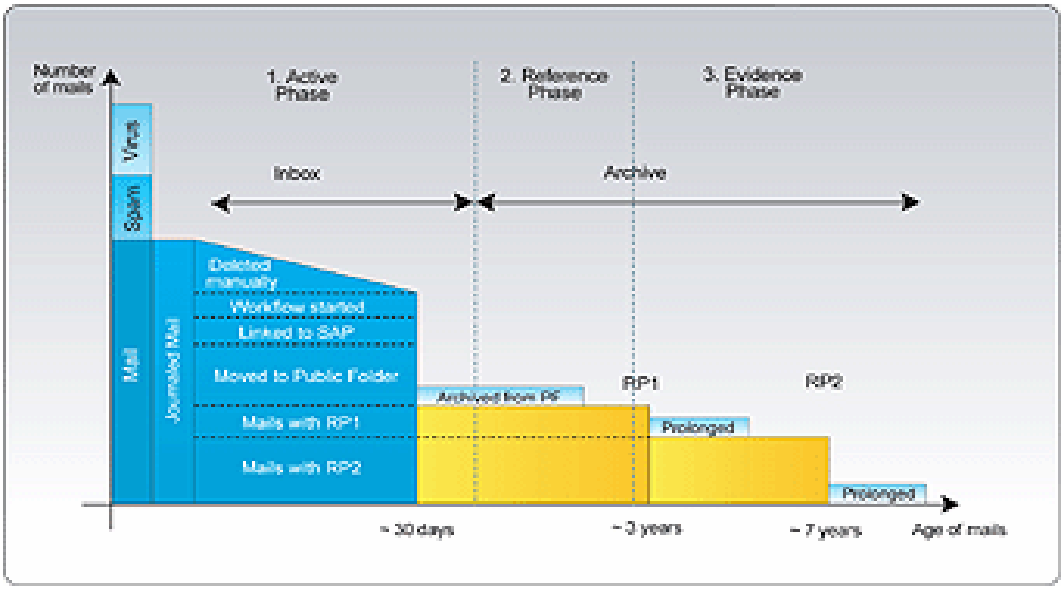
Οι ιδιότητες αυτές περιλαμβάνουν τον τύπο του πελάτη, τον τύπο του προϊόντος ή της υπηρεσίας, την απαίτηση, το θέμα και τη στάση. Προκειμένου να γίνει η αναγνώριση και ο προσδιορισμός των ιδιοτήτων χρησιμοποιούνται από την *XIVA* διάφορες τεχνολογίες, όπως είναι θησαυροί, λεξιλόγια και μορφολογικοί αναλυτές.

Η αρχιτεκτονική του *EchoMail* βασίζεται στην *Trinity* που παρέχει μία ανοιχτή αρχιτεκτονική, βάση προτύπων για τη διαχείριση των ηλεκτρονικών μηνυμάτων. Βασικό πλεονέκτημα είναι ότι επιτρέπει την παράλληλη χρήση ενός ή περισσοτέρων server, όπου λειτουργούν διαφορετικά λειτουργικά συστήματα. Παράλληλα, ενσωματώνει το μοντέλο Βάσης Δεδομένων *INO,* που παρέχει την ικανότητα διαχείρισης τόσο των εισερχόμενων όσο και των εξερχόμενων ηλεκτρονικών μηνυμάτων.

Το *EchoMail* υποστηρίζει προγράμματα όπως *Windows NT, Lotus Domino, Oracle και Sun Solaris.*

2.2.2.2.5 TOWER

Η ***TOWER*** [k] εισάγει τη διαχείριση ηλεκτρονικών μηνυμάτων σε ένα μεικτό σύστημα διαχείρισης αρχείων (RMA). Το σύστημα, γνωστό ως ***TRIM Context,*** διαχειρίζεται τον κύκλο ζωής των αρχείων παραδοσιακής αλλά και ηλεκτρονικής μορφής, από τον εντοπισμό τους μέχρι την αποθήκευσή και διατήρησή τους. Ενδιάμεσες λειτουργίες περιλαμβάνουν την αυτόματη σύλληψή τους και κατηγοριοποίησή τους, την εξαγωγή μεταδεδομένων και την επιβολή κανόνων διατήρησης και διάθεσης.



Εικόνα 9-Κύκλος ζωής των ηλεκτρονικών μηνυμάτων και διαχείριση

Η αρχιτεκτονική του *TRIM Context* βασίζεται στο μοντέλο της Microsoft για κατανομή των συστατικών, γνωστό ως *DCOM* (*Distributed Component Object Model*) που παρέχει μία multi tier (πολλαπλής διαβάθμισης) αρχιτεκτονική. Βάση

της πολλαπλής διαβάθμισης -αντί για την αρχιτεκτονική διπλής διαβάθμισης, του τύπου πελάτη-εξυπηρετητή- επιτυγχάνεται η μείωση του βάρους εργασίας στους servers και η υπερφόρτωση του δικτύου μέσω του καταμερισμού διαδικασίας επεξεργασίας δεδομένων.

Σχετικά με τα ηλεκτρονικά μηνύματα, ο *TRIM Context* παρέχει ένα σύνολο υπηρεσιών όπως είναι η σύλληψη, η κατηγοριοποίηση, μετατροπή και η μεταφορά σε άλλους αποθηκευτικούς χώρους, η διάθεση και διατήρηση βάση κανόνων, η εγγραφή και τέλος, η παραγωγή εργαλείων ελέγχου.

Η σύλληψη των ηλεκτρονικών μηνυμάτων γίνεται αυτόματα, ενώ παράλληλα παρέχει και τη δυνατότητα **drag and drop** από τους *email clients*. Κατά την σύλληψη των μηνυμάτων, εξάγονται κάποια χαρακτηριστικά τους όπως είναι ο αποστολέας, η ημερομηνία, οι ηλεκτρονικές διευθύνσεις και το θέμα του μηνύματος, τα οποία χρησιμοποιούνται ως δεδομένα εισόδου στο προφίλ του αρχείου. Στην περίπτωση που ένα ηλεκτρονικό μήνυμα περιλαμβάνει και συνημμένα αρχεία, ο *TRIM Context* δίνει στον χρήστη τη δυνατότητα επιλογής διάσπασης του μηνύματος. Αυτό σημαίνει ότι ο χρήστης μπορεί να επιλέξει να αποθηκεύσει το μήνυμα και το συνημμένο αρχείο μαζί ή ξεχωριστά.

Η αποθήκευση των μηνυμάτων γίνεται σε ένα *NTFS* (*NT File System*) server. Κατά την αποθήκευση των μηνυμάτων αλλά και των άλλων αρχείων, προσδίδονται άδειες πρόσβασης σε αυτά, σε επίπεδο αρχείου, φακέλου και τεκμηρίου, που καθορίζουν ποιοι χρήστες έχουν τη δυνατότητα πρόσβασης αλλά και διαγραφής τους. Παράλληλα, το σύστημα παρέχει τη δυνατότητα «*εγγραφής*» (registration) , η οποία περιλαμβάνει στοιχεία σχετικά με τη δημιουργία ή μετατροπή του αρχείου.

Ο *TRIM Context* περιλαμβάνει ένα συστατικό, το *Record Plan* ή *File Classification* που επιτρέπει τη δημιουργία ενός συστήματος ταξινόμησης. Βάση των κατηγοριών του συστήματος αυτού γίνεται η κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων και των άλλων αρχείων. Βασικό χαρακτηριστικό του είναι ότι μπορεί να χτιστεί με τέτοιο τρόπο ώστε να αντανακλά τις λειτουργίες του οργανισμού. Επίσης, στις βασικές κατηγορίες του ταξινομικού συστήματος μπορούν να αποδοθούν και αριθμητικοί προσδιορισμοί, όπως στο δεκαδικό σύστημα

ταξινόμησης του Dewey. Παράλληλα, οι κατηγορίες αυτές, μπορούν να διασυνδεθούν με την υπηρεσία *Disposal Schedules,* που είναι μία προγραμματισμένη υπηρεσία διάθεσης των μηνυμάτων και των τεκμηρίων γενικότερα.

Η κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων γίνεται βάση μεταδεδομένων στις προκαθορισμένες κατηγορίες του File Plan και με παράλληλη χρήση θησαυρού. Τα στοιχεία μεταδεδομένων που θα εισαχθούν για την κατηγοριοποίηση των μηνυμάτων μπορούν να οριστούν με τέτοιο τρόπο ώστε να ικανοποιούν τις ανάγκες του οργανισμού. Μία καινοτομία που εισάγει ο *TRIM Context* είναι η χρήση του *a.k.a.* Το *a.k.a* είναι ένα λογισμικό δημιουργίας θησαυρού και ταξινομιών, που στηρίζεται στο πρότυπο ISO 15489. Η εισαγωγή του λογισμικού αυτού στο σύστημα του TRIM επιτρέπει την εύκολη διαχείριση των όρων του θησαυρού του και του File Plan.

Επίσης, ο *TRIM Context* περιλαμβάνει εργαλεία για τη μεταφορά γενικά των αρχείων και ειδικότερα των ηλεκτρονικών μηνυμάτων από τον κεντρικό server σε άλλα συστήματα σχεσιακών Βάσεων Δεδομένων. Η τελευταία υπηρεσία του *TRIM Context* είναι η δημιουργία εργαλείων ελέγχου, γνωστά ως *audit trails.* Τα εργαλεία αυτά δημιουργούν εγγραφές που περιλαμβάνουν τα στοιχεία των αρχείων, όπως είναι ο τίτλος και τα τεκμήρια που περιέχουν, ενώ επίσης δεδομένα σχετικά με τροποποιήσεις που έχουν πραγματοποιηθεί σε αυτά. Ο *TRIM Context* εισάγει δύο τύπων σχετικές εγγραφές ελέγχου που διαφοροποιούνται ως προς το είδος των πληροφοριών που περιέχουν. Η βασική εγγραφή ελέγχου, **core logging** περιλαμβάνει τα στοιχεία του αρχείου καθώς και διάφορες αρχειακές δραστηριότητες, όπως είναι η μετακίνηση αρχείων και η διαγραφή. Οι εγγραφές του δεύτερου τύπου, **full logging** περιλαμβάνουν τα στοιχεία των προηγούμενων ενώ προσθέτουν και στοιχεία αναφορικά με αλλαγές στην ασφάλεια των αρχείων, αλλαγές στους όρους του θησαυρού ή αλλαγές αναφορικά με τους όρους διάθεσης και διατήρησης των αρχείων.

Για τη διάθεση και διατήρηση των αρχείων, εισάγονται στο σύστημα κανόνες συμβατοί με διάφορα πρότυπα, όπως είναι το DoD 5015.2 για τις Η.Π.Α., οι απαιτήσεις του UK National Electronic Records Management για τις ευρωπαϊκές χώρες και για τις χώρες της Αφρικής και οι κανόνες του Victorian Electronic

Records Strategy για τις ασιατικές χώρες. Τέλος, στον παρακάτω πίνακα εμφανίζονται τα προγράμματα με τα οποία το *TRIM Context* είναι συμβατό.

Λειτουργικό Σύστημα Server	MS Windows UNIX	2000 Server Solaris 8.0
Λειτουργικό Σύστημα Σταθμών Εργασίας	MS Windows	2000 Professional XP Professional
Σύστημα Διαχείρισης Βάσης Δεδομένων	MS SQL Oracle	2000 9i
E-mail Servers	MS Exchange Lotus	2000/ 5.5 6
E-mail	MS Outlook Lotus Notes	2000/XP 6
Αυτοματοποίηση Γραφείου	MS Office	2000 Professional XP Professional

2.2.2.2.6 Documentum

Η ***Documentum [I]*** είναι μία εταιρία η οποία επικεντρώνεται στην παροχή συστημάτων λογισμικού διαχείρισης αρχείων (RMS). Σε ένα από τα συστήματα που παρέχει στην αγορά, το ***Enterprise Records Management Edition*** εισάγει και τη διαχείριση αρχείων ηλεκτρονικών μηνυμάτων στο πλαίσιο της διαχείρισης των αρχείων της εταιρίας. Βάση της κοινής διαχείρισης τους, είναι δυνατός ο καθορισμός κοινών πολιτικών και διαδικασιών για τον προσδιορισμό και την προστασία των σημαντικών αρχείων, για τη σύλληψη και αποθήκευσή τους, καθώς και για την ανάκτηση, διάχυση και καταστροφή τους.

