Μέθοδοι Κατάταξης και Ομαδοποίησης Βιβλιογραφικών και Διαδικτυακών Δεδομένων

Αντώνης Σιδηρόπουλος

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
ΕΓΚΡΙΘΕΙΣΑ ΑΠΟ ΤΟ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΟΥ ΑΡΙΣΤΟΤΕΛΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΘΕΣΣΑΛΟΝΙΚΗΣ

Ιούνιος 2006

Η διατριβή αφιερώνεται
στους γονείς μου
Νίκο και Διαμάντω
και
στη γυναίκα μου
Ζωή

ΑΝΤΩΝΗΣ ΣΙΔΗΡΟΠΟΥΛΟΣ

Μέθοδοι Κατάταξης και Ομαδοποίησης Βιβλιογραφικών και Διαδικτυακών Δεδομένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Εγκρίθηκε από το Τμήμα Πληροφορικής του Α.Π.Θ.
Ημερομηνία Προφορικής Εξέτασης: 22 Ιουνίου 2006
Ημερομηνία Ορκομωσίας: 29 Ιουνίου 2006

Εξεταστική Επιτροπή
Καθηγητής Α.Π.Θ., Ιωάννη Μανωλόπουλος, Επιβλέπων
Καθηγητής Α.Π.Θ., Ιωάννης Βλαχάβας, Μέλος Τριμελούς Συμβουλ. Επιτροπής
Καθηγητής Ε.Κ.Π.Α., Ιωάννης Ιωαννίδης, Μέλος Τριμελούς Συμβουλ. Επιτροπής
Επ. Καθηγητής Α.Π.Θ., Ελευθέριος Αγγελής
Επ. Καθηγήτρια Α.Π.Θ., Αθηνά Βακάλη
Επ. Καθηγητής Α.Π.Θ., Νικόλαος Βασιλειάδης
Επ. Καθηγητής Α.Π.Θ., Ιωάννη Σταμέλο

Σύνοψη και κυριότερα επιτεύγματα της διατριβής

Η παρούσα διατριβή διαπραγματεύεται τα ζητήματα της κατάταξης και ομαδοποίησης δεδομένων στο περιβάλλον του Παγκόσμιου Ιστού καθώς και σε Ψηφιακές Βιβλιοθήκες βιβλιογραφικού περιεχομένου, με κίνητρο την ποιοτική βελτίωση των μεθόδων που χρησιμοποιούνται αυτή τη στιγμή, αλλά και την πρακτική εφαρμογή αλγορίθμων κατάταξης και ομαδοποίησης με τελικό σκοπό την αποτελεσματικότερη αναζήτηση. Το αντικείμενο της διατριβής ανήκει στον επιστημονικό τομέα της Πληροφοριομετρίας, όπου συμπεριλαμβάνεται η Βιβλιομετρία και η Ιστομετρία.

Οι κυριότερες συνεισφορές της διατριβής συνοψίζονται στα ακόλουθα:

- Η επινόηση μιας μεθόδου κατάταξης συλλογών από δημοσιεύσεις, όπως περιοδικά και συνέδρια. Στην νέα αυτή μέθοδο έχουμε ως είσοδο το πλήθος των αναφορών που γίνονται από μια συλλογή σε μια άλλη. Η νέα αυτή μέθοδος λαμβάνει υπ' όψη τη "σημαντικότητα" κάθε συλλογής. Η νέα μέθοδος υπολογίζει την κατάταξη-αξιολόγηση με λίγα επίπεδα επαναλήψεων και συνεπώς είναι ιδιαίτερα αποτελεσματική χρονικά.

- Η επινόηση μιας μεθόδου κατάταξης βιβλιογραφικών δεδομένων ή αντικειμένων του Παγκόσμιου Ιστού. Η μέθοδος μπορεί να θεωρηθεί και ως γενίκευση ή βελτίωση του αλγορίθμου `PageRank`. Έχει σχεδιασθεί ειδικά για βιβλιογραφικά δεδομένα, τα πειραματικά αποτελέσματα όμως δείχνουν ότι μπορεί να εφαρμοσθεί και σε γράφους που αναπαριστούν τον Παγκόσμιο Ιστό. Πλεονεκτεί απέναντι στις ανταγωνιστικές μεθόδους στην ποιότητα και στην ταχύτητα.

- Η σχεδίαση ενός αλγορίθμου για ομαδοποίηση δεδομένων που αναπαριστώνται με μη-κατευθυνόμενους χωρίς βάρη γράφους. Η μέθοδος βασίζεται σε γραφο-θεωρητικές ιδιότητες και υπολογίζει την ομαδοποίηση η οποία είναι NP-πλήρες πρόβλημα, σε γραμμικό χρόνο ώς προς το πλήθος των ακμών και κορυφών.

- Η γενίκευση της νέας μεθόδου *h-index* για την κατάταξη συγγραφέων και η προσθήκη της χρονικής διάστασης στην μετρική. Τέλος γίνεται εφορμογή της για την κατάταξη περιοδικών και συνεδρίων.

Abstract and Contribution

The present thesis negotiates the questions of ranking and clustering in the environment of the Web data, as well as in the digital libraries of bibliographic content, motivated by the qualitative improvement of the methods that are currently used, but also the practical application of the ranking and clustering algorithms with final aim the more effective search. The subject of the thesis belongs in the scientific domain of Informetrics, where both Bibliometrics and Webometrics are included.

The major contributions of the thesis are summarized as follows:

- The invention of a ranking method for collections of publications such as journals and conferences. This new method uses as input the number of citations from one collection to another. In this new method the "importance" of each collection is taken into account. The method computes the ranking with few iterations and consequently is specially effective temporally.

- The invention of a ranking method of bibliographic or Web objects. The method can be considered as a generalization or improvement of the `PageRank` algorithm. It has been designed specifically for bibliographic data. However, the experimental results show that it can be also applied in webgraphs. It has the advantage of the quality and computation speed compared to the competitive methods.

- The designing of an algorithm for identification of communities in unweighted and undirected graphs. The method is based on Graph-theoretical attributes and computes the clustering in O(nm) time. The identification of the optimal communities is proved to be an NP-hard problem.

- The generalization of the new ranking method named *h-index* for authors and researchers and the addition of the time dimension in the metrics. Finally, the methods are used for conference and journal ranking.

Σύντομο Βιογραφικό

Ο Αντώνης Σιδηρόπουλος γεννήθηκε στην Κομοτηνή στις 10 Ιουλίου 1973. Είναι μόνιμος κάτοικος Θεσσαλονίκης. Έλαβε το Πτυχίο Πληροφορικής το Μάρτιο του 1996 από το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης, και μεταπτυχιακό στην Επιστήμη Υπολογιστών τον Μάρτιο του 1999 από το ίδιο τμήμα. Υπηρέτησε τη στρατιωτική του θητεία το 2000-2001. Εκπονεί τη Διδακτορική Διατριβή του από το Σεπτέμβρη του 1999 μέχρι σήμερα στο Τμήμα Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Έχει εργασθεί ως προγραμματιστής στη βιομηχανία της πληροφορικής και ως διδάσκων στα ΤΕΙ Θεσσαλονίκης και άλλους φορείς.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές, ειλικρινείς ευχαριστίες μου στον κ. Ιωάννη Μανωλόπουλο, καθηγητή του τμήματος Πληροφορικής του Α.Π.Θ., και κύριο επιβλέποντα της διατριβής μου για την επιστημονική του διορατικότητα, η οποία με οδήγησε στη συγκεκριμένη ερευνητική κατεύθυνση. Θα ήθελα επίσης να ευχαριστήσω τον κ. Ιωάννη Βλαχάβα, καθηγητή του Τμήματος Πληροφορικής του Α.Π.Θ., και τον κ. Ιωάννη Ιωαννίδη, καθηγητή στο Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, που διετέλεσαν μέλη της τριμελούς επιτροπής επίβλεψης της διατριβής μου. Επίσης ευχαριστίες οφείλονται στους κ. Ελευθέριο Αγγελή, επ. καθηγητή του τμήματος Πληροφορικής του Α.Π.Θ., κ. Αθηνά Βακάλη, επ. καθηγήτρια του τμήματος Πληροφορικής του Α.Π.Θ., κ. Νικόλαο Βασιλειάδη, επ. καθηγητή του τμήματος Πληροφορικής του Α.Π.Θ., κ. Ιωάννη Σταμέλο, επ. καθηγητή του τμήματος Πληροφορικής του Α.Π.Θ., οι οποίοι διετέλεσαν μέλη της εξεταστικής επιτροπής της διατριβής μου.

Θα ήθελα να ευχαριστήσω το φίλο και συν-συγγραφέα μου σε κάποιες ερευνητικές εργασίες, Δημήτρη Κατσαρό, για την πολύτιμη συνεργασία του. Επίσης θα ήθελα να ευχαριστήσω όλα τα μέλη του Εργαστηρίου Τεχνολογίας και Επεξεργασίας Δεδομένων του Τμήματος Πληροφορικής για την άριστη συνεργασία που διατηρήσαμε. Ξεχωριστές ευχαριστίες θέλω να απευθύνω στους καλούς φίλους που απέκτησα εκεί, τους Απόστολο Παπαδόπουλο, Αλέξη Νανόπουλο, Ιωάννη Καρύδη, Μαρία Κοντάκη.

Πάνω απ' όλα όμως θέλω να ευχαριστήσω τους γονείς μου Νίκο και Διαμάντω, και την σύζηγό μου Τζούλια. Μου συμπαραστάθηκαν ηθικά και υλικά σε κάθε μου προσπάθεια κατά την εκπόνηση της παρούσας διατριβής. Πραγματικά ένα μεγάλο τμήμα της διατριβής αυτής ανήκει σε αυτούς. Οφείλω ένα μεγάλο ευχαριστώ στον πατέρα μου ο οποίος με παρότρυνε να ξεκινήσω αυτό το μεγάλο εγχείρημα. Επίσης, ιδιαίτερα ευχαριστώ τη σύζηγό μου Τζούλια, η οποία συνετέλεσε κυριολεκτικά στην παρουσίαση της διατριβής.

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Περιεχόμενα

1.1 Πληροφοριομετρία

Με τον όρο *Πληροφοριομετρία* (Informetrics) αναφερόμαστε στο σύνολο των μετρικών μεθόδων που σχετίζονται με την Επιστήμη της Πληροφορίας (information science) [22]. Στην έννοια αυτή συμπεριλαμβάνονται οι τομείς:

1. Βιβλιομετρία (bibliometrics): αναφέρεται σε βιβλιογραφικές πληροφορίες, ψηφιακές βιβλιοθήκες κ.τ.λ.

2. Επιστημομετρία (scientometrics): αναφέρεται στην περιοχή που αφορά στην ανάλυση αναφορών, αξιολόγηση έρευνας κ.τ.λ., και

3. Ιστομετρία (webometrics): σχετίζεται με μετρικές για τον Παγκόσμιο Ιστό (World Wide Web - WWW) ή άλλα κοινωνικά δίκτυα (social networks) όπως δίκτυα αναφορών ή συνεργασιών.

Ο όρος Πληροφοριομετρία προτάθηκε το 1979 από τους Blackert και Siegel [11] και παράλληλα από τον Nacke [68]. Σχετικά αργότερα και εν πάσει περιπτώσει μετά το 1987 άρχισε να γίνεται δημοφιλής με τη διοργάνωση του αντίστοιχου συνεδρίου για την Πληροφοριομετρία [23]. Ωστόσο, ο ερευνητικός τομέας της Πληροφοριομετρίας θεμελιώθηκε σχεδόν από τις αρχές του εικοστού αιώνα από τους Lotka, Bradford και Zipf [64, 13, 97]. Ο όρος *Βιβλιομετρία* εισήχθη το 1969 από τον Pritchard [78]. Την ίδια χρονιά χρησιμοποιήθηκε και ο όρος *Επιστημομετρία* από τους Nalimov και Mulcenko [69].

Η περιοχή της Πληροφοριομετρίας αναπτύχθηκε μεν τον εικοστό αιώνα, αλλά τις τελευταίες δεκαετίες έχει γνωρίσει ιδιαίτερα μεγάλη ανάπτυξη. Στην εργασία [63] ο Lipetz αποδεικνύει τη γιγάντωση του περιοδικού JASIS (πλέον JASIST - Journal of the American Society for Information Science) μελετώντας το πλήθος των δημοσιεύσεων στο περιοδικό, καθώς και το πλήθος των συγγραφέων που ασχολούνται με το αντικείμενο. Σήμερα, μπορούμε να ισχυρισθούμε ότι η περιοχή της Πληροφοριομετρίας καλύπτει τους τομείς της Βιβλιομετρίας, Επιστημομετρίας, Ιστομετρίας και των Κοινωνικών Δικτύων. Όλες οι περιοχές αυτές έχουν κάποια κοινά χαρακτηριστικά, καθώς προφανώς και κάποια διαφορετικά χαρακτηριστικά. Το πιο σημαντικό *κοινό* χαρακτηριστικό είναι ότι στους τομείς αυτούς χρησιμοποιούνται γράφοι για την αναπαράσταση των δεδομένων. Για παράδειγμα, στη Βιβλιομετρία στηριζόμαστε σε γράφους αναφορών, δηλαδή, γράφους όπου οι οντότητες που αναπαριστώνται με τις κορυφές είναι δημοσιεύσεις (publications), ενώ οι ακμές αναπαριστούν αναφορές (citations) από μια δημοσίευση σε μια άλλη. Αντίστοιχα, στον Παγκόσμιο Ιστό και πάλι χρησιμοποιείται ένας γράφος για την αναπαράσταση των δεδομένων. Πιο συγκεκριμένα, οι κορυφές αναπαριστούν ιστοσελίδες (webpages), ενώ οι ακμές αναπαριστούν υπερσυνδέσμους (hyperlinks) μεταξύ των ιστοσελίδων. Σε άλλες περιπτώσεις μπορούμε επίσης να κατασκευάσουμε γράφους αναφορών συγγραφέων (author citation graphs) ή γράφους συνεργασιών συγγραφέων (co-authorship graphs). Αυτή η αναπαράσταση της πληροφορίας με τη βοήθεια γράφων μπορεί να δώσει πολλαπλές επιπρόσθετες πληροφορίες που προφανώς δεν είναι άμεσα αντιληπτές αν δεν μελετήσουμε το γράφο στο σύνολό του.

