Γ' ΚΟΙΝΟΤΙΚΟ ΠΛΑΙΣΙΟ ΣΤΗΡΙΞΗΣ 2000-2006

ΑΝΑΛΥΤΙΚΗ ΕΚΘΕΣΗ ΠΡΟΟΔΟΥ ΦΥΣΙΚΟΥ ΑΝΤΙΚΕΙΜΕΝΟΥ ΕΡΓΟΥ

ΕΝΔΙΑΜΕΣΗ [ ]

ΤΕΛΙΚΗ [X]

Τίτλος Υποέργου : Αλγόριθμοι και στρατηγικές επιλογής πηγών πληροφοριών και σύνθεσης αποτελεσμάτων (collection fusion algorithms) σε κατανεμημένα συστήματα αναζήτησης πληροφοριών

Α/Α ΕΚΘΕΣΗΣ: 1

Κωδικός υποέργου	:	8.3.1
Μέτρο	:	8.3
Έργο/Δράση	:	03ΕΔ404
Αρμόδια Διεύθυνση ΓΓΕΤ	:	ΔΙΕΥΘΥΝΣΗ ΥΠΟΣΤΗΡΙΞΗΣ ΕΡΕΥΝΗΤΙΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ

1. ΣΤΟΙΧΕΙΑ ΥΠΟΕΡΓΟΥ

1.1. Τίτλος υποέργου

Αλγόριθμοι και στρατηγικές επιλογής πηγών πληροφοριών και σύνθεσης αποτελεσμάτων (collection fusion algorithms) σε κατανεμημένα συστήματα αναζήτησης πληροφοριών

1.2. Επιστημονικός Υπεύθυνος Υποέργου

Ονοματεπώνυμο	**Μιχάλης Σαλαμπασης**
Φορέας:	**ΤΕΙ Θεσσαλονίκης, Τμήμα Πληροφορικής**
Θέση:	Αναπληρωτής καθηγητής
Διεύθυνση:	Σίνδος, ΤΚ: 54101
Τηλ.:	2310 791595
E-mail:	cs1msa@it.teithe.gr

1.3. Ανάδοχος

Επωνυμία:	ΤΕΙ Θεσσαλονίκης, Τμήμα Πληροφορικής, Τομέας ανάλυσης & προγραμματισμού
Διεύθυνση:	Σίνδος, ΤΚ: 54101
Τηλ.:	2310 791101
E-mail:	protei@teithe.gr

1.4. Διάρκεια υποέργου:

		Προβλεπόμενη (αρχική σύμβαση και τροποποιήσεις αυτής)
Μήνες	:	36
Ημερομηνία έναρξης	:	28/11/2005
Ημερομηνία λήξης	:	28/11/2008
Χρονική παράταση (συνολικά σε σχέση με την αρχική σύμβαση)	:	Μέχρι 30/6/2009

1.5. Απόφαση έγκρισης εκτέλεσης υποέργου και τροποποιήσεις αυτής

	ΑΡ. ΠΡΩΤΟΚΟΛΛΟΥ	ΗΜΕΡΟΜΗΝΙΑ	ΑΙΤΙΑ ΤΡΟΠΟΠΟΙΗΣΗΣ
ΑΡΧΙΚΗ ΑΠΟΦΑΣΗ	20175	29/11/2005	
1Η ΤΡΟΠΟΠ. ΑΠΟΦΑΣΗΣ	5648(ΕΡΕ)815	24/7/2008	Αύξηση κατηγορίας δαπανών για ταξίδια

Comment [A1]: ΔΩΣΑΤΕ ΕΠΙΓΡΑΜΜΑΤΙΚΗ ΠΕΡΙΓΡΑΦΗ . ΟΠΩΣ π.χ. *ΑΛΛΑΓΗ ΜΕΛΟΥΣ ΟΜΑΔΑΣ ΕΡΓΟΥ, ΑΝΑΚΑΤΑΝΟΜΗ Π/Υ, ΑΛΛΑΓΗ Π/Υ, ΤΡΟΠΟΠΟΙΗΣΗ ΑΙΤΙΟΛΟΓΗΣΗΣ ΚΑΤΗΓΟΡΙΑΣ ΔΑΠΑΝΗΣ*

2. ΦΥΣΙΚΟ ΑΝΤΙΚΕΙΜΕΝΟ

2.1. Περίληψη πραγματοποιηθεισών εργασιών

Στην πρώτη φάση του έργου δόθηκε μεγάλη έμφαση στην εκπαίδευση του νέου ερευνητή. Δόθηκε βαρύτητα στην εξοικείωσή του με την σχετική βιβλιογραφία στον τομέα της ανάκτησης πληροφοριών, τόσο με κεντρικοποιημένα μοντέλα όσο και με κατανεμημένα. Προς αυτό τον σκοπό έγιναν δύο σειρές σεμιναρίων από τους καθηγητές κ. Γιώργο Ευαγγελίδη και κ. Μαρία Σατρατζέμη. Παράλληλα, έγινε η εξοικείωση του ερευνητή με την έννοια των πειραμάτων στον τομέα. Έτσι, αναπτύχθηκαν οι ήδη υπάρχουσες και διαδεδομένες συλλογές TREC και WEBTrec ενώ παράλληλα χρειάστηκε και η πρόσθετη ανάπτυξη 2 νέων πειραματικών συλλογών για την εξομοίωση ενός όσο το δυνατόν πιο ρεαλιστικού πειραματικού περιβάλλοντος. Στις πειραματικές αυτές συλλογές, ο νέος ερευνητής μελέτησε την απόδοση και λειτουργία μεγάλου αριθμού μοντέλων ανάκτησης πληροφοριών, τόσο σε κεντρικοποιημένες προσεγγίσεις όσο και σε κατανεμημένες με σκοπό να εμβαθύνει στον τρόπο λειτουργίας τους και να εντοπίσει τυχόν μειονεκτήματα και προβλήματά τους.

Στη συνέχεια ο νέος ερευνητής, έχοντας μελετήσει τις ήδη υπάρχουσες προσεγγίσεις προχώρησε στην ανάπτυξη δύο νέων αλγορίθμων κατανεμημένης ανάκτησης πληροφοριών. Ο πρώτος αλγόριθμος επιλογής πηγών, ο οποίος λειτουργεί αποκλειστικά με την χρήση υπερσυνδέσμων ανάμεσα σε κείμενα του διαδικτύου, καλείται να καλύψει το κενό που υπάρχει στην βιβλιογραφία για την χρήση υπερσυνδέσμων στον τομέα αυτό. Στη συνέχεια, ο νέος ερευνητής προχώρησε στην ανάπτυξη και ενός νέου αλγορίθμου σύνθεσης αποτελεσμάτων, ο οποίος κάνει χρήση πολλαπλών μοντέλων παλινδρόμησης. Ο αλγόριθμος αυτός μπορεί να λειτουργήσει αποδοτικά σε περιβάλλοντα όπου οι απομακρυσμένες συλλογές δεν επιστρέφουν σκορ σχετικότητας μαζί με τα κείμενα, αλλά μονάχα καταταγμένα κείμενα (όπως συμβαίνει κατά κόρον στο διαδίκτυο) καθιστώντας τον εφαρμόσιμο και χρήσιμο σε περιβάλλοντα μη-συνεργατικά, όπως είναι αυτά που βρίσκονται στο διαδίκτυο. Τα αποτελέσματα αυτής της σειράς ερευνητικών εργασιών δημοσιευθήκαν στα συνέδρια ECIR 07, PCI07 καθώς και στο περιοδικό Information Processing & Management.

Ό ΥΔ στη συνέχεια ανέπτυξε ένα νέο υβριδικό αλγόριθμο σύνθεσης αποτελεσμάτων (Hybrid Results Merging) και ο οποίος δημοσιεύθηκε στο συνέδριο CIKM 07.

Ο ΥΔ στη συνέχεια ανέπτυξε ένα νέο αλγόριθμο για επιλογή συλλογών με βάση τη μοντελοποίηση των κατανεμημένων συλλογών και της σχετικότητας των κειμένων που αυτά έχουν με τη χρήση ολοκληρωμάτων. Έγινε μία σειρά πειραμάτων που δημοσιεύτηκαν στο CIKM08 (LSDS-IR'08 workshop) καθώς επίσης και στα περιοδικά Information Processing & Management και Information Sciences (οι δημοσιεύσεις σε αυτά τα δύο περιοδικά θα γίνουν το 2009).

Επίσης έγινε μία μελέτη σχετικά με την απόδοση κεντρικοποιημένων και κατανεμημένων συστημάτων ανάκτησης που δημοσιεύτηκε στο συνέδριο PCI 08 καθώς και μία εργασία σχετικά με την ανάλυση και δημιουργία ευρετηρίων ελληνικών κειμένων που δημοσιεύτηκε στο 2nd International ACM Workshop Improving Non-English Web Searching (iNEWS08) του CIKM08.

Τέλος έγινε και μία εργασία σχετικά με την επιλογή πηγών (source selection) με βάση data fusion algorithms. Αυτή η εργασία δημοσιεύτηκε στο συνέδριο ECIR 09.

Τον Φεβρουάριο του 2009 ο υποψήφιος διδάκτορας υπέβαλλε την διδακτορική του διατριβή (στα Αγγλικά κάνοντας χρήση της σχετικής νομοθεσίας) και ακολούθως συγκλήθηκε επταμελής επιτροπή. Η διατριβή του ΥΔ εξετάστηκε τον Απρίλιο και ομόφωνα βαθμολογήθηκε με άριστα.

Αμέσως παρακάτω ακολουθεί η περίληψη της διδακτορικής διατριβής στα ελληνικά.

ΠΕΡΙΛΗΨΗ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ του ΥΔ ΣΤΑ ΕΛΛΗΝΙΚΑ

Οι γενικές μηχανές αναζήτησης, όπως η Google και η Yahoo!, παρέχουν ένα εύκολο μηχανισμό για τους χρήστες τους για να βρίσκουν πληροφορίες στο Διαδίκτυο. Πέραν των φανερών πλεονεκτημάτων τους όμως, έχουν ένα σημαντικό αριθμό περιορισμών, επειδή δεν μπορούν να προσεγγίσουν και να αναλύσουν ένα σημαντικό μέρος της πληροφορίας που είναι διαθέσιμη. Τα Κατανεμημένα Συστήματα Αναζήτησης Πληροφοριών, κάνοντας χρήση αλγορίθμων συγχώνευσης συλλογών, παρέχουν μία λύση στο παραπάνω πρόβλημα, επιτρέποντας στους χρήστες τους να υποβάλλουν ερωτήματα συγχρόνως σε πολλαπλές πηγές πληροφόρησης, παρέχοντας μία πολύ μεγαλύτερη κάλυψη της διαθέσιμης πληροφορίας.

Αυτή η διατριβή ασχολείται με δύο από τα βασικά προβλήματα που αφορούν στον σχεδιασμό και στην υλοποίηση αποτελεσματικών και αποδοτικών Κατανεμημένων Συστημάτων Αναζήτησης Πληροφοριών: την επιλογή πηγών και την σύνθεση αποτελεσμάτων. Το πρώτο πρόβλημα ασχολείται με την ικανότητα του συστήματος να επιλέγει τις πιο κατάλληλες πηγές πληροφόρησης για να μεταβιβάσει το ερώτημα του χρήστη και το δεύτερο αποβλέπει στο να παράξει την καλύτερη δυνατή τελική λίστα κειμένων μέσω της σύνθεσης των επιμέρους ανακτημένων κειμένων από τις επιλεγμένες πηγές.