Ειδικά για τα ηλεκτρονικά μηνύματα, το σύστημα αυτοματοποιεί οποιαδήποτε διαδικασία σχετίζεται με αυτά. Σημαντικό είναι ότι οι σχετικές διαδικασίες μπορούν να πραγματοποιηθούν τόσο από το ίδιο το προϊόν της *Documentum* ή από εξωτερικές πηγές μέσω διασύνδεσης των σχετικών συστημάτων. Οι διαδικασίες διαχείρισης, δηλαδή, μπορούν να πραγματοποιηθούν από την εφαρμογή που καθεαυτού διαχειρίζεται τα ηλεκτρονικά μηνύματα και στην συνέχεια να εισαχθούν στο *Enterprise Records Management Edition*. Το βήμα αυτό πραγματοποιείται με την ενεργοποίηση της υπηρεσίας, *Records Activators.* Αναφορικά με τις δυνατότητες του ίδιου του συστήματος, η *Enterprise Records Management Edition* συλλαμβάνει τόσο τα εισερχόμενα όσο και τα εξερχόμενα ηλεκτρονικά μηνύματα και τα αποθηκεύει σε on-line αλλά και φυσικούς χώρους. Στην περίπτωση που τα ηλεκτρονικά μηνύματα περιλαμβάνουν και συνημμένα αρχεία, τότε πραγματοποιείται διάσπαση του βασικού σώματος από το συνημμένο αρχείο και στην συνέχεια αποθηκεύονται και κατηγοριοποιούνται ξεχωριστά.

Κατά τη διαδικασία της σύλληψης των ηλεκτρονικών μηνυμάτων πραγματοποιείται η εξαγωγή των βασικών τους ιδιοτήτων, όπως είναι η ημερομηνία, ο αποστολέας, ο χρόνος εισαγωγής ή εξαγωγής και ο παραλήπτης, ενώ παράλληλα, προσδιορίζει το περιεχόμενο του μηνύματος και το κατηγοριοποιεί. Η κατηγοριοποίηση των μηνυμάτων όπως και των άλλων αρχείων γίνεται βάση του file plan με χρήση του ***CIS***.

Το *CIS* (*Content Intelligence Services*) είναι μία υπηρεσία που αυτοματοποιεί και ελέγχει τον εμπλουτισμό και την οργάνωση του περιεχομένου όλων των τεκμηρίων που διακινούνται στα πλαίσια μίας εταιρίας βάση των εξής

διαδικασιών: *εξαγωγή πληροφοριών*, *νοηματική κατηγοριοποίηση*, *ανάλυση της εταιρίας* και τέλος βάση των δυνατοτήτων *διαχείρισης μεταδεδομένων και ταξινομίας*. Η *εξαγωγή πληροφοριών* βασίζεται στην εξαγωγή μεταδεδομένων από το περιεχόμενο με σκοπό τον προσδιορισμό, την κατηγοριοποίηση και την ένταξη των τεκμηρίων στη Βάση Δεδομένων. Η *νοηματική ταξινόμηση* αναφέρεται στον προσδιορισμό συγκεκριμένου τμήματος του περιεχομένου, το οποίο είναι ικανό να παράσχει πλουσιότερες συσχετίσεις και να διευρύνει την αναζήτηση. Το περιεχόμενο των ηλεκτρονικών μηνυμάτων κατά τον προσδιορισμό του και την κατηγοριοποίησή του συσχετίζεται με το περιεχόμενο άλλων σχετικών αρχείων, προκειμένου να δημιουργηθεί ένας αποθηκευτικός χώρος που να παρέχει ευρύτερες δυνατότητες αναζήτησης. Η νοηματική κατηγοριοποίηση βασίζεται αλγόριθμους που αναλύουν το περιεχόμενο των ηλεκτρονικών μηνυμάτων και των άλλων αρχείων βάση του file plan. Η διαδικασία *ανάλυση εταιρίας* περιλαμβάνει την αυτόματη ανάλυση του περιεχομένου της εταιρίας ώστε να αντικατοπτρίζει τις δραστηριότητές της. Τέλος, η *διαχείριση μεταδεδομένων και ταξινομίας* αναφέρεται στη δημιουργία, τροποποίηση και διατήρηση ταξινομιών σε XML διάταξη.

Με την εξαγωγή των βασικών ιδιοτήτων των ηλεκτρονικών μηνυμάτων και με τη νοηματική ταξινόμηση του περιεχομένου τους, η *Enterprise Records Management Edition* παρέχει ένα διττό τρόπο κατηγοριοποίησης των μηνυμάτων. Μετά την κατηγοριοποίηση τους, τα μηνύματα εισάγονται στον αποθηκευτικό χώρο, από όπου γίνονται αντικείμενα επεξεργασίας άλλων υπηρεσιών. Μέσω ενός ειδικού web client πραγματοποιείται η πρόσβαση στα δεδομένα, η αναζήτηση και η ανάκτηση, ενώ παράλληλα επιτρέπεται η επισκόπηση των file plan καθώς και ο ορισμός ή τροποποίηση πολιτικών που σχετίζονται με τη διάθεση και την διατήρησή τους.

2.2.2.2.7 IBM

Το **DB2 Content Management [m]** είναι η συλλογική λύση της **IBM** και της **iLumin** για τη διαχείριση της ηλεκτρονικής αλληλογραφίας των χρηστών. Τα

συστατικά τμήματα του προϊόντος περιλαμβάνουν το **DB@ CommonStore** της **IBM** για τη διατήρηση των ηλεκτρονικών μηνυμάτων (που περιλαμβάνει τα **DB2 Content Manager, CommonStore & Tivoli Storage Manager ExtendedExtension**) και την εφαρμογή **Assentor** της **iLumin.**

Μετά την σύλληψη των ηλεκτρονικών μηνυμάτων, ο *Content Manager* χρησιμοποιείται για να αναλύσει το περιεχόμενο τους βάση της τεχνολογίας *Natural Language Content Analysis (Ανάλυση Περιεχομένου Φυσικής Γλώσσας).* Με την ανάλυση του περιεχομένου είναι δυνατή η ευρετηρίαση όλων των πεδίων των μηνυμάτων, δηλαδή των πεδίων της επικεφαλίδας και του σώματος του μηνύματος. Στη διαδικασία της ευρετηρίασης υπόκεινται και τα συνημμένα αρχεία. Αποτέλεσμα της ευρετηρίασης είναι η εξαγωγή οντοτήτων, όπως ονόματα ανθρώπων, εταιριών, ημερομηνίες και σύμβολα. Τα στοιχεία αυτά χρησιμοποιούνται από τον **DB2 Record Manager** για την κατηγοριοποίηση των μηνυμάτων και των σχετικών τους εγγράφων. Παράλληλα, τα σημεία αυτά μπορούν να χρησιμοποιηθούν ως σημεία αναζήτησης και πρόσβασης στα μηνύματα. Σημαντική είναι η δυνατότητα που παρέχεται στους χρήστες να επιλέξουν αν θέλουν να διατηρήσουν και να κατηγοριοποιήσουν το σύνολο του μηνύματος ή μόνο το συνημμένο αρχείο.

Το συστατικό **Tivoli Storage Manager ExtendedExtension,** χρησιμοποιείται για τη δυνατότητα παροχής πρόσβασης σε ετερογενείς αποθηκευτικούς χώρους , τη συλλογή των δεδομένων και τη συγκέντρωσή τους υπό τον *content Manager.* Με αυτό τον τρόπο επιτυγχάνεται η δημιουργία ενός κεντρικού σημείου διαχείρισης των μηνυμάτων και εφαρμογής γενικών κανόνων για τη διάθεση και διατήρησή τους. Το συστατικό της *iLumin,* **Assentor Message Manager,** χρησιμοποιείται για την τροποποίηση των ηλεκτρονικών μηνυμάτων, όπως είναι η ανάλυση στα συστατικά του και η μετατροπή του σε κείμενο. Αφού, πραγματοποιηθούν οι σχετικές τροποποιήσεις, οι πληροφορίες συγκεντρώνονται στον *Content Manager,* από όπου γίνονται προσβάσιμες και διαχειρίσιμες από άλλες εφαρμογές.

Τέλος, ο *DB2 Content Management* υποστηρίζει τον *MS Exchange, τον Lotus Notes και Novel GroupWise* και άλλα προϊόντα διαχείρισης πληροφοριών όπως είναι τα *IM Logic, IM Age, Hub IM, Akonix* και *Omnipad.*

Αξιολόγηση

Η αξιολόγηση των συστημάτων που παρουσιάστηκαν, θα πραγματοποιηθεί βάση δύο βασικών αξόνων: συγκριτικά με τις υπηρεσίες που παρέχουν και αναφορικά με άλλα λειτουργικά συστήματα, που είναι συμβατά. Τα κριτήρια, αυτά επιλέχθηκαν, ακριβώς επειδή σχετίζονται με την χρηστικότητα των συστημάτων τόσο όσον αφορά τις διευκολύνσεις που παρέχουν στους ίδιους τους χρήστες αλλά και όσον αφορά την ευκολία εγκατάστασής τους.

Ο πρώτος πίνακας αξιολόγησης – τμήμα α & β- αναφέρεται στις υπηρεσίες που παρέχουν τα συστήματα προς τους χρήστες τους. Στην συνέχεια περιγράφονται περιληπτικά οι διάφορες υπηρεσίες:

- **Σύλληψη**: Η υπηρεσία αυτή αναφέρεται στη δυνατότητα του συστήματος να συλλαμβάνει τα ηλεκτρονικά μηνύματα- εισερχόμενα και εξερχόμενα- αυτόματα και σε πραγματικό χρόνο καθώς αυτά μεταδίδονται.

- **Κατηγοριοποίηση:** Αναφέρεται στην υπηρεσία που παρέχουν τα συστήματα εισάγοντας αυτόματα τα ηλεκτρονικά μηνύματα –εισερχόμενα και εξερχόμενα- στις κατάλληλες θεματικές κατηγορίες.

- **Αποθήκευση:** Αναφέρεται στην υπηρεσία που παρέχουν τα συστήματα στους χρήστες να αποθηκεύουν τοπικά ή κεντρικά την ηλεκτρονική τους αλληλογραφία. Κάποια από τα συστήματα- αυτά που συνοδεύονται με * - αποθηκεύουν τα μηνύματα σε ειδικούς χώρους.

- **Διατήρηση/ διάθεση:** Βάση αυτής της υπηρεσίας τα συστήματα διαχειρίζονται τα ηλεκτρονικά μηνύματα ως ηλεκτρονικά αρχεία, δηλαδή εφαρμόζουν σε αυτά κανόνες που αφορούν τον κύκλο ζωής τους. Συνήθως είναι κανόνες που ορίζονται από τον ίδιο τον ενδιαφέροντα οργανισμό ή βάση νόμων π.χ. DoD 50151 για την Αμερική. Τα συστήματα που υποσημειώνονται με το σύμβολο του αστερίσκου, επιτελούν αρχειοθέτηση των μηνυμάτων.

- **Αναζήτηση:** Αναφέρεται στην υπηρεσία που παρέχουν τα συστήματα βάση της οποίας ο χρήστης μπορεί να εκτελέσει αναζήτηση στο σώμα τν ηλεκτρονικών του μηνυμάτων.

- **Μεταφορά:** Αναφέρεται στην υπηρεσία των συστημάτων βάση της οποίας τα μηνύματα μεταφέρονται από το mail server ή από τοπικούς αποθηκευτικούς χώρους σε έναν κεντρικό, όπου συγκεντρώνονται όλα τα μηνύματα των χρηστών. Τα συστήματα που υποσημειώνονται επιτελούν και διαγραφή των μηνυμάτων από τον server κατά τη μεταφορά τους, προκειμένου να μειώνεται το βάρος αποθήκευσης από τον mail server.