Στην παρούσα διατριβή μελετώνται δυο βασικές διεργασίες που εφαρμόζονται σε γράφους των τύπων που αναφέρθηκαν. Η πρώτη είναι η *κατάταξη* (ranking) των αντικειμένων μας (δηλαδή των κορυφών του γράφου) με βάση την *αξία* τους. Έτσι μπορούμε να κατατάξουμε οντότητες που αναπαριστούν:

- δημοσιεύσεις - ώστε να ανιχνεύσουμε τις καλύτερες,

- συγγραφείς-ερευνητές - ώστε να αναγνωρίσουμε τους σημαντικότερους ή αυτούς που συνεισέφεραν στην επιστήμη περισσότερο
- περιοδικά ή συνέδρια - ώστε να τα αξιολογήσουμε και
- σελίδες στον Παγκόσμιο Ιστό - ώστε να μπορέσουμε να βελτιώσουμε τη διαδικασία της αναζήτησης πληροφορίας στο διαδίκτυο.

Η δεύτερη διεργασία που εξετάζει η παρούσα διατριβή είναι η *ομαδοποίηση* (clustering) των προηγουμένων οντοτήτων με βάση το γράφο αναπαράστασης. Η τεχνική αυτή μπορεί να χρησιμοποιηθεί για την:

- ομαδοποίηση αντικειμένων του Παγκόσμιου Ιστού,
- ομαδοποίηση-εύρεση κοινοτήτων ερευνητών,
- ομαδοποίηση συναφών δημοσιεύσεων κ.ο.κ.

1.2 Δομή και Συνεισφορά της Διατριβής

Στη συνέχεια του παρόντος κεφαλαίου παρουσιάζονται διάφορες εισαγωγικές έννοιες της περιοχής της Πληροφοριομετρίας και της Ιστομετρίας, με σκοπό την καλύτερη κατανόηση των ζητημάτων που θα αναπτυχθούν στα επόμενα κεφάλαια της παρούσας διατριβής.

Στο Κεφάλαιο 2 εξετάζεται το ζήτημα της αξιολόγησης-κατάταξης συνεδρίων και γενικότερα συλλογών δημοσιεύσεων. Η αξιολόγηση αυτή γίνεται σε επίπεδο γράφου αναφορών συλλογών. Δηλαδή, οι κορυφές του αντίστοιχου γράφου είναι οι συλλογές (τα συνέδρια), ενώ οι ακμές χαρακτηρίζονται από ένα βάρος που αντιστοιχεί στο πλήθος των αναφορών που έχουν γίνει από το ένα συνέδριο στο άλλο. Αντίστοιχη μέθοδος, με την έννοια ότι στηρίζεται στο ίδιο σύνολο δεδομένων εισόδου, είναι ο *Παράγοντας Αντίκτυπου ISI* (ISI Impact Factor) [31, 32]. Το κύριο μειονέκτημα του Παράγοντα Αντίκτυπου ISI είναι ότι δεν λαμβάνει υπ'όψη τη σπουδαιότητα των συνεδρίων από τα οποία γίνεται η αναφορά. Βέβαια στην πράξη ο Παράγοντας Αντίκτυπου ISI εφαρμόζεται μόνο σε περιοδικά και όχι σε συνέδρια. Θα μπορούσε όμως να εφαρμοσθεί και σε συνέδρια; Για να καλύψουμε αυτό το κενό ορίσαμε νέες μετρικές μεθόδους που λαμβάνουν υπ'όψη και τη σπουδαιότητα του κάθε συνεδρίου που κάνει αναφορές. Έτσι, η βαθμολογία-κατάταξη των συνεδρίων υπολογίζεται αναδρομικά. Σε κάθε βήμα λαμβάνουμε υπ'όψη τις *σπουδαιότητες* (δηλαδή, βάρη) που υπολογίσθηκαν στο προηγούμενο. Ένα άλλο

μεγάλο μειονέκτημα του Παράγοντα Αντίκτυπου ISI είναι ότι αξιολογεί ένα περιοδικό (ή ένα συνέδριο) για ένα χρονικό παράθυρο. Υπολογίζοντας δηλαδή τον Παράγοντα Αντίκτυπου ISI για το συνέδριο VLDB και για το έτος 1999, ουσιαστικά αξιολογούμε το συνέδριο αυτό για τα 5 (ή 2) προηγούμενα έτη. Κάτι τέτοιο δεν είναι πάντα επιθυμητό. Για παράδειγμα, συχνά θα προτιμούσαμε να αξιολογήσουμε μία μεμονωμένη διοργάνωση του συνεδρίου VLDB, π.χ. του 1995. Για να το επιτύχουμε αυτό ορίσαμε τον Αντεστραμμένο Παράγοντα Αντίκτυπου

Στο Κεφάλαιο 3 μελετώνται πάλι οι γράφοι αναφορών, αλλά κατά ένα επίπεδο λεπτομερέστερης πληροφορίας σε σχέση με τους γράφους του Κεφαλαίου 2. Πιο συγκεκριμένα, στο κεφάλαιο αυτό οι κορυφές στους γράφους αναφορών αναπαριστούν δημοσιεύσεις, ενώ οι ακμές αναπαριστούν αναφορές από μια δημοσίευση προς κάποια άλλη. Στο επίπεδο αυτό οι ακμές δεν έχουν βάρος ή μπορεί να θεωρηθεί ότι όλες οι ακμές είναι ισοβαρείς (με βάρος 1). Η οντότητα που αξιολογείται-βαθμολογείται με βάση έναν τέτοιο γράφο είναι οι ίδιες οι δημοσιεύσεις. Προφανώς, σε αυτή την περίπτωση δεν μπορούν να εφαρμοσθούν οι αλγόριθμοι που αναπτύχθηκαν στο Κεφάλαιο 2. Παρουσιάζουμε λοιπόν, μια νέα οικογένεια αλγορίθμων που ως σύνολο δεδομένων εισόδου δέχονται κατευθυνόμενους γράφους αναφορών χωρίς βάρη. Άλλοι γνωστοί αλγόριθμοι που χρησιμοποιούν ίδια σύνολα δεδομένων εισόδου είναι οι `PageRank`, `HITS`, `Prestige`, `SALSA`. Όλοι αυτοί οι αλγόριθμοι αναπτύχθηκαν κυρίως για εφαρμογή στην κατάταξη-αξιολόγηση σελίδων του Παγκόσμιου Ιστού και όχι για γράφους αναφορών δημοσιεύσεων. Η εφαρμογή τους σε γράφους αναφορών δημοσιεύσεων παρουσιάζει αρκετές προβληματικές συμπεριφορές. Προκειμένου να τις παρακάμψουμε προτείνουμε τις παραλλαγές `B-HITS` και `B-SALSA`, καθώς επίσης εισάγουμε μία νέα οικογένεια αλγορίθμων, την οποία ονομάζουμε `SCEASRank`. Χρησιμοποιώντας αυτούς τους αλγορίθμους μπορούμε να επιτύχουμε την αξιολόγηση-κατάταξη των δημοσιεύσεων ή μιας ομάδας δημοσιεύσεων. Έτσι μπορούμε να εντοπίσουμε αξιόλογες δημοσιεύσεις ή να προτείνουμε δημοσιεύσεις προς βράβευση ή τέλος, να προβλέψουμε ποιες δημοσιεύσεις είναι πιθανό να βραβευθούν στο μέλλον.

Στο δεύτερο τμήμα του ίδιου κεφαλαίου υπολογίζεται η βαθμολογία-κατάταξη για συγγραφείς κάνοντας χρήση των βαθμολογιών που υπολογίσθηκαν προηγουμένως. Η βαθμολογία ενός συγγραφέα προκύπτει από τις βαθμολογίες των δημοσιεύσεών του χρησιμοποιώντας μια συνάρτηση συνάθροισης. Αντίστοιχα, μπορούμε να εντοπίσουμε αξιόλογους συγγραφείς-ερευνητές ή να προτείνουμε συγγραφείς προς βράβευση ή να προβλέψουμε ποιοι ερευνητές είναι πιθανό να βραβευθούν στο μέλλον.

Στο Κεφάλαιο 4 κάνουμε χρήση των μεθόδων της οικογένειας SCEASRank

σε γράφους που αναπαριστούν τον Παγκόσμιο Ιστό. Ταυτόχρονα συγκρίνουμε με τους ήδη υπάρχοντες (PageRank, HITS, Prestige, SALSA).

Στο Κεφάλαιο 5 εστιάζουμε και πάλι στον τομέα της Ιστομετρίας προσπαθώντας να ομαδοποιήσουμε οντότητες του διαδικτύου. Το σύνολο δεδομένων εισόδου είναι και σε αυτήν την περίπτωση ένας γράφος που αναπαριστά τον Παγκόσμιο Ιστό. Οι υπάρχουσες μέθοδοι έχουν ένα κοινό μεγάλο μειονέκτημα: απαιτούν πάρα πολλούς υπολογιστικούς πόρους: χρόνο και μνήμη. Με τη νέα προτεινόμενη μέθοδο υπολογίζουμε ομάδες-κοινότητες αντικειμένων, έτσι ώστε η σχέση κόστους/απόδοσης να είναι κατά πολύ βελτιωμένη. Η νέα αυτή τεχνική εφαρμόσθηκε ήδη σε αλγορίθμους έξυπνης προ-τοποθέτησης (caching) αντικειμένων σε CDNs (Content Delivery Networks). Ο ίδιος αλγόριθμος μπορεί να χρησιμοποιηθεί και σε γράφους αναφορών, ώστε να βρεθούν ομάδες δημοσιεύσεων ή κοινότητες (communities) ερευνητών.

Στο Κεφάλαιο 6 γίνεται μελέτη της τεχνικής *h-index*, που παρουσιάσθηκε πρόσφατα, μόλις το Νοέμβριο του 2005 από τον J. E. Hirsch [44]. Η μέθοδος αυτή κάνει χρήση του γράφου αναφορών σε επίπεδο δημοσιεύσεων. Λαμβάνοντας όμως παράλληλα υπ'όψη και τη σχέση δημοσίευση-συγγραφέας υπολογίζει κατ'ευθείαν τη βαθμολογία για τους συγγραφείς. Στο παρόν κεφάλαιο προτείνουμε νέες εκδοχές της τεχνικής h-index για συγγραφείς σε μία προσπάθεια για ανάδειξη περισσότερο ραφιναρισμένης πληροφορίας. Επίσης, μεταφέρουμε την έννοια του h-index και σε ανώτερα συγκεντρωτικά επίπεδα, δηλαδή στο επίπεδο των συλλογών δημοσιεύσεων (όπως περιοδικά και συνέδρια) ορίζοντας τις κατάλληλες παραλλαγές.

Τέλος, στο Κεφάλαιο 7 κάνουμε μια σύνοψη των επιτευγμάτων της διατριβής και δίνουμε έναυσμα για περαιτέρω έρευνα.

1.3 Εισαγωγικές Έννοιες

Σε αυτό το κεφάλαιο θα εισαγάγουμε βασικές έννοιες σχετικά με τον τομέα της Πληροφοριομετρίας. Ειδικότερα θα ασχοληθούμε με το χώρο της Βιβλιομετρίας και της Ιστομετρίας. Ειδικότερα, οι μέθοδοι που μας ενδιαφέρουν είναι μέθοδοι κατάταξης καθώς και μέθοδοι ομαδοποίησης στους τομείς αυτούς.

1.3.1 Βιβλιομετρία

Βιβλιομετρία είναι ένας τύπος ανάλυσης πληροφοριών και κυρίως υποστηρίζει τον τομέα των Βιβλιοθηκονομίας και της Επιστήμης της Πληροφορίας. Πιο συγκεκρι-

μένα, Βιβλιομετρία είναι η περιοχή που ασχολείται με τεχνικές μετρικής κειμένων και πληροφοριών [74]. Βιβλιομετρικές μέθοδοι χρησιμοποιούνται για να βρεθούν *σχέσεις* μεταξύ επιστημονικών περιοδικών[1], συνεδρίων, δημοσιεύσεων και συγγραφέων. Στην έννοια της Βιβλιομετρίας εμπεριέχεται η έννοια της Ανάλυσης Αναφορών (citation analysis), ενώ σχεδόν όλες οι μέθοδοι που αναπτύσσονται στα πλαίσια της Βιβλιομετρίας αφορούν στην Ανάλυση Αναφορών. Η Ανάλυση Αναφορών στηρίζεται στην απεικόνιση των πληροφοριών μας με ένα γράφο, όπου οι ακμές αντιστοιχούν σε *αναφορές*, ενώ οι κορυφές του γράφου μπορεί να αντιστοιχούν σε δημοσιεύσεις, συγγραφείς, περιοδικά ή συνέδρια.

Η Ανάλυση Αναφορών χρησιμοποιείται κυρίως για την αξιολόγηση-κατάταξη οντοτήτων σε μια ψηφιακή βιβλιοθήκη. Έτσι, μπορεί να γίνει αξιολόγηση περιοδικών, συνεδρίων, δημοσιεύσεων και συγγραφέων. Όσο περισσότερες αναφορές δέχεται μια οντότητα, τόσο *καλύτερη* είναι. Το *καλύτερη* σημαίνει ότι είτε έχει συνεισφέρει στην επιστήμη τόσο ώστε να γίνεται σημείο αναφοράς από πολλούς άλλους ερευνητές, είτε γενικώς έχει απασχολήσει την επιστημονική κοινότητα, είτε έχει δώσει έναυσμα για περαιτέρω έρευνα κ.τ.λ.

Μια ψηφιακή βιβλιοθήκη, η οποία περιέχει πληροφορίες αναφορών συνήθως αναφέρεται στη βιβλιογραφία ως *Ευρετήριο Αναφορών* (Citation Index). Το πρώτο οργανωμένο Ευρετήριο Αναφορών είναι αυτό του Ινστιτούτου για την Επιστημονική Πληροφορία (Institute for Scientific Information - ISI [30]). Το ISI ιδρύθηκε το 1960 από τον Eugene Garfield[2] και αργότερα (το 1992) εξαγοράσθηκε από την Thomson Scientific. Οι ψηφιακές βιβλιοθήκες του ISI είναι γνωστές ως SCI (Science Citation Index) και SSCI (Social Science Citation Index).