Οι νέοι αλγόριθμοι που παρουσιάζονται σε αυτή τη διατριβή έχουν σχεδιαστεί ώστε να λειτουργούν αποτελεσματικά σε περιβάλλοντα όπου οι πηγές δεν παρέχουν καμία συνεργασία, με αποτέλεσμα να είναι εφαρμόσιμη στην μεγαλύτερο δυνατό σύνολο περιβαλλόντων και συνθηκών. Ο αλγόριθμος επιλογής πηγών που προάγεται παρέχει έναν καινοτόμο τρόπο μοντελοποίησης των πηγών ως περιοχές σε ένα χώρο που παράγεται από τα κείμενα τα οποία περιέχουν. Διατυπώνει ένα πλήρες θεωρητικό πλαίσιο επίλυσης του προβλήματος της επιλογής πηγών, ενώ παράλληλα αποτελεσματικά συλλαμβάνει πειραματικές παρατηρήσεις και γενικά αποδεκτές αντιλήψεις στον τομέα της Ανάκτησης Πληροφοριών. Εκτεταμένα πειράματα επιδεικνύουν ότι είναι ικανός να διασφαλίσει μία απόδοση που είναι τουλάχιστον τόσο καλή όσο άλλες μεθοδολογίες αιχμής και συχνότερα καλύτερη.

Οι νέοι αλγόριθμοι σύνθεσης αποτελεσμάτων που παρουσιάζονται είναι βασισμένη στην υπόθεση ότι οι μηχανές αναζήτησης επιστρέφουν μονάχα κατατάξεις κειμένων, χωρίς σκορ σχετικότητας, ένα σενάριο που είναι συνήθης πρακτική σε σύγχρονα συστήματα ανάκτησης πληροφοριών. Και οι δύο επιλύουν το πρόβλημα της έλλειψης πληροφόρησης πολύ αποτελεσματικά, επιδεικνύοντας σημαντικά οφέλη στην απόδοση συγκρίσει με άλλους αλγορίθμους αιχμής. Επιπροσθέτως, ο δεύτερος αλγόριθμος ενοποιεί τις δύο γενικές κατευθύνσεις από τις οποίες έχει προσεγγιστεί το πρόβλημα στην έρευνα, συνδυάζοντας τα πλεονεκτήματά τους, ενώ συγχρόνως ελαχιστοποιώντας τα μειονεκτήματά τους.

2.2. Αναλυτική Περιγραφή των Ενοτήτων Εργασιών (Ε.Ε.) που υλοποιήθηκαν

Στην πρώτη φάση αυτή του έργου, δόθηκε μεγάλο βάρος στην εκπαίδευση του νέου ερευνητή σε θέματα εισαγωγής και εμβάθυνσης στον τομέα της «Ανάκτησης Πληροφοριών» αρχικά και της «Κατανεμημένης Ανάκτησης Πληροφοριών» σε δεύτερη φάση.
Ο νέος ερευνητής είχε την δυνατότητα να μελετήσει την διεθνή βιβλιογραφία στον τομέα, ώστε να αποκτήσει την απαραίτητη εξοικείωση που θα του επέτρεπε να συνεχίσει με καινοτόμα έρευνα. Δόθηκε μεγάλη βαρύτητα στην κατανόηση των ήδη ανεπτυγμένων αλγορίθμων «Ανάκτησης Πληροφοριών» σε κεντρικοποιημένα και κατανεμημένα συστήματα, ώστε να εντοπιστούν οι τυχόν ελλείψεις και τα μειονεκτήματά τους και να επικεντρωθεί η ερευνητική προσπάθεια στην κάλυψη αυτών των κενών.

Παράλληλα, δόθηκε μεγάλη βαρύτητα στην πειραματική διαδικασία και στην εξοικείωση του νέου ερευνητή με τις πειραματικές διαδικασίες που ακολουθούνται στον τομέα. Κρίθηκε απαραίτητη η αναπαραγωγή πειραμάτων από την διεθνή βιβλιογραφία, ώστε να επιβεβαιωθεί η εγκυρότητα του πειραματικού περιβάλλοντος που αναπτύχθηκε και παράλληλα να εξοικειωθεί ο νέος ερευνητής με την έννοια και τις απαιτήσεις των πειραμάτων στον τομέα Information Retrieval.
Παράλληλα, έγιναν δύο σειρές σεμιναρίων κατά την διάρκεια του προγράμματος ως τώρα. Η πρώτη σειρά, συνολικής διάρκειας 6 ωρών έγινε τον Ιανουάριο 2006 από τον κ. Ευαγγελίδη και η δεύτερη, συνολικής διάρκειας 6 ωρών, τον Φεβρουάριο 2006 από την κ. Σατρατζέμη.
Η πρώτη σειρά σεμιναρίων είχε ως σκοπό την εξοικείωση του ερευνητή σε θέματα ανάκτησης πληροφοριών από βάσεις δεδομένων και εξόρυξης γνώσης, τομείς οι οποίοι είναι κριτικής σημασίας στην επιτυχημένη (σε θέματα απόδοσης και ακρίβειας) διαδικασία collection fusion.
Η δεύτερη σειρά σεμιναρίων από την κ. Σατρατζέμη είχε ως σκοπό την εξοικείωση του νέου ερευνητή με θεωρίες γραφημάτων, οι οποίες κρίνονται ως απαραίτητες για την εκμετάλλευση των υπερσυνδέσμων σε περιβάλλοντα διαδικτύου.

Έχοντας εξοικειωθεί με την διεθνή βιβλιογραφία και τα ποίκιλλα μοντέλα κεντρικοποιημένης ανάκτησης πληροφοριών σε θεωρητικό επίπεδο, το επόμενη βήμα για τον νέο ερευνητή ήταν να αναπαράγει κάποια από τα δημοσιευμένα πειράματα ώστε να εξοικειωθεί σε πρακτικό επίπεδο με την έννοια των ad-hoc πειραμάτων.

Τα πειράματα τα οποία έγιναν βασίστηκαν σε τρεις πειραματικές συλλογές, διεθνούς κύρους, οι οποίες έχουν χρησιμοποιηθεί σε πληθώρα ερευνών. Συγκεκριμένα οι συλλογές ήταν οι: TREC (η οποία αποτελείται αποκλειστικά από ειδησεογραφικά άρθρα), WEBTREC WT10g (η οποία αποτελείται από ένα κομμάτι – crawl του προσβάσιμου διαδικτύου το έτος 1998) και GOV (η οποία αποτελείται από ιστοσελίδες παρμένες από τους κρατικούς ιστοτόπους των ΗΠΑ).

Η κάθε μία συλλογή έχει τα δικά της χαρακτηριστικά και επικεντρώνεται σε διαφορετικά στοιχεία του διαδικτύου όπως αυτό έχει αναπτυχθεί σήμερα, εξ'ού και κρίθηκε σημαντικό η διενέργεια πειραμάτων σε όλες τις συλλογές. Συγκεκριμένα, η συλλογή TREC αποτελούμενη από ειδησεογραφικά άρθρα, χαρακτηρίζεται από υψηλής ποιότητας περιεχόμενό. Τα στοιχεία αυτά είναι όμοια με αυτά που υπάρχουν σε ιστοτόπους hidden web (κρυμμένους από τις γενικές μηχανές αναζήτησης), οι οποίοι είναι ανεπτυγμένοι με βάση θεματικές ενότητες, δημιουργώντας πολύ υψηλής ποιότητας περιεχόμενο. Η WT10g συλλογή αποτελεί μία αρκετά πιστή αναπαράσταση του ευρέως διαδεδομένου διαδικτύου (όσον αφορά ποικίλη θεματολογία και ύπαρξη υπερσυνδέσμων) και τέλος η GOV συλλογή αποτελούμενη από ιστοσελίδες παρμένες από κρατικούς ιστοτόπους, έχει αποκλειστικά ποιοτικό πολιτικό-οικονομικό περιεχόμενο, ενώ παράλληλα συνδυάζει και στοιχεία διαδικτύου.

Τα μοντέλα κεντρικοποιημένης ανάκτησης πληροφοριών που δοκιμάστηκαν βασίζονταν σε τρεις παραλλαγές γλωσσολογικών στατιστικών μοντέλων (language statistical models), δίκτυα εμπιστοσύνης (inference networks) και στατιστικά ευρετικά (heuristic) μοντέλα ανάκτησης πληροφοριών τα οποία αποτελούν και το state-of-the-art στον τομέα. Η απόδοση των μοντέλων αυτών καταγράφηκε και στην συνέχεια έγινε μία ανάλυση των αποτελεσμάτων ώστε να εντοπιστούν τυχόν αποκλίσεις από την διεθνή βιβλιογραφία. Παράλληλα εντοπίστηκαν τα μειονεκτήματα αυτών των μοντέλων και επισημάνθηκαν οι τομείς στους οποίους μία κατανεμημένη προσέγγιση θα τα αναιρούσε. Τα συμπεράσματα που αναπτυχθήκαν συνοψίζονται στα εξής:

1. Αυξημένη ύπαρξη θορύβου (μη-σχετικών σελίδων) στα τελικά αποτελέσματα.
2. Ασταθής απόδοση αναλόγως της «ευκολίας» ή μη του ερωτήματος. Ερωτήματα τα οποία περιείχαν συχνές λέξεις επέστρεφαν πολλά κείμενα, ενώ ερωτήματα τα οποία περιείχαν σπάνιες λέξεις, πολύ λιγότερα.
3. Αυξημένοι χρόνοι επιστροφής αποτελεσμάτων, αναλόγως του μεγέθους της συλλογής.
4. Κείμενα τα οποία είχαν μεγαλύτερο πλήθος λέξεων ήταν πιο ευνοημένα από άλλα μικρότερου μεγέθους.

Στην συνέχεια μελετήθηκαν κατανεμημένα μοντέλα ανάκτησης πληροφοριών για να διαπιστωθεί κατά πόσο αναιρούσαν τα προαναφερθέν προβλήματα. Κρίθηκε και πάλι σκόπιμο η τέλεση πειραμάτων με τα μοντέλα αυτά ώστε να κατανοηθεί σε βάθος ο τρόπος λειτουργίας τους, τα πλεονεκτήματα που προσφέρουν σε σχέση με τις κεντρικοποιημένες προσεγγίσεις και τέλος τα τυχόν μειονεκτήματα που έχουν οι ήδη ανεπτυγμένοι αλγόριθμοι κατανεμημένης ανάκτησης.
Συγκεκριμένα, μελετήθηκαν και υλοποιήθηκαν δύο από τους πιο γνωστούς αλγορίθμους «επιλογής πηγών» (Source selection): CORI και KL-Divergence.

Ο πρώτος είναι βασισμένος σε δίκτυα εμπιστοσύνης (inference networks) και αποτελεί την βάση σύγκρισης (baseline) με την οποία γίνονται πειράματα επιλογής πηγών. Αποτελεί παραλλαγή του Inquiry αλγόριθμου για ανάκτηση πληροφοριών σε κεντρικοποιημένα συστήματα. Παρόμοια, ο KL-Divergence είναι βασισμένος σε γλωσσολογικά μοντέλα ανάκτησης πληροφοριών. Τα μειονεκτήματα των δύο αυτών προσεγγίσεων είναι ότι προσομοιώνουν την κάθε απομακρυσμένη συλλογή του κατανεμημένου συστήματος ως το άθροισμα των κειμένων που την αποτελούν (bag of words approach). Η προσέγγιση αυτή, αν και είναι αποδοτική σε κάποιο βαθμό, αποτυγχάνει να συλλάβει υπόψη τις ιδιαιτερότητες τις κατανεμημένης ανάκτησης πληροφοριών, ότι δηλαδή πέρα από την σχετικότητα της κάθε συλλογές με βάση το ερώτημα που θέτει ο χρήστης (collection relevancy), αυτό που πρέπει επίσης να ληφθεί σοβαρά υπόψη είναι ο αριθμός των σχετικών κειμένων που μπορεί να επιστρέψει η συλλογή. Η παρατήρηση αυτή είναι η αφορμή της ανάπτυξης ενός νέου αλγορίθμου ο οποίος θα λαμβάνει υπόψη και θα επιδιώκει συγκεκριμένα να μεγιστοποιήσει τον αριθμό των επιστρεφόμενων σχετικών κειμένων από κάθε συλλογή. Επισημάνθηκε η μη-χρήση υπερσυνδέσμων (hyperlinks) από αυτούς τους αλγορίθμους, ένα στοιχείο το οποίο θα μπορούσε να αυξήσει την απόδοσή τους.