- **Εργαλεία Ελέγχου:** Το σύστημα που παρέχει μία τέτοια υπηρεσία, ουσιαστικά δημιουργεί εγγραφές οι οποίες περιλαμβάνουν στοιχεία σχετικά με το περιεχόμενο των κατηγοριών και των μηνυμάτων που περιλαμβάνουν, διαχειριστικά στοιχεία για αυτά καθώς και στοιχεία που αναφέρονται σε τυχόν τροποποιήσεις που έχουν πραγματοποιηθεί στα μηνύματα.

- **Συγχώνευση:** Αναφέρεται στην υπηρεσία κατά την οποία τα μηνύματα κα τα τη μεταφορά τους και την αποθήκευσή τους σε ένα κοινό αποθηκευτικό χώρο, συγχωνεύονται με άλλα ηλεκτρονικά ή κι παραδοσιακά τεκμήρια. Το βήμα αυτό αποτελεί μεγάλο πλεονέκτημα προς τους χρήστες καθώς με αυτό τον τρόπο παρέχεται ένα κοινό σημείο πρόσβασης και διαχείρισης παντός τύπου ηλεκτρονικά τεκμήρια.

- **Συμπίεση:** Αναφέρεται στην υπηρεσία των συστημάτων να μειώνουν το μέγεθος του ηλεκτρονικού μηνύματος καθώς αυτό μεταφέρεται στον κοινό αποθηκευτικό χώρο.

- **Δημιουργία Περιλήψεων:** Αναφέρεται σε μία υποβοηθητική υπηρεσία προς την υπηρεσία της αναζήτησης που παρέχεται από διάφορα συστήματα. Με τη δημιουργία περιλήψεων ο χρήστης μπορεί να γίνει πιο γρήγορα γνώστης του περιεχομένου του μηνύματος.

- **Τροποποίηση Διάταξης:** αναφέρεται στην υπηρεσία βάση της οποίας το ηλεκτρονικό μήνυμα μετατρέπεται σε διάταξη πλήρους κειμένου ή άλλη, ώστε να είναι πιο εύκολη η διαχείρισή του.

Προϊόντα/ Υπηρεσίες	MetaTagger CIS	ViewDirect	eManage	EchoMail
Σύλληψη	*Drag & Drop*	*Αυτόματη*	*Αυτόματη*	*Αυτόματη*
Κατηγοριοποίηση[38]	√	√	√	√
Αποθήκευση	√	√*	√*	√
Διατήρηση\ Διάθεση	–	√ Αρχειοθέτ.	√ Αρχειοθέτ.	√ *Αρχειοθέτ.*
Αναζήτηση	√	√	√	√
Μεταφορά	√	√*	√	√
Εργαλεία Ελέγχου	–	–	–	–
Άλλες	*1.Συγχώνευση με άλλα τεκμήρια 2.Δημιουργία Περιλήψεων*	*1.Διαγραφή 2.Αποθήκευση σε οπτικά μέσα 2.Εναλλακτική πρόσβαση*	*1.Συγχώνευση 2. Συμπίεση 3.Περιλήψεις 4.Διαγραφή 5.Πολλαπλή κατηγοριοποίηση*	*1.Δημιουργία Mailing list*

[38] Τα σχετικά συστήματα επιτελούν κατηγοριοποίηση βάση μεταδεδομένων και με χρήση θησαυρού

Πίνακας 2-Συγκριτικός πίνακας υπηρεσιών (μέρος ΄α)

Προϊόντα/ Υπηρεσίες	Trim	Enterprise RME	DB2 Content Management
Σύλληψη	*Αυτόματη / d&d*	*Αυτόματη*	*Αυτόματη*
Κατηγοριοποίηση	√	√	√
Αποθήκευση	√	√	√
Διατήρηση\ Διάθεση	√ *Αρχειοθέτηση*	√ Αρχειοθέτηση	√ Αρχειοθέτηση
Αναζήτηση	√	√	√
Μεταφορά	√	√	√
Εργαλεία Ελέγχου	√	–	–
Άλλες	*1.Χρήση a.k.a για δημιουργία θησαυρού 2.Συγχώνευση*	*1.On-line αποθήκευση 2.Διάσπαση μηνυμάτων 3.Πολλαπλός τρόπος κατηγοριοποίησης 4.Συγχώνευση*	*1.Τροποποίηση μηνυμάτων σε κείμενο 2.Συγχώνευση*

Πίνακας 3-δεύτερος συγκριτικός πίνακας υπηρεσιών (μέρος ΄β)

Πίνακας Συμβατότητας Συστημάτων με Λειτουργική Συστήματα

Λειτουργικά Συστήματα/ Προϊόντα	*Λειτουργικό Σύστημα Server*	*Λειτουργικό Σύστημα Σταθμών Εργασίας*	*Σύστημα διαχείρισης Βάσεων Δεδομένων*	*E-mail Servers/ E-mail Clients*	*Αυτοματοποίηση Γραφείου*
MetaTagger CIS	MS Windows	MS Windows	SQL	MS Exchange Lotus/ MS Outlook Lotus Notes	MS Office Novel GroupWise WordPerfect
ViewDirect	MS Windows (200,2003, XP,98)	MS Windows	SQL Sybase IBM DB2 Oracle 9i MSDE	MS Exchange (2000,2003, 5.5)/ MS Outlook	MS Office (XP,2000,2003) (διάφορες γλώσσες)
eManage	MS Windows	MS Windows (95/98/NT/ 2000)	MS SQL Oracle IBM DB2 Sybase Sybase Anywhere	MS Exchange/ MS Outlook (98/2000)	MS Office
EchoMail	MS Windows Unix	MS Windows NT	Lotus Domino Oracle Sun Solaris	MS Exchange Lotus/ MS Outlook Lotus Notes	MS Windows

Trim	MS Windows 2000 Server UNIX Solaris 8.0	MS Windows (2000/XP)	MS SQL 2000 Oracle 9i	MS Exchange (2000/ 5.5) Lotus 6/ MS Outlook 2000/XP Lotus Notes 6	MS Office (2000/ XP Professional)
Enterprise RME	(δεν είναι γνωστά)				
DB2 Content Management	MS Windows	MS Windows	IBM2 Lotus MS SQL	MS Exchange Lotus Novel GroupWise / MS Outlook Lotus Notes	MS Office Novel GroupWise

Πίνακας 4- Συμβατότητα μεταξύ συστημάτων

Από τους παραπάνω πίνακες μπορούν να εξαχθούν συμπεράσματα αναφορικά με την χρηστικότητα των άνωθεν συστημάτων. Αρχικά, οι δύο πρώτοι πίνακες αναφέρονται στις υπηρεσίες, τις οποίες τα συστήματα εκτελούν. Είναι φανερό, ότι τα σχετικά πακέτα λογισμικού επιτελούν σχεδόν τις ίδιες λειτουργίες όσον αφορά τον κύκλο «σύλληψη-αποθήκευση». Το πρώτο σημείο στο οποίο θεωρητικά διαφοροποιούνται, είναι σχετικά με την υπηρεσία «διάθεση-διατήρηση». Το μόνο σύστημα, που φαίνεται να μην επιτελεί την εν λόγω εργασία είναι της Interwoven, **MetaTagger CIS**, που εντάσσεται στους clients Work/Desk Site. Τα υπόλοιπα συστήματα, επιτελούν την σχετική εργασία βάση κανόνων, που μπορούν να εισαχθούν από την ίδια την εταιρία ή τους χρήστες της ή βάση κανόνων που να είναι συμβατοί με τα σχετικά πρότυπα διατήρησης ηλεκτρονικών αρχείων.

Ένα δεύτερο σημείο, που τα συστήματα διαφοροποιούνται είναι οι εργασίες στις οποίες υπόκεινται αντικείμενα τα ηλεκτρονικά μηνύματα κατά ή μετά τη μεταφορά τους στον κεντρικό server του συστήματος. Όλα τα συστήματα,

μεταφέρουν τα ηλεκτρονικά μηνύματα από τα ηλεκτρονικά γραμματοκιβώτια των χρηστών ή από τον κεντρικό server της εταιρίας σε ένα κοινό αποθηκευτικό χώρο δεδομένων, που αποτελεί κοινό σημείο πρόσβασης και αναζήτησης όλων των δεδομένων της εταιρίας. Κατά τη μεταφορά των μηνυμάτων, αυτόματα πραγματοποιείται και η διαγραφή τους από τον κεντρικό server, προκειμένου να επιτευχθεί αποσυμφόρηση του. Το **eManage,** προχωρά ένα βήμα περισσότερο σε αυτή τη διαδικασία και κατά τη μεταφορά των μηνυμάτων, τα συμπιέζει κατά 80% του αρχικού τους μεγέθους. Παράλληλα, το **ViewDirect,** τα αποθηκεύει παράλληλα και σε οπτικά μέσα, ενώ, η **Enterprise RME** τα αποθηκεύει on-line. Τέλος, το **DB2 Content Management** κατά τη μεταφορά των μηνυμάτων, τα τροποποιεί σε απλό κείμενο και αποθηκεύει τη νέα διάταξη τους στον κοινό αποθηκευτικό χώρο.

Από τους πίνακες, παράλληλα, είναι εμφανές ότι όλα τα συστήματα οργανώνουν τον κοινό αποθηκευτικό χώρο των δεδομένων, οργανώνοντάς τον σε κατηγορίες. Οι σχετικές κατηγορίες μπορούν να δημιουργηθούν αυτόματα ανάλογα με το περιεχόμενο των δεδομένων ή να οριστούν εκ των προτέρων από τους χρήστες. Η κατηγοριοποίηση των μηνυμάτων, στα συστήματα που εξετάστηκαν, βασίζεται στην εξαγωγή μεταδεδομένων από τα πεδία της επικεφαλίδας αλλά και από το περιεχόμενου του σώματος του ηλεκτρονικού μηνύματος. Η **Enterprise RME**, για παράδειγμα, υποστηρίζει πολλαπλούς, παράλληλους τρόπους κατηγοριοποίησης των μηνυμάτων, τόσο βάση των πεδίων της επικεφαλίδας, όσο και σε γενικότερες κατηγορίες, όπως προσωπικά, εργασία κλπ. Επίσης, το ίδιο προϊόν, διασπά, τα μηνύματα από τα συνημμένα τους αρχεία και τα κατηγοριοποιεί ξεχωριστά.

Ένα άλλο πλεονέκτημα, που παρουσιάζει η συγχώνευση των μηνυμάτων με τα λοιπά ηλεκτρονικά και μη τεκμήρια της εταιρίας, είναι ότι οι χρήστες μπορούν εκτελέσουν την αναζήτησή τους σε όλα τα εταιρικά δεδομένα. Ο **MetaTagger** και το **eManage** διευκολύνουν ακόμα περισσότερο τη διαδικασία της αναζήτησης και ανάκτησης των μηνυμάτων, παράγοντας αυτόματα περιλήψεις του περιεχομένου τους. Με αυτό τον τρόπο, ο χρήστης που φυλλομετρά, το σύνολο των δεδομένων, μπορεί να ανατρέξει στην περίληψη του περιεχομένου για να κρίνει αν το σχετικό μήνυμα είναι κατάλληλο για αυτόν. Τα μόνα συστήματα, που δεν επιτελούν συγχώνευση του περιεχομένου είναι το **EchoMail.**

Τέλος, σχετικά με τις υπηρεσίες, που προσφέρουν στους χρήστες τους, σημαντική είναι η υπηρεσία του **Trim Context,** «*audit trails*», με την οποία παρέχει εργαλεία ελέγχου των αρχείων ηλεκτρονικών μηνυμάτων.

