

Ένα Μοντέλο για τη Διαλογή Ηλεκτρονικών Εγγράφων Βάσει Περιεχομένου με Χρήση Ευφυών Πρακτόρων

Νικόλαος Σ. Παπασπύρου*
(nickie@softlab.ntua.gr)

Κλειώ Η. Σγουροπούλου*
(csgouro@softlab.ntua.gr)

Δημήτριος Ν. Πέτρογλου*
(dpetro@softlab.ntua.gr)

Αλέξανδρος Β. Γκερμπεσιώτης†
(alexg@comlab.ox.ac.uk)

Πάνος Λιβαδάς‡
(pel@cs.ufl.edu)

Εμμανουήλ Σ. Σκορδαλάκης*
(skordala@softlab.ntua.gr)

Περίληψη

Λόγω του τεράστιου όγκου ετερόκλητων πληροφοριών που είναι σήμερα διαθέσιμες σε ηλεκτρονική μορφή, πολλοί άνθρωποι καταναλώνουν σημαντικό ποσοστό του χρόνου τους αναζητώντας μέσα σε ένα κυκλώνα άχρηστων για αυτούς πληροφοριών εκείνες που τους είναι χρήσιμες. Επιπλέον, η διαλογή ηλεκτρονικών εγγράφων είναι δύσκολο να γίνει με αυτόματο τρόπο, αφενός λόγω της μεγάλης ανομοιογένειας στο είδος και τη μορφή των ηλεκτρονικών εγγράφων, αφετέρου λόγω της αδυναμίας του ηλεκτρονικού υπολογιστή να κατανοήσει το περιεχόμενο των εγγράφων και να το κρίνει με βάση τα υποκειμενικά κριτήρια του ανθρώπου. Στην εργασία αυτή προτείνεται ένα μοντέλο για την αυτοματοποίηση της διαλογής ηλεκτρονικών εγγράφων βάσει περιεχομένου, που βασίζεται στη σύγχρονη τεχνολογία των ευφυών πρακτόρων.

1 Εισαγωγή

Η *διαχείριση εγγράφων* (document management) είναι σήμερα μια από τις σπουδαιότερες εργασίες σε κάθε γραφείο και μια από τις κυριότερες εφαρμογές των ηλεκτρονικών υπολογιστών. Τα ηλεκτρονικά έγγραφα καταλαμβάνουν πλέον το μεγαλύτερο ποσοστό των πληροφοριών που υπάρχουν διαθέσιμες σε ηλεκτρονική μορφή και δε νοούνται πλέον ως τα ηλεκτρονικά ανάλογα των εγγράφων σε χαρτί, αλλά ως δυναμικές οντότητες σε ποικιλία μέσων (multimedia). Μπορούν να περιέχουν πληροφορίες σχετικές με την προέλευσή τους καθώς και εκτελέσιμο κώδικα που αναλαμβάνει τη διαχείρισή τους. Τα έγγραφα γίνονται σιγά-σιγά το κέντρο της προσοχής όσον αφορά στο σχεδιασμό λογισμικού, καθώς γίνεται εξαιρετικά σημαντική η εύκολη και φιλική προς το χρήστη διαχείρισή τους, και απαιτούνται νέα εργαλεία για αυτό το σκοπό.

Η αυτοματοποίηση της εκτέλεσης εργασιών σχετικών με τη διαχείριση εγγράφων αποτελεί σήμερα τον κύριο στόχο της πληροφορικής για τα περιβάλλοντα γραφείου. Στο *γραφείο χωρίς χαρτιά* (paperless office) που είναι ο στόχος του μέλλοντος, τα έγγραφα θα φυλάσσονται σε ηλεκτρονική μορφή και θα προσπελούνται με τρόπο εύκολο και “προσωπικό” για κάθε χρήστη. Πολυετής έρευνα προς αυτή την κατεύθυνση δεν έχει αποφέρει ως τώρα ιδιαίτερα θετικά αποτελέσματα.

Μια λύση στο πρόβλημα της διαχείρισης εγγράφων που έχει προταθεί στην πρόσφατη βιβλιογραφία είναι οι *ευφείς πράκτορες* (intelligent agents). Μπορεί κανείς να θεωρήσει ότι ένας ευφυής πράκτορας παρεμβαίνει μεταξύ ενός χρήστη και ενός συστήματος λογισμικού και αναλαμβάνει τη διεκπεραίωση εργασιών που μόνο του το σύστημα λογισμικού δεν είναι σε θέση να επιτελέσει. Η χρήση του πράκτορα ως μεσολαβητή υποβοηθεί και απλοποιεί το έργο του χρήστη και κατά συνέπεια αυξάνει την παραγωγικότητά του.

*Εθνικό Μετσόβιο Πολυτεχνείο, Τμήμα Ηλεκτρολόγων Μηχ. και Μηχ. Υπολογιστών
Τομέας Πληροφορικής, Εργαστήριο Λογισμικού, Πολυτεχνειούπολη, 15780 Ζωγράφου, Αθήνα.

†Oxford University Computing Laboratory, Parks Rd., Oxford OX1 3QD, UK.

‡University of Florida, Computer and Information Sciences Department, Gainesville, FL 32611, USA.

Οι ευφυείς πράκτορες μπορούν να δώσουν λύση στο πρόβλημα της συνένωσης των ετερογενών μορφών πληροφορίας και των ασύμβατων εφαρμογών διακίνησης και διαχείρισης εγγράφων [1, 2]. Ερευνητικά αποτελέσματα, που προκύπτουν από την κατασκευή και πειραματική χρήση ευφύων πρακτόρων, αποδεικνύουν ότι αυτοί διευκολύνουν σημαντικά την αυτοματοποίηση των πληροφορικών εργασιών ενός γραφείου (π.χ. διαχείριση ηλεκτρονικού ταχυδρομείου [3] και ηλεκτρονικών νέων [4], αυτοματοποίηση προγραμματισμού συναντήσεων [5], διαχείριση αρχείων και βιβλιοθηκών, κ.α.), όπως επίσης και την εκπαίδευση [6]. Σημαντικό μέρος της σχετικής βιβλιογραφίας ασχολείται με την προδιαγραφή και υλοποίηση πρωτοκόλλων επικοινωνίας μεταξύ ευφύων πρακτόρων [7].