Παράλληλα μελετήθηκαν και υλοποιήθηκαν και δύο αλγόριθμοι σύνθεσης αποτελεσμάτων (Results Merging): CORI Results Merging και Semi-Supervised Learning, η οποίοι θεωρούνται state-of-the-art. Μελετήθηκε η απόδοσή τους και εντοπίστηκαν τα μειονεκτήματά τους και τα σημεία στα οποία θα μπορούσε να επικεντρωθεί νέα καινοτόμα έρευνα. Ένα από τα μειονεκτήματα τα οποίο εντοπίστηκε είναι ότι για να λειτουργήσουν αποδοτικά αυτοί οι αλγόριθμοι πρέπει οι απομακρυσμένες συλλογές του κατανεμημένου συστήματος ανάκτησης πληροφοριών να αναφέρουν και την σχετικότητα κάθε κειμένου που επιστρέφουν σε ποσοτικοποιημένη μορφή (relevance score). Καθώς όμως στα σύγχρονα περιβάλλοντα ανάκτησης πληροφοριών (π.χ. γενικές μηχανές αναζήτησης κλπ) τα σκορ αυτά δεν αναφέρονται, αυτοί οι αλγόριθμοι παρουσιάζουν σημαντικά προβλήματα εφαρμογής σε ρεαλιστικές συνθήκες.

Ένα δεύτερο μειονέκτημα το οποίο επίσης εντοπίστηκε είναι ότι ενώ οι αλγόριθμοι σύνθεσης αποτελεσμάτων που κάνουν χρήση μεθοδολογιών παλινδρόμησης (linear regression), όπως είναι ο Semi-Supervised Learning, απαιτούν από τις κατανεμημένες συλλογές να επιστρέφουν έναν σημαντικό αριθμό κειμένων, της τάξης των 300-1000, για να εκπαιδεύσουν ένα επαρκώς ακριβές μοντέλο. Καθώς όμως οι περισσότερες μηχανές αναζήτησης έχουν ένα μέγιστο αριθμό 100 αποτελεσμάτων ανά σελίδα, αυτό σημαίνει πως απαιτείται πολλαπλή αλληλεπίδραση με την μηχανή αναζήτησης, που μεταφράζεται σε σημαντική καθυστέρηση της διαδικασίας σύνθεσης αποτελεσμάτων. Με βάση τις παρατηρήσεις αυτές αναπτύχθηκαν δύο νέοι αλγόριθμοι σύνθεσης αποτελεσμάτων. Ο πρώτος αναιρεί την ανάγκη επιστροφής relevance scores από τις απομακρυσμένες συλλογές ενώ ο δεύτερος, ο οποίος αποτελεί προέκταση του προηγούμενου, είναι σχεδιασμένος ώστε να λειτουργεί με επάρκεια σε ρεαλιστικά περιβάλλοντα ανάκτησης πληροφοριών, δηλαδή σε περιβάλλοντα όπου οι απομακρυσμένες πηγές επιστρέφουν ένα μικρό αριθμό κειμένων, της τάξης του 10-20 κείμενα ανά ερώτημα. Ο πρώτο

αλγόριθμος έχει ήδη υλοποιηθεί και περιγράφεται παρακάτω ενώ ο δεύτερος αλγόριθμος βρίσκεται στην φάση σχεδιασμού γι' αυτό και δεν περιγράφεται εδώ.

Τα πειράματα έγιναν στις συλλογές που έχουν ήδη αναφερθεί, σε διάφορες κατανομές (τρόπους διαχωρισμού). Ο νέος ερευνητής προχώρησε στην ανάπτυξη κατανεμημένων συλλογών με απώτερο στόχο την υλοποίηση και τον πειραματισμό με κατανεμημένα μοντέλα αναζήτησης. Από τις συλλογές που αναπαράχθηκαν, κάποιες έχουν ήδη χρησιμοποιηθεί κατά κόρον στην διεθνή βιβλιογραφία, ενώ 2 καινούργιες συλλογές δημιουργήθηκαν για τις ανάγκες απεικόνισης ενός κατά το δυνατόν πιο ρεαλιστικού πειραματικού δείγματος.

Η TREC συλλογή διαχωρίστηκε με δύο μεθοδολογίες. Σύμφωνα με την πρώτη μεθοδολογία (trec123-100col-bysource), τα κείμενα χωρίστηκαν σε συλλογές με βάση το ειδησεογραφικό πρακτορείο έκδοσης του άρθρου και της ημερομηνίας έκδοσης. Έτσι π.χ. τα άρθρα από το Wall Street Journal χωρίστηκαν σε 15 επιμέρους συλλογές με βάση την ημερομηνία. Σύμφωνα με την δεύτερη μεθοδολογία (trec4-kmeans), εφαρμόστηκε ένας αλγόριθμος clustering (k-means) ο οποίος χώρισε τα άρθρα με βάση το περιεχόμενό τους. Οι διαφορές των δύο μεθοδολογιών είναι πολύ σημαντικές και οδηγούν σε πολλά συμπεράσματα:

1. Η πρώτη μεθοδολογία οδήγησε σε ομοιογενείς συλλογές όσον αφορά το μέγεθός τους. Εν αντιθέσει, τα μεγέθη των συλλογών που έγιναν σύμφωνα με την δεύτερη μεθοδολογία ποικίλλουν σημαντικά σε μέγεθος.
2. Η δεύτερη μεθοδολογία οδήγησε σε συλλογές πολύ πιο «θεματικά» ομοιογενείς. Όλα τα κείμενα που έχουν να κάνουν με συγκεκριμένα θέματα έχουν συγκεντρωθεί σε περιορισμένο αριθμό συλλογών, δημιουργώντας πολύ πιο επικεντρωμένες και συγκεκριμένες συλλογές. Αντίθετα, στην πρώτη συλλογή η θεματογραφία είναι πιο ομοιόμορφα κατανεμημένη.
3. Η κατανομή των λέξεων στην δεύτερη μεθοδολογία είναι πολύ ασύμμετρη, δηλαδή λέξεις που υπάρχουν σε πληθώρα σε μία συλλογή δεν είναι τόσο κοινές σε άλλες συλλογές. Ακόμα περισσότερο, λέξεις που δεν είναι συχνές (ορολογίες π.χ.) εμφανίζουν πολύ ασύμμετρη κατανομή.

Η συλλογή WebTREC διαχωρίστηκε κατά δύο τρόπους επίσης, παράγοντας δύο νέες συλλογές, η κάθε μία από τις οποίες έχει ως στόχο να δημιουργήσει μία όσο το δυνατόν πιο πιστή απεικόνιση συγκεκριμένων ιδιοτήτων του διαδικτύου.

Η πρώτη από αυτές τις συλλογές αυτές είναι βασισμένη σε μία παραλλαγή της WebTREC (WT-Dense), η οποία χρησιμοποιήθηκε σε πειράματα που έκαναν χρήση των υπερσυνδέσμων (hyperlinks) ανάμεσα στις ιστοσελίδες. Με την χρήση μεθοδολογιών ομαδοποίησης (clustering) ο νέος ερευνητής ήταν σε θέση να διαχωρίσει την άνωθεν πειραματική συλλογή σε επιμέρους υποσύνολα, κατάλληλα για την τέλεση πειραμάτων κατανεμημένης αναζήτησης. Συγκεκριμένα, η συλλογή χωρίστηκε αρχικά σε υποσύνολα με βάση τους ιστοτόπους από τους οποίους προέρχονταν. Στην συνέχεια οι «αρχικές σελίδες» αυτών των ιστοτόπων ομαδοποιήθηκαν με βάση το περιεχόμενό τους με ένα kmeans clustering αλγόριθμο. Τέλος, ολόκληροι οι ιστοτόποι ομαδοποιήθηκαν με βάση των διαχωρισμό των αρχικών τους σελίδων. Η διαδικασία αυτή είχε ως αποτέλεσμα την δημιουργία υποσυνόλων με αρκετά ομογενή περιεχόμενο, χωρίς όμως να παραβιάζεται η ολότητα του ιστοτόπου. Η διαδικασία αυτή επαναλήφθηκε αρκετές φορές με διαφορετικό αριθμό υποσυνόλων κάθε φορά.

Η δεύτερη συλλογή είναι βασισμένη και πάλι στην WebTREC και αποτελεί μία απόπειρα εξέλιξης των στάνταρντ TREC πειραματικών συλλογών για κατανεμημένα πειράματα. Αν και οι συλλογές αυτές έχουν χρησιμοποιηθεί κατά κόρον στην βιβλιογραφία, παρόλα αυτά παρουσιάζουν σημαντικά μειονεκτήματα: πρώτον δεν είναι βασισμένες στο διαδίκτυο (είναι άρθρα από ειδησεογραφικά πρακτορεία και όχι ιστοσελίδες), είναι διαχωρισμένες σε επιμέρους συλλογές με τεχνητό τρόπο και προσφέρουν τον ίδιο βαθμό διαχωρισμού (100 συλλογές). Η νέα συλλογή που αναπτύχθηκε αναιρεί τα μειονεκτήματα αυτά καθώς είναι βασισμένη σε ιστοσελίδες του διαδικτύου, είναι φυσικά διαχωρισμένες σε ιστοτόπους, όπως αυτοί αναπτύχθηκαν από τους δημιουργούς τους και τέλος προσφέρει πολύ μεγαλύτερη κατανομή (1000 συλλογές). Ο νέος αλγόριθμος που σχεδιάζεται αναμένεται να δοκιμαστεί στην συλλογή αυτή.

Θα πρέπει στο σημείο αυτό να επισημανθούνε κάποιες γενικές παρατηρήσεις που αφορούν στην κατανεμημένη ανάκτηση πληροφοριών γενικότερα και στην σύγκριση αυτής με κεντρικοποιημένες προσεγγίσεις.

1. Η απόδοση των κατανεμημένων μοντέλων σε σύγκριση με κεντρικοποιημένα μοντέλα γενικά ποικίλλει. Συγκεκριμένα, σε συλλογές όπου ο διαχωρισμός των κειμένων έχει γίνει με αλγορίθμους ομαδοποίησης (clustering), η απόδοση των κατανεμημένων προσεγγίσεων είναι σταθερά πάνω από τις κεντρικοποιημένες. Σε συλλογές όπου ο διαχωρισμός των κειμένων είναι αυθαίρετος ή έστω λιγότερο ομογενείς, η επίδοση των κατανεμημένων συστημάτων είναι γενικά χαμηλότερη από αυτή των κεντρικοποιημένων προσεγγίσεων.
2. Η απόδοση των συστημάτων κατανεμημένης ανάκτησης πληροφοριών επηρεάζεται από όλους τους παράγοντες και τις διαδικασίες που εμπλέκονται. Τόσο η επιλογή των καλύτερων πηγών όσο και η τελική σύνθεση των αποτελεσμάτων είναι κριτικής σημασίας για την επίτευξη του καλύτερου αποτελέσματος. Αν έστω και ένα κομμάτι της διαδικασία αποτύχει ή έστω δεν έχει την αναμενόμενη απόδοση, τότε το σύνολο της διαδικασίας έχει μειωμένη απόδοση.
3. Η συνεργασία των απομακρυσμένων συλλογών δεν είναι απαραίτητη για την επίτευξη των καλύτερων αποτελεσμάτων. Αν και σε συνθήκες όπου υπάρχει συνεργασία, η επίδοση των κατανεμημένων συστημάτων είναι γενικά καλύτερη, η διαφορά είναι μικρή. Εξάλλου, έχουν αναπτυχθεί μεθοδολογίες οι οποίες έχουν την δυνατότητα να εξάγουν από μη-συνεργάσιμες πηγές σημαντικά στοιχεία, όπως τον αριθμό των κειμένων που περιέχουν, την γενική θεματολογία τους κλπ.