Ο τελευταίος πίνακας των συστημάτων που αναφέρθηκαν, περιέχει στοιχεία, σχετικά με την συμβατότητά τους με άλλα λειτουργικά συστήματα του οργανισμού. Η σύγκριση αυτή είναι σημαντική καθώς σχετίζεται με την επιλογή εγκατάστασής του αλλά και με τη διαλειτουργικότητά του με τα άλλα συστήματα διαχείρισης τεκμηρίων. Ένα σύστημα, δηλαδή, που δεν είναι συμβατό με το λογισμικό της Notes, είναι φυσιολογικό να μην ανταποκρίνεται στις απαιτήσεις της εταιρίας, που χρησιμοποιεί το Lotus Notes, ως client των ηλεκτρονικών της μηνυμάτων.

Τα περισσότερα συστήματα που εξετάστηκαν μέχρι τώρα είναι συμβατά μόνο με προϊόντα λογισμικού της Microsoft (MS Exchange, MS Outlook), με εξαίρεση τα **MetaTagger CIS, Trim Context, DB2** και **EchoMail** που είναι συμβατά και με τα προϊόντα της Lotus, ενώ τα δύο τελευταία και με το λογισμικό της Novel GroupWise. Όσον αφορά τη συμβατότητα με Βάσεις Δεδομένων διαφόρων εταιριών, τις περισσότερες επιλογές παρέχει, το **ViewDirect,** το **eManage** και το **EchoMail.** Τέλος, το **EchoMail** και το **Trim Context** είναι συμβατά και με το λειτουργικό σύστημα server Unix.

2.2.2.3 Κατηγοριοποίηση Ηλεκτρονικών Μηνυμάτων σε Μη-πραγματικό χρόνο

Εκτός από τα διάφορα συστήματα διαχείρισης ηλεκτρονικών μηνυμάτων που αναφέρθηκαν, υπάρχει και ένα σύνολο προϊόντων λογισμικού, τα οποία διαχειρίζονται τα ηλεκτρονικά μηνύματα ως τμήμα των δεδομένων της εταιρίας. Οι βασικές λειτουργίες, που επιτελούν τα σχετικά προϊόντα είναι η δημιουργία ταξινομιών, η συγκέντρωση και οργάνωση των δεδομένων και τέλος η κατηγοριοποίησή τους υπό τις κατάλληλες κατηγορίες. Οι ταξινομίες είναι ιεραρχικές ή επίπεδες οντολογίες των δεδομένων μίας εταιρίας ή ενός οργανισμού. Οι κατηγορίες που περιλαμβάνουν μπορούν να οριστούν από τους χρήστες ή από κάποιο ειδικό ή πραγματοποιείται αυτόματα από το σύστημα. Κοινό χαρακτηριστικό τους είναι ότι συχνά αντικατοπτρίζουν τη δομή και τις ανάγκες της εν λόγω εταιρίας. Η χρήση τους διευκολύνει την οργάνωση των δεδομένων, καθώς με αυτό τον τρόπο, διευκολύνεται η πρόσβαση, η αναζήτηση και η ανάκτηση αυτών. Τα ηλεκτρονικά μηνύματα αποτελούν αντικείμενα διαχείρισης των συγκεκριμένων προϊόντων καθώς αποτελούν σημαντικές πηγές δεδομένων.

Ο πίνακας που ακολουθεί αναφέρει τις εταιρίες και τα προϊόντα που θα περιγραφτούν στη συνέχεια, ενώ παράλληλα, δίνει συνοπτική περιγραφή της τεχνολογίας στην οποία στηρίζονται και τις γενικές ιδιότητές τους:

Εταιρία	**Προϊόν**	**Τεχνολογία**	**Ιδιότητες**
Autonomy	**Categorizer\ IDOL server**	Pattern matching, Naïve Bayes & αρχή Shannon Δημιουργία ταξινομίας μέσω clustering.	1. Συγχώνευση Δεδομένων & Διαχείριση 2. Δημιουργία Ταξινομίας
Inxight	**LinguistiX**	Entity Extraction Metadata	1. Υποστήριξη γλωσσολογικής ανάλυσης 2. Δημιουργία Περιλήψεων 3. Εξαγωγή 25 διαφορετικών οντοτήτων 4. Δημιουργία & Διαχείριση ταξινομιών 5. Υποστήριξη 12 γλωσσών & 70 διατάξεων τεκμηρίων
Microsoft	**SharePoint Πύλη Server**	Βαρύτητα όρων Accelerated SMV	1. Ανακάλυψη γνώσης 2. Διαχείριση Δεδομένων Διαφόρων διατάξεων 3. Υποστήριξη διαφορετικών Γλωσσών
Teragram	**Categorization engine**	Linguistic analysis Statistical & Rule-based Algorithms Entity Extraction Incremental Learning	1. Δημιουργία & διαχείριση ταξινομιών 2. Διττός τρόπος κατηγοριοποίησης 3. Ειδοποίηση για εισαγωγή νέων θεμάτων 4. Υποστήριξη Unicode γλωσσών
Stratify	**Classification Server**	Naïve Bayes & keyword rule classifier	1. Ανακάλυψη Γνώσης 2. Υποστήριξη 200

		Δημιουργία ταξινομίας βάση clustering & Εκμάθηση Μηχανής Τεχνικών	διατάξεων δεδομένων 3. Δημιουργία ταξινομίας 4. Εξαγωγή Μεταδεδομένων σε XML διάταξη 5. Υποστήριξη Unicode γλωσσών
Verity	**K2 Enterprise**	Rules & logistic regression Thematic Mapping	1, Ανακάλυψη γνώσης 2. Δημιουργία & Διαχείριση Ταξινομιών
Xerox	**CategoriX**	Γλωσσολογική ανάλυση Probabilistic Model Incremental Learning	1. Υποστήριξη διαφόρων Διατάξεων

Πίνακας 5- Αναφορά συστημάτων και περιγραφή τεχνολογίας.

2.2.2.3.1 Autonomy

Ο ***IDOL Server*** [n] είναι η απάντηση της ***Autonomy*** στο πρόβλημα της κατηγοριοποίησης. Ο *server* είναι ένα το βασικό συστατικό της γενικής υποδομής *IDOL (Intelligent Data Operating Layer)* που παρέχεται από την εταιρία. Βάση αυτής της υποδομής επιτυγχάνεται η συγκέντρωση δομημένων, ημι-δομημένων και αδόμητων πληροφοριών από διάφορους αποθηκευτικούς χώρους στον server και η διαχείριση τους βάση του περιεχομένου τους. Παράλληλα, καθίσταται δυνατή η κατανόηση του περιεχομένου σε επίπεδο θέματος ή έννοιας για κάθε διακριτό τμήμα, υποστηρίζοντας με αυτό τον τρόπο την κατηγοριοποίηση των πληροφοριών. Τέλος, μέσω της υποδομής *IDOL* είναι δυνατός ο αυτόματος συσχετισμός του περιεχομένου με τις προκαθορισμένες κατηγορίες ή με τη δομή άλλων εφαρμογών.

Ο *IDOL Server* παρέχει αυτοματοποιημένη πρόσβαση σε όλων των ειδών τις πληροφορίες που διακινούνται στα πλαίσια ενός οργανισμού. Βάση της κατανόησης και διαχείρισης του περιεχομένου των πληροφοριών, ο *classification server* οργανώνει την ταξινομία, που μπορεί να δημιουργηθεί χειρονακτικά,

αυτόνομα ή βάση ενός συνδυασμού και των δύο. Παράλληλα, παρέχει και ένα άλλο μεγάλο σύνολο υπηρεσιών, όπως είναι αυτές που διακρίνονται στο παρακάτω σχήμα. Από τις υπηρεσίες που παρέχονται, ιδιαίτερη έμφαση θα δοθεί σε αυτή της δημιουργίας ταξινομιών και της κατηγοριοποίησης.

Βάση του χαρακτηριστικού, *taxonomy generation,* ο IDOL *server* μπορεί να κατανοήσει αυτόματα το περιεχόμενο και να δημιουργήσει μία ιεραρχική, νοηματική ταξινομία των πληροφοριών. Η διαδικασία αυτή πραγματοποιείται μέσω clustering, δηλαδή μέσω ομαδοποίησης παρόμοιων δεδομένων από διάφορες πηγές, όπως είναι μη-δομημένα δεδομένα, τα προφίλ των χρηστών και οι agents. Η ταξινομία που δημιουργείται έχει ένα τριπλό ρόλο: παρέχει ταυτόχρονα τη βάση για επισκόπηση των πληροφοριών και εξειδίκευση του περιεχομένου αλλά παράλληλα αποτελεί και την βάση για κατηγοριοποίηση. Αφού πραγματοποιηθεί η δημιουργία των ταξινομιών, ο *server* κατανοεί αυτόματα το είδος των πληροφοριών που πρέπει να διαχειριστεί. Στο σημείο αυτό επιδρά ένα άλλο χαρακτηριστικό, το *automatic taxonomy to category generation* που χρησιμοποιεί τα αποτελέσματα της ταξινομίας για τη δημιουργία κατηγοριών.

Η κατηγοριοποίηση των τεκμηρίων στον IDOL *server* πραγματοποιείται συνδυάζοντας τεχνικές pattern matching, το μοντέλο Naïve Bayes και τις αρχές του θεωρήματος Shannon. Το θεώρημα Shannon αναφέρεται στην ιεράρχηση της σημασίας των ιδεών. Για παράδειγμα, αν σε ένα κείμενο αναφέρονται οι όροι «μαύρο», «θάλασσα», «πουλί» και «δεν πετάει», υπάρχει μεγάλη πιθανότητα, οι όροι αυτοί και συνεπώς η φράση που σχηματίζεται να αναφέρεται σε πιγκουΐνους. Το θεώρημα του Shannon στηρίζεται στην αρχή ότι η παρουσία ή η απουσία διακριτών λέξεων δεν αλλάζει ιδιαίτερα τη πιθανότητα της ιδέας που εκφράζεται .Η αρχή αυτή καθιστά πιο έγκυρη την ταύτιση ιδεών από την ταύτιση λέξεων. Αφού πραγματοποιηθεί η διαδικασία της pattern matching, εισάγονται αυτόματα ετικέτες προσδιορισμού στο σύνολο δεδομένων. Η τελική διαδικασία στα πλαίσια της κατηγοριοποίησης περιλαμβάνει τη δρομολόγηση του περιεχομένου στον χρήστη ή την ειδοποίησή του, στην περίπτωση που εισαχθεί στο σύστημα υλικό που τον ενδιαφέρει βάση του προφίλ του.

Η τρέχουσα έκδοση του IDOL server είναι ο *IDOL server 5.* Οι πλατφόρμες που υποστηρίζει το σύστημα περιλαμβάνουν Microsoft Windows NT4, 2000, XP και

2003, Linux kernel 2.2, 2.4 και 2.6, San Solaris 5-9, AIX 4.3, 5, 5.1, HP-UX 10, 11 και 11.i και τέλος Tru64 5.1.

2.2.2.3.2 Inxight

Η ***Inxight [o]*** παρέχει μία σειρά μηχανών λογισμικού για την ανάλυση και διαχείριση μεγάλων αποθηκευτήριων χώρων όπως είναι το Διαδίκτυο, οι βιβλιοθήκες αρχείων ειδήσεων, ηλεκτρονικών μηνυμάτων και τεκμηρίων. Τα προϊόντα αυτά μπορούν να εισαχθούν σε διάφορες εφαρμογές όπως είναι μηχανές αναζήτησής, συστήματα διαχείρισης περιεχομένου ή σχέσεων πελατών και άλλα συστήματα που σχετίζονται με τη διαχείριση μεγάλων ποσοτήτων δεδομένων σε μορφή κειμένου. Η ***LinguistiX Platform,*** είναι μία μηχανή επεξεργασίας φυσικής γλώσσας, που εισαγόμενη σε άλλες εφαρμογές, μπορεί να προσφέρει τη βάση για την ανάπτυξη πιο έξυπνων λύσεων για αναζήτηση πλήρους κειμένου, για την αυτόματη δρομολόγηση και απάντηση σε συστήματα ηλεκτρονικών μηνυμάτων, για την ηλεκτρονική δημοσίευση αλλά και τη μεταφορά περιεχομένου σε wireless συσκευές. Ο ***Summarizer SDK*** είναι μία άλλη μηχανή λογισμικού που παρέχει τη δυνατότητα παραγωγής περιλήψεων για on-line τεκμήρια σε κλάσματα δευτερολέπτων, ενώ το μέγεθος της περίληψης μπορεί να τροποποιηθεί ώστε να ανταποκρίνεται στις ανάγκες του χρήστη. Δύο βασικά χαρακτηριστικά και των δύο μηχανών είναι ότι στηρίζονται στη μέθοδο εξαγωγής οντοτήτων για την επιτέλεση των λειτουργιών τους, ενώ παρέχουν ένα μοναδικό API για την υποστήριξη 16 διαφορετικών γλωσσών.