Τη δεκαετία του 1990 εμφανίσθηκε το CiteSeer[3] του Ινστιτούτου Έρευνας της NEC, το οποίο αποτελεί επίσης ένα Ευρετήριο Αναφορών. Είναι ένα σύγχρονο σύστημα που κατασκευάζει το γράφο αναφορών από δημοσιεύσεις που αντλούνται από τον Παγκόσμιο Ιστό [58]. Πιο συγκεκριμένα, πρόκειται για ένα σύστημα που συλλέγει δημοσιεύσεις σχετικές με την Πληροφορική σαρώνοντας τον Παγκόσμιο Ιστό. Κατόπιν, ανιχνεύονται και εξάγονται οι βιβλιογραφικές πληροφορίες της δημοσίευσης (π.χ. τίτλος, συγγραφείς κ.τ.λ.), όπως επίσης και οι αναφορές που συμπεριλαμβάνονται σε αυτήν, ώστε να κατασκευασθεί ο αντίστοιχος γράφος αναφορών [57].

Αντίστοιχο σύστημα με το CiteSeer αναπτύχθηκε πλέον (το 2004) και από

[1] Εφεξής για λόγους συντομίας θα λέμε απλώς *περιοδικά* αντί *Επιστημονικά περιοδικά*.

[2] Ο Eugene Garfield γεννήθηκε στη Νέα Υόρκη το 1925 και θεωρείται ένας από τους θεμελιωτές των επιστημονικών τομέων την Βιβλιομετρίας και Επιστημομετρίας.

[3] `http://citeseer.ist.psu.edu/`

την Google[4], το οποίο αυτή τη στιγμή αποτελεί το μεγαλύτερο και ευρύτερο Ευρετήριο Αναφορών. Το ευρετήριο αυτό καλύπτει όλες τις επιστημονικές περιοχές. Προφανώς βέβαια τα συστήματα όπως το CiteSeer και το Google Scholar περιέχουν και *άκυρη* πληροφορία ή εν πάσει περιπτώσει μη απολύτως έγκυρη πληροφορία, δεδομένου ότι συλλέγουν δεδομένα από το Διαδίκτυο. Για παράδειγμα, αναγνωρίζουν ως δημοσιεύσεις και κείμενα που δεν είναι δημοσιευμένα σε κάποιο περιοδικό ή συνέδριο. Επίσης, όταν τα ονόματα των συγγραφέων είναι διατυπωμένα με διαφορετικό τρόπο δεν μπορούν να κάνουν σωστά την ταυτοποίηση των συγγραφέων. Προφανώς, η ερευνητική και η βιομηχανική κοινότητα ασχολείται και προσπαθεί να επιλύσει τέτοιου είδους προβλήματα.

Πλέον οι περισσότερες ψηφιακές βιβλιοθήκες που περιέχουν επιστημονικές δημοσιεύσεις συμπεριλαμβάνουν και Ευρετήριο Αναφορών, όπως για παράδειγμα η Ψηφιακή Βιβλιοθήκη της ACM[5], η βιβλιοθήκη του εκδοτικού οίκου Elsevier[6] καθώς και πολλά άλλα συστήματα εκδοτικών οίκων. Αυτό δείχνει ότι πλέον θεωρείται σημαντική η πληροφορία των αναφορών, αλλά και ότι έχει ωριμάσει αρκετά η τεχνολογία ώστε να είναι εφικτή η εύρεση της πληροφορίας αυτής.

1.3.2 Παγκόσμιος Ιστός

Ο Παγκόσμιος Ιστός [7] έχει γίνει το περισσότερο δημοφιλές δίκτυο κατά την τελευταία δεκαετία. Είναι πλέον η κύρια πηγή πληροφορίας στην καθημερινή, αλλά και στην επαγγελματική ζωή. Ακόμη και σε επιστημονικές δημοσιεύσεις συχνά γίνονται αναφορές σε πηγές από τον Παγκόσμιο Ιστό [43, 56, 89, 96]. Οι πληροφορίες που μπορούν να βρεθούν στον Παγκόσμιο Ιστό είναι τεράστιες. Επί πλέον ο Παγκόσμιος Ιστός μεγαλώνει με γρήγορους ρυθμούς. Στην αρχή της δεκαετίας του 1990 εμφανίσθηκαν οι πρώτες μηχανές αναζήτησης, όπως η altavista[7] και το lycos[8]. Στις μηχανές αυτές, ο χρήστης μπορούσε (μπορεί) να δώσει κάποιες λέξεις-κλειδιά (keywords) και η μηχανή αναζήτησης εμφάνιζε στο χρήστη μια λίστα από σελίδες (URLs - Uniform Resource Locator), οι οποίες περιείχαν τις λέξεις αυτές. Η σειρά εμφάνισης των σελίδων στη λίστα εξαρτώνταν από το πλήθος ή τη συχνότητα εμφάνισης των λέξεων μέσα στην κάθε σελίδα. Κάποιες περισσότερο *έξυπνες* μηχανές λάμβαναν υπ' όψη και τη *σημαντικότητα* (semantics) της κάθε λέξης-κλειδί.

[4] `http://scholar.google.com`
[5] The ACM Digital Library - `http://portal.acm.org/dl.cfm`
[6] `http://www.sciencedirect.com`
[7] `http://www.altavista.com/`
[8] `http://www.lycos.com/`

Καθώς ο Παγκόσμιος Ιστός μεγαλώνει, μια αναζήτηση με βάση κάποια λέξη-κλειδί μπορεί να δώσει ως αποτέλεσμα χιλιάδες, ίσως και εκατομμύρια σελίδες. Ένας άνθρωπος (χρήστης του Παγκόσμιου Ιστού) μπορεί να ελέγξει τις πρώτες είκοσι ή έστω λίγες περισσότερες σελίδες [88, 90]. Άρα πλέον, είναι πολύ σημαντική η σειρά εμφάνισης των σελίδων σε μια λίστα αποτελεσμάτων από μια μηχανή αναζήτησης. Οι περισσότερο σημαντικές σελίδες καθώς και οι σχετικότερες σελίδες με τις λέξεις-κλειδιά πρέπει να κατατάσσονται πρώτες στη λίστα αποτελεσμάτων.

Η παλιά μέθοδος με βάση τη συχνότητα των λέξεων-κλειδιά κατέστη αναποτελεσματική καθώς οι δημιουργοί των ιστοχώρων (website) βρήκαν τρόπους ώστε να *ξεγελούν* τις μηχανές αναζήτησης περιλαμβάνοντας μέσα στη σελίδα πληθώρα λέξεων-κλειδιά. Συνεπώς, προέκυψε η ανάγκη για μια περισσότερο αντικειμενική μέθοδο κατάταξης. Η Κατάταξη που προκύπτει από την Κατάταξη με Ανάλυση Αναφορών (Link Analysis Ranking - LAR) είναι ένας αντικειμενικός τρόπος για την ταξινόμηση των αποτελεσμάτων μιας αναζήτησης στον Παγκόσμιο Ιστό. Σήμερα, οι περισσότερες μηχανές αναζήτησης χρησιμοποιούν Κατάταξη με Ανάλυση Αναφορών (LAR) σε συνδυασμό με άλλες τεχνικές. Η περισσότερο δημοφιλής μηχανή αναζήτησης είναι η Google[9][72]. Αν και πλέον οι μηχανές αναζήτησης κρατούν κρυφές τις μεθόδους κατάταξης, γνωρίζουμε ότι ανάμεσα στα 100 κριτήρια που χρησιμοποιεί η Google συμπεριλαμβάνεται και ο αλγόριθμος PageRank[10][16].

1.4 Υπόβαθρο

Είδαμε στις προηγούμενες παραγράφους μια συνοπτική παρουσίαση του χώρου της Βιβλιομετρίας καθώς και της Ιστομετρίας. Όπως αναφέραμε και προηγουμένως, στην παρούσα διατριβή θα ασχοληθούμε με δυο βασικές λειτουργίες που είναι απαραίτητες για τους ανωτέρω χώρους: την Κατάταξη-Αξιολόγηση οντοτήτων καθώς και την Ομαδοποίηση αυτών.

1.4.1 Κατάταξη - Αξιολόγηση

Όπως αναφέρθηκε και στην προηγούμενη παράγραφο, η κατάταξη-αξιολόγηση σελίδων στον Παγκόσμιο Ιστό είναι μια κρίσιμη λειτουργία. Η πρωτεύουσα χρήση

[9] `http://www.google.com/`

[10] Η ονομασία PageRank προέρχεται από το όνομα του εμπνευστή του Lawrence Page και ιδρυτή της Google.

της είναι από τις μηχανές αναζήτησης για έξυπνη εμφάνιση και ταξινόμηση των αποτελεσμάτων αναζήτησης των χρηστών του Παγκόσμιου Ιστού. Από την άλλη μεριά η αξιολόγηση ενός ιστοχώρου είναι χρήσιμη, εφόσον κάτι τέτοιο διαμορφώνει την εμπορική τιμή του ιστοχώρου σε επιχειρηματικό επίπεδο, καθώς επίσης και το κόστος των καταχωριζόμενων διαφημίσεων (banners)[11].

Αντίστοιχα, στο χώρο της Βιβλιομετρίας, η αξιολόγηση της ερευνητικής δουλειάς ενός ιδρύματος ή πανεπιστημίου ή εκδοτικού οίκου μπορεί ακόμη και να κρίνει τη χρηματοδότηση προς αυτό. Προφανώς ισχύει και στο χώρο της Βιβλιομετρίας, ότι η κατάταξη των δημοσιεύσεων διευκολύνει τη διαδικασία της αναζήτησης μέσω των ψηφιακών βιβλιοθηκών.

Οι αλγόριθμοι αξιολόγησης-κατάταξης μπορούν γενικά να χωρισθούν σε δύο κατηγορίες. Θα αναφερόμαστε στην πρώτη κατηγορία με την ονομασία *Αλγόριθμοι Αξιολόγησης σε επίπεδο Συλλογών*. Σε αυτήν την κατηγορία χρησιμοποιείται ένας γράφος αναφορών με βάρη. Οι κόμβοι του γράφου αναπαριστούν συλλογές, ενώ οι έχουσες βάρη ακμές αναπαριστούν το συνολικό αριθμό αναφορών που έγιναν από μία συλλογή σε μία άλλη. Οι συλλογές μπορεί να είναι πρακτικά συνεδρίων, περιοδικά, τεχνικές αναφορές πανεπιστημίων και ιδρυμάτων, ή οποιοδήποτε άλλο εννοιολογικό σύνολο δημοσιεύσεων. Υπό αυτή την έννοια, θα μπορούσαν να θεωρηθούν ως συλλογές και οι δημοσιεύσεις ενός συγγραφέα. Επομένως, θα μπορούσαμε να αξιολογήσουμε συγγραφείς, χρησιμοποιώντας μεθόδους που εντάσσονται σε αυτήν την κατηγορία αξιολόγησης. Σε επίπεδο Παγκόσμιου Ιστού, ομοίως μπορούν να ορισθούν συλλογές αντικειμένων. Ο απλούστερος εννοιολογικός ορισμός μιας συλλογής στον Παγκόσμιο Ιστό είναι η έννοια ενός ιστοχώρου.

Τη δεύτερη κατηγορία εφεξής θα ονομάζουμε *Αλγόριθμοι Αξιολόγησης σε επίπεδο αντικειμένων*, όπου τα αντικείμενα μπορεί να είναι είτε δημοσιεύσεις (για τη Βιβλιομετρία), είτε ιστοσελίδες (για την Ιστομετρία). Σύμφωνα με αυτή την προσέγγιση, οι κόμβοι του γράφου αναπαριστούν δημοσιεύσεις (ή ιστοσελίδες), ενώ μία ακμή από τον κόμβο x στον κόμβο y αναπαριστά μία αναφορά (ή έναν υπερσύνδεσμο) από τη δημοσίευση (ή ιστοσελίδα) x στην y. Οι ακμές δεν έχουν βάρη. Ο υπολογισμός της κατάταξης-βαθμολογίας στο επίπεδο των δημοσιεύσεων έχει το πλεονέκτημα ότι εκτελείται μία μοναδική διαδικασία για να κάνει ταυτόχρονη αξιολόγηση περισσότερων της μίας οντότητας: την ίδια την εργασία, τη συλλογή όπου ανήκει και τέλος τους συγγραφείς. Οι τελευταίοι

[11] Για παράδειγμα, το σύστημα `http://www.alexa.com` αξιολογεί ιστοχώρους υπολογίζοντας την κατάταξής τους με ανάλυση αναφορών και παρακολούθηση (tracking) της επισκεψιμότητάς τους από τους χρήστες τους.

δύο υπολογισμοί μπορούν να γίνουν από έναν αθροιστικό μέσο όρο του πρώτου υπολογισμού ή χρησιμοποιώντας μία πιο εξελιγμένη συνάρτηση συνάθροισης.

Στην πρώτη κατηγορία μεθόδων ανήκει η μέθοδος του *Παράγοντα Αντικτύπου ISI* (ISI impact factor) [31, 32] για τον οποίο θα μιλήσουμε εκτενέστερα στο Κεφάλαιο 2. Επίσης, μπορούμε να πούμε ότι και η μετρική *h-index* για συγγραφείς ανήκει σε αυτήν την κατηγορία (Κεφάλαιο 6). Στη δεύτερη κατηγορία ανήκουν οι αλγόριθμοι `CitationCount`, `PageRank`, `Prestige`, `HITS` και `SALSA`, οι οποίοι περιγράφονται εκτενώς στο Κεφάλαιο 3 της παρούσας διατριβής.

Το κίνητρα της παρούσας διατριβής, όσον αφορά στην κατάταξη-αξιολόγηση, είναι:

1. Για την πρώτη κατηγορία μεθόδων, η Μέθοδος του *Παράγοντα Αντικτύπου ISI* δεν λαμβάνει υπ' όψη της τη σημαντικότητα (semantics) ή μη των αναφορών. Έτσι όλες οι αναφορές μετρούν το ίδιο, είτε προέρχονται από κορυφαία συνέδρια είτε όχι, είτε προέρχονται από κορυφαίους ερευνητές είτε όχι. Υποστηρίζουμε ότι η "συμπεριφορά" αυτή δεν είναι δίκαια. Έτσι, προσθέτουμε την έννοια του βάρους στις αναφορές. Περισσότερα αναλύονται στο Κεφάλαιο 2.