Ένα από τα επιμέρους προβλήματα που παρουσιάζονται στη διαχείριση ηλεκτρονικών εγγράφων είναι το πρόβλημα της *διαλογής εγγράφων βάσει περιεχομένου*. Μια κατηγορία ευφύων πρακτόρων που σχετίζεται με αυτό το πρόβλημα είναι αυτή των *πρακτόρων διαλογής* (filtering agents). Οι πράκτορες αυτοί έχουν ως σκοπό να φιλτράρουν τις πληροφορίες που φθάνουν με τη μορφή ηλεκτρονικών εγγράφων και να παρουσιάζουν στο χρήστη μόνο εκείνες που ο τελευταίος θεωρεί ενδιαφέροντες, απορρίπτοντας τυχόν άχρηστες πληροφορίες. Συνήθως συνεργάζονται με βάσεις γνώσης που περιέχουν τα κριτήρια επιλογής του χρήστη, τα οποία χρησιμοποιούν για τη διαλογή των πληροφοριών. Χαρακτηριστικό της ευφυΐας μερικών πρακτόρων αυτής της κατηγορίας είναι ότι δεν είναι απαραίτητος ο καθορισμός των κριτηρίων από το χρήστη. Κατά τη διάρκεια μιας περιόδου εκπαίδευσης, ο πράκτορας ενημερώνει αυτόματα τα κριτήρια επιλογής με την έμμεση καθοδήγηση του χρήστη. Δεδομένης της αύξησης στην ποσότητα των πληροφοριών που διακινούνται με ηλεκτρονικό τρόπο, καθώς και της αυξανόμενης εξειδίκευσης των πληροφοριών αυτών, οι πράκτορες διαλογής αναμένεται να αποτελέσουν στο μέλλον σημαντικό εργαλείο για την αποδοτική προσπέλαση των πληροφοριών και των πηγών τους.

Στην εργασία αυτή προτείνεται ένα μοντέλο για τη διαλογή ηλεκτρονικών εγγράφων βάσει περιεχομένου, χρησιμοποιώντας ένα σύνολο ευφύων πρακτόρων διαλογής. Στην ενότητα 2 δίνεται ένας ορισμός του προβλήματος και περιγράφονται οι λειτουργικές απαιτήσεις από ένα τέτοιο σύστημα. Στην ενότητα 3 περιγράφεται λεπτομερώς το προτεινόμενο μοντέλο, ενώ στην ενότητα 4 παρουσιάζεται μια αρχική πειραματική υλοποίηση του μοντέλου και σχολιάζονται τα αρχικά αποτελέσματα από τη χρήση της. Τέλος, στην ενότητα 5 μελετώνται κατευθύνσεις για μελλοντική έρευνα, ενώ στην ενότητα 6 γίνεται η σύνοψη της εργασίας και παρατίθενται κάποια τελικά συμπεράσματα.

2 Ορισμός του προβλήματος

Στη σημερινή εποχή εξέχουσα θέση κατέχει η πληροφορία. Ένας σημαντικός αριθμός εργαζομένων αφιερώνει μεγάλο ποσοστό του χρόνου του στον εντοπισμό και τη συγκέντρωση χρήσιμων πληροφοριών, πολλές από τις οποίες έχουν τη μορφή ηλεκτρονικών εγγράφων. Η εξειδίκευση, οι ανάγκες και τα ενδιαφέροντα των ανθρώπων διαφέρουν σε τέτοιο βαθμό που πληροφορίες πολύτιμες για μια κατηγορία ανθρώπων είναι άχρηστες για μια άλλη κατηγορία. Ήδη, λόγω του τεράστιου όγκου ανομοιογενών πληροφοριών που υπάρχουν σε ηλεκτρονική μορφή, η *διαλογή* των ηλεκτρονικών εγγράφων σε χρήσιμες και μη είναι μια προσωπική, χρονοβόρα, ανιαρή και κουραστική εργασία.

Οι δυσκολίες στην επίλυση του προβλήματος της αυτόματης διαλογής ηλεκτρονικών εγγράφων οφείλονται στο γεγονός ότι η διαλογή στηρίζεται συνήθως στο *περιεχόμενο* των εγγράφων και όχι σε εξωτερικά χαρακτηριστικά τους, π.χ. στο χρώμα και την υφή του χαρτιού ή στον αριθμό των σελίδων. Από τεχνικής πλευράς, το περιεχόμενο ενός εγγράφου δεν είναι παρά μια πεπερασμένη ακολουθία δυαδικών αριθμών και είναι αδύνατο να γίνει αντιληπτό σε υψηλό επίπεδο χωρίς τη διαμεσολάβηση κατάλληλου λογισμικού. Όμως, ακόμα κι αν ξεπερασθεί αυτό το εμπόδιο, το περιεχόμενο της πλειοψηφίας των χρήσιμων για τον άνθρωπο ηλεκτρονικών εγγράφων παραμένει εντελώς ακατάληπτο για τον ηλεκτρονικό υπολογιστή, που λόγω έλλειψης ευφυΐας δεν είναι σε

θέσει να αντιληφθεί την ουσία. Για τους λόγους αυτούς, θεωρούμε σκόπιμο να διευκρινίσουμε ότι ερευνητικό ενδιαφέρον παρουσιάζει το πρόβλημα της *διαλογής εγγράφων βάσει περιεχομένου*.

Το μοντέλο που προτείνεται στην επόμενη ενότητα επιχειρεί να επιλύσει το πρόβλημα της αυτόματης διαλογής ηλεκτρονικών εγγράφων βάσει περιεχομένου. Κρίνεται σκόπιμο να αντιμετωπισθούν συγχρόνως και άλλα προβλήματα διαχείρισης που σχετίζονται άμεσα με αυτό, όπως η *αναζήτηση* και η *συλλογή πληροφοριών*, η *αρχειοθέτησή τους* και η *εκμετάλλευση υπαρχόντων αρχείων*.

Μια από τις βασικές προϋποθέσεις που πρέπει να πληρεί ένα σύστημα αυτόματης διαλογής είναι η *ανεξαρτησία* από τη μορφή των εγγράφων. Απόλυτη ανεξαρτησία φυσικά δεν είναι εφικτή λόγω της μεγάλης ανομοιογένειας στο είδος αλλά και τη μορφή των ηλεκτρονικών εγγράφων. Είναι όμως απαραίτητο η διαδικασία της διαλογής να γίνεται αντιληπτή από τους χρήστες με τρόπο ενιαίο.