Η εξαγωγή οντοτήτων πραγματοποιείται από ένα εργαλείο ανάλυσης κειμένου, το ***SmartDiscovery.*** Το εργαλείο αυτό είναι ικανό να εξάγει 25 διαφορετικές οντότητες χωρίς προηγούμενη εκπαίδευση, βάση της τεχνικής ταύτισης μοτίβων. Οι οντότητες που εξάγονται περιλαμβάνουν ονόματα ανθρώπων, εταιριών, ημερομηνίες, οικονομικά σύμβολα και διευθύνσεις ηλεκτρονικών μηνυμάτων. Συνακόλουθο της εξαγωγής οντοτήτων από το κείμενο είναι η αυτόματη παραγωγή μεταδεδομένων, που αντικατοπτρίζουν το περιεχόμενο των αρχείων του οργανισμού. Στην συνέχεια, τα μεταδεδομένα μπορούν να χρησιμοποιηθούν για ένα πλήθος άλλων εργασιών όπως είναι η αναζήτηση ή κατηγοριοποίηση

τεκμηρίων – συμπεριλαμβανομένων των ηλεκτρονικών μηνυμάτων. Πέρα από την εξαγωγή οντοτήτων, το *SmartDiscovery* παρέχει και ένα πλήθος άλλων υπηρεσιών, όπως είναι η δημιουργία και η διαχείριση ταξινομιών και η δρομολόγηση τεκμηρίων βάση του τύπου, του θέματος ή του περιεχομένου του τεκμηρίου. Επίσης, παρέχει τη δυνατότητα αναζήτησης και ανάκτησης τεκμηρίων, της επισκόπησης των αποτελεσμάτων αναζήτησης και την φυλλομέτρηση και πλοήγηση σε σύνολα δεδομένων.

Η κατηγοριοποίηση των ηλεκτρονικών τεκμηρίων και μηνυμάτων πραγματοποιείται στην ταξινομία που έχει δημιουργηθεί από το *SmartDiscovery* και βασίζεται στα μεταδεδομένα που έχουν δημιουργηθεί από την εξαγωγή οντοτήτων. Τα μεταδεδομένα, δεδομένου ότι έχουν εξαχθεί από το ίδιο το κείμενο που κατηγοριοποιείται, παρέχουν μεγάλο ποσοστό ακρίβειας στην σχετική διαδικασία.

Το εργαλείο που πραγματοποιεί την κατηγοριοποίηση στα πλαίσια του *SmartDiscovery* είναι ο *Categorizer.* Στη νέα του έκδοση *Categorizer* υποστηρίζει 12 διαφορετικές γλώσσες και 70 διαφορετικές διατάξεις τεκμηρίων, όπως είναι τα HTML, PDF και τα ηλεκτρονικά μηνύματα. Παράλληλα, υποστηρίζει μία ποικιλία λειτουργικών συστημάτων, όπως είναι τα MS Windows NT & XP, το Sun Solaris 2.6, 2.7 και το Red Hat Linux 2.7. Τέλος, οι βάσεις δεδομένων που υποστηρίζονται είναι η Oracle, Sybase και η βάση δεδομένων του MS SQL server.

2.2.2.3.3 Microsoft

Ο **SharePoint Πύλη Server** της **Microsoft** [p] παρέχει στους χρήστες του ένα γενικό περιβάλλον για τη διαχείριση των πληροφοριών και την ανακάλυψη γνώσης, περιλαμβάνοντας πτυχές όπως η αποθήκευση, η αναζήτηση, η επισκόπηση, η δημοσίευση και η διανομή των πληροφοριών. Στα πλαίσια της αρχιτεκτονικής του περιλαμβάνεται μία μηχανή αναζήτησης, ικανή να αναζητήσει και να ανακτήσει πληροφορίες από μία πληθώρα μορφών διατάξεων των

δεδομένων (συμπεριλαμβανομένων των ηλεκτρονικών μηνυμάτων) που μπορεί να είναι αποθηκευμένα σε ένα ή περισσότερους servers. Για να επιτευχθεί η βελτίωση της αποτελεσματικότητάς της όσον αφορά τις δύο τελευταίες λειτουργίες εφαρμόζει διάφορες μαθηματικές και υπολογιστικές τεχνικές, όπως είναι το *υπολογισμός ομοιότητας* και η *αυτόματη κατηγοριοποίηση.*

Ο *υπολογισμός ομοιότητας* ή *relevance scoring* είναι η διαδικασία ανάλυσης του περιεχομένου ενός τεκμηρίου προκειμένου να καθοριστεί το ποσοστό ομοιότητας του με το ερώτημα, που έθεσε ο χρήστης. Ο καθορισμός της ομοιότητας πραγματοποιείται βάση του υπολογισμού πιθανοτήτων από τον αλγόριθμο «Okapi». Ο σχετικός αλγόριθμος αναπτύχθηκε στα πλαίσια ερευνών της ίδιας της Microsoft και θεωρήθηκε ένας από τους πλέον αποτελεσματικούς στην TREC (Text Retrieval Conference) 2002. Για να επιλεχθεί το κατάλληλο τεκμήριο από ένα γενικό σύνολο, ο αλγόριθμός χρησιμοποιεί την τεχνική «βαρύτητα όρων» στους όρους που έθεσε ο χρήστης στο πεδίο της αναζήτησης. Σε κάθε ένα από αυτούς τους όρους εφαρμόζεται ένα «βάρος» και κατά την φυλλομέτρηση των τεκμηρίων από τη μηχανή, κάθε όρος αναζήτησης «ζυγίζεται» στο τεκμήριο βάση διαφορετικών μέτρων (συχνότητα συλλογής, συχνότητα όρου, μέγεθος τεκμηρίου, θέση στο τεκμήριο). Στο πέρας της αναζήτησης κάθε ανακτημένο έγγραφο ιεραρχείται βάση καταλληλότητας ανάλογα με το σύνολο των «βαρών».

Η *κατηγοριοποίηση* πραγματοποιείται όταν το τεκμήριο καταχωρείται στον αποθηκευτικό χώρο του *SharePoint Πύλη Server* και βασίζεται σε ένα τροποποιημένο μοντέλο των *Support Vector Machines*, ενώ διευκολύνεται από την χρήση μίας ειδικής διεπιφάνειας χρήστη. Η αυτόματη κατηγοριοποίηση πραγματοποιείται μέσω της εξής διαδικασίας: αρχικά ο χρήστης ή ο διαχειριστής του συστήματος ορίζει μία ιεραρχία. Στην ιεραρχία αυτή μπορεί να ορίσει τις επιθυμητές για αυτόν κατηγορίες, όπως είναι π.χ. δουλειά, προσωπικά ή πεδία της επικεφαλίδας των ηλεκτρονικών μηνυμάτων. Στην συνέχεια το σύστημα επεξεργάζεται ένα μικρό σύνολο προ-ταξινομημένων τεκμηρίων ή μηνυμάτων για κάθε κατηγορία. Από τα τεκμήρια αυτά, ο SharePoint Πύλη Server δημιουργεί μία ταξινομία, που χρησιμοποιείται για να προβλέψει βάση των Support Vector Machines την κατηγορία στην οποία εντάσσονται τα νέα τεκμήρια. Κάθε τεκμήριο, μπορεί να ανήκει σε μία, σε πολλαπλές ή σε καμία κατηγορία. Μετά την σχετική

εκπαίδευση, ο SharePoint Πύλη Server είναι ικανός να κατηγοριοποιεί τα τεκμήρια αυτόματα.

Το τελικό αποτέλεσμα της διαδικασίας κατηγοριοποίησης είναι μία ιεραρχία τεκμηρίων- περιλαμβάνοντας ηλεκτρονικά τεκμήρια, μηνύματα, εικόνες, video clips- από όπου ο χρήστης μπορεί να βρει εύκολα και γρήγορα το κατάλληλο τεκμήριο. Θα πρέπει να σημειωθεί ότι η κατηγοριοποίηση πραγματοποιείται βάση του περιεχομένου ή των ιδιοτήτων του τεκμηρίου. Τέλος, το μοντέλο των Support Vector Machines που χρησιμοποιείται, αναπτύχθηκε, επίσης στα πλαίσια των ερευνών της Microsoft, ονομάζεται *accelerated SVM* και έχει απονεμηθεί ως μία βασικές πατέντες (αριθμός US6327851) του 2002 από το M.I.T περιοδικό.

Η αναζήτηση στον *SharePoint Πύλη Server* πραγματοποιείται μέσω *Protocol Handlers* και *I-Filters (Information filters).* Οι *Protocol Handlers* υποστηρίζουν την πρόσβαση σε διάφορα συστήματα, όπως είναι δικτυακοί ιστότοποι του Intranet/Internet, ο Exchange server ή Lotus Notes και FTP ιστότοποι. Παράλληλα, τα *I-Filters,*είναι φίλτρα που εξάγουν πληροφορίες από συγκεκριμένες διατάξεις τεκμηρίων, όπως είναι τα έγγραφα Word, αρχεία TIFF, XTML τεκμήρια και τα ηλεκτρονικά μηνύματα. Οι εξαγόμενες πληροφορίες οργανώνονται σε ένα ανεστραμμένο ευρετήριο που διασυνδέει τις πληροφορίες με τα τεκμήρια από πού εξήχθησαν. Επίσης, γίνεται και χρήση θησαυρού, όπου πραγματοποιείται η συγκέντρωση των όρων που χρησιμοποιούνται στα ερωτήματα των χρηστών. Παράλληλα, ο *SharePoint Πύλη Server* υποστηρίζει πληθώρα γλωσσών όπως είναι τα γαλλικά, γερμανικά, ιταλικά, ισπανικά, Ιαπωνικά, ταϊλανδέζικα κ.α. , χωρίς, όμως να γίνεται κάποια αναφορά για ελληνικά.

Η τρέχουσα έκδοση του *SharePoint Πύλη Server* είναι η *SharePoint Πύλη Server 2003.* Η νέα έκδοση παρουσιάζει διάφορες βελτιώσεις συγκριτικά με την έκδοση του 2001 και διάφορες παρατηρούνται στο σύστημα κατηγοριοποίησης, οι οποίες διαφαίνονται στον παρακάτω πίνακα []. (Ο πίνακας αναφέρεται μόνο στις τροποποιήσεις που πραγματοποιήθηκαν στην κατηγοριοποίηση.)

Συστατικό	(N)[39] ή (B)	Όνομα	Περιγραφή
Categories	(B)	**Category structure and site navigation**	Εύκολη διαχείριση των κατηγοριών της πύλης[40] και των θεμάτων του φυλλομετρητή. Δημιουργία, μετακίνηση, μετονομασία και διαγραφή κατηγοριών χρησιμοποιώντας τον site map της πύλης.
	(N)	**Rich category areas**	Ο ιστότοπος της πύλης είναι μία ιεραρχία υπό-ιστοτόπων που επιτρέπουν στους διαχειριστές των κατηγοριών να προσθέτουν λίστες, εικόνες και τεκμήρια στις κατηγορίες. Οι διαχειριστές έχουν τον έλεγχο της ασφάλειας και του περιεχομένου.
	(N)	**Advanced category security**	Εφαρμογή επιπλέον ασφάλειας στις κατηγορίες.
	(B)	**Category Assistant**	Η πύλη προτείνει αντικείμενα για να εισαχθούν υπό μία κατηγορία, που ο διαχειριστής μπορεί να επιλέξει ή να απορρίψει. Καθώς, κατηγορίες εισάγονται στον ιστότοπο, ο Category Assistant μαθαίνει τις κατηγορίες και μπορεί να προτείνει.
	(B)	**New page templates**	Δημιουργία νέων φορμών κατηγοριών με όλα τα χαρακτηριστικά των υπηρεσιών του Windows Share Point
	(N)	**Category object model**	Δημιουργία κατηγοριών προγραμματικά χρησιμοποιώντας ένα τροποποιήσιμο μοντέλο αντικειμένου.
	(N)	**Suggested links for πύλη categories**	Οι χρήστες μπορούν να προτείνουν περιεχόμενο για τις κατηγορίες, που ο διαχειριστής μπορεί να αποδεχτεί ή να απορρίψει.