2. Για τη δεύτερη κατηγορία μεθόδων, από τη μια πλευρά η μέθοδος *Καταμέτρησης Αναφορών* (`CitationCount`) "ισοπεδώνει" όλες τις αναφορές. Από την άλλη πλευρά, οι μέθοδοι `PageRank`, `Prestige`, `HITS`, `SALSA`, οι οποίες δίνουν βάρη στις αναφορές, έχουν δημιουργηθεί ειδικά για τον Παγκόσμιο Ιστό. Έτσι, εννοιολογικά δεν ταιριάζουν απόλυτα στο χώρο της Βιβλιομετρίας, οπότε όταν τις εφαρμόζουμε στο χώρο αυτό σε πολλές περιπτώσεις δεν μας δίνουν τα επιθυμητά αποτελέσματα. Περαιτέρω ανάλυση για την κακή συμπεριφορά των αλγορίθμων αυτών γίνεται στο Κεφάλαιο 3 της παρούσας διατριβής.

Συνεπώς, στα δυο αυτά κεφάλαια που αναφέραμε βελτιώνουμε και καλύπτουμε τα κενά των ανωτέρω περιπτώσεων. Για τη δεύτερη κατηγορία αλγορίθμων αξιολόγησης βιβλιογραφικών δεδομένων ορίζουμε μια νέα προτεινόμενη οικογένεια με την ονομασία `SCEASRank`. Από την άλλη μεριά, η νέα οικογένεια αλγορίθμων που ορίσαμε φαίνεται στο Κεφάλαιο 4 να δίνει πολύ καλά αποτελέσματα και στο χώρο της Ιστομετρίας, παρότι αρχικώς ορίσθηκε ειδικά για το χώρο της Βιβλιομετρίας. Στην προκειμένη περίπτωση, τα αποτελέσματα είναι αντίστοιχα με τους υπάρχοντες αλγορίθμους, αλλά η οικογένεια `SCEASRank` υπερτερεί κατά πολύ στην ταχύτητα υπολογισμού.

1.4.2 Ομαδοποίηση

Συγχρόνως με την ανάγκη για ορθή κατάταξη-αξιολόγηση των δεδομένων, δημιουργήθηκε και η ανάγκη για ομαδοποίηση των δεδομένων αυτών. Για παράδειγμα, στον Παγκόσμιο Ιστό οι ιστοχώροι μοιράζουν το περιεχόμενό τους σε διάφορους διακομιστές ανά τον κόσμο με σκοπό την ταχύτερη πρόσβαση από τους χρήστες. Χαρακτηριστικό παράδειγμα είναι η AKAMAI[12], που διαθέτει περίπου 18,000 διακομιστές σε 69 χώρες. Έτσι, ο χρήστης του Παγκόσμιου Ιστού μπορεί να έχει πρόσβαση στο αρχείο που θέλει κάνοντας προσπέλαση του αρχείου στον κοντινότερό του διακομιστή, χωρίς να είναι αναγκασμένος να αναφερθεί στο διακομιστή-πηγή. Έτσι γίνεται καταμερισμός του φόρτου του δικτύου και ο χρήστης απολαμβάνει γρήγορη πρόσβαση στους ιστοχώρους που τον ενδιαφέρουν. Η δημιουργία αντιγράφων σε χιλιάδες διακομιστές έχει ένα πολύ σημαντικό μειονέκτημα. Απαιτείται μεγάλο οικονομικό κόστος για τα αποθηκευτικά μέσα. Με σκοπό λοιπόν τη βελτίωση του χρόνου πρόσβασης αυξάνεται το κόστος αποθήκευσης. Το κόστος αποθήκευσης μπορεί να μειωθεί μόνο εάν γίνεται έξυπνη αντιστοιχία αντικειμένων-διακομιστών, προκειμένου να μην δημιουργήσουμε αντίγραφα σε όλους τους διακομιστές, αλλά σε όσους το δυνατόν λιγότερους. Για να επιτευχθεί αυτή η διαδικασία απαιτείται ομαδοποίηση των δεδομένων, έτσι ώστε κάθε ομάδα να αντιγραφεί σε επιλεγμένο σύνολο διακομιστών. Συνεπώς υπάρχει η ανάγκη για ομαδοποίηση των δεδομένων.

Από την άλλη, η ομαδοποίηση δεδομένων του Παγκόσμιου Ιστού μπορεί να βελτιώσει και να αναβαθμίσει τη διαδικασία αναζήτησης στο διαδίκτυο. Δεδομένης μιας πληροφορίας ή ενός κειμένου μπορούν να αναζητηθούν και άλλα που ανήκουν στην ίδια ομάδα με αυτό. Έτσι δίνεται στη διαδικασία αναζήτησης και η δυνατότητα για εξόρυξη πληροφορίας (data mining).

Το ίδιο ισχύει και κατά την αναζήτηση βιβλιογραφικών δεδομένων σε ψηφιακές βιβλιοθήκες. Έχοντας τη δυνατότητα για ομαδοποίηση, πολύ εύκολα είναι δυνατόν να βρεθούν κείμενα σχετικά με κάποιο αρχικό, δηλαδή, αντικείμενα που ανήκουν στην ίδια ομάδα, χωρίς να είναι απαραίτητο αυτά να συνδέονται μεταξύ τους ούτε με άμεσες αναφορές, αλλά ούτε με κοινές λέξεις-κλειδιά. Υπάρχουν και άλλες εφαρμογές της ομαδοποίησης για τους χώρους που μελετούμε, οι οποίες περιγράφονται στο Κεφάλαιο 5.

Όπως αναφέραμε και πριν, τα δεδομένα με τα οποία ασχολούμαστε αναπαριστώνται με γράφους. Έχοντας τέτοιας μορφής οντότητες είναι δύσκολη η αναπαράστασή τους σε ένα χώρο λίγων διαστάσεων, ώστε να εφαρμόσουμε τους

[12] `http://www.akamai.com/`

παραδοσιακούς αλγορίθμους ομαδοποίησης. Ουσιαστικά χρειαζόμαστε τόσες διαστάσεις όσες είναι και τα αντικείμενά μας. Η ομαδοποίηση τέτοιων δεδομένων, τα οποία δεν μπορούν να αναπαρασταθούν σε ένα δισδιάστατο ή πολυδιάστατο χώρο έχει απασχολήσει για πολλά χρόνια την ερευνητική κοινότητα. Μέθοδοι που μεταφέρουν τα δεδομένα σε ένα χώρο με λιγότερες διαστάσεις, οδηγούν σε απώλεια πληροφορίας από την αρχική και προφανώς δεν είναι δυνατό να δώσουν αποδεκτά αποτελέσματα.

Μελετώντας την σχετική βιβλιογραφία βρίσκουμε 4 κατηγορίες μεθόδων για ομαδοποίηση σε γράφους: βιβλιομετρικές, φασματικές, μέγιστης ροής και γραφο-θεωρητικές. Η πρώτη από αυτές μπορεί να εφαρμοσθεί μόνο σε βιβλιογραφικά δεδομένα διότι λαμβάνει υπ' όψη τα χαρακτηριστικά των βιβλιογραφικών δεδομένων. Από την άλλη, οι φασματικές μέθοδοι έχουν ορισθεί για την περίπτωση του Παγκόσμιου Ιστού και επίσης χρειάζονται επιπλέον δεδομένα όπως αναλύεται στο Κεφάλαιο 5. Άρα, οι δυο πρώτες οικογένειες αλγορίθμων δεν είναι γενικές για γράφους. Οι μέθοδοι μέγιστης ροής και γραφο-θεωρητικές μπορούν να εφαρμοσθούν σε έναν οποιοδήποτε γράφο διότι απαιτούν για τη λειτουργία τους μόνο τα χαρακτηριστικά-ιδιότητες του γράφου και καμιά επιπλέον πληροφορία. Οι μέθοδοι μέγιστης ροής βασίζονται στην ιδιότητα για ροές σε δίκτυα που ορίσθηκε το 1961 από τους Gomory και Hu [37]. Με την οικογένεια αυτή των αλγορίθμων έχουν ασχοληθεί μεταξύ άλλων οι Goldberg, Flake, Tarjan και Tsioutsiouklis [36, 28, 27, 29]. Το κύριο πρόβλημα αυτής της οικογένειας, αν και οι αλγόριθμοι βασίζονται στις ιδιότητες γράφων, είναι το υπολογιστικό κόστος. Τέλος, ίσως οι εξυπνότεροι αλγόριθμοι είναι της οικογένειας των γραφο-θεωρητικών μεθόδων με περισσότερο ενδιαφέρουσα την εργασία των Girvan και Newman [35]. Η μέθοδος που προτείνουμε στο Κεφάλαιο 5 ανήκει στην οικογένεια των γραφο-θεωρητικών μεθόδων και βασίζεται στις ιδιότητες που ορίζονται στο [35].

1.5 Ανάπτυξη Υποδομής

Σε αυτό το κεφάλαιο θα παρουσιάσουμε σε συντομία την υποδομή που αναπτύξαμε, προκειμένου να υλοποιηθούν και να δοκιμασθούν οι μέθοδοι και αλγόριθμοι που παρουσιάζονται στις επόμενες σελίδες της παρούσης διατριβής.

1.5.1 Σύστημα SCEAS

Για τη μελέτη των διάφορων μεθόδων που προτάθηκαν και διερευνήθηκαν για τη Βιβλιομετρία, αναπτύχθηκε ένα σύστημα Ψηφιακής Βιβλιοθήκης, το οποίο ονομάσθηκε SCEAS (Scientific Collection Evaluator using Advanced Scoring). Το σύστημα είναι web-based, δηλαδή προσβάσιμο μέσω του Παγκόσμιου Ιστού[13]. Το σύστημα αυτό αποτελεί ένα αυτόνομο σύστημα που κάνει εισαγωγή βιβλιογραφικών δεδομένων από το διαδίκτυο[14], υπολογίζει τα κριτήρια που έχουν ορισθεί και κάνει παρουσίαση των αποτελεσμάτων στον Παγκόσμιο Ιστό. Το εν λόγω σύστημα επεκτάθηκε και χρησιμοποιήθηκε σε όλες τις μελέτες σε σχέση με αξιολόγηση-κατάταξη δημοσιεύσεων, συγγραφέων, συνεδρίων και περιοδικών.

Αναλυτικότερα, το σύστημά μας εισάγει εγγραφές XML από την ψηφιακή βιβλιοθήκη DBLP σε ένα σύστημα διαχείρισης βάσεων δεδομένων MySQL. Χρησιμοποιήσαμε το συγκεκριμένο λογισμικό, επειδή είναι ελαφρύ, γρήγορο και ταιριάζει στις ανάγκες μας, καθώς οι συναλλαγές που αποστέλλονται στον εξυπηρετητή είναι κυρίως αναγνώσεις και όχι ανανεώσεις. Το σύστημα SCEAS είναι διαθέσιμο στο Διαδίκτυο και επομένως ο χρήστης είναι εύκολο να το προσπελάσει, να στείλει ερωτήματα, να λάβει απαντήσεις και να εξάγει χρήσιμες πληροφορίες. Στο μοντέλο που υλοποιήσαμε οι κύριες οντότητες είναι:

Δημοσιεύσεις που μπορεί να είναι άρθρα, δημοσιεύσεις συνεδρίων κ.τ.λ.

Συλλογές όπως συνέδρια, βιβλία, περιοδικά κ.τ.λ.

Άτομα όπως συγγραφείς (authors) ή επιμελητές εκδόσεων (editors).

Κάθε δημοσίευση ανήκει σε μία συλλογή (ή σε περισσότερες, π.χ. οι δημοσιεύσεις των συνεδρίων ανήκουν σε μία συλλογή του συνεδρίου και σε μία συλλογή των καταγεγραμμένων πρακτικών αυτού). Μία συλλογή μπορεί να είναι τμήμα μιας άλλης, π.χ η VLDB'97 είναι μία συλλογή και είναι μέρος της συλλογής VLDB. Τα άτομα μπορεί να σχετίζονται με τις δημοσιεύσεις ως συγγραφείς ή με τις συλλογές ως επιμελητές εκδόσεων. Τέλος, οι δημοσιεύσεις μπορεί να σχετίζονται μεταξύ τους με τη σχέση "αναφορά".

Βασιζόμενοι στην ψηφιακή βιβλιοθήκη DBLP, κατασκευάσαμε το γράφο αναφορών της συλλογής, η οποία περιλαμβάνει άρθρα σε περιοδικά όπως και δημοσιεύσεις σε συνέδρια. Χρησιμοποιώντας αυτό το γράφο μπορούμε να κατασκευάσουμε συγκεντρωτικούς Γράφους Αναφορών, το Γράφο Αναφορών Συνεδρίων,

[13] `http://delab.csd.auth.gr/sceas`

[14] αυτή τη στιγμή μόνο από το σύστημα DBLP - Data Bases and Logic Programming του Πανεπιστημίου του Trier

Γράφο Αναφορών Περιοδικών και Γράφο Αναφορών Συγγραφέων. Με τον ίδιο τρόπο, μπορεί να χρησιμοποιηθεί οποιοσδήποτε άλλος τύπος σημασιολογικής ομαδοποίησης των δημοσιεύσεων για την δημιουργία ανάλογων γράφων αναφορών (π.χ Γράφος Αναφορών Βιβλίων). Σε ένα συγκεντρωτικό Γράφο Αναφορών, το βάρος μιας ακμής από έναν κόμβο i σε έναν κόμβο j αναπαριστά το πλήθος των αναφορών από όλες τις δημοσιεύσεις του πρώτου κόμβου σε όλες τις δημοσιεύσεις του τελευταίου.

1.5.2 Συλλογή Ιστογράφων

Για τη μελέτη των διάφορων μεθόδων που προτάθηκαν και διερευνήθηκαν για τον Παγκόσμιο Ιστό, συλλέχθηκαν δεδομένα χρησιμοποιώντας την τεχνολογία για Robots του w3c[15] καθώς και με τη βοήθεια της μηχανής αναζήτησης της Google. Στο Κεφάλαιο 4 περιγράφεται αναλυτικά η διαδικασία συλλογής ιστογράφων, καθώς και ποιά τμήματα του Παγκόσμιου Ιστού αφορούν. Σε αυτό το σημείο ας αναφέρουμε ότι έχουν συλλεχθεί μετα-δεδομένα για 76M σελίδες του Παγκόσμιου Ιστού με 405M συνδέσμους μεταξύ τους. Σε αυτές τις σελίδες δεν περιλαμβάνουμε εικόνες ή άλλου τύπου αρχεία, παρά μόνο αρχεία τύπου html με υπερκείμενο (hypertext).