Η ανάγκη *εξατομίκευσης* (personalization) του συστήματος για κάθε χρήστη έχει ως αποτέλεσμα την αναζήτηση ενός μοντέλου πολύ διαφορετικού από αυτό των τυπικών συστημάτων ανάκτησης πληροφοριών. Τα κριτήρια διαλογής πρέπει να απορρέουν από μια σειρά επαναλαμβανόμενων χρήσεων του συστήματος και όχι από ένα μεμονωμένο γεγονός αναζήτησης πληροφοριών. Το σύστημα διαλογής πρέπει να αντιμετωπίζει μακροπρόθεσμες μεταβολές, βασιζόμενο σε μια αλυσίδα τέτοιων γεγονότων. Επομένως, τα θέματα της *εκμάθησης* και της *προσαρμογής* είναι ιδιαίτερης σημασίας για την ανάπτυξη του συστήματος. Η χρήση ευφών πρακτόρων είναι η πλέον ενδεδειγμένη για την αντιμετώπιση όλων αυτών των ιδιαιτεροτήτων.

Κατά τον ορισμό του μοντέλου, ιδιαίτερη προσοχή πρέπει να δοθεί στο θέμα της *ικανότητας* (competence) των πρακτόρων να αυτοματοποιούν σωστά τις απαιτούμενες εργασίες και να εμπνέουν *εμπιστοσύνη* (trust, believability) στους χρήστες. Πολύ σημαντικός είναι επίσης ο *τρόπος επικοινωνίας* μεταξύ πρακτόρων και χρηστών. Κύρια προϋπόθεση θεωρείται η εξασφάλιση ενός περιβάλλοντος εργασίας ευχάριστου και ευέλικτου με φιλική διαπροσωπεία, ικανή να προσαρμόζεται δυναμικά στις προσωπικές απαιτήσεις των χρηστών. Εξέχουσας σημασίας επίσης είναι η δυνατότητα *αυτόματης εκμάθησης* από μέρους των πρακτόρων, δια μέσου της παρακολούθησης των ενεργειών των χρηστών, με χρήση παραδειγμάτων ή με άμεση υπόδειξη (feedback) από τους χρήστες.

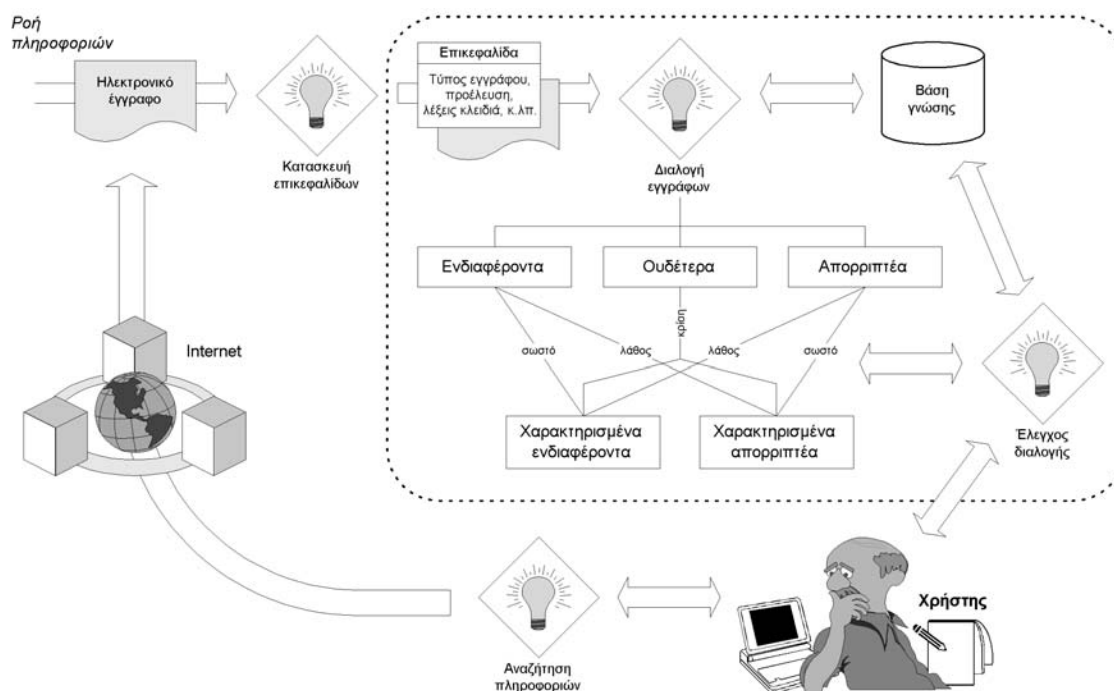
3 Περιγραφή του μοντέλου

Για την επίλυση του προβλήματος της διαλογής ηλεκτρονικών εγγράφων βάσει περιεχομένου προτείνεται ένα μοντέλο συστήματος αποτελούμενου κατά κύριο λόγο από τέσσερις ευφυείς πρακτόρες και μια βάση γνώσης. Μια σχηματική παράσταση του μοντέλου φαίνεται στο Σχήμα 1. Το μέρος του σχήματος που περικλείεται από τη διακεκομμένη γραμμή είναι το τμήμα του μοντέλου που ασχολείται αποκλειστικά με τη διαλογή των εγγράφων. Το υπόλοιπο του σχήματος ασχολείται με την διεκπεραίωση απαραίτητων βοηθητικών λειτουργιών.

Στα δυο άκρα του προτεινόμενου μοντέλου βρίσκονται μια πηγή πληροφοριών και ο ενδιαφερόμενος χρήστης. Το μοντέλο προδιαγράφει ένα σύστημα διαμεσολάβησης μεταξύ αυτών των δυο. Κυρίαρχες θέσεις κατέχουν οι τέσσερις ευφυείς πράκτορες, που παριστάνονται ως ρόμβοι με φωτεινή λυχνία. Οι πράκτορες αυτοί είναι εφαρμογές λογισμικού που αναλαμβάνουν τη διεκπεραίωση συγκεκριμένων λειτουργιών. Ο τίτλος *πράκτορες* που τους αποδίδεται δικαιολογείται από την απαίτηση ευφυΐας καθώς και από τον τρόπο με τον οποίο επικοινωνούν μεταξύ τους και με το χρήστη.