Στα πειράματα που έγιναν έγινε χρήση κυρίως μη συνεργατικών συλλογών, οι οποίες αποτελούν και την πλειοψηφία των συλλογών που υπάρχουν σε ρεαλιστικά δεδομένα (όπως π.χ. στο διαδίκτυο).

Έχοντας μελετήσει τις υπάρχουσες προσεγγίσεις σε θέματα κατανεμημένης αναζήτησης και έχοντας εντοπίσεις τις αδυναμίες τους, ο νέος ερευνητής στην συνέχεια προχώρησε στην ανάπτυξη νέων αλγορίθμων. Στην πρώτη φάση της έρευνάς του επικεντρώθηκε σε δύο κομμάτια. Αφ'ενός αναπτύχθηκε ένας αλγόριθμος επιλογής πηγών (source selection) ο οποίος κάνει χρήση των υπερσυνδέσμων ανάμεσα στα κείμενα (Link-Based source selection) και αφ'ετέρου, στο κομμάτι της σύνθεσης αποτελεσμάτων (results merging), αναπτύχθηκε ένας νέος αλγόριθμος για να αναπληρώσει τα κενά που υπήρχαν και επισημάνθηκαν παραπάνω (δηλαδή, την χρήση relevancy scores από τις κατανεμημένες συλλογές). Οι αλγόριθμοι αυτοί εξετάστηκαν στο κατάλληλο πειραματικό περιβάλλον που περιγράφηκε παραπάνω. Επισημαίνεται και πάλι ότι ο δεύτερος αλγόριθμος σύνθεσης αποτελεσμάτων βρίσκεται ακόμα σε φάση σχεδιασμού οπότε και δεν κατέστη δυνατόν να περιγραφεί εδώ.

ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΠΗΓΩΝ ΜΕ ΧΡΗΣΗ ΥΠΕΡΣΥΝΔΕΣΜΩΝ (LINK-BASED SOURCE SELECTION)

Ο αλγόριθμος επιλογής πηγών με χρήση υπερσυνδέσμων που προτείνεται, προσπαθεί να προσεγγίσει τη διανομή των σχετικών εγγράφων μεταξύ των κατανεμημένων συλλογών. Η προτεινόμενη στρατηγική επιλογής πηγών συνδυάζει δύο χαρακτηριστικά γνωρίσματα που δεν βρίσκονται συνήθως σε άλλες μεθόδους που καθιστούν τον αλγόριθμο εφαρμόσιμο σε δυναμικά περιβάλλοντα ανάκτησης πληροφοριών. Κατ' αρχάς, λύνει το πρόβλημα επιλογής πηγής αποκλειστικά με την χρήση υπερσυνδέσμων ανάμεσα σε κείμενα. Δεύτερον, δεν απαιτεί οποιαδήποτε φάση εκμάθησης (training phase) προτού να μπορέσει να χρησιμοποιηθεί.

Η συγκεκριμένη στρατηγική είναι βασισμένη στην αποκαλούμενη link-hypothesis (δηλ. συνδεμένα έγγραφα τείνουν να είναι σχετικά με την ίδια πληροφοριακή ανάγκη). Ενσωματώνει την έννοια της «αυθεντίας» που βρίσκεται σε άλλες προσεγγίσεις κεντρικοποιημένης ανάκτησης και την εφαρμόζει σε ένα κατανεμημένο περιβάλλον. Συγκεκριμένα, ο αλγόριθμος υποθέτει ότι για κάθε πληροφοριακή ανάγκη, που εκφράζεται ως ερώτημα Q, υπάρχουν ορισμένες συλλογές που είναι αυθεντίες (authoritative), οι οποίες προσδιορίζονται από τον αριθμό σχετικών εγγράφων που περιέχουν (συλλογές αυθεντίες αναμένεται να περιέχουν πολλά σχετικά κείμενα). Ο στόχος του αλγορίθμου είναι να βρεθούν αυτές ακριβώς οι συλλογές, χρησιμοποιώντας την τοπολογία γραφικών παραστάσεων του διαδικτύου.

Η έκδοση του αλγορίθμου που παρουσιάζεται εδώ χρησιμοποιεί τις εξερχόμενες υπερσυνδέσεις (outlinks) για να εντοπίσει εκείνες τις συλλογές «αυθεντίες». Συγκεκριμένα, ο αλγόριθμος λειτουργεί σε δύο βήματα. Κατ' αρχάς, λαμβάνοντας υπόψη μια πληροφοριακή ανάγκη που εκφράζεται από ένα ερώτημα Q, μια αρχική αναζήτηση γίνεται χρησιμοποιώντας το Q για να ανακτηθούν m έγγραφα από μία δειγματοληπτική συλλογή (sampling collection). Η χρησιμοποίηση της δειγματοληπτικής συλλογής είναι μια πολύ συνήθης πρακτική στους περισσότερους αλγορίθμους κατανεμημένης ανάκτησης πληροφοριών και εξετάζεται σε περισσότερο βάθος στη συνέχεια. Τα έγγραφα m που ανακτώνται από τη δειγματοληπτική συλλογή υποβάλλονται σε επεξεργασία ώστε να εξαχθούν οι εξερχόμενες συνδέσεις (outlinks).

Δεύτερον, οι πληροφορίες αυτές περνούν σε μια συνάρτηση μεγιστοποίησης που εκτιμά τη διανομή των σχετικών εγγράφων στις απομακρυσμένες συλλογές. Η βασική έκδοση του αλγορίθμου χρησιμοποιεί αναλογικές μεθοδολογίες για να προσεγγίσει τη σχετικότητα κάθε συλλογής. Αυτή η εκτίμηση μαζί με το συνολικό αριθμό εγγράφων που ανακτώνται (δηλ. το συνολικό αριθμό κειμένων που διευκρινίζεται από το χρήστη), χρησιμοποιείται με τη μέθοδο επιλογής για να υπολογίσει τον αριθμό εγγράφων που θα ζητηθεί από κάθε συλλογή.
Πιο αναλυτικά, συμβολίζουμε ως $R_s = \{d_1, d_2, d_3, \ldots, d_m\}$ το αποτέλεσμα αναζήτησης στην δειγματοληπτική συλλογή χρησιμοποιώντας το ερώτημα Q. Ορίζουμε ως το Εισερχόμενο Σκορ (Inbound Score) $I(d^c)$ οποιουδήποτε εγγράφου d^c που ανήκει σε συλλογή c, ως τον αριθμό των εγγράφων που επιστρέφονται από τη δειγματοληπτική συλλογή, τα οποία δεν ανήκουν στο c, τα οποία δείχνουν προς το έγγραφο d^c μέσω ενός εξερχόμενου υπερσυνδέσμου. Ο ορισμός του Εισερχόμενου Σκορ και του περιορισμού που εισάγει είναι μεγάλης σπουδαιότητας, δεδομένου ότι αναγκάζει τον αλγόριθμο να εξετάσει μόνο τους υπερσύνδεσης σε έγγραφα εκτός του ίδιου ιστοτόπου. Σε ένα διαδικτυακό περιβάλλον, αυτός ο περιορισμός θα μεταφραζότανε σε off-site outlinks (δηλαδή υπερσυνδέσμους σε άλλους ιστοτόπους). Αυτή η έμφαση δόθηκε προκειμένου να αποφευχθούν οι σύνδεσμοι για λόγους πλοήγησης (δηλ. ιστοσελίδων που δείχνουν στην αρχική σελίδα) και να φιλτραριστεί η πληροφορία η οποία έχει να κάνει αποκλειστικά με ομοιότητα περιεχομένου. Επομένως, εάν $d_i \rightarrow d^c$ δηλώνει ότι το έγγραφο d_i έχει ένα υπερσύνδεσμο προς το κείμενο d^c:

$$I(d^c) = \sum_{d_i \rightarrow d^c} d_i, \quad where\ d_i \notin c\ and\ d^c \in c$$

Το Εισερχόμενο Σκορ $I(d^c)$ κάθε εγγράφου είναι δυναμικό και συγκεκριμένο για το ερώτημα και την πληροφοριακή ανάγκη που έχει τεθεί κάθε φορά. Ένα έγγραφο μπορεί να έχει ένα υψηλό σκορ εάν ένας μεγάλος αριθμός εγγράφων που ανακτώνται από τη δειγματοληπτική συλλογή «δείχνουν» προς αυτό μέσω υπερσυνδέσμων, και ένα πολύ χαμηλό αποτέλεσμα εάν λίγα ή κανένα έγγραφο δεν δείχνουν προς αυτό. Θεωρώντας ότι κάθε έγγραφο ανήκει σε μια συλλογή, η Σχετικότητα $R(c)$ κάθε συλλογής c υπολογίζεται σαν:

$$\mathrm{R}(\mathrm{c}) = \sum_{d^c \in c} I(d^c)$$

Διαισθητικά, μπορούμε να καθορίσουμε τη Σχετικότητα κάθε συλλογής ως τον αριθμό των outlinks που εξάγονται από τα αποτελέσματα της δειγματοληπτικής συλλογής από τα έγγραφα που ανήκουν σε άλλες συλλογές που δείχνουν προς έγγραφα αυτής της συλλογής. Επομένως, υποθέτοντας ότι ο χρήστης ζητά k σε πλήθος έγγραφα να ανακτηθούν από όλες τις συλλογές, ο συνολικός αριθμός εγγράφων που θα ζητηθεί από κάθε συλλογή $N(c_j)$ υπολογίζεται σαν:

$$N(c_j) = \frac{R(c_j)}{\sum_{c_i} R(c_i)} * k$$

Πρακτικά, όταν $N(c_j)=0$ τότε η συλλογή δεν απαιτείται να επιστρέψει έγγραφα. Όπως αναφέρθηκε ήδη, μόνο outlinks των εγγράφων που επιστρέφονται από τη δειγματοληπτική συλλογή εξετάζεται, καθώς αυτό θεωρήθηκε πιο ρεαλιστικό. Οι inlink υπερσύνδεσμοι θα μπορούσαν επίσης να ληφθούν υπόψη αλλά θα καθιστούσαν τον αλγόριθμο μη ρεαλιστικό δεδομένου ότι θα απαιτούσαν γνώση που δεν είναι εύκολα διαθέσιμη, ιδιαίτερα σε μη-συνεργατικά περιβάλλοντα.

Τα πειράματα που έγιναν βασίστηκαν στην WT-Dense, καθώς είναι από τις λίγες πειραματικές συλλογές οι οποίες έχουν δειχθεί στην διεθνή βιβλιογραφία ότι παρέχουν κατανομή υπερσυνδέσμων η οποίο προσομοιάζει αυτή που υπάρχει στο διαδίκτυο. Περισσότερα στοιχεία για την συλλογή παρέχονται στον πίνακα:

Ιδιότητες	WT10g	WT-Dense
Αριθμός εγγράφων	1.692.096	120.494
Αριθμός offsite υπερσυνδέσεων	171.740	171.740
Μέσος όρος του offsite indegree	0.10	1.43
Αριθμός μοναδικών ιστοτόπων	11.680	11.611
Γενικότητα	0.15%	0.21%
Αριθμός ερωτημάτων με σχετικά έγγραφα	50	36
Μέσος αριθμός σχετικών εγγράφων ανά ερώτημα	52	7

Τρία σετ πειραμάτων πραγματοποιήθηκαν. Αυξήσαμε βαθμιαία το επίπεδο διανομής προκειμένου να εξεταστεί η απόδοση του αλγορίθμου σε μια σειρά κατανεμημένων περιβαλλόντων. Σε κάθε σετ, η συλλογή διαιρέθηκε σε 30, 100 και 1000 συλλογές χωρίς κοινά έγγραφα χρησιμοποιώντας την μεθοδολογία που περιγράφηκε παραπάνω για τον διαχωρισμό της WT-Dense.