1.6 Συνεισφορά σε Δημοσιεύσεις

Το Κεφάλαιο 2 εμπεριέχει το υλικό από τις εργασίες [86, 84]. Το Κεφάλαιο 3 εμπεριέχει υλικό από τις εργασίες [83, 82, 85]. Στο Κεφάλαιο 5 περιλαμβάνεται υλικό από την εργασία [80]. Τέλος, στο Κεφάλαιο 6 περιέχει υλικό από την εργασία [81]. Ο πλήρης κατάλογος των ερευνητικών εργασιών βρίσκεται στο Παράρτημα Α της παρούσας διατριβής.

[15] `http://www.w3c.org/Robot/`.

ΚΕΦΑΛΑΙΟ 2

Κατάταξη Συνεδρίων

Περιεχόμενα

2.1 Εισαγωγή

Σήμερα υπάρχουν δύο μεγάλα συστήματα που πραγματοποιούν Ανάλυση Αναφορών : το Science Citation Index (SCI) του Ινστιτούτου για την Επιστημονική Πληροφορία (ISI) και το CiteSeer του Ινστιτούτου Έρευνας της NEC. Η ιδέα για την πραγματοποίηση της ανάλυσης αναφορών αναπτύχθηκε στις αρχές της δεκαετίας του '60. Αργότερα, το 1972 άρχισε να χρησιμοποιείται από το SCI για την αξιολόγηση περιοδικών που κάλυπταν πολλούς επιστημονικούς τομείς, μεταξύ των οποίων και η πληροφορική. Ο *Παράγοντας Αντικτύπου ISI* [31, 32] ήταν η βασική μετρική μέθοδος που χρησιμοποιήθηκε από το SCI για την αξιολόγηση και κατάταξη περιοδικών, μία απαραίτητη εργασία ώστε να ληφθούν αποφάσεις για τη διατήρηση θέσεων, χρηματοδοτήσεων, επιπέδων μισθών κ.ο.κ. Αυτή η έννοια υπολογίζεται για κάθε έτος κατά τη διάρκεια του χρόνου ζωής ενός περιοδικού βασιζόμενη στον αριθμό των αναφορών που έγιναν την εκάστοτε χρονιά

σε εργασίες που δημοσιεύτηκαν σε αυτό τα προηγούμενα k χρόνια, όπου το k είναι συνήθως ίσο με 2 ή 5. (Ωστόσο, ο υπολογισμός μπορεί να γίνει για κάθε τιμή του k, αν αυτό είναι απαραίτητο για κάποια ειδική επιστημονική περιοχή ή για στατιστικούς λόγους.) Από την άλλη μεριά, το σύστημα CiteSeer είναι ένα σύγχρονο σύστημα και κατασκευάζει το γράφο αναφορών από δημοσιεύσεις που αποκτήθηκαν από τον Παγκόσμιο Ιστό [58]. Το CiteSeer επίσης βασίζεται στον Παράγοντα Αντικτύπου ISI για την αξιολόγηση συνεδρίων και περιοδικών [38] – αν και δεν είναι αυτός ο κύριος στόχος του.

Η Ανάλυση Αναφορών βασίζεται στην έννοια των γράφων αναφορών που αναπαρισθούν τις δημοσιεύσεις ως κόμβους, ενώ μία ακμή από τον κόμβο x στον κόμβο y αναπαριστά μία αναφορά από τη δημοσίευση x στη δημοσίευση y. Οι γράφοι αναφορών μπορούν να χρησιμοποιηθούν για να αποκομίσουμε χρήσιμες στατιστικές πληροφορίες σχετικά με την αξιολόγηση και την κατάταξη πολλών οντοτήτων, όπως συγγραφείς, δημοσιεύσεις σε επιστημονικά συνέδρια και περιοδικά, όπως και συνέδρια και περιοδικά υπό τη μορφή επιστημονικών συλλογών.

Συγκεκριμένα, η ανάλυση ενός γράφου αναφορών είναι παρόμοια με την ανάλυση ενός Ιστογράφου (webgraph). Ο Ιστογράφος αναπαριστά τμήμα του Παγκόσμιου Ιστού, και κάθε κόμβος αντιστοιχεί σε μια σελίδα ενώ οι ακμές αντιστοιχούν σε υπερσυνδέσμους. Αξιοσημείωτος είναι ο αλγόριθμος `PageRank` των Brin και Page [16], ο οποίος χρησιμοποιείται από τη μηχανή αναζήτησης Google. Αυτός ο αλγόριθμος υπολογίζει τη βαθμολογία μιας σελίδας ως το άθροισμα των βαθμολογιών των σελίδων που κάνουν αναφορά σε αυτήν. Επομένως, αξιολογεί τις σελίδες που επιστρέφονται στο χρήστη ανάλογα με τη σχετικότητά τους ως προς το ερώτημα του χρήστη. Έχει εξαχθεί ότι η στατιστική κατανομή της μετρικής μεθόδου του `PageRank` ακολουθεί το γνωστό αντίστροφο πολυωνυμικό νόμο που έχει αναφερθεί για βαθμούς ιστοσελίδων [18]. Επίσης, έχει γίνει περαιτέρω ανάλυση στον τύπο του `PageRank` [77].

Εκτός από την αξιολόγηση, μπορούν να εφαρμοσθούν κι άλλες λειτουργίες στους γράφους αναφορών χρησιμοποιώντας τεχνικές θεωρίας των γράφων και εξόρυξης δεδομένων [3]. Για παράδειγμα, ας υποθέσουμε ένα σύνολο από επιστημονικές συλλογές, σχετικά βιβλία, συνέδρια, περιοδικά ή/και τεχνικές αναφορές - αναφερόμενα σε μία συγκεκριμένη περιοχή - μπορούν να κατηγοριοποιηθούν χρησιμοποιώντας κάποια ομαδοποίηση. Με ανάλογο τρόπο, οι συγγραφείς μπορούν να τοποθετηθούν σε ομάδες με σκοπό να βρεθούν και να ιδρυθούν οι κοινότητές τους, όπως π.χ. οι συγγραφείς που συνεργάζονται και αλληλοαναφέρονται, και αντίστοιχα να βρούμε τα *εστιακά σημεία* (hubs) και τις *αυθεντίες* (authorities), όπως π.χ. τις ομάδες των συγγραφέων που αντίστοιχα αναφέρουν και

αναφέρονται περισσότερο. Δύο σημαντικές εργασίες σε αυτή την περιοχή είναι ο αλγόριθμος HITS (Hyperlink Induced Topic Search) του Kleinberg [52, 53], ο οποίος υπολογίζει μία βαθμολογία με βάρη για τις προηγούμενες έννοιες. Μία ακόμη μελέτη για τον Παγκόσμιο Ιστό ως γράφο και τις έννοιες των εστιακών σημείων/αυθεντιών έχει εμφανισθεί στο [65].

Η δομή του παρόντος κεφαλαίου έχει ως εξής: Στην επόμενη παράγραφο κάνουμε μία ανασκόπηση των συστημάτων SCI και CiteSeer και συγκρίνουμε τα πλεονεκτήματα και τα μειονεκτήματά τους. Επίσης εξετάζουμε τη σχετική βιβλιογραφία πολλών προσπαθειών με στόχο την ανάλυση αναφορών και την αξιολόγηση συγκεκριμένων επιστημονικών συλλογών. Στην Παράγραφο 2.3 ερευνούμε νέες εναλλακτικές ιδέες εκτός του Παράγοντα Αντικτύπου, ώστε να παρέχουμε μία καινοτόμα προσέγγιση που στοχεύει στην αξιολόγηση και την κατάταξη επιστημονικών περιοδικών και συνεδρίων. Επίσης, παρουσιάζουμε τις βασικές λειτουργίες ενός συστήματος βασιζόμενο στον Παγκόσμιο Ιστό που ονομάζεται *Αξιολογητής Επιστημονικών Συλλογών με χρήση Προηγμένης Βαθμολογίας* (Scientific Collection Evaluator by using Advanced Scoring - SCEAS). Το σύστημά μας έχει κατασκευασθεί εισάγοντας δεδομένα από τον ιστοχώρο για Βάσεις Δεδομένων και Λογικό Προγραμματισμό (Data Bases and Logic Programming - DBLP) του Πανεπιστημίου του Trier. Το σύστημα αυτό, χρησιμοποιώντας τις νέες μετρικές μεθόδους αναφορών, αναδεικνύεται σε χρήσιμο εργαλείο για την αξιολόγηση επιστημονικών συλλογών, όπως συνέδρια ή/και περιοδικά. Κάτω από αυτό το πρίσμα, στην Παράγραφο 2.4, παραθέτονται κάποια πρώτα σχόλια, π.χ. η αξιολόγηση των συνεδρίων που σχετίζονται με την περιοχή των Βάσεων Δεδομένων. Η τελευταία παράγραφος είναι συμπερασματική του παρόντος κεφαλαίου.

2.2 Τα Σημαντικότερα Συστήματα Ανάλυσης Αναφορών

Όπως αναφέρθηκε, σήμερα υπάρχουν δύο μεγάλα συστήματα που πραγματοποιούν ανάλυση αναφορών: το SCI και το CiteSeer. Εδώ θα εξετάσουμε βαθύτερα αυτά τα συστήματα ώστε να διαπιστώσουμε τα *αδύνατά* τους σημεία και να υποκινήσουμε την παρούσα έρευνα. Προηγουμένως όμως είναι σημαντικό να σημειώσουμε ότι το SCI έχει υπηρετήσει ολόκληρη την ακαδημαϊκή κοινότητα για πολλές δεκαετίες παρέχοντας χρήσιμες πληροφορίες, ελλείψει εναλλακτικής λύσης. Ωστόσο σήμερα τα μειονεκτήματα και οι περιορισμοί αυτού του συστήματος είναι εμφανή. Για παράδειγμα, τα κύρια μειονεκτήματα του SCI είναι:

1. Κάθε επιστημονικός τομέας είναι χωρισμένος σε συγκεκριμένες περιοχές, οι οποίες παραμένουν στατικές όλα αυτά τα χρόνια και δεν αντανακλούν την επιστημονική εξέλιξη και πιο συγκεκριμένα τη ραγδαία εξέλιξη της Πληροφορικής.

2. Σε κάθε περιοχή επιλέγεται μόνο ένα σύνολο περιοδικών για αξιολόγηση. Επομένως, η αντιπροσωπευτική αξία των επιλεγμένων περιοδικών είναι υπό αμφισβήτηση.

3. Αν και κάθε σύνολο είναι δυναμικό και ανανεώνεται περιοδικά, αυτή η ανανέωση γίνεται με έναν υποκειμενικό ή εμπορικό τρόπο, που μπορεί επίσης να πυροδοτήσει ερωτήσεις του τύπου πότε, γιατί, πώς και από ποιον.

4. Σε μερικές περιπτώσεις, άσχετα περιοδικά (π.χ. τεχνικά έναντι δημοφιλών-popular) ομαδοποιούνται σε μία συγκεκριμένη περιοχή οδηγώντας σε εσφαλμένα αποτελέσματα.

5. Επιστημονικά συνέδρια, βιβλία και τεχνικές αναφορές δεν λαμβάνονται υπ'όψη.

6. Κατασκευάζεται χειροκίνητα (δεν είναι αυτοματοποιημένο) τουλάχιστον μέχρι πρόσφατα, και επομένως είναι ένα ακριβό σύστημα για να κατασκευασθεί και να συντηρηθεί.

7. Δεν είναι δωρεάν στην πλήρη του έκδοση ούτε για βιβλιοθήκες αλλά ούτε και για μεμονωμένους χρήστες.

Από την άλλη πλευρά, το CiteSeer είναι ένα σύγχρονο σύστημα που κατασκευάζει το γράφο αναφορών από δημοσιεύσεις που αποκτώνται από τον Παγκόσμιο Ιστό [58]. Πιο συγκεκριμένα, πρόκειται για ένα αυτόνομο σύστημα το οποίο συλλέγει δημοσιεύσεις σχετικές με την Πληροφορική, σαρώνοντας τον Παγκόσμιο ιστό. Κατόπιν, ανάλογα με τον τρόπο αποθήκευσής της δημοσίευσης (δηλαδή, postscript ή pdf), το σύστημα ανιχνεύει και εξάγει τις βιβλιογραφικές πληροφορίες τους (π.χ. τίτλο, συγγραφείς κ.τ.λ.) όπως επίσης και τις συμπεριλαμβανόμενες αναφορές, ώστε να κατασκευάσει τον αντίστοιχο γράφο αναφορών [57]. Τα πλεονεκτήματα του συστήματος CiteSeer είναι:

1. Είναι αυτοματοποιημένο και διαφανές, επομένως είναι αντικειμενικό καθώς η παρέμβαση του ανθρώπου είναι περιορισμένη.

2. Λαμβάνει υπ'όψη του όλα τα είδη επιστημονικών δημοσιεύσεων, όπως βιβλία και τεχνικές αναφορές. Υπό τις παρούσες συνθήκες, αυτό είναι πολύ σημαντικό καθώς αντιμετωπίζουμε μία βαθμιαία και σταδιακή αύξηση της ποσότητας επιστημονικών πληροφοριών υψηλής ποιότητας που διαχέονται μέσω συνεδρίων, συμποσίων και λευκών τεχνικών δημοσιεύσεων.

3. Είναι ευαίσθητο στο γεγονός ότι υπάρχουν μερικά συνέδρια με υψηλό ανταγωνισμό, των οποίων το ποσοστό αποδοχής είναι πολύ υψηλότερο από αυτό πολλών σχετικών περιοδικών. Για παράδειγμα, στην περιοχή της Διαχείρισης Δεδομένων (Management of Data), τα συνέδρια SIGMOD (ACM Conference of the Special Interest Group on Management of Data) και VLDB (International Conference on Very Large Databases), όπου κατά τη διάρκεια των τελευταίων 20 ετών παρατηρούμε ένα τυπικό ποσοστό αποδοχής 1:5 έως 1:7.