3.1 Κατασκευή επικεφαλίδων

Προκειμένου να αντιμετωπισθεί το πρόβλημα της ανομοιογένειας των ηλεκτρονικών εγγράφων στη διαδικασία της διαλογής, κρίνεται απαραίτητη η ύπαρξη ενός προκαταρκτικού σταδίου επε-



Σχήμα 1: Σχηματική παράσταση του μοντέλου.

ξεργασίας, κατά τη διάρκεια του οποίου για κάθε ηλεκτρονικό έγγραφο κατασκευάζεται ένα *αρχείο επικεφαλίδας* (header file). Οι επικεφαλίδες αυτές έχουν συγκεκριμένη και ομοιογενή μορφή για όλα τα είδη ηλεκτρονικών εγγράφων. Η κατασκευή των επικεφαλίδων παρουσιάζει αρκετές ομοιότητες με τον καθορισμό και την εξαγωγή κατάλληλων *μετα-πληροφοριών* (metadata) από ηλεκτρονικά αντικείμενα [8]. Στο προτεινόμενο μοντέλο επιτελείται από τον πράκτορα κατασκευής επικεφαλίδων, ο οποίος δε χρειάζεται να αλληλεπιδρά με το χρήστη και βρίσκεται διαρκώς σε λειτουργία. Στη συνέχεια, η διαδικασία διαλογής γίνεται με την επεξεργασία μόνο των επικεφαλίδων.

Η μορφή των επικεφαλίδων οφείλει να είναι πολύ γενική, προκειμένου αυτές να είναι σε θέση να περιγράφουν με ακρίβεια όλα τα στοιχεία του περιεχομένου των ηλεκτρονικών εγγράφων που είναι χρήσιμα για τη διαλογή. Για το λόγο αυτό, κάθε επικεφαλίδα περιέχει ένα σύνολο από ζεύγη πεδίων-τιμών. Ακολουθεί παράδειγμα επικεφαλίδας που περιγράφει, χωρίς πολλές λεπτομέρειες, μια ηλεκτρονική δημοσιογραφική είδηση.

προέλευση	=	"Πρακτορείο Reuter"
συγγραφέας	=	"Alan Smith"
τίτλος	=	"Νέα έκρηξη στο μετρό του Λονδίνου"
ημερομηνία	=	"12 Δεκεμβρίου 1996, 17:55 GMT"
κατηγορία-είδηση	=	"Ξένη επικαιρότητα"
λέξεις-κλειδιά	=	"IRA, έκρηξη, μετρό, τρομοκρατία, νεκροί, δηλώσεις"
γλώσσα	=	"Αγγλικά"
αριθμός-λέξεων	=	"1340"

Ο τρόπος κατασκευής των επικεφαλίδων εξαρτάται σε πολύ μεγάλο βαθμό από το πεδίο εφαρμογής. Τόσο τα ονόματα των πεδίων όσο και οι τιμές τους δεν είναι δυνατό να καθορισθούν εκ των προτέρων με ενιαίο τρόπο.

Η αυτόματη κατασκευή επικεφαλίδων από ανομοιογενή ηλεκτρονικά έγγραφα είναι ένα πολύ

δύσκολο πρόβλημα και δεν αναμένεται να λυθεί στη γενική του μορφή στο άμεσο μέλλον. Όμως, τρέχουσα έρευνα στον χώρο των *ημι-δομημένων πληροφοριών* (semi-structured data) έχει οδηγήσει ήδη σε ικανοποιητικές προσεγγίσεις για την εξαγωγή πληροφοριών και την αυτόματη ταξινόμηση εγγράφων [9, 10, 11, 12]. Η παρούσα εργασία διαφέρει κύρια στην ελευθερία μορφής των εξαγόμενων μετα-πληροφοριών και στη χρήση αυτόματης εκμάθησης κατά τη διαδικασία εξαγωγής τους.

3.2 Διαλογή εγγράφων

Η διαλογή των εγγράφων βασίζεται σε ένα σύνολο υποκειμενικών κριτηρίων που καθορίζονται από το χρήστη. Υποθέτουμε ότι οι επικεφαλίδες των νέων εγγράφων, που πρόκειται να περάσουν από τη διαδικασία διαλογής, βρίσκονται τοποθετημένες στην *περιοχή άφιξης*. Κάθε μια τέτοια επικεφαλίδα εξετάζεται από τον *πράκτορα διαλογής*, ο οποίος τελικά της αντιστοιχεί ένα βαθμό ενδιαφέροντος, που στο εξής θα ονομάζεται *βαθμολογία*. Η βαθμολογία είναι μια πραγματική τιμή στο διάστημα $[-1, +1]$. Η τιμή $+1$ αντιστοιχεί σε έγγραφα που χαρακτηρίζονται ενδιαφέροντα χωρίς αμφιβολία, και κατά συνέπεια γίνονται δεκτά, ενώ η τιμή -1 σε έγγραφα που απορρίπτονται χωρίς αμφιβολία. Οι ενδιάμεσες τιμές παριστάνουν με ομοιόμορφο τρόπο το βαθμό αμφιβολίας του πράκτορα. Η τιμή 0 αντιστοιχεί σε έγγραφα για τα οποία δεν μπορεί να γίνει χαρακτηρισμός.

Αφού ο πράκτορας διαλογής αποφασίσει τη βαθμολογία για ένα έγγραφο, βασιζόμενος στην αντίστοιχη επικεφαλίδα και με ένα τρόπο που θα περιγραφεί αναλυτικά στην ενότητα 3.4, κατατάσσει το έγγραφο σε μια από τρεις δυνατές περιοχές: *ενδιαφέροντα έγγραφα*, *ουδέτερα έγγραφα* και *απορριπτέα έγγραφα*. Η κατάταξη γίνεται με βάση τις τιμές των κατωφλίων που διαχωρίζουν τις τρεις αυτές περιοχές και που καθορίζονται, άμεσα ή έμμεσα από το χρήστη.