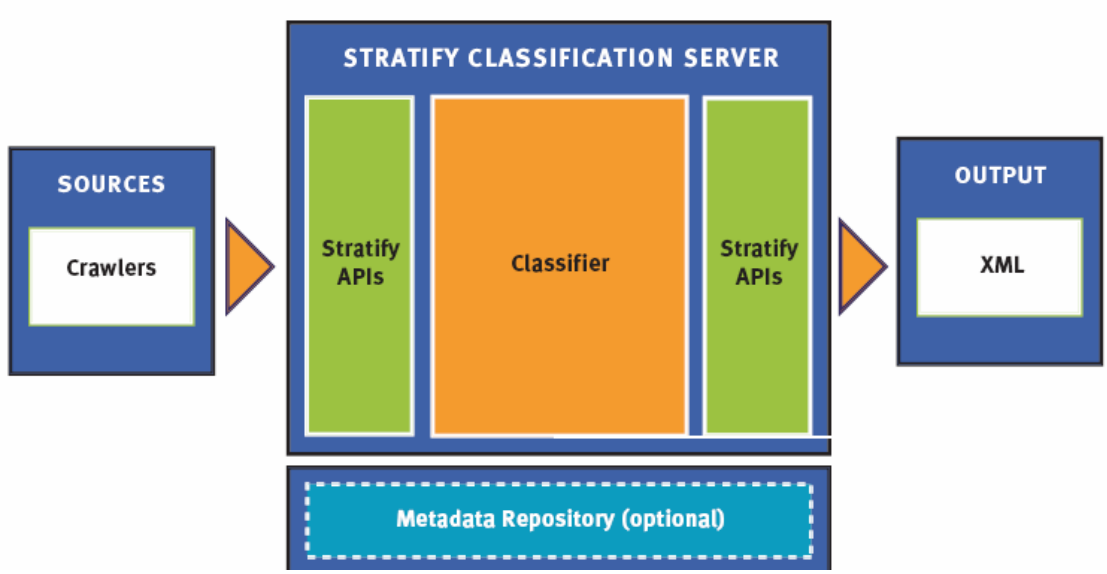
Πίνακας 6-Διαφοροποιήσεις στη νέα έκδοση του SharePoint Πύλη Server

[39] Τα γράμματα (N) & (B), χρησιμοποιούνται ως αντικατάσταση των όρων Νέο & Βελτιωμένο, αντίστοιχα.
[40] portal

2.2.2.3.4 Stratify

Ο ***Classification Server*** της ***Stratify*** [r] είναι ένα από τα συστατικά του συστήματος ανακάλυψης πληροφοριών της εταιρίας (discovery system) αλλά μπορεί να εισαχθεί και σε υπάρχοντα συστήματα διαχείρισης πληροφοριών και τεκμηρίων. Η εισαγωγή του διευρύνει τις ικανότητες των σχετικών εφαρμογών, όπως είναι οι μηχανές αναζήτησης, τα συστήματα διαχείρισης περιεχομένου και τεκμηρίων, οι εταιρικές πύλες κ.α. και τις καθιστά ικανές τόσο να διαχειριστούν μεγαλύτερο ποσοστό μη-δομημένων δεδομένων όσο και να το παρουσιάσουν στον χρήστη. Η εισαγωγή του σε τέτοια συστήματα γίνεται βάση της ανοιχτής, αρθρωτής αρχιτεκτονικής του και η διασύνδεση του πραγματοποιείται μέσω Java ή C++ APIs. Ειδικά για εφαρμογές που σχετίζονται με το διαδίκτυο, η διασύνδεση γίνεται μέσω APIs που υποστηρίζουν το πρωτόκολλο SOAP και WSDL.

Ο *Classification Server* είναι ικανός να κατηγοριοποιήσει τεκμήρια από περίπου 200 διαφορετικές διατάξεις, όπως είναι τεκμήρια σε HTML, PDF, Microsoft Office αλλά και ηλεκτρονικά μηνύματα ή άλλα τεκμήρια απλού κειμένου. Η διαδικασία της κατηγοριοποίησης βασίζεται στην συγκέντρωση των τεκμηρίων - προαιρετικά με την χρήση crawlers- στο σύστημα του server και σε διάφορες τεχνικές κατηγοριοποίησης που χρησιμοποιούνται. Συγκεκριμένα, ο *Classification Server* χρησιμοποιεί ένα συνδυασμό Naïve Bayes και keyword-rule ταξινομητές. Οι crawlers συγκεντρώνουν τεκμήρια οποιασδήποτε διάταξης από τους αποθηκευτικούς χώρους ή από το διαδίκτυο και εξάγουν το κείμενο από αυτά. Τα δεδομένα αυτά εισάγονται στον *classification server,* ο οποίος προσδιορίζει τις βασικές ιδέες στο κείμενο και ταξινομεί τα τεκμήρια και εξωτερικές πληροφορίες σε μία ιεραρχία. Τα δεδομένα εξόδου του *classification server* είναι μεταδεδομένα σε XML διάταξη. Τα μεταδεδομένα αυτά, που περιγράφουν τα τεκμήρια που κατηγοριοποιήθηκαν, αποθηκεύονται σε ένα metadata repository, που είναι μία SQL Βάση Δεδομένων.



Εικόνα 10- Λειτουργία επεξεργασίας του Classification server

Ο Classification server της Stratify μπορεί να κατηγοριοποιήσει τεκμήρια σε όλες τις Ινδοευρωπαϊκές γλώσσες και στα Αραβικά και είναι συμβατός με το πρότυπο Unicode. Παράλληλα, συλλέγει τεκμήρια από συστήματα αρχείων, από Lotus Notes και Microsoft Exchange servers, από το intranet και από το internet, ενώ είναι συμβατός και με SQL και Oracle Βάσεις Δεδομένων.

2.2.2.3.5 Teragram

Η λύση της **Teragram** [q] για τη δημιουργία και τη διαχείριση ταξινομιών συνίσταται από την παροχή διαφόρων εργαλείων λογισμικού, τα οποία μπορούν να εισαχθούν σε ένα οργανισμό ή σε μία εταιρία, τόσο ως μία ενιαία λύση ή ως διακριτά εργαλεία. Τα προϊόντα αυτά, αναφορικά είναι: **TK240**, **Teragram Taxonomy Manager**, **Teragram Editor Workbench**, **Teragram Entities Extractor** και **Teragram Real time Alerts**.

Το **Teragram TK240** είναι ένα επιπρόσθετο συστατικό στο λογισμικό διαχείρισης ταξινομιών και στη μηχανή αναζήτησης της Teragram. Το *TK240* υιοθετεί την τεχνολογία κατηγοριοποίησης, που έχει αναπτυχθεί στα πλαίσια της εταιρίας, για την ανάλυση του περιεχομένου των τεκμηρίων και την οργάνωσή τους σε

επίπεδες ή ιεραρχικές ταξινομίες προκειμένου να διευκολυνθεί και να επιταχυνθεί η ανάκτηση πληροφοριών. Η κατηγοριοποίηση βασίζεται στη διαδικασία εξαγωγής οντοτήτων. Η νέα έκδοση του *Teragram TK240 v5 [],* που πραγματοποιήθηκε το 2005, παρέχει ένα νέο οπτικό περιβάλλον, όπου οι διαχειριστές μπορούν εύκολα και γρήγορα να εισάγουν νέες κατηγορίες ή έννοιες στην ταξινομία, ώστε να είναι πάντα ανανεωμένη σύμφωνα με τις απαιτήσεις των χρηστών.

Ο ***Teragram Taxonomy Manager*** και ο **Categorizer Administrator** είναι δύο εργαλεία τα οποία παρέχονται στα πλαίσια του λογισμικού κατηγοριοποίησης. Ο *Taxonomy Manager* είναι μία διεπιφάνεια client/server και μία υποκείμενη μηχανή αναζήτησης για πλοήγηση στη δομή της ταξινομίας. Με το εργαλείο αυτό, οι χρήστες έχουν τη δυνατότητα να πλοηγηθούν στις πληροφορίες που βρίσκονται οργανωμένες σε κατηγορίες και υπό-κατηγορίες μίας ιεραρχικής δομής. Με την πλοήγηση στην ιεραρχία είναι προσβάσιμες τόσο οι κατηγορίες και οι υποδιαιρέσεις τους αλλά και το περιεχόμενό τους. Το περιεχόμενο των κατηγοριών μπορεί να περιλαμβάνει πληροφορίες, ηλεκτρονικά τεκμήρια ή μηνύματα αλλά και web sites. Παράλληλα, ο *Categorizer Administrator* παρέχει τα αναγκαία εργαλεία για τη δημιουργία, τη διαγραφή και την τροποποίηση κατηγοριών για τον Automatic και τον Rule-based Categorizer. Το λογισμικό αυτό, παρέχεται ως Windows εφαρμογή ή ως σειρά εργαλείων βάση του web ή της Java.

Στα πλαίσια της μηχανής αναζήτησης δρα ένα άλλο εργαλείο, ο **Categorizer**, που ταξινομεί τα τεκμήρια και οργανώνει τις πληροφορίες σε κατηγορίες που ικανοποιούν τη δομή του οργανισμού. Το σχετικό εργαλείο κατηγοριοποίησης εμφανίζεται σε δύο μορφές ως ***Automatic Categorizer*** για την αυτόματη κατηγοριοποίηση των τεκμηρίων και ως ***Teragram rule-based categorizer*** για την κατηγοριοποίηση βάση κανόνων. Η λύση της αυτόματης κατηγοριοποίησης χρησιμοποιείται γιατί είναι ο πιο γρήγορος τρόπος οργάνωσης των τεκμηρίων του οργανισμού. Ο *Automatic Categorizer* μαθαίνει αυτόματα να κατηγοριοποιεί τεκμήρια από παραδείγματα τεκμηρίων που έχουν ήδη κατηγοριοποιηθεί. Η τεχνική που χρησιμοποιείται για την κατανόηση του περιεχομένου και την συνακόλουθη ταξινόμησή του είναι βάση γλωσσολογικής και σημασιολογικής ανάλυσης των νοημάτων και της δομής των λέξεων και φράσεων. Παράλληλα, είναι ενσωματωμένα λεξικά όρων και εγχειρίδια γραμματικής των διαφόρων

γλωσσών που υποστηρίζονται. Η διαδικασία της εξαγωγής σχετικών γλωσσολογικών χαρακτηριστικών από τα τεκμήρια γίνεται αυτόματα και στην συνέχεια επακολουθεί διασύνδεση τον χαρακτηριστικών με τις κατηγορίες που σχετίζονται.

Ο *Teragram rule-based Categorizer* χρησιμοποιείται για την καλύτερη οργάνωση των τεκμηρίων βάση κανόνων των ειδικών, παρέχοντας με αυτό τον τρόπο, μεγαλύτερη ακρίβεια. Βάση αυτού του εργαλείου η κατηγοριοποίηση βασίζεται σε στατιστικούς και βάση κανόνων αλγόριθμους. Κάθε κατηγορία με ένα σύνολο κανόνων που περιγράφουν τα τεκμήρια που εντάσσονται σε αυτήν. Οι κανόνες μπορούν να οριστούν στο επίπεδο λέξεων αλλά και να αναφέρονται σε γλωσσολογικά χαρακτηριστικά των λέξεων. Παράλληλα, οι κανόνες διευκολύνουν την συνεχή και σταδιακή (incremental) δημιουργία και τροποποίηση των κατηγοριών. Τέλος, παρέχονται από την Teragram προκαθορισμένοι κανόνες και κατηγορίες, οι οποίες μπορούν να τροποποιηθούν ή να εισαχθούν νέες.