Σε κάθε σετ, 4 στρατηγικές επιλογής συλλογών εξετάστηκαν: *ομοιόμορφη, τυχαία, βέλτιστη* και link-based και 1000 έγγραφα ζητήθηκαν από κάθε στρατηγική. Η σύγκριση αυτή έγινε καθαρά με σκοπό να επιδείξει κατά πόσο ο νέος αλγόριθμος παρέχει χρήσιμες πληροφορίες για την επιλογή πηγών. Σε ρεαλιστικά περιβάλλοντα, η χρήση του αλγορίθμου θα γινότανε μαζί με κάποιον αλγόριθμο ο οποίος βασίζεται στο περιεχόμενο των συλλογών και όχι μόνο στην ύπαρξη υπερσυνδέσμων (π.χ. CORI ή KL Divergence). Σκοπός των πειραμάτων δεν ήτανε τόσο η ανάπτυξη ενός νέου απομονωμένου αλγορίθμου επιλογής πηγών αλλά η διαπίστωση κατά πόσο οι υπερσύνδεσμοι μπορούν να βοηθήσουν στην καλύτερη επιλογή πηγών, παράλληλα με την χρήση άλλων αλγορίθμων οι οποίοι δεν κάνουν χρήση υπερσυνδέσμων (συμπληρωματική δηλαδή λειτουργία προς αυτούς).

Η ομοιόμορφη προσέγγιση μεταχειρίζεται κάθε συλλογή εξίσου και ζητά τον ίδιο αριθμό εγγράφων από κάθε μία. Εάν συμβολίσουμε ως $Docs(c_j)$ τον αριθμό εγγράφων που θα ζητηθεί από τη συλλογή cj, τότε:

$$Docs(c_j) = \frac{1000}{number\ of\ collections},\quad for\ each\ c_j \qquad \textbf{(4)}$$

Ο τυχαίος αλγόριθμος ζητά έναν τυχαίο αριθμό εγγράφων από τυχαίες συλλογές έως ότου να επιστραφούν συνολικά 1000 έγγραφα. Ο βέλτιστος αλγόριθμος είναι ένας αναδρομικός αλγόριθμος και απαιτεί τη γνώση της διανομής των σχετικών εγγράφων σε όλες τις συλλογές. Χρησιμοποιώντας τη άνωθεν σημειωγραφία και εάν $Rel(c_i)$ είναι ο αριθμός σχετικών εγγράφων στη συλλογή cj, τότε:

$$Docs(c_j) = \frac{Rel(c_j)}{\sum_{c_i} Rel(c_i)} * 1000 \qquad \textbf{(5)}$$

Τα αποτελέσματα από κάθε συλλογή εγγράφων συγχωνεύθηκαν (results merging) χρησιμοποιώντας έναν προκατειλημμένο c-faced die αλγόριθμο. Τα τελικά αποτελέσματα συγκρίθηκαν με τα αποτελέσματα όταν αναζητήθηκε η WT-Dense ως κεντρικοποιημένο σύστημα (*ενιαία προσέγγιση)*.

Όλα τα αποτελέσματα αξιολογήθηκαν χρησιμοποιώντας το επίσημο trec_eval εργαλείο, που παρέχεται από την NIST. Η μηχανή ανάκτησης στις μεμονωμένες συλλογές σε όλα τα πειράματα ήταν η ίδια. Η αποτελεσματικότητα της μηχανής IR δεν είναι σημαντική σε απόλυτους όρους, δεδομένου ότι ο στόχος του πειράματος είναι να αξιολογηθεί η σχετική αποτελεσματικότητα μεταξύ των διαφορετικών στρατηγικών επιλογής πηγών. Με άλλα λόγια, η μόνη παράμετρος που ποίκιλε είναι η μέθοδος επιλογής πηγών και οποιαδήποτε διαφορά στην απόδοση πρέπει να αποδοθεί μόνο στις διαφορές στις στρατηγικές.

Η ιδέα πίσω από τη δημιουργία μιας δειγματοληπτικής συλλογής (sampling collection) είναι να δημιουργηθεί ένας αντιπρόσωπος για κάθε συλλογή. Η ποιότητα της συλλογής αυτής είναι κάποιας σπουδαιότητας στο αλγόριθμο που παρουσιάστηκε. Πολλή σκέψη τέθηκε στη μεθοδολογία που επρόκειτο να ακολουθηθεί για τη δημιουργία της.

Τελικά, αποφασίστηκε ότι η δειγματοληπτική συλλογή θα δημιουργούταν με τυχαία επιλογή εγγράφων από κάθε συλλογή και την ανάλυση του περιεχομένου τους. Μια τέτοια προσέγγιση θεωρήθηκε ότι παράγει αρκετά καλό αντιπροσωπευτικό δείγμα των συλλογών, ενώ παράλληλα δεν παραβιάζει τη μη-συνεργατική φύση των πειραμάτων. Ο αλγόριθμος χρησιμοποίησε για την εξαγωγή συνδέσμων τα 100 πρώτα αποτελέσματα από μια συλλογή δειγματοληψίας, αποτελούμενη από το 20% της WT-Dense συλλογής.

Παρακάτω αναφέρονται μέρος των αποτελεσμάτων. Για τα πλήρη πειράματα και αποτελέσματα, δείτε την δημοσίευσή μας στο 11^{o} Panhellenic Conference on Informatics (PCI 2007).

Πειραματικά αποτελέσματα

Σε πειράματα επιλογής πηγών, συνηθίζεται να μετράται η ανάκληση (Recall) των αλγορίθμων. Ως ανάκληση ορίζεται ο λόγος των σχετικών κειμένων που έχουν ανακτηθεί έναντι όλων των σχετικών κειμένων που υπάρχουν για κάποιο ερώτημα. Συγκεκριμένα:

$$\text{Ανάκληση} = \frac{\text{Αριθμός σχετικών κειμένων που έχουν ανακτηθεί}}{\text{Σύνολο σχετικών κειμένων}}$$

Τα αποτελέσματα των πειραμάτων δίνονται στον παρακάτω πίνακα.

Αριθμός Συλλογών	**Μέθοδος επιλογής πηγών**	**Ανάκληση (Recall)**	**Ποσοστό συλλογών που χρησιμοποιή-θηκαν**
30 συλλογές	Link-based	0.6831[a]	84%[a]
	Βέλτιστος	0.8148[a]	16%[b]
	Τυχαίος	0.1371[b]	23%[b]
	Ενιαίος	0.7279[a]	NA
	Ομοιόμορφος	0.6543[a]	100%[c]
100 συλλογές	Link-based	0.7845[a, c]	72%[a]
	Βέλτιστος	0.9382[a]	5%[b]
	Τυχαίος	0.0786[b]	7%[b]
	Ενιαίος	0.7279[c]	NA
	Ομοιόμορφος	0.7442[a, c]	100%[c]
1000 συλλογές	Link-based	0.2573[a]	15%[a]
	Βέλτιστος	0.8412[b]	0.50%[b]
	Τυχαίος	0.0786[c]	0.70%[b]
	Ενιαίος	0.7279[b]	NA
	Ομοιόμορφος	0.3977[a]	100%[c]

Το ποσοστό των συλλογών που χρησιμοποιήθηκαν αναφέρεται στο ποσοστό των συλλογών που επιλέχτηκαν από κάθε αλγόριθμο για να συμβάλουν στην διαδικασία.

Ένα από τα πρώτα συμπεράσματα που μπορούν να συναχθούν είναι το γεγονός που ο νέος αλγόριθμος αποδίδει πολύ καλύτερα από τον τυχαίο σε όλες τις συνθήκες, επικυρώνοντας κατά συνέπεια την υπόθεση ότι οι υπερσύνδεσμοι δεν επιλέγουν τις πηγές με τυχαίο τρόπο και μπορούν πράγματι να παρέχουν πολύτιμες πληροφορίες στην επιλογή πηγής. Πρέπει να τονιστεί ότι η σύγκριση του αλγορίθμου με την τυχαία προσέγγιση γίνεται μόνο για να επικυρώσει ότι η προτεινόμενη μέθοδος έχει μια λογική και είναι βασισμένη σε μια έγκυρη αρχή. Στο πρώτο περιβάλλον των 30 συλλογών, ο ομοιόμορφος αλγόριθμος και ο link-based αποδίδουν παρόμοια με τον ενιαίο. Το γεγονός ότι η νέος αλγόριθμος αποδίδει όμοια με τον ενιαίο σημαίνει ότι οι περισσότερες από τις σχετικές πηγές ανακαλύφθηκαν επιτυχώς. Ο νέος αλγόριθμος ξεπερνά την ανάκληση του ομοιόμορφου, με 16% λιγότερη χρησιμοποίηση συλλογών, το οποίο σημαίνει ότι ο αλγόριθμος χρησιμοποίησε κατά μέσον όρο 25 από τις 30

διαθέσιμες συλλογές. Ο βέλτιστος αλγόριθμος, αν και αποδίδει καλύτερα, δεν παράγει μια στατιστικά σημαντική διαφορά.

Τα αποτελέσματα ακολουθούν το ίδιο σχέδιο στο δεύτερο περιβάλλον με τις 100 συλλογές. Ο νέος αλγόριθμος αποδίδει σταθερά καλύτερα από τον ομοιόμορφο, με 28% λιγότερη χρησιμοποίηση συλλογών. Αυτό που είναι μεγάλης σπουδαιότητας εδώ είναι ότι και οι δύο αλγόριθμοι έχουν καλύτερη απόδοση από την ενιαία προσέγγιση. Οι επιπτώσεις αυτού του αποτελέσματος συζητούνται στη συνέχεια.

Στο τελευταίο περιβάλλον, η απόδοση του νέου αλγορίθμου μειώνεται, αλλά παραμένει στα ίδια επίπεδα απόδοσης με τον ομοιόμορφο, με 85% λιγότερη χρησιμοποίηση κεντρικών υπολογιστών. Τα αποτελέσματα καταδεικνύουν την ικανότητα του αλγορίθμου να διακρίνει μεταξύ όλων των πηγών πληροφοριών τις πιο σημαντικές από τις τετριμμένες.
Εν κατακλείδι, τα πειράματα που παρέχονται όπως και άλλα που έγιναν δείχνουν ότι οι υπερσύνδεσμοι ανάμεσα σε ιστοσελίδες παρέχουν χρήσιμη πληροφορία για την επιλογή πηγών σε κατανεμημένα περιβάλλοντα. Η χρήση του νέου αλγορίθμου μαζί με κάποια άλλη μέθοδο επιλογής πηγών που βασίζεται σε περιεχόμενο αναμένεται ότι μπορεί να αυξήσει την απόδοσή των κατανεμημένων προσεγγίσεων.

Ένα άλλο σημαντικό αποτέλεσμα είναι ότι η (αναδρομική) βέλτιστη στρατηγική τήξης αποδίδει σταθερά καλύτερα από οποιαδήποτε άλλη στρατηγική, συμπεριλαμβανομένης της ενιαίας συλλογής. Επίσης, σε μερικές περιπτώσεις (100 συλλογές) η νέα μέθοδος αποδίδει επίσης καλύτερα από την ενιαία συλλογή. Αυτά τα αποτελέσματα επιβεβαιώνουν τη δυνατότητα των κατανεμημένων συστημάτων ανάκτησης πληροφοριών να επιτυγχάνουν απόδοση που είναι καλύτερη από αυτή των ενιαίων συστημάτων. Περισσότερες πληροφορίες για τον αλγόριθμο καθώς και για τα πειράματα που έγιναν μπορούν να βρεθούν στο άρθρο που δημοσιεύτηκε στα πρακτικά του συνεδρίου «11th Panhellenic Conference on Informatics (PCI 2007)» υπό τον τίτλο «Using linkage information to approximate the distribution of relevant documents in DIR».