Ένα απαραίτητο χαρακτηριστικό του πράκτορα διαλογής είναι η αυτογνωσία. Ο πράκτορας πρέπει να είναι σε θέση να εκτιμήσει τις κριτικές του ικανότητες, βάσει του χρόνου εκμάθησης και των αποτελεσμάτων προηγούμενων κρίσεων, όπως αναλυτικά περιγράφεται στην ενότητα 3.5. Το χαρακτηριστικό αυτό μπορεί να παρασταθεί στο μοντέλο με μια πραγματική τιμή στο διάστημα $[0, 1]$, που παριστάνει το βαθμό *εμπιστοσύνης* του πράκτορα στις ικανότητές του. Προκειμένου να λαμβάνεται υπόψη αυτή η “διστακτικότητα” του πράκτορα διαλογής, η βαθμολογία κάθε εγγράφου πολλαπλασιάζεται με τον εκάστοτε βαθμό εμπιστοσύνης και η κατάταξη γίνεται βάσει του γινομένου.

3.3 Βάση γνώσης

Η βάση γνώσης περιέχει τα κριτήρια με τα οποία γίνεται η διαλογή των εγγράφων, κωδικοποιημένα με κατάλληλο τρόπο. Ο καθορισμός των κριτηρίων γίνεται από το χρήστη σταδιακά και, ως επι το πλείστον, με έμμεσο τρόπο. Η κωδικοποίηση των κριτηρίων πρέπει να είναι αόρατη από το χρήστη, προκειμένου να εξασφαλισθεί η ευχρηστία του συστήματος. Επίσης, η μορφή των κριτηρίων πρέπει να είναι εξίσου γενική με τη μορφή των επικεφαλίδων, μια και σε αυτές θα εφαρμοσθούν.

Στο προτεινόμενο μοντέλο, τα κριτήρια διαλογής αποθηκεύονται στη βάση γνώσης με τη μορφή *κανόνων*. Κάθε κανόνας είναι της μορφής:

αν συνθήκη τότε τιμή-ενδιαφέροντος

δηλαδή οι κανόνες είναι ζεύγη αποτελούμενα από μια συνθήκη και μια τιμή ενδιαφέροντος. Σε γενικές γραμμές, προκειμένου να εφαρμοσθεί ένας κανόνας για κάποια εγγραφή, εξετάζεται πρώτα αν ισχύει η συνθήκη. Αν συμβαίνει αυτό, τότε αποδίδεται στην εγγραφή η τιμή ενδιαφέροντος που αντιστοιχεί στον κανόνα, αλλιώς δε γίνεται τίποτα. Η διαδικασία αυτή περιγράφεται αναλυτικά στην ενότητα 3.4.

Η *συνθήκη* ενός κανόνα είναι μια λογική έκφραση που μπορεί να περιέχει ονόματα πεδίων, που εμφανίζονται στην επικεφαλίδα του εγγράφου, καθώς και σταθερές τιμές. Παραδείγματα χρήσιμων περιπτώσεων απλών συνθηκών είναι τα ακόλουθα:

- *πεδίο ίσο με τιμή*
- *πεδίο περιέχει τιμή*

Επίσης, χρήσιμη είναι η δυνατότητα συνένωσης απλών συνθηκών σε πιο σύνθετες, με χρήση των τελεστών **και**, **ή** και **όχι**.

Το αποτέλεσμα της εφαρμογής ενός κανόνα είναι η απόδοση μιας τιμής *ενδιαφέροντος* σε ένα έγγραφο. Οι τιμές ενδιαφέροντος ανήκουν σε ένα σύνολο που στο εξής θα ονομάζεται $\mathcal{E}$. Οι ιδιότητες και η μορφή των στοιχείων αυτού του συνόλου θα περιγραφούν αναλυτικά στην ενότητα 3.4. Η περιγραφή του τρόπου κωδικοποίησης των κανόνων στη βάση γνώσης ξεφεύγει από τους στόχους του προτεινόμενου μοντέλου διαλογής. Αξίζει, όμως, να σημειωθεί ότι ρεαλιστικές υλοποιήσεις του μοντέλου θα πρέπει να δίνουν ιδιαίτερη έμφαση στην ταχύτητα με την οποία ανακτώνται και ενημερώνονται τα περιεχόμενα της βάσης γνώσης, ο όγκος των οποίων εξαρτάται μόνο από τις ανάγκες και τα ενδιαφέροντα των χρηστών.

3.4 Υπολογισμός της βαθμολογίας

Ιδιαίτερης σημασίας για το προτεινόμενο μοντέλο διαλογής είναι ο αλγόριθμος υπολογισμού της βαθμολογίας. Τα δεδομένα που χρησιμοποιεί είναι η επικεφαλίδα του εγγράφου και η βάση γνώσης, που περιέχει τα κριτήρια διαλογής. Το αποτέλεσμα του υπολογισμού είναι η βαθμολογία που θα αποδοθεί στο έγγραφο. Αυτή προκύπτει συνολικά από την εφαρμογή όλων των κανόνων που περιέχονται στη βάση γνώσης. Κάθε κανόνας του οποίου η συνθήκη αληθεύει για ένα συγκεκριμένο έγγραφο συνεισφέρει μια τιμή ενδιαφέροντος στη συνολική βαθμολογία. Η εφαρμογή όλων των κανόνων οδηγεί σε ένα πλήθος συνεισφερομένων τιμών, που πρέπει να συνδυασθούν με κατάλληλο τρόπο. Το προτεινόμενο μοντέλο διακρίνει δυο μορφές συνδυασμού των συνεισφερομένων τιμών ενδιαφέροντος, που αντιστοιχούν τελικά στη διάκριση των κανόνων σε δυο κατηγορίες.

Ο προφανής τρόπος συνδυασμού λαμβάνει υπόψη όλες τις συνεισφερόμενες τιμές. Αν από την εφαρμογή όλων των κανόνων για μια συγκεκριμένη εγγραφή προκύπτει σύνολο συνεισφερομένων τιμών $\{a_1, a_2, \dots, a_m\}$, όπου $a_i \in \mathcal{E}$, η συνολική βαθμολογία δίνεται από τον ακόλουθο τύπο:

$$\text{βαθμολογία} = \| a_1 \oplus a_2 \oplus \dots \oplus a_m \|$$

Ο τελεστής $\oplus$ υλοποιεί το συνδυασμό δυο τιμών ενδιαφέροντος σε μια τρίτη, ενώ ο τελεστής $\| \cdot \|$ απεικονίζει μια τιμή ενδιαφέροντος στο διάστημα $[-1, +1]$ των πραγματικών αριθμών. Απαραίτητη επίσης είναι η ύπαρξη ενός τελεστή άρνησης ενδιαφέροντος, που θα απεικονίζει μια τιμή ενδιαφέροντος a στην *αντίθετή* της $\bar{a}$. Η επιλογή των τελεστών $\oplus$ και $\| \cdot \|$, καθώς και του τελεστή άρνησης, πρέπει να γίνει με κατάλληλο τρόπο ώστε να πληρούνται οι ιδιότητες που περιγράφονται στο Σχήμα 2.