Ο ***Teragram Entity Extraction*** είναι ένα λογισμικό εξαγωγής οντοτήτων από μη-δομημένα δεδομένα, όπως είναι το σώμα των ηλεκτρονικών μηνυμάτων. Οι οντότητες που μπορούν να εξαχθούν είναι ονόματα ανθρώπων, τοποθεσιών, εταιριών, προϊόντων αλλά και γεγονότα ή πράγματα. Παράλληλα, παρέχονται τρεις τόποι προσδιορισμού των οντοτήτων, που περιλαμβάνουν τον προσδιορισμό μέσω κανόνων, μέσω γραμματικής ή βάση αρχείου καθιερωμένων όρων.

Ο ***Teragram Real Time Alerts*** είναι μία μηχανή ειδοποίησης για την εισαγωγή νέων τεκμηρίων ή νέων ηλεκτρονικών μηνυμάτων στο σύστημα που αφορούν ένα συγκεκριμένο θέμα. Το τελευταίο εργαλείο της σειράς, ο ***Teragram Editor Workbench*** είναι μία web εφαρμογή, στην διεπιφάνεια της οποίας υποστηρίζεται η αναζήτηση και η αυτόματη περίληψη πληροφοριών, ενώ παράλληλα υποδεικνύει προτεινόμενες ή σχετικές κατηγορίες.

Τέλος, τα εργαλεία της σειράς κατηγοριοποίησης της *Teragram* υποστηρίζουν 25 διαφορετικές γλώσσες (ασιατικές και ευρωπαϊκές), οι οποίες είναι συμβατές με το Unicode. Παράλληλα, τρέχουν σε πλατφόρμες Windows, Macintosh και Unix. Η εισαγωγή των εργαλείων ως διακριτά συστατικά και η εφαρμογή τους με άλλα

συστήματα γίνεται μέσω API της C και της Java, ενώ σε κάποιες περιπτώσεις και με την Perl ή τη JavaScript.

2.2.2.3.6 Verity

Η ***K2 Enterprise*** της ***Verity*** [s] παρουσιάζει ένα σύνολο τεχνολογικών ικανοτήτων για την εφαρμογή στρατηγικών ταξινόμησης σε έναν οργανισμό. Τα συστατικά της στοιχεία περιλαμβάνουν υπηρεσίες ανακάλυψης περιεχομένου, προκαθορισμένες δομές ταξινομιών, εργαλεία δημιουργίας ταξινομιών, τεχνολογίες ανακάλυψης και χαρτογράφησης θεμάτων, ταξινόμηση βάση κανόνων, ανάπτυξη πολλαπλών σχεσιακών ταξινομιών, οθόνες δυναμικών ταξινομιών καθώς και δυνατότητες διατήρησης και διαχείρισης ταξινομιών.

Παράλληλα, η *K2 Enterprise* υποστηρίζει διάφορες τεχνικές δημιουργίας ταξινομιών, όπως είναι η χειρονακτική δημιουργία, η αυτόματη δημιουργία και ο συνδυασμός αυτόματης και χειρονακτικής δημιουργίας ταξινομιών. Πριν όμως εισαχθούν τα τεκμήρια στις διάφορες κατηγορίες της ταξινομίας, θα πρέπει να δημιουργηθεί ένα μοντέλο προσδιορισμού του χαρακτήρα κάθε κατηγορίας. Τα μοντέλα αυτά μπορούν να δημιουργηθούν με διάφορους τρόπους, όπως είναι η αυτόματη δημιουργία κανόνων, η εισαγωγή κανόνων από προϋπάρχοντα μοντέλα ταξινομιών καθώς και προσδιορισμός των κανόνων από ειδικούς. Η ένταξη των τεκμηρίων στις κατηγορίες γίνεται αυτόματα και ταυτόχρονα με την ευρετηρίαση των τεκμηρίων.

Για την πραγματοποίηση της ταξινόμησης περιλαμβάνεται στην *K2 Enterprise* ένα επιπλέον συστατικό, που υποστηρίζει τις δυνατότητές της όσον αφορά την οργάνωση τεκμηρίων. Το συστατικό αυτό είναι ο ***Verity Intelligent Classifier.*** Βάση του ταξινομητή αυτού η διαδικασία της κατηγοριοποίησης μπορεί να επιτευχθεί με τρεις τρόπους:

- Αυτόματη Κατηγοριοποίηση- Θετικά και αρνητικά παραδείγματα μπορούν να χρησιμοποιηθούν για να δημιουργήσουν αυτόματα τους κανόνες που θα προσδιορίσουν τις κατηγορίες. Η Verity χρησιμοποιεί για τη μέθοδο αυτή το μοντέλο της *logistic regression*.

- Business Rules- Ειδικοί μπορούν να δημιουργήσουν χειρονακτικά νέους κανόνες ή να τροποποιήσουν εισαγόμενους κανόνες ή κανόνες που δημιουργήθηκαν αυτόματα, προκειμένου να βελτιωθεί η ακρίβεια ή να καλυφθούν οι ανάγκες του οργανισμού.

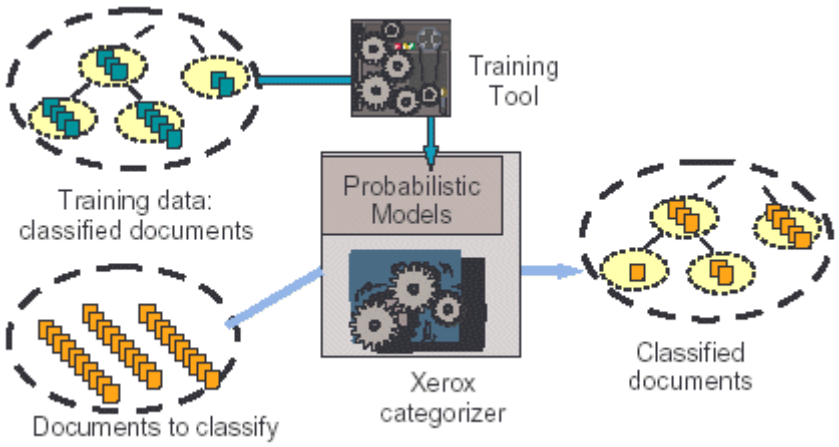
- Εξαγωγή νοημάτων- Το συστατικό *Thematic Mapping* μπορεί να χρησιμοποιηθεί για να αναλύσει το περιεχόμενο ενός συνόλου τεκμηρίων, προκειμένου να ανακαλύψει θέματα και νοήματα. Αυτό μπορεί να χρησιμοποιηθεί για να δημιουργήσει μία νέα ταξινομία ή για να διασπάσει μία κατηγορία σε υποδιαιρέσεις. Μόλις το συστατικό αυτό προσδιορίζει τα νοήματα, τα προσδιορίζει με τέτοιο τρόπο ώστε να μπορούν να γίνουν κατανοητά από ανθρώπους.

2.2.2.3.7 Xerox

Ο ***CategoriX*** είναι η πρόταση της ***Xerox [t]*** για διαχείριση των ηλεκτρονικών τεκμηρίων, με βασικό σκοπό την κατηγοριοποίησή τους. Παράλληλα, παρέχει και άλλες υπηρεσίες όπως είναι η φυλλομέτρηση, η αναζήτηση και το φιλτράρισμα των πληροφοριών σε μεγάλες συλλογές δεδομένων. Πρόκειται για ένα λογισμικού, το οποίο μπορεί να ενσωματωθεί σε οποιοδήποτε σύστημα διαχείρισης τεκμηρίων, ενώ οι διατάξεις των τεκμηρίων, που μπορεί να διαχειριστεί περιλαμβάνουν την XTML, XML, απλού κειμένου και τέλος την διάταξη, σύνολο λέξεων- με την οποία συχνά αναπαρίστανται τα ηλεκτρονικά τεκμήρια, προκειμένου να γίνουν κατανοητά από τους ταξινομητές.

Ουσιαστικά, ο *CategoriX* αποτελείται από δύο διακριτά εργαλεία, ένα training εργαλείο, το οποίο εκπαιδεύεται στην εκμάθηση μοντέλων από τεκμήρια που

έχουν κατηγοριοποιηθεί εκ των προτέρων και από το βασικό εργαλείο κατηγοριοποίησης, το οποίο συγκρίνει τα νέα τεκμήρια με τα αρχικά μοντέλα, προκειμένου να καθορίσει την κατηγορία υπό την οποία θα ενταχθούν.



Εικόνα 11- Λειτουργία του Xerox Categorizer

Η διαδικασία της κατηγοριοποίησης πραγματοποιείται βάση της διαδικασίας που διαφαίνεται στο παραπάνω σχήμα. Η τεχνολογία που χρησιμοποιείται βασίζεται στην τεχνική της γλωσσολογικής ανάλυσης, που αναπτύχθηκε στα πλαίσια ερευνών της *Xerox*. Οι κατηγορίες βρίσκονται ιεραρχημένες βάση μίας ταξινομίας, που μπορεί να δημιουργηθεί από τον οργανισμό ή να εισαχθεί, ενώ παράλληλα, μπορεί να είναι επίπεδη ή ιεραρχική. Ο *CategoriX* ορίζει μία τιμή εμπιστοσύνης σε κάθε τεκμήριο που κατηγοριοποιεί, όπως και οι agents, ανάλογα με την ικανότητα του να ταυτίσει το συγκεκριμένο τεκμήριο με κάποιο από τα μοντέλα. Στην περίπτωση που το εμπιστοσύνης του είναι χαμηλό, επιτρέπει στον χρήστη να ελέγξει και να διορθώσει την κατηγοριοποίηση. Παράλληλα, υποστηρίζεται το incremental learning που καθιστά τον *CategoriX* ιδιαίτερα αποτελεσματικό, ενώ επιτρέπει την εύκολη εισαγωγή και τροποποίηση των κατηγοριών.

Το σύστημα του *CategoriX* είναι γραμμένο σε Java και μπορεί να εφαρμοσθεί σε διάφορες πλατφόρμες, όπως είναι τα Windows, το UNIX, Linux και σε Java 1.4.2 η μεγαλύτερη.

Αξιολόγηση

Η αξιολόγηση των τελευταίων συστημάτων της συγκεκριμένης έρευνας, θα πραγματοποιηθεί πάνω στα δεδομένα του πίνακα 4. Τα στοιχεία αυτά αναφέρονται πάνω στις μεθόδους κατηγοριοποίησης και δημιουργίας ταξινομιών που υιοθετούνται από κάθε προσέγγιση αλλά και στις δυνατότητες κάθε συστήματος. Σκοπός της σύγκρισης, δεν είναι η επιλογή του καλύτερου συστήματος αλλά η σύγκρισή τους, στις κοινές ιδιότητες που παρουσιάζουν.

Τα συγκεκριμένα συστήματα στο σύνολό τους υποστηρίζουν μία νέα υπηρεσία, την «ανακάλυψη γνώσης», η οποία συνίσταται στην συγκέντρωση διαφόρων διατάξεων υλικού από διαφορετικές και κατανεμημένες πηγές δεδομένων σε ένα κεντρικό αποθηκευτικό χώρο, οργανωμένο υπό κατηγορίες. Στα πλαίσια της υπηρεσίας αυτής εφαρμόζονται οι σχετικές διαδικασίες διαχείρισης των τεκμηρίων και η δημιουργία ταξινομιών, υπό τις κατηγορίες της οποίας θα ενταχθεί το υλικό.