ΑΛΓΟΡΙΘΜΟΣ ΣΥΝΘΕΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΒΑΣΙΣΜΕΝΟΣ ΣΤΗΝ ΧΡΗΣΗ ΠΟΛΛΑΠΛΩΝ ΜΟΝΤΕΛΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
(Multiple Regression Rank Merging – MRRM)

Στην συνέχεια θα περιγράψουμε τον νέο αλγόριθμο σύνθεσης αποτελεσμάτων που αναπτύχθηκε στα πλαίσια του προγράμματος.

Οι γνωστότεροι αλγόριθμοι σύνθεσης αποτελεσμάτων στηρίζονται στην προϋπόθεση ότι οι απομακρυσμένες συλλογές επιστρέφουν τα σκορ σχετικότητας (relevance scores) μαζί με τα επιστρεφόμενα κείμενα. Εντούτοις, στα περισσότερα σύγχρονα περιβάλλοντα, αυτό δεν συμβαίνει. Συχνότερα, οι μηχανές αναζήτησης επιστρέφουν μονάχα κείμενα σε φθίνουσα σειρά σχετικότητας χωρίς σκορ γιατί στηρίζονται στο γεγονός ότι ο μέσος χρήστης δεν έχει καμία ανάγκη για τα αποτελέσματα σχετικότητας, δεδομένου ότι δεν μπορεί να ερμηνευθούν άμεσα. Δυστυχώς, πολύ λιγότερη πληροφορία μεταβιβάζεται σε απλές λίστες χωρίς σκορ. Ακόμα κι αν μια συλλογή είναι ελάχιστα σχετική με ένα ερώτημα και τα επιστρεφόμενα έγγραφα είναι μόνο λίγο σχετικά τα ίδια, η λίστα είναι η ίδια με μία πολύ σχετική συλλογή που επιστρέφει πολύ σχετικά κείμενα, δηλαδή θα επιστρέφεται απλά μία λίστα κειμένων.

Το κίνητρο πίσω από τον νέο αλγόριθμο είναι να λειτουργήσει αποδοτικά σε τέτοιου είδους περιβάλλοντα όπου οι συλλογές επιστρέφουν μονάχα λίστες με αποτελέσματα χωρίς σκορ σχετικότητας. Για να το επιτύχει αυτό, μεγιστοποιεί την χρήση των ήδη υπαρχόντων πόρων. Η δειγματοληπτική συλλογή που αναφέρθηκε και παραπάνω αποτελεί ένα τέτοιο παράδειγμα. Αυτή αποτελείται από όλα τα κείμενα τα οποία έχουν παρθεί με τρόπο τυχαίο από κάθε συλλογή ώστε το σύνολό τους να αντιπροσωπεύει μία κεντρικοποιημένη συλλογή. Υπό το ίδιο πρίσμα, τα κείμενα από κάθε συλλογή ξεχωριστά μπορούν να λειτουργήσουν σαν αντιπρόσωποι της ίδιας της συλλογής. Έτσι σημαντικά στατιστικά στοιχεία ανάμεσα στην απομακρυσμένη συλλογή και τον αντιπρόσωπό της θεωρούνται ότι είναι ίδια (κοινές λέξεις, σπουδαιότητα λέξεων κλπ).

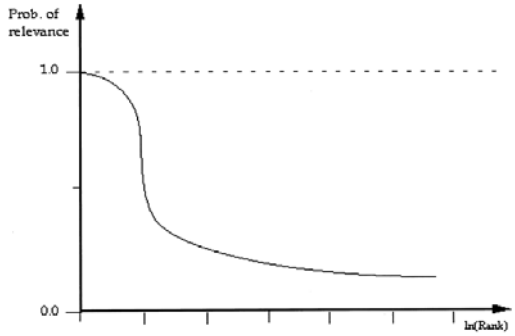
Προηγούμενες εργασίες προσπάθησαν να αντισταθμίσουν την έλλειψη σκορ σχετικότητας υποθέτοντας ότι υπάρχει μια γραμμική συσχέτιση μεταξύ της κατάταξης των εγγράφων και των σκορ σχετικότητας. Συγκεκριμένα, σε περιπτώσεις όπου οι απομακρυσμένες συλλογές δεν

επέστρεφαν σκορ σχετικότητας, τεχνητά σκορ ορίζονται στα επιστρεφόμενα έγγραφα, δίνοντας ένα σκορ 0.6 στο 1^{ο} ταξινομημένο έγγραφο και μειώνοντας το σκορ σε σταθερά διαστήματα μέχρι να δοθεί ένα σκορ 0.4 στο τελευταίο έγγραφο.

Έχει επισημανθεί εντούτοις, ότι η μείωση στη σχετικότητα δεν συνδέεται γραμμικά με την κατάταξη. Συγκεκριμένα, αποδείχθηκε ότι μια logistic συνάρτηση, με b<0, θα παρείχε μια καλύτερη συσχέτιση. Δηλαδή, η πιθανότητα της σχετικότητας του κειμένου D_i λαμβάνοντας υπόψη την κατάταξή του x_i, δίνεται από την εξίσωση:

$$\text{Prob}\left[D_i \text{ is Rel}/x_i\right] = \frac{e^{a+bx_i}}{1+e^{a+bx_i}}$$

Το παρακάτω σχήμα καταδεικνύει τον αναμενόμενο συσχετισμό μεταξύ της σχετικότητας και της κατάταξης.



Με βάση τα ανωτέρω, η παρούσα εργασία απομακρύνεται από να ορίσει τεχνητά σκορ σχετικότητας γραμμικά και προσπαθεί να υπολογίσει την πραγματική γραφική παράσταση για κάθε μεμονωμένη συλλογή προκειμένου να παραχθούν τα ακριβή αποτελέσματα σχετικότητας.

Υπολογισμός τοπικών σκορ σχετικότητας από κατατάξεις

Μετά από το στάδιο επιλογής πηγών, το ερώτημα εκτελείται στις απομακρυσμένες συλλογές και παράλληλα στα τοπικά αποθηκευμένα δείγματα των επιλεγμένων συλλογών και στην κεντρική δειγματοληπτική συλλογή. Έχουμε χρησιμοποιήσει τον αλγόριθμο ανάκτησης inquiry για τις τοπικές συλλογές και τη κεντρική δειγματοληπτική συλλογή αλλά οποιοσδήποτε αποτελεσματικός αλγόριθμος ανάκτησης θα λειτουργούσε επίσης (kl divergence, okapi κ.λπ.). Για κάθε συλλογή, δύο λίστες εγγράφων επιστρέφονται, μία από τη απομακρυσμένη συλλογή, που περιέχει μόνο μία κατάταξη εγγράφων (χωρίς σκορ σχετικότητας) και ένας από το τοπικό δείγμα, που περιέχει σκορ σχετικότητας. Ο κατάλογος αποτελέσματος από τη κεντρική δειγματοληπτική συλλογή δεν λαμβάνεται υπόψη προς το παρόν, αλλά ενσωματώνεται στον αλγόριθμο σε ένα μεταγενέστερο στάδιο. Για κάθε συλλογή, οι δύο λίστες εγγράφων συγκρίνονται και όλα τα κοινά έγγραφα αποθηκεύονται μαζί με την κατάταξη που έλαβαν στη μακρινή συλλογή και το σκορ σχετικότητας στο τοπικό δείγμα.

Υπολογισμός της καμπύλης για κάθε συλλογή

Λαμβάνοντας υπόψη τα κοινά έγγραφα που βρίσκονται μεταξύ των μακρινών και επιλεγμένων συλλογών, ο αλγόριθμος υπολογίζει την καμπύλη για κάθε συλλογή, ορίζοντας κατά συνέπεια σκορ σχετικότητας στα μη κοινά έγγραφα που επιστρέφονται από τις μακρινές συλλογές. Επηρεασμένοι από προγενέστερη δουλειά, υποθέτουμε ότι ο συσχετισμός μεταξύ της κατάταξης Χ ενός εγγράφου και του σκορ σχετικότητας Υ δίνεται από μια logistic συνάρτηση:

$$Y = \frac{e^{a+b*X}}{1+e^{a+b*X}}$$

Εφαρμόζοντας τους ακόλουθους μετασχηματισμούς, είμαστε σε θέση να τροποποιήσουμε την ανωτέρω εξίσωση σε μια γραμμική:

$$\frac{Y}{1-Y}=e^{\alpha+\beta*X}$$

$$\ln\left(\frac{Y}{1-Y}\right)=\alpha+\beta*X$$

$$\text{logit}[Y]=\alpha+\beta*X$$

Πρέπει να υπολογίσουμε τις παραμέτρους α, β του ανωτέρω μοντέλου προκειμένου να υπολογιστεί η καμπύλη για κάθε συλλογή. Δεδομένου ότι η εξίσωση είναι γραμμική, αυτή η εκτίμηση μπορεί να ολοκληρωθεί μέσω γραμμικής παλινδρόμησης.

Γραμμική Παλινδρόμηση

Ένα μοντέλο γραμμικής συμμεταβολής μπορεί να δηλωθεί τυπικά σαν:

$$y=a + b*x + \varepsilon$$

όπου το X είναι η ανεξάρτητη μεταβλητή (στην περίπτωσή μας, την κατάταξη του εγγράφου στη μακρινή συλλογή), το Y είναι η εξαρτώμενη μεταβλητή (το σκορ σχετικότητας του εγγράφου στη τοπική δειγματοληπτική συλλογή), το α και το β είναι οι παράμετροι του μοντέλους και το ε είναι το λάθος (δείτε κατωτέρω).

Οι παρατηρήσεις που χρησιμοποιούνται για την εκτίμηση του μοντέλου είναι ζευγάρια (x_i, y_i) i=1,…, ν όπου x_i είναι η κατάταξη του κοινού εγγράφου i, y_i είναι το σκορ σχετικότητας του εγγράφου στη δειγματοληπτική συλλογή και το ν είναι ο αριθμός κοινών εγγράφων που εντοπίζονται. Ο στόχος του μοντέλου είναι να υπολογιστούν οι παράμετροι α και β που ελαχιστοποιούν το λάθος ε που αντιπροσωπεύει τη διαφορά μεταξύ των τιμών του Y που έχουν παρατηρηθεί και αυτών που υπολογίζονται μέσω του μοντέλου. Ο καλύτερος τρόπος να γίνει αυτό είναι μέσω της μεθόδου ελάχιστων τετραγώνων. Συγκεκριμένα, ο αλγόριθμος στοχεύει στην ελαχιστοποίηση του ποσού S:

$$S=\sum_{i=1}^{n}\left[y_i-\hat{y}_i\right]^2$$

όπου y_i είναι το σκορ σχετικότητας του κοινού εγγράφου i και του $\hat{y}_i$ είναι αυτό που υπολογίζεται από το μοντέλο.

Το πρόβλημα μπορεί να τυποποιηθεί χρησιμοποιώντας την ορολογία πινάκων ως εξής:

$$Y=X*B + \varepsilon$$

όπου

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ .... \\ y_n \end{bmatrix},\ X=\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ ..... & \\ 1 & x_n \end{bmatrix},\ B=\begin{bmatrix} a \\ b \end{bmatrix},\ e=\begin{bmatrix} e_1 \\ e_2 \\ .... \\ e_n \end{bmatrix}$$

όπου το ν είναι ο αριθμός κοινών εγγράφων που εντοπίζονται. Η βέλτιστη λύση για τις παραμέτρους α και β είναι αυτή που ελαχιστοποιεί το S και δίνεται από:

$$B=\left(X'X\right)^{-1}X'Y$$

Στα πειράματα που πραγματοποιήθηκαν χρησιμοποιήσαμε μόνο τα πρώτα 10 κοινά έγγραφα και δεν λάβαμε υπόψη τα υπόλοιπα. Επίσης, προκειμένου να προσομοιώσουμε καλύτερα την κλίση στο τέλος της γραφικής παράστασης, παρεμβάλαμε ένα «τεχνητό» κοινό έγγραφο στην θέση κατάταξης 3000, με σκορ σχετικότητας 0.001. Αν και αυτές οι δύο ρυθμίσεις δεν ήταν απαραίτητες, βρέθηκαν να αυξάνουν την αποτελεσματικότητα του αλγορίθμου.