Η προφανής επιλογή για τον ορισμό του συνόλου $\mathcal{E}$ είναι το διάστημα $[-1, +1]$ των πραγματικών αριθμών, οπότε η ύπαρξη του τελεστή $\| \cdot \|$ είναι περιττή. Δυστυχώς όμως, με αυτή την επιλογή είναι αδύνατο να ορισθούν οι υπόλοιποι τελεστές, έτσι ώστε να ικανοποιούνται όλες οι προαναφερθείσες συνθήκες. Μια πιο πρόσφορη επιλογή είναι η ακόλουθη. Τα στοιχεία $a \in \mathcal{E}$ ορίζονται ως τα ζεύγη της μορφής $\langle x, n \rangle$, όπου x ένας πραγματικός αριθμός στο διάστημα $[-1, +1]$ και n ένας φυσικός αριθμός. Η τιμή του x παριστάνει το βαθμό ενδιαφέροντος με τον ίδιο τρόπο όπως και η βαθμολογία. Η τιμή του n ονομάζεται *πολλαπλότητα* και εκφράζει το βάρος που φέρει αυτή η τιμή. Μια τιμή ενδιαφέροντος με μεγαλύτερη πολλαπλότητα επηρεάζει περισσότερο το αποτέλεσμα του τελεστή

Σχήμα 2: Επιθυμητές ιδιότητες τελεστών για τον προσθετικό συνδυασμό τιμών ενδιαφέροντος.

$a \oplus (b \oplus c) = (a \oplus b) \oplus c$ $a \oplus b = b \oplus a$ $a \oplus \mathbf{O} = a$ $-1 \leq \ a\ \leq +1$ $\ \mathbf{O}\ = 0$ $\ a_1\ \leq \ a_2\ \Rightarrow \ a_1 \oplus b\ \leq \ a_2 \oplus b\ $	$\ a\ = \ b\ \Rightarrow \ a \oplus b\ = \ a\ = \ b\ $ $\ a \oplus \bar{a}\ = 0$ $\overline{\mathbf{O}} = \mathbf{O}$ $\overline{a \oplus b} = \bar{a} \oplus \bar{b}$ $\ \bar{a}\ = -\ a\ $
---	--

$\oplus$ απ' όσο μια τιμή με μικρότερη πολλαπλότητα. Είναι εύκολο να διαπιστωθεί ότι οι τελεστές που ορίζονται παρακάτω ικανοποιούν τις προαναφερθείσες συνθήκες.

$$\langle x_1, n_1 \rangle \oplus \langle x_2, n_2 \rangle = \begin{cases} \left\langle \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}, n_1 + n_2 \right\rangle, & \text{αν } n_1 + n_2 \neq 0 \\ \langle 0, 0 \rangle, & \text{αν } n_1 + n_2 = 0 \end{cases}$$

$$\begin{aligned} \|\langle x, n \rangle\| &= x \\ \overline{\langle x, n \rangle} &= \langle -x, n \rangle \\ \mathbf{O} &= \langle 0, 0 \rangle \end{aligned}$$

δηλαδή ο τελεστής $\oplus$ είναι ουσιαστικά ο μέσος όρος των δυο τιμών, λαμβανομένων υπόψη των βαρών.

Μερικές φορές όμως, είναι χρήσιμο κάποιες τιμές ενδιαφέροντος να υπερισχύουν των άλλων. Στην περίπτωση αυτή, μόνο οι “προνομιούχες” τιμές συνεισφέρουν στον υπολογισμό της τελικής βαθμολογίας. Ένα παράδειγμα εφαρμογής κανόνων με τέτοιες προνομιούχες τιμές είναι η απαίτηση να απορρίπτονται έγγραφα κάποιου συγκεκριμένου συγγραφέα ως αναξιόπιστα, ακόμα κι αν αυτά έχουν εξαιρετικά ενδιαφέρον περιεχόμενο.

Θα ήταν αρκετά δύσκολο να συμπεριληφθεί αυτή η απαίτηση στο μαθηματικό μοντέλο για τις τιμές ενδιαφέροντος. Για το λόγο αυτό, στον υπολογισμό της συνολικής βαθμολογίας υιοθετείται μια διαφορετική προσέγγιση. Οι τιμές ενδιαφέροντος χαρακτηρίζονται βάσει των κανόνων στους οποίους εμφανίζονται και οι κανόνες διακρίνονται σε δυο κατηγορίες: τους *προσθετικούς κανόνες*, που συνεισφέρουν μη προνομιούχες τιμές, και τους *ανταρχικούς κανόνες*, που συνεισφέρουν προνομιούχες τιμές. Προκειμένου η διάκριση αυτή να είναι σαφής, οι δυο κατηγορίες κανόνων μπορούν να τοποθετηθούν σε δυο διαφορετικές βάσεις. Ο πλήρης αλγόριθμος υπολογισμού της βαθμολογίας φαίνεται στο Σχήμα 3.

3.5 Έλεγχος διαλογής

Η λειτουργία του ελέγχου της διαλογής επιτελείται από το χρήστη μέσω ενός εξειδικευμένου πράκτορα. Ο χρήστης είναι ο μόνος αρμόδιος να αποφασίσει τελικά αν ένα έγγραφο είναι ενδιαφέρον για εκείνον ή όχι. Για το λόγο αυτό, καλείται να ελέγξει τις κρίσεις του πράκτορα διαλογής και να κατατάξει τελικά τα έγγραφα, που έχουν καταταγεί στις τρεις προαναφερθείσες περιοχές, σε δυο νέες περιοχές: *τελικά ενδιαφέροντα έγγραφα* και *τελικά απορριπτέα έγγραφα*. Με τη νέα κατάταξη είναι δυνατό να υπάρξουν μεταβάσεις εγγράφων από/προς όλους τους δυνατούς συνδυασμούς περιοχών. Με την πάροδο του χρόνου και όσο η βάση γνώσεων εμπλουτίζεται, εκφράζεται η ελπίδα ότι στη φάση ελέγχου δε θα απαιτείται μεγάλος αριθμός μεταβάσεων μεταξύ περιοχών αντίθετου βαθμού ενδιαφέροντος.