Όπως φαίνεται και από την στήλη του πίνακα «Τεχνολογία», τα διάφορα συστήματα χρησιμοποιούν διαφορετικές προσεγγίσεις για τη δημιουργία ταξινομιών. Η **Autonomy** και η **Stratify** υιοθετούν τη μέθοδο «clustering», γεγονός που προσδίδει μεγάλο ποσοστό αυτοματοποίησης της διαδικασίας. Αντίθετα, στον **SharePoint** της **Microsoft,** απαιτείται η χειρωνακτική δημιουργία των κατηγοριών, προτού το σύστημα να είναι ικανό να δημιουργήσει την ταξινομία. Σε ένα ενδιάμεσο πλαίσιο, κινείται η **K2 Enterprise** της **Verity,** που παρέχει τρεις τρόπους για τη δημιουργία των ταξινομιών, αυτόματα, χειρονακτικά ή μία υβριδική προσέγγιση των δύο προηγούμενων. Τέλος, η δημιουργία ταξινομιών στην **Categorization Engine** της **Teragram** γίνεται αυτόματα ή βάση κανόνων, παράλληλα με τη διαδικασία της κατηγοριοποίησης. Παράλληλα, τα συστήματα ακολουθούν διαφορετικές προσεγγίσεις για την εκτέλεση της διαδικασίας της κατηγοριοποίησης. Για να πραγματοποιηθεί μία σύγκριση στις χρησιμοποιούμενες μεθόδους κατηγοριοποίησης, υιοθετούνται από την συγκεκριμένη έρευνα, τα αποτελέσματα των ερευνητικών πειραμάτων του *Sebastiani*, όπως αυτά αναφέρονται στο άρθρο του «*Εκμάθηση Μηχανής in*

Automated Text Categorization». Βέβαια, ο *Sebastiani* στην έρευνά του, συγκρίνει μόνο τους ταξινομητές βάση της Εκμάθηση Μηχανής τεχνικής μεταξύ τους και όχι με άλλες μεθόδους, όπως είναι η κατηγοριοποίηση βάση μεταδεδομένων. Όπως έχει ήδη αναφερθεί, ο ίδιος είναι ενάντια στην χρήση μεταδεδομένων και γενικά εξωτερικών παραγόντων για την κατηγοριοποίηση κειμένου. Παρόλα αυτά, η χρήση μεταδεδομένων στη σχετική διαδικασία είναι μία πολλά υποσχόμενη μέθοδος, που χρησιμοποιείται ευρέως για τη διαχείριση τεκμηρίων.

Σύμφωνα, με τα αποτελέσματα του *Sebastiani*, όπως έχει ήδη αναφερθεί την καλύτερη επίδοση όσον αφορά την ακρίβεια και την ανάκληση παρουσιάζει το μοντέλο των **Support Vector Machines** και η μέθοδος **Regression**, που υιοθετούνται αντίστοιχα από τον **SharePoint Πύλη Server** και την **K2 Enterprise.** Αντίθετα, το μοντέλο **Naïve Bayes,** που υιοθετείται από την **Autonomy** και από την **Stratify,** είναι αυτό που στα σχετικά πειράματα, αποδίδει λιγότερο καλά. Σημαντικό είναι να σημειωθούν και οι προσεγγίσεις που ακολουθούν το μοντέλο του **Incremental Learning,** όπως αυτή της **Teragram** και του **Xerox.** Τα πλεονεκτήματα της σχετικής μεθόδου έχουν ήδη αναφερθεί και αφορούν τη συνεχή εκπαίδευση και εκμάθηση του ταξινομητή στις αλλαγές που πραγματοποιούνται στις κατηγορίες και στο περιεχόμενό τους. Όσον αφορά, τις μεθόδους της γλωσσολογικής ανάλυσης, της αναγνώρισης μοτίβων και της εξαγωγής μεταδεδομένων δεν μπορεί να πραγματοποιηθεί κάποια σύγκριση με τις μεθόδους της Εκμάθηση Μηχανής όσον αφορά την ακρίβεια και την ανάκληση αλλά όσον αφορά το ποσοστό της αυτοματοποίησης. Δεδομένου ότι οι συγκεκριμένοι μέθοδοι εντάσσονται υπό την γενική κατηγορία «μη-Με επίβλεψη

κατηγοριοποίηση», συνεπάγεται ότι προσφέρουν μεγαλύτερο επίπεδο αυτοματοποίησης της διαδικασίας από τις μεθόδους της Εκμάθηση Μηχανής.

Παράλληλα, αναφορικά με τα ίδια τα συστήματα, θα πρέπει να υπογραμμιστούν αρχικά αυτά, τα οποία είναι συμβατά με το πρότυπο **Unicode** και συνεπώς παρέχουν τις υπηρεσίες τους και για γλώσσες πέρα της αγγλικής, όπως είναι η **LinguistiX,** ο **SharePoint Πύλη Server,** ο **Classification Server** και η **Categorization Engine** της **Teragram.** Τέλος, τα ίδια συστήματα παρέχουν υποστήριξη των υπηρεσιών τους και για δεδομένα διαφορετικών διατάξεων, όπως ο **SharePoint Πύλη Server** που υποστηρίζει τα αρχεία τύπου **.tif** και ο **Classification Server της Teragram** που υποστηρίζει 200 διαφορετικές διατάξεις αρχείων.

Επίλογος

Στο πλαίσιο της σχετικής εργασίας επιχειρείται η αποτύπωση της τρέχουσας τεχνολογικής πραγματικότητας όσον αφορά τη διαχείριση της ηλεκτρονικής αλληλογραφίας των χρηστών. Η παρουσίαση των πλέον εξελιγμένων συστημάτων του σχετικού τομέα κρίθηκε απαραίτητη λόγω της αύξουσας χρήσης των ηλεκτρονικών μηνυμάτων από παντός είδους χρήστες χάρη στην χρηστικότητα και τις διευκολύνσεις που παρέχουν.

Βασικό κοινό χαρακτηριστικό των συστημάτων στα πλαίσια των υπηρεσιών που προσφέρουν είναι η κατηγοριοποίησή των μηνυμάτων βάση προεπιλεγμένων ρυθμίσεων των χρηστών. Στην υπηρεσία αυτή δίνεται ιδιαίτερη σημασία και έκταση όσον αφορά το γενικό περιεχόμενο της εργασίας για δύο βασικούς λόγους. Αρχικά, επειδή η σχετική διαδικασία αποτελεί τον ακρογωνιαίο λίθο των συστημάτων όσον αφορά τις διαδικασίες οργάνωσης του περιεχομένου, που είναι και η βασική υπηρεσία που προσφέρουν τα συστήματα. Όσο πιο εξελιγμένες είναι οι τεχνικές που υποστηρίζουν την κατηγοριοποίηση και όσο μεγαλύτερο είναι το επίπεδο αυτοματοποίησης που παρέχεται, ανάλογα υψηλό είναι και το επίπεδο και ο βαθμός οργάνωσης που προσφέρεται στους χρήστες, με τα συνεπάγοντα πλεονεκτήματα. Κατά δεύτερο λόγο, η κατηγοριοποίηση ηλεκτρονικών μηνυμάτων και ηλεκτρονικών τεκμηρίων γενικότερα είναι ένα τρέχον ερευνητικό πεδίο, που τις τελευταίες δεκαετίες γνωρίζει μεγάλο ενδιαφέρον και αύξηση. Οι

ερευνητικές προσεγγίσεις πληθαίνουν με την πάροδο του χρόνου και αποσκοπούν στην πλήρως αποτελεσματική και αυτοματοποιημένη μορφή της διαδικασίας.

Τέλος, μέσω της συνεχούς εξέλιξης και βελτίωσης των υπηρεσιών που παρέχουν τα συστήματα διαχείρισης ηλεκτρονικών μηνυμάτων, παρέχονται στους χρήστες διάφορα «εργαλεία», με τα οποία μπορούν να εκμεταλλευτούν επαρκώς την πληροφορία που διακινείται ηλεκτρονικά. Αυτό είναι ένα πολύ σημαντικό βήμα, δεδομένου ότι η τρέχουσα κοινωνική αλλά και τεχνολογική πραγματικότητα οδεύει σε μία ηλεκτρονική μορφή της κοινωνίας αλλά και της πληροφόρησης. Συνεπώς, η δυνατότητα οργάνωσης των πληροφοριών που διακινούνται ηλεκτρονικά με ένα εύκολα κατανοητό και χρηστικό τρόπο, αποτελεί πλεονέκτημα και ουσιαστικά ένα πολύ δυνατό όπλο για αυτούς που το κατέχουν.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Balter, Olle. Bifrost Inbox Organizer: Giving Control Over the Inbox. Language Technologies Institute Carnegie Mellon University.**

2. **Bergman, Ruth. A Personal Email Assistant. 2002. Hewlett Packard Laboratories.**

3. **Boone, Gary. (1998). Concept features in Re: Agent, an intelligent email agent. Proceedings of the Second International Conference on Autonomous Agents. New York: ACM Press, 141-148.**

4. **Clarc, James. Email classification: A hybrid approach combining genetic algorithms with neural networks.**

5. **Crawford, J. Kay, and E. McCreath: Automatic Induction of Rules for e-mail Classification. In ADCS2001 Proceedings of the Sixth Australasian Document Computing Symposium, pages 13-20, Cos Harbour, NSW Australia, 2001.**

6. **Delphi Group. Information Intelligence: Content Classification and the Enterprise Taxonomy Practice.**

7. **Griffiths, Richard T. History of the Internet, Internet for Historians.**

8. **Han, Eui-Hong. (1999). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. Technical Report #99-019, Department of Computer Science and Engineering, University of Minnesota**

9. **Kiritchenko, Svetlana. Email Classification with Co-Training.**

10. **Klimt, Bryan, Yang, Y. The Enron Corpus: A New Dataset For Email Classification Research. ECML 2004: 217-226**

11. **Ko, J , Sco, J. Automatic Categorization by Unsupervised Learning.**

12. **Ma, Liping. Document Classification via Structure Synopses.**

13. **Maes,P. Agents that Reduce Work and Information Overload Communications of the ACM July 1987, Vol.37, No. 7, pp.30-40**

14. **McCallum, Nigam (1998).A Comparison of Event Models for Naive Bayes Text Classification.**

15. **Segal, Richard, Kephart, J. MailCat: An Intelligent Assistant for Organizing Email. In Proceedings of the Third International Conference on Autonomous Agents, May 1999**
16. **Segal, R. and Kephart, J. Incremental Learning in SwiftFile. In Proceedings of the Seventh International Conference on Εκμάθηση Μηχανής. June, 2000**

17. **Sebastiani, Fabrizio. (1999) Εκμάθηση Μηχανής in Automated Text Classification**

18. **Takkinen, Juha. An adaptive approach to Text Categorization and Understanding- A preliminary Study**

19. **Yang, Yiming, and J.P. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the Fourteenth International Conference on Εκμάθηση Μηχανής (ICML'97).**

20. **Φλωρινά, Ν. «Intelligent Agents». URL:www.dide.flo.sch.gr/Plinet/Tutorials/ Tutorials-IntelligentAgents.html**

ΙΣΤΟΤΟΠΟΙ

a. **www.vicomsoft.com/knowledge/reference/email.history.htm (The history of email)**

b. www.dmreview.com/article_sub.cfm?articleId=6501 **(Automated classification: Moving to the Mainstream)**

c. **www.cs.ucl.ac.uk/staff/a.hunter/tradepress/fuzzy.html (Fuzzy Sets Classification)**

d. **www.cat.pdx.edu/~shervais/research.interests.neural.html (Artificial Neural Network)**

e. **www.popfile.sourceforge.net**

f. **www.nexor.com**

g. **www.interwoven.com**

h. **www.mobius.com/mobius/index_mobius.jsp**

i. **www.bytequest.com**

j. **www.echomail.com**

k. **www.towersoft.com/ap/**

l. **www.documentum.com**

m. **www-306.ibm.com/software/data/cm**

n. **www.autonomy.com**

o. **www.inxight.com**

p. **www.microsoft.com**

q. **www.teragram.com**

r. **www.stratify.com**

s. **www.verity.com**

t. **www.xerox.com**