Η ανωτέρω διαδικασία επαναλαμβάνεται για κάθε συλλογή που επιλέγεται από τον αλγόριθμο επιλογής πηγών. Εφαρμόζοντας την εξίσωση με τις κατ' εκτίμηση παραμέτρους για κάθε

μακρινή συλλογή, ο αλγόριθμος ορίζει τοπικά σκορ σχετικότητας σε όλα τα επιστρεφόμενα έγγραφα.

Υπολογισμός των τελικών αποτελεσμάτων σχετικότητας
Υπολογίζοντας ένα τοπικό σκορ για κάθε έγγραφο που επιστρέφεται από τις μακρινές συλλογές, ο αλγόριθμος προχωρά στη δεύτερη φάση, η οποία είναι να υπολογιστούν τα τελικά αποτελέσματα για τα επιστρεφόμενα έγγραφα. Είναι σε αυτή τη φάση που η λίστα αποτελεσμάτων που επιστρέφεται την κεντρική δειγματοληπτική συλλογή μπαίνει σε χρήση. Όπως προηγουμένως, ο αλγόριθμος εντοπίζει τα κοινά έγγραφα που επιστρέφονται από τις δειγματοληπτικές συλλογές και την κεντρική δειγματοληπτική συλλογή και αποθηκεύει τα αντίστοιχα αποτελέσματά τους και υπολογίζει και πάλι ένα μοντέλο γραμμικής παλινδρόμησης ανάμεσα στα δύο σκορ.

Μια πιθανή ερώτηση είναι εάν η γραμμική παλινδρόμηση είναι η καλύτερη επιλογή, ή εάν μια πολυωνυμική (μη γραμμική) παλινδρόμηση θα ήταν καλύτερη επιλογή. Διάφοροι λόγοι πρότειναν την ανωτέρω απόφαση. Εκτενή πειράματα έδειξαν ότι το όφελος από τη μετάβαση από ένα γραμμικό σε ένα μη γραμμικό μοντέλο θα ήταν ελάχιστο.

Τα ζευγάρια (x_i, y_i) i=1,…, μ που θα βοηθήσουν στην εκτίμηση των παραμέτρων α και β είναι σε αυτήν την περίπτωση x_i (το σκορ που ορίζεται στο κοινό έγγραφο i από τη δειγματοληπτική συλλογή) και y_i (το σκορ που ορίζεται στο κοινό έγγραφο i από την κεντρική δειγματοληπτική συλλογή). Πάλι, η προτιμημένη μεθοδολογία για τον υπολογισμό των παραμέτρων είναι η μέθοδος ελάχιστων τετραγώνων, για την οποία η βέλτιστη λύση δίνεται από την εξίσωση παραπάνω. Υπολογίζοντας τις παραμέτρους α και β για κάθε συλλογή, ο αλγόριθμος εφαρμόζει την αντίστοιχη γραμμική συνάρτηση σε όλα τα έγγραφα που επιστρέφονται από τις μακρινές συλλογές, χρησιμοποιώντας ως ανεξάρτητη μεταβλητή το τοπικό αποτέλεσμα (x) που αποδόθηκε σε τους κατά τη διάρκεια της πρώτης φάσης του αλγορίθμου και έτσι υπολογίζει τα τελικά σκορ των κειμένων.

Πειράματα
Στα πειράματα που έγιναν χρησιμοποιήθηκαν οι συλλογές TREC123 και TREC4, που είναι υποσύνολα της TREC. Παρακάτω δίνονται τα στοιχεία για τις συλλογές

Όνομα	Αριθμός συλλογών	Μέγεθος GB	Αριθμός Εγγράφων		
			Ελάχιστο	Μέγιστο	Μέσος όρος
Trec123	100	3.2	752	39713	10782
Trec4	100	2.0	300	82700	5700

Τα αποτελέσματα των πειραμάτων δίνονται παρακάτω. Ο αλγόριθμος ονομάστηκε Multiple Regression Rank Merging, λόγω του τρόπου με τον οποίο λειτουργεί. Σε πειράματα σύνθεσης αποτελεσμάτων είναι σύνηθες να μετράται η ακρίβεια (Precision) των τελικών λιστών. Συγκεκριμένα ως ακρίβεια σε κάποιον αριθμό κειμένων k, ορίζεται ως:

$$P@k = \frac{\textit{πλήθος των σχετικών κειμένων στα k κείμενα}}{k}$$

Ακρίβεια κάνοντας χρήση 10 συλλογών, όπου η κάθε μία επιστρέφει 1000 αποτελέσματα

	Trec123-100col		
Precision	**CORI**	**SSL**	**MRRM**
P@5:	0.1460	0.1560	0.1600(+9.6%) (+2.6%)
P@10:	0.1370	0.1350	0.1400(+2.2%) (+3.7%)
P@15:	0.1260	0.1267	0.1300(+3.2%) (+2.6%)
P@20:	0.1180	0.1205	0.1250(+5.9%) (+3.7%)
P@30:	0.1113	0.1103	0.1120(+0.1%) (+1.5%)

Ακρίβεια κάνοντας χρήση 10 συλλογών, όπου η κάθε μία επιστρέφει 300 αποτελέσματα

	Trec123-100col		
Precision	**CORI**	**SSL**	**MRRM**
P@5:	0.1340	0.1600	0.1800(+34.3%) (+12.5%)

P@10:	0.1340	0.1470	0.1510(+12.7%) (+2.7%)
P@15:	0.1333	0.1373	0.1387(+4.0%) (+1.0%)
P@20:	0.1300	0.1290	0.1320(+1.5%) (+2.3%)
P@30:	0.1260	0.1177	0.1197(-5%) (+1.7%)

Ακρίβεια κάνοντας χρήση 3 συλλογών, όπου η κάθε μία επιστρέφει 300 αποτελέσματα

	Trec123-100col		
Precision	**CORI**	**SSL**	**MRRM**
P@5:	0.1360	0.1580	0.1720 (+26.5%)(+8.9%)
P@10:	0.1360	0.1300	0.1430 (+5.1%) (+10.0%)
P@15:	0.1293	0.1187	0.1340 (+3.6%) (+12.9%)
P@20:	0.1205	0.1075	0.1225 (+1.7%) (+14.0%)
P@30:	0.1153	0.0947	0.1080 (-6.3%) (+14.0%)

Ακρίβεια κάνοντας χρήση 10 συλλογών, όπου η κάθε μία επιστρέφει 1000 αποτελέσματα

	Trec4-kmeans			
Precision	**CORI**	**SSL**	**MRRM**	
P@5:	0.2640	0.3520	0.3880(+47.0%)(+10.2%)	
P@10:	0.2220	0.3180	0.3320(+49.5%)(+4.4%)	
P@15:	0.2133	0.2867	0.3093(+45.0%)	(+7.9%)
P@20:	0.2010	0.2740	0.2830(+40.8%)	(+3.2%)
P@30:	0.1793	0.2580	0.2467(+37.6%)	(-4.4%)

Ακρίβεια κάνοντας χρήση 10 συλλογών, όπου η κάθε μία επιστρέφει 300 αποτελέσματα

	Trec4-kmeans		
Precision	**CORI**	**SSL**	**MRRM**
P@5:	0.2800	0.3400	0.3840(+37.1%) (+12.9%)
P@10:	0.2340	0.3140	0.3360(+43.6%) (+7.0%)
P@15:	0.2160	0.2840	0.3093(+43.2%) (+8.9%)
P@20:	0.2050	0.2740	0.2840(+30.9%) (+3.6%)
P@30:	0.1833	0.2540	0.2527(+37.9%) (-0.1%)

Ακρίβεια κάνοντας χρήση 3 συλλογών, όπου η κάθε μία επιστρέψει 300 αποτελέσματα

	Trec4-kmeans		
Precision	**CORI**	**SSL**	**MRRM**
P@5:	0.2800	0.3120	0.3480 (+24.3%)(+11.5%)
P@10:	0.2360	0.2800	0.3100 (+31.3%)(+10.7%)
P@15:	0.2133	0.2587	0.2827 (+32.5%)(+9.3%)
P@20:	0.2030	0.2430	0.2600 (+28.1%)(+7.0%)
P@30:	0.1807	0.2233	0.2340 (+29.5%)(+4.8%)

Τα ποσοστά στις παρενθέσεις εκφράζουν την διαφορά στην απόδοση από τους αλγορίθμους CORI και SSL αντίστοιχα. Όπως είναι εμφανές από τους πίνακες ο νέος αλγόριθμος πάντα επιτυγχάνει επίδοση καλύτερη από τους προηγούμενους state-of-the-art αλγορίθμους στην πλειοψηφία των περιπτώσεων, χωρίς να κάνει χρήση σκορ σχετικότητας από τις απομακρυσμένες πηγές. Για περισσότερες πληροφορίες, δείτε την δημοσίευσή μας στο συνέδριο «29th European Conference on Information Retrieval (ECIR- 2007)» υπό τον τίτλο «Results Merging Algorithm using multiple regression models».

ΥΒΡΙΔΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΥΝΘΕΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Ο αλγόριθμος αυτός (hybrid) προσπαθεί να συνδυάσει τον αλγόριθμο MRMM με άλλους αλγορίθμους και πρακτικές σύνθεσης αποτελεσμάτων που βασίζονται στο «κατέβασμα» των κειμένων των συλλογών. Ο αλγόριθμος ενοποιεί τις δύο γενικές κατευθύνσεις (download/don't) από τις οποίες έχει προσεγγιστεί το πρόβλημα στην έρευνα, συνδυάζοντας τα πλεονεκτήματά τους, ενώ συγχρόνως ελαχιστοποιώντας τα μειονεκτήματά τους.

Το άρθρο που παρουσιάζει αυτή την εργασία είναι το «Georgios Paltoglou, Michail Salampasis, Satratzemi Maria. **Hybrid Results Merging**. In *Proceedings of* Sixteenth ACM Conference on Information and Knowledge Management (CIKM07), Lisbon, 6-9 November, 2007.». Αντίγραφο του άρθρου υπάρχει στα συνημμένα αυτής της τελικής έκθεσης.

ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΠΗΓΩΝ ΜΕ ΧΡΗΣΗ ΟΛΟΚΛΗΡΩΜΑΤΩΝ

Ο αλγόριθμος επιλογής πηγών που προάγεται παρέχει έναν καινοτόμο τρόπο μοντελοποίησης των πηγών ως περιοχές σε ένα χώρο που παράγεται από τα κείμενα τα οποία περιέχουν. Διατυπώνει ένα πλήρες θεωρητικό πλαίσιο επίλυσης του προβλήματος της επιλογής πηγών, ενώ παράλληλα αποτελεσματικά συλλαμβάνει πειραματικές παρατηρήσεις και γενικά αποδεκτές αντιλήψεις στον τομέα της Ανάκτησης Πληροφοριών.

Εκτεταμένα πειράματα επιδεικνύουν ότι είναι ικανός να διασφαλίσει μία απόδοση που είναι τουλάχιστον τόσο καλή όσο άλλες μεθοδολογίες αιχμής και συχνότερα καλύτερη. Τα πειράματα αυτά δημοσιεύτηκαν σε μία σειρά από συνέδρια και περιοδικά. Αναλυτικά οι εργασίες που σχετίζονται με αυτό τον αλγόριθμο επιλογής πηγών πληροφοριών είναι:

- Georgios Paltoglou, Michail Salampasis, Maria Satratzemi. Integral based Source Selection for uncooperative Distributed Information Retrieval environments. In Proceedings of the 6th ACM Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'08), CIKM08, Napa Valey, 2008.