Σχήμα 3: Αλγόριθμος διαλογής βάσει περιεχομένου.

για κάθε αρχείο επικεφαλίδας H στην περιοχή άφιξης **κάνε**
 βαθμολογία $:=$ ακαθόριστη
 για κάθε κανόνα K στη βάση αυταρχικών κανόνων **κάνε**
 αν η συνθήκη του K είναι αληθής για το H **τότε**
 βαθμολογία $:= \max \{ \text{βαθμολογία}, \text{τιμή ενδιαφέροντος του } K \}$
 αν η βαθμολογία είναι ακόμα ακαθόριστη **τότε**
 βαθμολογία $:= 0$ (αδιάφορο)
 για κάθε κανόνα K στη βάση προσθετικών κανόνων **κάνε**
 αν η συνθήκη του K είναι αληθής για το H **τότε**
 βαθμολογία $:= \text{βαθμολογία} \oplus \text{τιμή ενδιαφέροντος του } K$

Ο έλεγχος διαλογής χωρίζεται σε δυο επιμέρους εργασίες: Η διαδικασία εκμάθησης αποσκοπεί στην ενημέρωση και τον εμπλουτισμό της βάσης γνώσης με τα πραγματικά κριτήρια διαλογής του χρήστη. Στη φάση αυτή ο χρήστης κατατάσσει τα έγγραφα που βρίσκονται στην περιοχή ουδέτερων εγγράφων, λόγω αδυναμίας του πράκτορα διαλογής να τα κατατάξει σωστά. Παράλληλα, υποδεικνύει έμμεσα τα κριτήρια με τα οποία αυτός κάνει την κατάταξη, μέσω μιας φιλικής διαπροσωπείας. Κατ'αυτό τον τρόπο, ο χρήστης επηρεάζει τη βάση γνώσης, προσθέτοντας ή ενημερώνοντας κανόνες, χωρίς να ασχοληθεί με τη μορφή και την κωδικοποίηση των κανόνων. Η διαδικασία διόρθωσης σφαλμάτων, στη διάρκεια της οποίας ο χρήστης κατατάσσει έγγραφα που βρίσκονται στις άλλες δυο περιοχές, αποσκοπεί στη διόρθωση των σφαλμάτων που γίνονται από τον πράκτορα διαλογής και έχει αντίκτυπο στο βαθμό εμπιστοσύνης του συστήματος.

3.6 Αναζήτηση πληροφοριών

Η ανάγκη διαλογής εγγράφων βάσει περιεχομένου δε θα ήταν τόσο επιτακτική αν εξ αρχής η ροή πληροφοριών περιείχε ενδιαφέρουσες πληροφορίες με αρκετά μεγάλη πιθανότητα. Για να συμβεί κάτι τέτοιο, πρέπει οι χρήστες να έχουν τη δυνατότητα να επισημαίνουν στο σύστημα τις πηγές που παράγουν ενδιαφέρουσες για αυτούς πληροφορίες. Τη λειτουργία αυτή αναλαμβάνει ο *πράκτορας αναζήτησης εγγράφων*, που αλληλεπιδρά με το χρήστη και, μέσω παραδειγμάτων, παρατήρησης ή άμεσης υπόδειξης, ενημερώνεται για τις πηγές πληροφοριών που εκείνος βρίσκει ενδιαφέρουσες και αξιόπιστες. Στη συνέχεια, ο ίδιος πράκτορας αναλαμβάνει τη συλλογή πληροφοριών από αυτές τις πηγές σε τακτά χρονικά διαστήματα.

4 Υλοποίηση

Στην περιγραφή του προτεινόμενου μοντέλου διαλογής εγγράφων βάσει περιεχομένου, έμφαση έχει δοθεί στην απλότητα, ούτως ώστε το μοντέλο να είναι σαφώς ορισμένο και ελέγξιμο με βάση τις προδιαγραφές του. Για το λόγο αυτό έχουν γίνει οι ακόλουθες απλοποιήσεις:

- (A1) Το μοντέλο ασχολείται με ένα χρήστη και μια κατηγορία ενδιαφερόντων αυτού.
- (A2) Η διαδικασία κατασκευής των επικεφαλίδων είναι πάγια και ανεξάρτητη των περιεχομένων της βάσης γνώσης.

Το προτεινόμενο μοντέλο έχει χρησιμοποιηθεί για τη μερική υλοποίηση ενός πειραματικού συστήματος διαλογής ηλεκτρονικών εγγράφων σε περιβάλλον δημοσιογραφικού γραφείου, σε περιβάλλον δικτύου προσωπικών υπολογιστών με λειτουργικό σύστημα Microsoft Windows. Η ανάπτυξη

του συστήματος έχει γίνει σε Visual C++. Εκτός των απλοποιητικών παραδοχών που μόλις περιγράφηκαν, η μερική υλοποίηση έχει τους εξής επιπλέον περιορισμούς:

(A3) Ο πράκτορας κατασκευής επικεφαλίδων χειρίζεται μόνο αρχεία κειμένου και HTML.

(A4) Οι μορφές συνθηκών των κανόνων της βάσης γνώσης είναι οι απλούστερες δυνατές.

(A5) Ο πράκτορας αναζήτησης πληροφοριών δεν έχει υλοποιηθεί.

Η μελέτη της πειραματικής υλοποίησης δεν έχει ολοκληρωθεί ακόμα και δεν υπάρχουν αξιόπιστα και ικανοποιητικά σε όγκο αποτελέσματα από τη χρήση της. Στόχος μελλοντικής έρευνας, όσον αφορά στην υλοποίηση, θα είναι η αφενός η βελτίωσή της, αφετέρου η αξιολόγησή της σε πραγματικό περιβάλλον δημοσιογραφικού γραφείου.