4. Χρησιμοποιεί το γράφο αναφορών για να εκτελέσει πολλές προηγμένες λειτουργίες, όπως την κατάταξη αποτελεσμάτων αναζήτησης σελίδων (βασιζόμενο στην τιμή έσω-βαθμού (in-degree) της κάθε εύρεσης), ερευνά σχετικές δημοσιεύσεις (βασιζόμενο στην ανάλυση συναναφορών – co-citation analysis), κ.τ.λ.

5. Είναι προσβάσιμο δωρεάν μέσω του Διαδικτύου, βοηθώντας έτσι την ερευνητική και διοικητική εργασία όλης της ακαδημαϊκής κοινότητας της Πληροφορικής.

Ωστόσο, οι περιορισμοί του CiteSeer είναι:

1. Αναφέρεται μόνο στην Πληροφορική και επομένως δεν είναι χρήσιμο πέραν αυτής στη γενικότερη ακαδημαϊκή κοινότητα.

2. Στην πραγματικότητα δεν εστιάζει στην αξιολόγηση/κατάταξη συνεδρίων ή περιοδικών. Στο CiteSeer[1] υπάρχει μόνο μία κατάταξη, που

   (a) περιέχει δεδομένα μόνο από την ψηφιακή βιβλιοθήκη DBLP,

   (b) κάνει μίξη συνεδρίων και περιοδικών, και

   (c) κάνει κοινή ομαδοποίηση διακριτών επιστημονικών περιοχών.

3. Τέλος, βασίζεται στην έννοια του Παράγοντα Αντικτύπου, που αν και για μεγάλο χρονικό διάστημα έπαιξε πολύ σημαντικό ρόλο στην αξιολόγηση

[1] `http://citeseer.nj.nec.com/impact.html`

των περιοδικών (και κατ'επέκταση των ακαδημαϊκών συγγραφέων), είναι αρκετά δύσκαμπτη μέθοδος και δεν μπορεί να εφαρμοσθεί σε βαθύτερη ποιοτική ανάλυση.

Το παρόν κεφάλαιο έχει ως κίνητρο από το τελευταίο σημείο. Για να καταλάβουμε τους περιορισμούς του Παράγοντα Αντικτύπου, είναι απαραίτητο να παρατηρήσουμε ότι αυτή η έννοια δεν μπορεί να εφαρμοσθεί στην περίπτωση που θέλουμε να κατατάξουμε/αξιολογήσουμε συνέδρια ενός συγκεκριμένου έτους. Πιο συγκεκριμένα, ένα περιοδικό εκδίδεται πολλές φορές κατά τη διάρκεια ενός χρόνου. Από την άλλη μεριά, ένα συνέδριο οργανώνεται μία φορά το χρόνο ή και σπανιότερα. Σε αυτή την περίπτωση, ποιά θα ήταν η τιμή k του Παράγοντα Αντικτύπου για ένα συνέδριο; Χρησιμοποιώντας την τιμή k, ουσιαστικά αξιολογούμε το συνέδριο για την περίοδο των k προηγούμενων ετών. Όμως πως θα μπορούσαμε να αξιολογήσουμε όλα τα συνέδρια που πραγματοποιήθηκαν για παράδειγμα το 1996; Επομένως, είναι προφανές ότι δεν είναι ασφαλής μέθοδος να χρησιμοποιήσουμε τον Παράγοντα Αντικτύπου για την αξιολόγηση συνεδρίων. Επιπρόσθετα, κάποιος θα μπορούσε να διαφωνήσει με την *επίπεδη* φύση του Παράγοντα Αντικτύπου. Για παράδειγμα,

- Είναι δίκαιο να μετρήσουμε μία αναφορά από τον Καθηγητή "Πολύ-γνωστός" ως ισοδύναμη με μία αναφορά από τον Καθηγητή"Άγνωστος";
- Είναι δίκαιο να μετρήσουμε μία αναφορά από το περιοδικό "Το καλύτερο" ως ισοδύναμη με μία αναφορά από το περιοδικό "Το χειρότερο"; Η τελικά
- Είναι δίκαιο να μετρήσουμε μία αναφορά από τη δημοσίευση "Η καλύτερη" ως ισοδύναμη με μία αναφορά από τη δημοσίευση "Η χειρότερη";

Από αυτές τις απλές ερωτήσεις, είναι φανερό ότι είναι απαραίτητο να ενσωματωθεί κάποιου είδους τεχνική με βάρη ώστε να απαντηθούν τέτοιες ερωτήσεις. Στο κεφάλαιο αυτό θα ερευνήσουμε μερικές νέες ιδέες για την αξιολόγηση επιστημονικών συλλογών και θα προσπαθήσουμε να τοποθετήσουμε το έργο της ανάλυσης αναφορών και της αξιολόγησης περιοδικών ή συνεδρίων σε μία πιο γενικευμένη προοπτική.

2.3 Επισκόπηση Βιβλιογραφίας

Εκτός από το SCI και το CiteSeer, υπάρχουν διάφορες άλλες ψηφιακές βιβλιοθήκες και συστήματα ευρετηρίων αναφορών, τα οποία εκτελούν κάποιου είδους

Ανάλυση Αναφορών. Όπως έχει ήδη αναφερθεί, ο ιστοχώρος DBLP, που διατηρείται από το 1993 από τον Michael Ley στο Πανεπιστήμιο του Trier είναι μία πλούσια ψηφιακή βιβλιοθήκη, που εστιάζει στην Πληροφορική και ειδικότερα στις περιοχές των Βάσεων Δεδομένων και του Λογικού Προγραμματισμού [60]. Πιο συγκεκριμένα, το Μάρτιο του 2003, ο ιστοχώρος DBLP περιείχε βιβλιογραφικά δεδομένα για περίπου 240.000 συγγραφείς, 1250 συνέδρια, 300 περιοδικά και 360.000 δημοσιεύσεις, άρθρα ή βιβλία, με συνδέσμους σε προσωπικές σελίδες, ερευνητικές ομάδες, εκδοτικούς οίκους, κ.τ.λ. Η πλοήγηση μέσω των περιεχομένων του DBLP καθιστά την έρευνα πληροφοριών μία εύκολη εργασία για τους ακαδημαϊκούς, τους ερευνητές και τους επαγγελματίες που εργάζονται στις περιοχές των Βάσεων Δεδομένων και του Λογικού Προγραμματισμού. Οι βασικές λειτουργίες του DBLP είναι η έρευνα με βάση το όνομα ενός συγγραφέα, ενός συνεδρίου, ενός περιοδικού ή μίας τεχνικής λέξης-κλειδί. Επιπλέον, για ένα μεγάλο τμήμα του ευρετηρίου δημοσιεύσεων παρέχει πλήρη ανάκτηση κειμένου και λίστας αναφορών. Αν και το DBLP είναι κυρίως ένα σύστημα ευρετηρίου και αναζήτησης, είναι ενδιαφέρον να σημειωθεί ότι παρέχει μία κατάταξη από τις πλέον αναφερόμενες δημοσιεύσεις στις δύο συγκεκριμένες περιοχές της Πληροφορικής. Ωστόσο, αυτή η κατάταξη βασίζεται μόνο σε δεδομένα του DBLP, δηλαδή, στις λίστες αναφορών ανά δημοσίευση, οι οποίες δεν είναι πλήρεις.

Πρόσφατα, έκανε την εμφάνισή του στη βιβλιογραφία ένα άλλο πρωτότυπο ευρετήριο [14], το οποίο ονομάζεται Rosetta, και είναι ένα σύστημα ψηφιακής βιβλιοθήκης για την επιστημονική βιβλιογραφία που επίσης σχετίζεται με την Πληροφορική. Το Rosetta ευρετηριάζει ερευνητικά άρθρα βασιζόμενο στον τρόπο περιγραφής τους όταν αναφέρθηκαν σε άλλα κείμενα. Η συνοπτική περιγραφή που συμβαίνει στις αναφορές είναι παρόμοια με τα σύντομα ερωτήματα που σχηματίζουν οι χρήστες όταν ψάχνουν για πληροφορίες. Το Rosetta παρέχει μία διεπιφάνεια χρήσης, η οποία παρέχει στους χρήστες ένα αυτόματα κατασκευασμένο κατάλογο από το χώρο πληροφοριών γύρω από το ερώτημα. Αναφέρεται ότι η συλλογή Rosetta περιέχει περισσότερο από 37.000 κείμενα καταχωρισμένα σε ευρετήρια χρησιμοποιώντας μόλις 450.000 αναφορές.

Ένα άλλο ενδιαφέρον σύστημα είναι το AuthorLink από τους Lin, White, και Buzydlowski [62], το οποίο είναι ένα εικονοποιημένο πρωτότυπο με σκοπό την βελτίωση της αναζήτησης συγγραφέων. Αυτό επιτυγχάνεται με την ανάλυση συν-αναφορών ανά συγγραφέα. Πιο συγκεκριμένα, δεδομένου ενός ερωτήματος για ένα συγκεκριμένο συγγραφέα, το σύστημα κατασκευάζει αλληλεπιδραστικούς χάρτες συγγραφέων σε πραγματικό χρόνο από μία βάση δεδομένων με 1.26 εκατομμύρια εγγραφές σχετικές με τις Τέχνες και τα Γράμματα, η οποία παρέχεται

από το Ινστιτούτο ISI. Αυτοί οι χάρτες περιέχουν τους 24 συγγραφείς που είναι περισσότερο συν-αναφερόμενοι (co-cited) με το υπό έρευνα όνομα, καθώς και μερικά δεδομένα σχετικά με μετρήσεις. Ο χρήστης διαλέγοντας ένα από τα 24 ονόματα μπορεί να προχωρήσει σε νέους χάρτες που κατασκευάζονται δυναμικά. Στην πραγματικότητα, αυτό το μέσο βοηθά σε πολλές περιπτώσεις έρευνας σε μία συγκεκριμένη, περιορισμένη περιοχή.

Το PubSearch είναι ένα σύστημα που αναπτύχθηκε από τους He και Hui [42] και στοχεύει στην απεικόνιση της ανάλυσης συν-αναφερόμενων συγγραφέων, χρησιμοποιώντας τη μεθοδολογία της εξόρυξης δεδομένων. Οι δημιουργοί του χρησιμοποιούν μία τεχνική παρόμοια με του CiteSeer, για να συλλέξουν βιβλιογραφικά δεδομένα σαρώνοντας τον Παγκόσμιο Ιστό. Οι εγγραφές που συλλέγονται τοποθετούνται σε μία αποθήκη δεδομένων, όπου πραγματοποιείται μία συσωρευτική ιεραρχική ομαδοποίηση (agglomerative hierarhical clustering), ώστε να κατασκευασθούν οι χάρτες των συγγραφέων, παρουσιάζοντας συγγραφείς με παρόμοια ενδιαφέροντα με το υπό αναζήτηση όνομα. Αυτό το σύστημα έχει ελεγχθεί πειραματικά με δεδομένα από το ISI Social Science Citation Index (SSCI). Πιο συγκεκριμένα, για την αξιολόγηση του συστήματος χρησιμοποιήθηκαν 1466 δημοσιεύσεις σχετικές με την Ανάκτηση Πληροφοριών, οι οποίες εμφανίσθηκαν κατά την περίοδο 1987-1997 σε 367 περιοδικά, με 44.836 αναφορές.

Τέλος, ένα σύστημα ανάλυσης συμφραζομένων παρουσιάσθηκε από τους Ding, Chowdhury και Foo [20]. Οι συγγραφείς αυτής της εργασίας συνέλεξαν από το SCI και το SSCI 2012 δημοσιεύσεις σχετικές με την Ανάκτηση Πληροφοριών, οι οποίες εμφανίσθηκαν κατά την περίοδο 1987-1997 και εξήγαγαν 193 λέξεις-κλειδιά, με 5.09 λέξεις-κλειδιά ανά δημοσίευση, ώστε να πραγματοποιήσουν ανάλυση συμφραζομένων και να αποκαλύψουν πρότυπα της εξέλιξης της συγκεκριμένης περιοχής με το χρόνο.

Όπως αναφέρθηκε, ούτε λίγο ούτε πολύ ο σκοπός των προηγούμενων συστημάτων είναι ο ευρετηριασμός, η Ανάλυση Αναφορών και η απεικόνιση, ωστόσο, ο σκοπός τους δεν είναι η αξιολόγηση. Σε σχέση με την αξιολόγηση συναντούμε πολλές προσπάθειες στη βιβλιογραφία, οι οποίες εκτείνονται σε πολλές περιοχές, όμως πέραν της Πληροφορικής.

Για παράδειγμα, μία μελέτη που βασίζεται στην Ανάλυση Αναφορών [5] αναφέρεται σε περιοδικά Μάρκετινγκ. Ουσιαστικά οι συγγραφείς πραγματοποίησαν δια χειρός μία εξαγωγή αναφορών από τα θέματα των ετών 1996-1997 από 49 περιοδικά μάρκετινγκ (26 από τα οποία δεν περιέχονται στο SSCI), όπου η συλλογή των 49 τίτλων βασιζόταν κυρίως σε μία δημοσκόπηση. Στο [45] αναφέρεται μία αξιολόγηση, που όμως δεν βασίζεται στην Ανάλυση Αναφορών, όπου οι συγγρα-

φείς αξιολόγησαν περιοδικά σχετικά με το Μάρκετινγκ σύμφωνα με την εισδοχή ερωτηθέντων (διδακτορικών έναντι μη διδακτορικών ινστιτούτων) βασιζόμενη και πάλι σε μία δημοσκόπηση.

Στο [91] το ενδιαφέρον των συγγραφέων ήταν γύρω από τα Λογιστικά. Απέκτησαν δεδομένα από το SSCI (πιο αναλυτικά, εξήγαγαν 351 άρθρα από 8 περιοδικά τα οποία εκδόθηκαν κατά την περίοδο 1992-1994 με 11.746 αναφορές), και έλαβαν υπ'όψη τους την έννοια του Παράγοντα Αντικτύπου και πρότειναν διαφοροποιήσεις αυτού βασιζόμενοι στη στατιστική κατανομή του αριθμού των αναφορών στο χρόνο. Στο [54] οι συγγραφείς ασκούν κριτική στη θεωρία και τη μεθοδολογία της αξιολόγησης περιοδικών, έχοντας κατά νου νομικά περιοδικά.