- Paltoglou, G., Salampasis, M., Satratzemi M. Collection-integral Source Selection for uncooperative distributed information retrieval environments, Information Sciences (accepted).
- Paltoglou, G., Salampasis, M., Satratzemi M. Modeling information sources as integrals for effective and efficient source selection, Information Processing and Management Journal (accepted).

ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΠΗΓΩΝ ΜΕ ΧΡΗΣΗ DATA FUSION ΤΕΧΝΙΚΩΝ

Ο αλγόριθμος αυτός χρησιμοποιεί τεχνικές που χρησιμοποιήθηκαν σε προβλήματα data fusion. Βασικά λειτουργεί ως ένα κλασσικός αλγόριθμος που βασίζεται σε ένα κεντρικό ευρετήριο που έχει δημιουργηθεί με τεχνικές sampling. Στη συνέχεια το κάθε κείμενο που ανακτάται από την κεντρική sampling συλλογή λειτουργεί ως «ψηφοφόρος» υπέρ της επιλογής της αντίστοιχης συλλογής όπου και ανήκει.

Διάφοροι αλγόριθμοι δοκιμάστηκαν και περιγράφονται συνοπτικά στον παρακάτω πίνακα:

Data Fusion technique	Collection Source Score
CombMAX	maximum of scores of docs in $Votes(S,Q)$
CombMIN	minimum of scores of docs in $Votes(S,Q)$
CombSUM	sum of scores of docs in $Votes(S,Q)$
CombMNZ	$\|Votes(S,Q)\| \cdot CombSUM(S,Q)$
CombANZ	arithmetic mean of scores of docs in $Votes(S,Q)$
CombMED	median of scores of docs in $Votes(S,Q)$
CombGMN	geometric mean of scores of docs in $Votes(S,Q)$
Count	$\|Votes(S,Q)\|$
Reciprocal Rank	sum of inverse of ranks of docs in $Votes(S,Q)$
Borda-fuse	sum of ($\|R(Q)\|$ - ranks of docs in $Votes(S,Q)$)
wBorda-fuse	$\frac{\hat{N}_S}{N_{S\_sample}} \cdot$ Borda-fuse(S,Q)
expCombSUM	sum of exp of scores of docs in $Votes(S,Q)$
expCombMNZ	$\|Votes(S,Q)\| \cdot expCombSUM(S,Q)$
expCombANZ	arithmetic mean of exp of scores of docs in $Votes(S,Q)$

Τα πειραματικά αποτελέσματα αυτής της εργασίας παρουσιάστηκαν στο συνέδριο ECIR 2009 με την εργασία G. Paltoglou, M. Salampasis, and M. Satratzemi. Simple adaptations of data fusion algorithms for source selection. In *Proceedings of* 31th European Conference on Information Retrieval, Toulouse, France, 2009.

2.3. Συνολικά αποτελέσματα και παραδοτέα του έργου

Η έρευνα που διεξήχθη επικεντρώθηκε στην πιο αποτελεσματική σύνθεση των αποτελεσμάτων (results merging) και επιλογής πηγών (source selection) σε εντελώς μη-συνεργατικά περιβάλλοντα, όπου οι κατανεμημένες συλλογές δεν επιστρέφουν σκορ σχετικότητας μαζί με τα κείμενα που επιστρέφουν. Η διεθνή έρευνα στον τομέα αυτό, θεωρεί δεδομένο ότι οι συλλογές επιστρέφουν σκορ, αλλά η πραγματικότητα είναι διαφορετική, καθώς η συντριπτική πλειοψηφία των μηχανών αναζήτησης επιστρέφουν μονάχα κείμενα. Σκοπός της έρευνας είναι να επιτευχθεί αρχικά η ίδια αποτελεσματικότητα με αυτή που επιτυγχάνεται όταν επιστρέφονται σκορ και σε δεύτερη φάση, σταθερά ανώτερη αποτελεσματικότητα από αυτή που επιτυγχάνεται όταν η ανάκτηση γίνεται από μία κεντρικοποιημένη βάση. Τα πειράματα που έγιναν απέδειξαν ότι ενώ η αναφορά σκορ είναι σημαντική για τους ήδη υπάρχοντας αλγόριθμους, ο νέος αλγόριθμος που σχεδιάστηκε μπορεί να λειτουργήσει αρκετά αποτελεσματικά χωρίς σκορ, επιτυγχάνοντας επιδόσεις που προσεγγίζουν και συχνά υπερβαίνουν αυτές των προσεγγίσεων που κάνουν χρήση σκορ.

Οι πρακτικές εφαρμογές των άνωθεν ερευνητικών διαδικασιών και οι επιπτώσεις που μπορούν να έχουν στην ανάκτηση πληροφοριών στο διαδίκτυο είναι πολλές. Ενδεικτικά αναφέρονται μερικές:

- Ανάπτυξη κατανεμημένων μηχανών αναζήτησης οι οποίες θα επιστρέφουν επίκαιρη και ενημερωμένη πληροφορία από τις καταλληλότερες και πιο αξιόπιστες πηγές.
- Ανάπτυξη αποτελεσματικότερων peer-to-peer μηχανών αναζήτησης, με παράλληλη μεγιστοποίηση της χρησιμότητας των ψηφιακών βιβλιοθηκών (digital libraries)
- Ανάπτυξη και διάδοση νέου τρόπου δημιουργίας, παροχής και διάχυσης ψηφιακού υλικού (οπτικού, ακουστικού, βασισμένου σε κείμενο) μέσω αποτελεσματικής αναζήτησης και διαχείρισης μέσω δικτύων p2p νέας γενιάς, που προσφέρουν αναζήτηση όχι μονάχα βάση ονόματος αλλά και περιεχομένου.
- Πολύ μεγαλύτερη ακρίβεια στην διαδικασία ανάκτησης πληροφοριών, με πολύ οικονομικότερα μέσα.
- Μεγαλύτερη κάλυψη της διαθέσιμης πληροφορίας στο διαδίκτυο, η οποία τώρα είναι μη-προσβάσιμη και κατανεμημένη σε hidden websites και εταιρικά δίκτυα.

Παραδοτέα του έργου

Το βασικό παραδοτέο του έργου είναι η διδακτορική διατριβή του νέου ερευνητή Παλτόγλου Γεώργιου.

Επίσης έγιναν μία σειρά από δημοσιεύσεις σε σημαντικά περιοδικά και συνέδρια και οι οποίες υπάρχουν αμέσως παρακάτω.

Δημοσιεύσεις του ΥΔ κατά τη διάρκεια του ερευνητικού έργου

Paltoglou, G., Salampasis, M., Satratzemi, M. Simple adaptations of data fusion algorithms for source selection, In Proc. 31st European Conference on Information Retrieval, ECIR '09, to appear.

Paltoglou, G., Salampasis, M., Satratzemi: Collection-integral Source Selection for uncooperative distributed information retrieval environments, Information Sciences, accepted.

Paltoglou, G., Salampasis, M., Satratzemi, Modeling information sources as integrals for effective and efficient source selection, Information Processing and Management Journal, accepted.

Paltoglou, G., Salampasis, M., Satratzemi, M. Integral Based Source Selection for Uncooperative Distributed Information Retrieval Environments, In Proc. Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'08), p. 67 - 74.

Paltoglou, G., Salampasis, F., Lazarinis , M. Indexing and Retrieval of a Greek Corpus, In Proc. Improving Non English Web Searching (iNEWS 2008), p. 47- 54.

Paltoglou, G., Salampasis, M., Satratzemi, M.: A comparison of Centralized and Distributed Information Retrieval approaches, In Proc. 12th Panhellenic Conference on Informatics (2008), p. 21-25.

Paltoglou, G., Salampasis, M., Satratzemi, M.: A results merging algorithm for distributed information retrieval environments that combines regression methodologies with a selective download phase, Information Processing and Management Journal 44(4): 1580-1599 (2008).

Paltoglou, G., Salampasis, M., Satratzemi, M.: Hybrid Results Merging, In Proc. 16th Conference on Information and Knowledge Management, CIKM 2007, p. 321-331.

Paltoglou, G., Salampasis, M., Satratzemi, M.: Results Merging Algorithm Using Multiple Regression Models, In Proc. 29th European Conference on Information Retrieval, ECIR 2007, p. 173-184.

Paltoglou G., Salampasis, M., Satratzemi, M., Evangelidis, G.: Using linkage information to approximate the distribution of relevant document in DIR, In Proc. 11th Panhellenic Conference on Informatics (2007),p. 234-244.

3. ΣΧΟΛΙΑ - ΠΡΟΒΛΗΜΑΤΑ - ΠΑΡΑΤΗΡΗΣΕΙΣ

3.1. Τεχνολογία / Τεχνογνωσία που αποκτήθηκε στα πλαίσια του έργου

- Η χρήση υπερσυνδέσμων μπορεί να βοηθήσει σημαντικά στην επιλογή πηγών σε κατανεμημένα περιβάλλοντα ανάκτησης πληροφοριών.
- Η σύνθεση αποτελεσμάτων μπορεί να γίνει επιτυχώς και με σημαντική ακρίβεια ακόμα και αν οι κατανεμημένες συλλογές δεν επιστρέφουν σκορ σχετικότητας μαζί με τα κείμενα.
- Το πρόβλημα της επιλογής πηγών πληροφοριών μπορεί να γίνει επιτυχώς και με σημαντική ακρίβεια ακόμα και αν οι κατανεμημένες συλλογές δεν επιστρέφουν σκορ σχετικότητας μαζί με τα κείμενα. Ιδιαίτερα η μοντελοποίηση του προβλήματος με τη χρήση ολοκληρωμάτων βοηθά στην καλύτερη επίλυση του.

3.2. Συνεργασία φορέων (οικονομικό και φυσικό αντικείμενο)

Η συνεργασία μεταξύ των φορέων υλοποίησης του έργου υπήρξε πολύ καλή. Ο ανάδοχος του έργου συνεργάστηκε σύμφωνα με το τεχνικό παράρτημα με τα πανεπιστήμια Μακεδονίας, Sunderland και Aberdeen. Στα αρχικά στάδια του έργου πραγματοποιήθηκε επίσκεψη των καθηγητών prof. John Tait & prof. David Harper με σκοπό την εκπαίδευση του νέου ερευνητή σε μεθοδολογίες έρευνας στο πεδίο της ανάκτησης πληροφοριών και ανταλλαγή απόψεων σχετικά με το ερευνητικό πρόγραμμα. Επίσης πολύ καλή ήτανη συνεργασία με το πανεπιστήμιο Μακεδονίας όπου πραγματοποιούνταν τακτικά συναντήσεις του νέου ερευνητή με την τριμελή επιτροπή επίβλεψης της διδακτορικής διατριβής και σεμινάρια έρευνας.

Η συνεργασία με τον φορέα συγχρηματοδότησης ALTEC ήταν επίσης πολύ καλή και ήταν πολύ σημαντική σε κατοπινά στάδια της έρευνας όπου δημιουργήθηκαν συστήματα αξιολόγησης των νέων αλγορίθμων σε όσο το δυνατό πιο ρεαλιστικά περιβάλλοντα.

3.3. Αιτιολόγηση αποκλίσεων δαπανών ανά φορέα και κατηγορία δαπάνης σε σχέση με την αρχική πρόβλεψη

Λόγω του αρκετά μεγάλου αριθμού των δημοσιεύσεων κατά τη διάρκεια του έργου προέκυψε ανάγκη για μεγαλύτερες δαπάνες όσο αφορά μετακινήσεις σε συνέδρια.

Μετά από αίτηση προς τη ΓΓΕΤ, υπήρξε μικρή αύξηση της σχετικής κατηγορίας δαπάνης.

3.4. Λοιπές παρατηρήσεις