5 Μελλοντικές κατευθύνσεις

Ο πρωταρχικός στόχος της μελλοντικής έρευνας είναι η αξιολόγηση του μοντέλου με την ανάπτυξη ρεαλιστικών υλοποιήσεων. Χρήσιμα στοιχεία για τη μελλοντική βελτίωση του μοντέλου μπορούν επίσης να προκύψουν από τα αποτελέσματα έρευνας στην περιοχή *ανάκτησης πλήρους κειμένου* (full text retrieval). Αν ο πράκτορας διαλογής είναι σε θέση να ερευνά το πλήρες κείμενο των ηλεκτρονικών εγγράφων, τότε είναι δυνατή η περιγραφή συνθηκών που βασίζονται στη σχετική θέση κάποιων λέξεων μέσα στο κείμενο. Η τεχνολογία εφαρμογών ανάκτησης πλήρους κειμένου χρησιμοποιείται ήδη από διάφορους πράκτορες αναζήτησης στο World Wide Web, με αρκετά ικανοποιητικά αποτελέσματα. Στο πλαίσιο του προτεινόμενου μοντέλου διαλογής, πάντως, δεν κρίνεται σκόπιμη η κατάργηση του στάδιου κατασκευής επικεφαλίδων.

Προκειμένου να ξεπερασθεί ο περιορισμός (A1), απαιτείται περαιτέρω έρευνα που θα αποσκοπεί στην υποστήριξη πολλών βάσεων γνώσης, που είτε θα ανήκουν σε διαφορετικούς χρήστες είτε θα εκφράζουν διαφορετικά ενδιαφέροντα των ίδιων χρηστών. Μια πιθανή προοπτική που θα εξετασθεί είναι αυτή της συνένωσης των διαφορετικών βάσεων γνώσης σε μια καθολική βάση (global knowledge base). Στην καθολική βάση γνώσης είναι επίσης δυνατό να υπάρχουν κοινώς αποδεκτά κριτήρια διαλογής, όπως και διάφορα κοινά χαρακτηριστικά του συστήματος.

Τέλος, προκειμένου να ξεπερασθεί ο περιορισμός (A2), στόχος μελλοντικής έρευνας θα είναι η αλληλεπίδραση του πράκτορα κατασκευής επικεφαλίδων με τη βάση γνώσης, προκειμένου ο τρόπος κατασκευής των επικεφαλίδων να αντανακλά τις ανάγκες των χρηστών. Θα εξετασθεί επίσης η δυνατότητα κατασκευής διαφορετικών επικεφαλίδων για κάθε χρήστη, ανάλογα με τα κριτήρια διαλογής του που περιέχονται στη βάση γνώσης.

6 Συμπεράσματα

Στην εργασία αυτή προτείνεται ένα μοντέλο για την αυτοματοποίηση της ανιάρης και χρονοβόρας διαδικασίας της διαλογής ηλεκτρονικών εγγράφων βάσει περιεχομένου. Το μοντέλο αυτό βασίζεται στη χρήση ευφών πρακτόρων, που αναλαμβάνουν τη διαμεσολάβηση μεταξύ των πηγών πληροφοριών και των αποδεκτών τους. Έμφαση δίνεται σε χαρακτηριστικά όπως η ομοιογενής αντιμετώπιση των ηλεκτρονικών εγγράφων, η εξατομίκευση, η αυτόματη εκμάθηση και προσαρμογή στις ανάγκες των χρηστών, η αίσθηση εμπιστοσύνης.

Παρά το γεγονός ότι η υλοποίηση του προτεινόμενου μοντέλου βρίσκεται σε πρώιμο στάδιο και δεν υπάρχουν ουσιαστικά και αξιόπιστα αποτελέσματα από τη χρήση της, οι πρώτες εμπειρίες από την εφαρμογή του στο χώρο ενός δημοσιογραφικού οργανισμού είναι ενθαρρυντικές. Εκφράζουμε την πεποίθηση ότι στο μέλλον η τεχνολογία των ευφών πρακτόρων θα δώσει ικανοποιητικές λύσεις

σε πολλά προβλήματα αυτοματοποίησης εργασιών που ταλαιπωρούν τον άνθρωπο, συμπεριλαμβανομένου του προβλήματος της διαλογής ηλεκτρονικών εγγράφων βάσει περιεχομένου.

Ευχαριστίες

Η ερευνητική αυτή εργασία χρηματοδοτείται από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας, στο πλαίσιο του Προγράμματος Ενίσχυσης Ερευνητικού Δυναμικού (ΠΕΝΕΔ 95), κωδικός έργου 463. Ευχαριστούμε επίσης τους ανώνυμους κριτές για τις εποικοδομητικές τους παρατηρήσεις.

Βιβλιογραφία

- [1] P. Maes, “Agents that reduce work and information overload,” *Communications of the ACM*, vol. 37, pp. 31–40, 146, July 1994.
- [2] M. R. Genesereth, “An agent-based approach to software interoperability,” in *Proceedings of the DARPA Software Technology Conference*, 1992.
- [3] Y. Lashkari, M. Metral, and P. Maes, *Collaborative Interface Agents*. Cambridge, MA: MIT Press, 1994.
- [4] B. Sheth and P. Maes, “Evolving agents for personalized information filtering,” in *Proceedings of the Ninth Conference on Artificial Intelligence for Applications*, pp. 345–352, IEEE Computer Society Press, 1993.
- [5] R. Kozierok and P. Maes, “A learning interface agent for scheduling meetings,” in *Proceedings of ACM SIGCHI International Workshop on Intelligent User Interfaces*, (New York, NY), pp. 81–88, ACM Press, 1993.
- [6] T. Selker, “Coach: A teaching agent that learns,” *Communications of the ACM*, vol. 37, pp. 92–99, July 1994.
- [7] Y. Shoham, “Agent-oriented programming,” *Artificial Intelligence*, vol. 60, no. 1, pp. 51–92, 1993.
- [8] “The Dublin Core Metadata Element Set.” URL: http://purl.org/metadata/dublin_core/.
- [9] “The Rufus System.” URL: <http://www.almaden.ibm.com/cs/showtell/rufus/>.
- [10] K. Shoen, A. Luniewski, P. Schwarz, J. Stamos, and J. Thomas, “The Rufus system: information organization for semi-structured data,” in *Proceedings of the 19th Conference on Very Large Databases (VLDB '93)*, (Dublin, Ireland), pp. 97–107, 1993.
- [11] M. Tresch, N. Palmer, and A. Luniewski, “Type classification of semi-structured documents,” in *Proceedings of the 21st Conference on Very Large Databases (VLDB '95)*, (Zurich, Switzerland), pp. 263–274, 1995.
- [12] D. Smith and M. Lopez, “Information extraction for semi-structured documents,” in *Proceedings of the Workshop on Management of Semistructured Data, in conjunction with PODS/SIGMOD*, (Tucson, AZ, USA), May 1997.