Οι Μυλωνόπουλος και Θεοχαράκης [67] επικεντρώθηκαν στην περιοχή των Πληροφοριακών Συστημάτων και πραγματοποίησαν μία δημοσκόπηση μέσω του Διαδικτύου για 87 περιοδικά με περίπου 1.000 ερωτηθέντες. Αν και αυτή η εργασία δεν είναι μία μελέτη αξιολόγησης βασιζόμενη στην Ανάλυση Αναφορών, παρουσιάζει ενδιαφέρον με την έννοια ότι πραγματοποιεί αξιολόγηση σύμφωνα με την αντίληψη των αναγνωστών (και των συγγραφέων) ως συνάρτηση της γεωγραφικής τοποθεσίας των ερωτηθέντων.

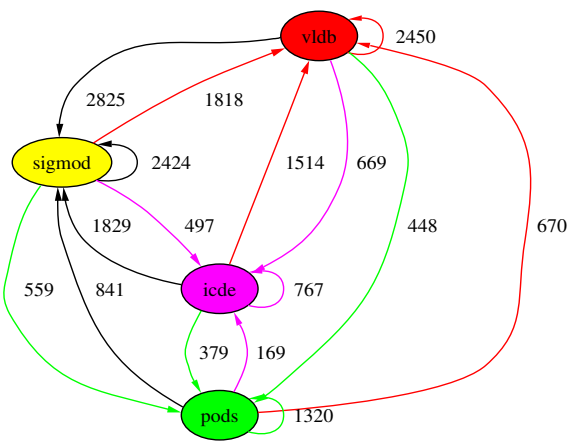
Έχοντας το ίδιο κίνητρο με τη δική μας εργασία, οι Keijnein και Groenendaal [51] εστιάζουν στην περιοχή των Πληροφοριακών Συστημάτων και προσπαθούν να αξιολογήσουν περιοδικά, συνέδρια και βιβλία, δηλαδή ένα μεγαλύτερο σύνολο εκδόσεων σε σχέση με την πρακτική του SCI. Χρησιμοποιώντας δειγματοληψία, εξερεύνηση και ταξινόμηση, ταξινομούν το σύνολο των δημοσιεύσεων σε έξι κατηγορίες αξιολόγησης σύμφωνα με τον αριθμό των αναφορών που έλαβαν. Ωστόσο, οι περιορισμοί αυτής της εργασίας είναι:

- περιορίζεται σε μία πολύ συγκεκριμένη περιοχή της Πληροφορικής,
- είναι μία χειροκίνητη μέθοδος, που χρησιμοποιεί ένα μικρό μόνο σύνολο δημοσιεύσεων (π.χ. μόνο 123 άρθρα περιοδικών και 82 άρθρα από πρακτικά συνεδρίων), οι οποίες συνολικά αναφέρουν 6.901 δημοσιεύσεις (3.128 άρθρα περιοδικών, 1.532 άρθρα από πρακτικά συνεδρίων, 1.577 βιβλία και 644 άλλες δημοσιεύσεις).
- όπως και το SCI αναμιγνύει ένα ευρύ φάσμα δημοσιεύσεων (π.χ. τεχνικών έναντι δημοφιλών).
- χρησιμοποιεί την ίδια έννοια με αυτήν του Παράγοντα Αντικτύπου ISI. Σημειώνεται ότι χρησιμοποιεί ένα παράθυρο άπειρου χρόνου (αντί για of k=2

ή 5 χρόνια που προτείνει το ISI) κατά τον υπολογισμό του Παράγοντα Αντικτύπου. Αυτό μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα, καθώς δεν συλλαμβάνει τη δυναμική φύση της εξέλιξης της επιστήμης και των προτιμήσεων των συγγραφέων, αφού ένα περιοδικό μπορεί να είχε καταταχθεί σε υψηλή θέση στο παρελθόν, αλλά να βρίσκεται πλέον χαμηλά ή και το αντίστροφο.

2.4 Το Σύστημα SCEAS

Ως πλατφόρμα εφαρμογής χρησιμοποιείται το σύστημα SCEAS, το οποίο Περιγράψαμε στην Παράγραφο 1.5.1. Βασιζόμενοι στα δεδομένα της ψηφιακής βιβλιοθήκης DBLP, κατασκευάσαμε το γράφο αναφορών της συλλογής, η οποία περιλαμβάνει περιοδικά όπως και δημοσιεύσεις συνεδρίων. Χρησιμοποιώντας αυτό το γράφο αποκομίσαμε δύο συγκεντρωτικούς Γράφους Αναφορών, το Γράφο Αναφορών Συνεδρίων και το Γράφο Αναφορών Περιοδικών. Με τον ίδιο τρόπο, μπορεί να χρησιμοποιηθεί οποιοσδήποτε άλλος τύπος σημασιολογικής ομαδοποίησης των δημοσιεύσεων για την αποκόμιση ανάλογων γράφων αναφορών (π.χ Γράφος Αναφορών Βιβλίων). Το Σχήμα 2.1 είναι ένα μικρό δείγμα του Γράφου Αναφορών Συνεδρίων. Αποτελείται από τέσσερις κόμβους, που παρισθούν τα συνέδρια SIGMOD, VLDB, PODS και ICDE όλων των ετών. Το βάρος της καμπύλης από τον κόμβο i στον κόμβο j αναπαριστά το πλήθος των αναφορών από όλες τις δημοσιεύσεις του πρώτου κόμβου σε όλες τις δημοσιεύσεις του τελευταίου.



Σχήμα 2.1. Γράφος αναφορών συνεδρίων.

Η αξιολόγηση των επιστημονικών συλλογών και πιο συγκεκριμένα η κατάταξη συνεδρίων, καθώς είναι η βασική μας προτεραιότητα, είναι ένα έργο για το οποίο προσπαθήσαμε να διερευνήσουμε εναλλακτικές μεθόδους. Η βασική ιδέα για την αξιολόγηση είναι ότι όλες οι αναφορές δεν θα πρέπει να έχουν το ίδιο βάρος. Για παράδειγμα, το βάρος θα πρέπει να εξαρτάται από δύο παράγοντες:

- την ποιότητα ενός συνεδρίου, όπου γίνεται αναφορά σε ένα άλλο συνέδριο, και
- την επιστημονική περιοχή των συνεδρίων, π.χ. αν ανήκουν στην ίδια περιοχή.

Επομένως υπάρχουν δύο έργα τα οποία πρέπει να επιτελεσθούν. Πρώτα, πρέπει να καθορίσουμε τις επιστημονικές περιοχές και μετά να εκτελέσουμε την κατάταξη. Πιο συγκεκριμένα, πρέπει να να επιτελέσουμε τα εξής:

1. Να **ομαδοποιήσουμε** (clustering) τα συνέδρια, βασιζόμενοι στο γράφο αναφορών συνεδρίων μετά από μία φάση προεπεξεργασίας, η οποία χρησιμοποιεί μερικές λέξεις-κλειδιά που εμφανίζονται στους τίτλους των συνεδρίων.
2. Να **εκκαθαρίσουμε** (cleansing), εφ'όσον τα δεδομένα από πολλά συνέδρια στην ψηφιακή βιβλιοθήκη DBLP δεν είναι ολοκληρωμένα. Επομένως, προσπαθούμε να αποκλείσουμε το υποσύνολο της συλλογής, το οποίο εισάγει θόρυβο στους αλγορίθμους μας. Και τέλος,
3. Να **αξιολογήσουμε/κατατάξουμε** (ranking) ξεχωριστά κάθε ομάδα συνεδρίων. Πραγματοποιήσαμε την αξιολόγηση λαμβάνοντας υπ'όψη ολόκληρη τη διάρκεια ζωής όλων των συνεδρίων και κάθε συγκεκριμένο έτος από κάθε συνέδριο. Η πρώτη μέθοδος δεν παράγει χρήσιμα αποτελέσματα από την άποψη των στατιστικών, αφού υπάρχουν πολλοί παράγοντες που επηρεάζουν την αξιολόγηση. Επομένως, εστιάσαμε στην τελευταία περίπτωση και για κάθε ξεχωριστό έτος πραγματοποιήσαμε κατατάξεις χρησιμοποιώντας:
   - Απλή Βαθμολογία (Plain Score),
   - Βαθμολογία Βαρών (Weighted Score)
   - τον Αντεστραμμένο Παράγοντα Αντίκτυπου (Inverted Impact Score),
   - Βαθμολογία Αντεστραμμένου Παράγοντα Αντικτύπου με Βάρη (Weighted Inverted Impact Score).

Όλες αυτές οι νέες έννοιες θα εξηγηθούν στη συνέχεια. Αξίζει να σημειωθεί ότι η αξιολόγηση πραγματοποιείται χρησιμοποιώντας διαφορετικούς αλγορίθμους, που επίσης θα παρουσιασθούν στη συνέχεια.

2.4.1 Ομαδοποίηση Συνεδρίων με Βάση τα Θέματα

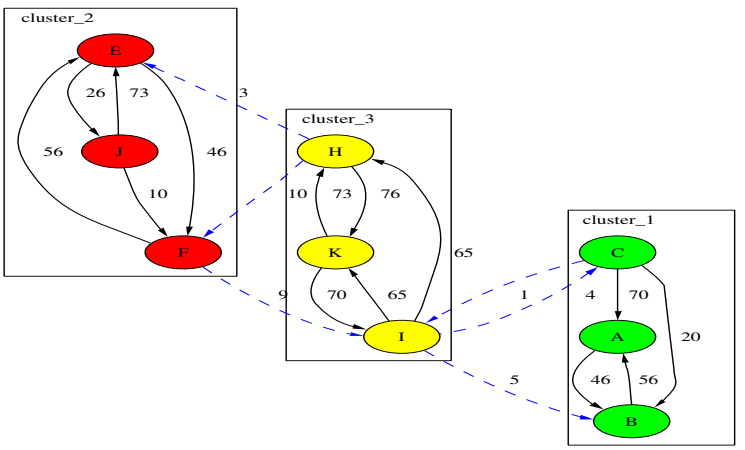
Βασιζόμενοι στο Γράφο Αναφορών Συνεδρίων (όπως στο παράδειγμα του Σχήματος 2.1), πρώτα εκτελέσαμε τη λειτουργία της ομαδοποίησης. Ως εργαλείο για την ομαδοποίηση των συνεδρίων χρησιμοποιήσαμε το hMetis [40, 48], το ηγετικό εργαλείο για την τμηματοποίηση υπεργράφων. Το hMetis χρησιμοποιείται με επιτυχία σε εφαρμογές που σχετίζονται με κυκλώματα VLSI, εξόρυξη δεδομένων και αριθμητική ανάλυση.

Το Σχήμα 2.2(α) παρουσιάζει ένα παράδειγμα ενός γράφου αναφορών συνεδρίων, όπου οι κόμβοι αναπαριστούν τα συνέδρια, ενώ οι ακμές αναπαριστούν τις αναφορές. Οι ακμές έχουν κατεύθυνση και βάρη. Μία ακμή από τον κόμβο A στον κόμβο B με βάρος w, σημαίνει ότι υπάρχουν w αναφορές από δημοσιεύσεις του συνεδρίου A σε δημοσιεύσεις του συνεδρίου B.

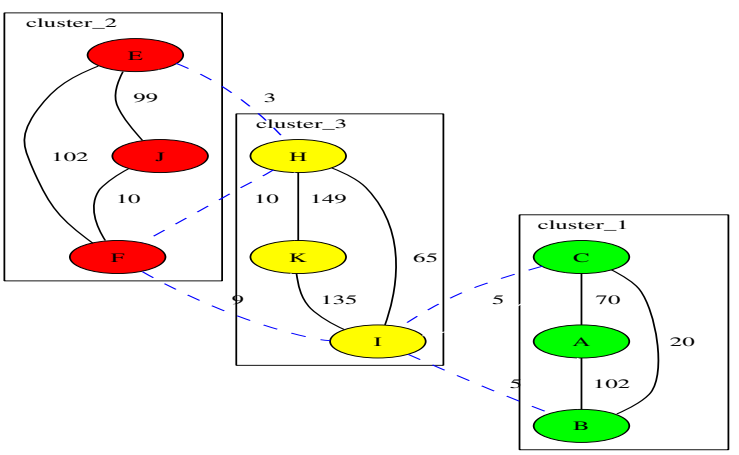
Για να πραγματοποιήσουμε την ομαδοποίηση των συνεδρίων, χρειάζεται να ελαχιστοποιήσουμε τα βάρη των ακμών, που διαπερνούν από μία ομάδα σε μία άλλη (αυτό υπολογίζεται από το hMetis). Για αυτό το σκοπό δεν χρειαζόμαστε κατευθυνόμενο γράφο. Συνεπώς, μετατρέπουμε τον τελευταίο γράφο σε γράφο χωρίς κατεύθυνση (χωρίς να αποφεύγουμε την απώλεια κάποιων πληροφοριών). Στο παράδειγμά μας, το Σχήμα 2.2(β) έχει προέλθει από το Σχήμα 2.2(α). Σε αυτό το γράφο, το βάρος μίας ακμής που συνδέει δύο κόμβους αναπαριστά το συνολικό αριθμό των αναφορών που κάνουν αυτά τα συνέδρια μεταξύ τους. Ο παραγόμενος γράφος δεν μπορεί να χρησιμοποιηθεί για την αξιολόγηση, που είναι ο κύριος στόχος μας, παρά μόνο για το βήμα της ομαδοποίησης. Στην πραγματικότητα, αυτός είναι ο τύπος του γράφου που εισάγουμε στο hMetis για την ομαδοποίηση.

Τα αποτελέσματα της ομαδοποίησης δεν ήταν τέλεια, αφού τμήμα του γράφου αναφορών δεν είναι ολοκληρωμένο (σε μερικές εγγραφές της ψηφιακής βιβλιοθήκης DBLP για δημοσιεύσεις συνεδρίων δεν περιλαμβάνονται αναφορές). Κατά συνέπεια, ομαδοποιούμε και πάλι τα συνέδρια, αφού πραγματοποιήσαμε κάποια προεπεξεργασία, η οποία βασίσθηκε σε λέξεις-κλειδιά που ταίριαζαν με τους τίτλους των συνεδρίων. Πιο συγκεκριμένα, προκαθορίζουμε 4 ομάδες:

- **Ομάδα 1: Βάσεις Δεδομένων (DataBases).** Οι λέξεις-κλειδιά που χρησιμοποιήθηκαν για να αναγνωρισθούν τα συνέδρια που ανήκουν σε αυτή



(α) Κατευθυνόμενος γράφος αναφορών συνεδρίων.



(β) Ισοδύναμος μη κατευθυνόμενος γράφος αναφορών συνεδρίων.

Σχήμα 2.2. Κατευθυνόμενος (α) και ισοδύναμος μη κατευθυνόμενος (β) γράφος αναφορών συνεδρίων.

την περιοχή ήταν: data (δεδομένα) base (βάση), database, information retrieval (ανάκτηση πληροφοριών), information system (πληροφοριακό σύστημα), mining (εξόρυξη) και geographic (γεωγραφικό).

- **Ομάδα 2: Λογικός Προγραμματισμός (Logic Programming).** Αναγνωρίσθηκαν χρησιμοποιώντας τις λέξεις-κλειδιά: AI, artificial (τεχνητή), intelligent (νοημοσύνη), knowledge (γνώση), logic (λογική), και algorithm (αλγόριθμος).

- **Ομάδα 3: Δίκτυα και Κατανεμημένα Συστήματα (Networks and Distributed Systems).** Αναγνωρίσθηκαν χρησιμοποιώντας τις λέξεις-κλειδιά: distributed (κατανεμημένος), network (δίκτυο), parallel (παράλληλο), web (ιστός), www, και w3c.

- **Ομάδα 4: Λειτουργικά Συστήματα (Operating Systems), Μηχανική Λογισμικού (Software Engineering), Γλώσσες και Μεταφραστές (Compilers και Languages)**, για τις αντίστοιχες περιοχές.

Αφού ορίσαμε αυτές τις τέσσερις ομάδες με κάποια μέλη συνεδρίων που περιλαμβάνονται σε αυτά, εισαγάγαμε αυτές τις προκαθορισμένες τμηματοποιήσεις στο hMetis για να συνεχίσει με τα συνέδρια που δεν είχαν ομαδοποιηθεί.

2.4.2 Εκκαθάριση των Ομάδων

Όπως αναφέρθηκε προηγουμένως, η ψηφιακή βιβλιοθήκη DBLP δεν είναι ολοκληρωμένη. Για παράδειγμα, για κάποια συνέδρια ή περιοδικά, δεν περιλαμβάνονται 1 ή 2 δημοσιεύσεις (οι πιο σημαντικές). Αυτό θα οδηγούσε τον αλγόριθμο αξιολόγησης να παράγει εσφαλμένα αποτελέσματα, αφού η κύρια μετρική μέθοδος είναι ο μέσος αριθμός αναφορών ανά δημοσίευση.

Σε αυτό το βήμα, αφαιρούμε από το σύνολο των συνεδρίων, το οποίο θα χρησιμοποιηθεί για την αξιολόγηση, τα συνέδρια που:

- περιέχουν λιγότερες από τρεις δημοσιεύσεις
- έχουν πραγματοποιηθεί μόνο μία φορά, και
- έχουν μέσο όρο δημοσιεύσεων ανά έτος λιγότερο από 0.5.

Για αυτά τα συνέδρια θέτουμε ένα δείκτη ότι δεν θα αξιολογηθούν, αλλά δεν τα σβήνουμε από τη βάση μας. Επομένως, μετρώνται όλες οι αναφορές που περιλαμβάνονται σε αυτά.

2.4.3 Ορισμός των Μετρικών Μεθόδων

Εδώ παρουσιάζουμε τις νέες μετρικές μεθόδους, προκειμένου να εγκαθιδρύσουμε μία νέα προοπτική για την αξιολόγηση των συνεδρίων και περιοδικών χρησιμοποιώντας το γράφο αναφορών. Αυτές οι μετρικές μέθοδοι ορίζονται ως ακολούθως:

2.4.3.1 Απλή Βαθμολογία (Plain Score)

Αν C είναι το σύνολο όλων των συνεδρίων, τότε η *Απλή Βαθμολογία (Plain Score)* Sc, που είναι η βαθμολογία για το συνέδριο c, ορίζεται ως:

$$S_c \ = \ \frac{1}{P_c} \sum_{\forall i \in C} N_{i \to c} \tag{2.1}$$

όπου $N_{i \to c}$ είναι ο αριθμός των αναφορών που έγιναν από το συνέδριο i στο συνέδριο c, όπου ο παράγοντας κανονικοποίησης P_c είναι ο αριθμός των δημοσιεύσεων στο συνέδριο c.

Η κατάταξη υπολογίζεται ταξινομώντας τις βαθμολογίες των συνεδρίων. Στην περίπτωση ισοβαθμίας προηγείται το συνέδριο με τις λιγότερες δημοσιεύσεις. Ουσιαστικά, η βαθμολογία είναι ακριβώς η ίδια με το εσω-βαθμού (in-degree) του αντίστοιχου κόμβου στο γράφο αναφορών συνεδρίων, διαιρούμενος με το πλήθος των δημοσιεύσεων που περιλαμβάνονται στο συνέδριο. Αυτή η βαθμολόγηση είναι η απλούστερη και βασικά χρησιμοποιείται σαν μία πρώτη προσέγγιση για την αξιολόγηση. Στην πραγματικότητα, αν και αυτή η μετρική μέθοδος δίνει πληροφορίες που αφορούν την αξιολόγηση, έχει το μειονέκτημα ότι συνέδρια με μεγαλύτερη παράδοση είναι πιθανότερο να έχουν περισσότερες αναφορές. Επομένως, μπορεί μόνο να χρησιμοποιηθεί για να συγκρίνει και να αξιολογήσει ένα σύνολο συνεδρίων που έχει ακριβώς την ίδια διάρκεια ζωής.

2.4.3.2 Βαθμολογία με Βάρη (Weighted Score)

Εδώ εισάγουμε την ιδέα της αξιολόγησης με χρήση βαρών. Αυτό σημαίνει ότι οι αναφορές δεν μετρούν το ίδιο. Η Εξίσωση 2.2 δείχνει αφαιρετικά πως μπορεί να υπολογισθεί η Βαθμολογία με Βάρη, που ορίζεται ως WS_c, για το συνέδριο c.

$$WS_c \ = \ \frac{1}{P_c} \frac{\sum_i W_i * N_{i \to c}}{\sum_i W_i} \tag{2.2}$$

Πώς μπορούμε να υπολογίσουμε τα βάρη; Ποια συνέδρια είναι "Τα-καλύτερα" που θα έπρεπε να έχουν μεγαλύτερα βάρη και ποια είναι "Τα-χειρότερα" συνέδρια; Εδώ προκύπτει η ανάγκη για αναδρομικό υπολογισμό. Αυτός ο υπολογισμός εκτελείται χρησιμοποιώντας τον εξής τύπο:

$$\begin{array}{lll} WS_{c,l} \ = \ \frac{1}{P_c} \frac{\sum_i W_{i,l-1} * N_{i \to c}}{\sum_i W_{i,l-1}} & & l \geq 1 \\ W_{i,0} \ = \ 1 & & \forall i \in C \end{array} \tag{2.3}$$

Αρχικά όλα τα βάρη τίθενται ίσα με 1 (στο επίπεδο 0). Έτσι, μπορούμε να υπολογίσουμε την κατάταξη για το επόμενο επίπεδο, βασιζόμενοι στα βάρη που

υπολογίσθηκαν στο προηγούμενο επίπεδο. Η κατάταξη που παίρνουμε στο επίπεδο 1, είναι ισοδύναμη με την κατάταξη της Απλής Βαθμολογίας (αφού χρησιμοποιήσαμε βάρη του 1 για όλες τις οντότητες). Αφού υπολογίσουμε τους βαθμούς του πρώτου επιπέδου, μπορούμε να υπολογίσουμε τα βάρη. Αυτό επιτυγχάνεται με έναν άλλο αλγόριθμο ομαδοποίησης. Μία λεπτομερέστερη συζήτηση για τον υπολογισμό των βαρών μπορεί να βρεθεί στην Παράγραφο 2.4.5.

Αφού υπολογίσουμε τα βάρη για το επίπεδο 1, συνεχίζουμε υπολογίζοντας τους βαθμούς για τα επόμενα επίπεδα, εφαρμόζοντας την ίδια διαδικασία έως ότου η κατάταξη παραμένει η ίδια. Αυτή είναι η συνθήκη τερματισμού μας. Ο υπολογισμός επαναλαμβάνεται $\forall l \geq 1$ έως L, όπου η κατάταξη για το επίπεδο L είναι ισοδύναμη με αυτήν του επιπέδου $L-1$. Εναλλακτικά, αν ενώ είμαστε στο επίπεδο L παίρνουμε τα ίδια βάρη με αυτά του επιπέδου $L-1$ ($W_{i,L} = W_{i,L-1}\ \forall i \in C$), τότε είναι φανερό ότι η κατάταξη για $L+1$ θα είναι η ίδια με αυτήν που υπολογίσθηκε στο επίπεδο L. Επομένως, μία εναλλακτική συνθήκη τερματισμού είναι η "καμία αλλαγή" στα υπολογιζόμενα βάρη. Τότε $\forall l \in \{L..\infty\}$ η συνθήκη: $WS_{i,l+1} = WS_{i,l}$ είναι αληθής, και θέτουμε:

$$WS_c \ = \ WS_{c,\infty} = WS_{c,L}$$

Αυτός ο τύπος αξιολόγησης, όπως και με την Απλή Βαθμολογία, δεν μπορεί να χρησιμοποιηθεί για την κατάταξη συνεδρίων χωρίς κίνδυνο. Παρά τη βελτίωση να υπολογίζουμε το μέσο βαθμό ανά δημοσίευση μέσω των βαρών των αναφορών, όλα τα συνέδρια δεν έχουν την ίδια διάρκεια ζωής, ενώ κάποια πραγματοποιούνται μία φορά ανά δύο ή τρία έτη. Επομένως, τα συνέδρια με παράδοση είναι πιθανότερο να έχουν περισσότερες αναφορές. Αυτή η κατάταξη μπορεί να χρησιμοποιηθεί μόνο για συνέδρια που έχουν ακριβώς την ίδια διάρκεια ζωής.

2.4.3.3 Απλή Βαθμολογία ανά Έτος

Υιοθετώντας την έννοια της Απλής Βαθμολογίας για να κατατάξουμε συνέδρια για κάθε διακριτό έτος, εισάγουμε τη μετρική μέθοδο Απλή Βαθμολογία ανά έτος ως:

$$SY_{c,y} \ = \ \frac{1}{P_{c,y}} \sum_{\forall i \in C} N_{i \to c,y} \tag{2.4}$$

όπου $SY_{c,y}$ είναι η βαθμολογία για το συνέδριο c το έτος y, $N_{i \to c,y}$ είναι ο αριθμός των αναφορών από το συνέδριο i στο συνέδριο c που πραγματοποιήθηκε το έτος y και $P_{c,y}$ είναι ο αριθμός των δημοσιεύσεων του συνεδρίου c κατά τη διάρκεια του έτους y. Πιο συγκεκριμένα, μία αναλυτικότερη έκφραση που χρησιμοποιήσαμε

για τον υπολογισμό μας είναι:

$$N_{i\to c,y} = \sum_{z=y}^{last\_year} N_{i,z\to c,y} \overset{(2.4)}{\Rightarrow} SY_{c,y} = \frac{1}{P_{c,y}} \sum_{i}^{\forall i\in C} \sum_{z=y}^{last\_year} N_{i,z\to c,y} \qquad (2.5)$$

όπου $N_{i,z\to c,y}$ είναι ο αριθμός των αναφορών που έγιναν από το συνέδριο i το έτος z στο συνέδριο c που πραγματοποιήθηκε στο έτος y. Η μεταβλητή $last\_year$ τίθεται στο μέγιστο έγκυρο έτος στη συλλογή μας (που κανονικά είναι το τρέχον έτος). Αυτή η κατάταξη μπορεί να χρησιμοποιηθεί για να συγκρίνουμε συνέδρια που πραγματοποιήθηκαν το ίδιο έτος.

2.4.3.4 Βαθμολογία με Βάρη ανά Έτος

Συνδυάζοντας το WS (Weighed Score - Βαθμολογία με Βάρη) με το SY (Plain Score per Year - Απλή Βαθμολογία ανά Έτος), για $l \geq 1$ παράγουμε το WSY, δηλαδή τη μετρική μέθοδο *Βαθμολογία με Βάρη ανά Έτος (Weighted Score per Year)*:

$$WSY_{c,y,l} = \frac{1}{P_{c,y}} \frac{\sum_{i}^{\forall i\in C} \left(W_{i,y,l-1} * N_{i,y\to c,y} + \sum_{z=y+1}^{last\_year} W_{i,z,\infty} * N_{i,z\to c,y} \right)}{\sum_{i}^{\forall i\in C} \left(W_{i,y,l-1} + \sum_{z=y+1}^{last\_year} W_{i,z,\infty} \right)} \qquad (2.6)$$

Κατά τον ίδιο τρόπο όπως προηγουμένως θέτουμε:

$$W_{i,z,0} = 1 \quad \forall i \in C \; and \; \forall z \in \{valid\_years\}$$

Ο υπολογισμός γίνεται για κάθε έτος ξεκινώντας από το τελευταίο έτος με αντίστροφη σειρά. Επομένως, όταν υπολογίζουμε τις βαθμολογίες για το έτος Y, όλα τα βάρη είναι γνωστά για τα έτη $\{Y+1...max\_year\}$. Για κάθε έτος, η διαδικασία επαναλαμβάνεται $\forall l \geq 1$ έως L, όπου η αξιολόγηση δεν αλλάζει ή η συνθήκη $W_{c,y,L} = W_{c,y,L-1}$ είναι αληθής $\forall c \in C$. Τότε $W_{c,y,\infty} = W_{c,y,L}$ και θέτουμε:

$$WSY_{c,y} = WSY_{c,y,\infty} = WSY_{c,y,L}$$

2.4.3.5 Βαθμολογία Αντεστραμμένου Αντικτύπου ανά Έτος

Ο Garfield [32] όρισε τον Παράγοντα Αντικτύπου ISI χρησιμοποιώντας το ακόλουθο παράδειγμα για το έτος 1992:

$\mathbb{A} =$ συνολικές αναφορές κατά το 1992
$\mathbb{B} =$ Αναφορές του 1992 σε άρθρα που δημοσιεύτηκαν το 1990-1991
$\mathbb{C} =$ ο αριθμός των άρθρων που δημοσιεύτηκαν το 1990-1991

τότε

$$\mathbb{D} = \mathbb{B}/\mathbb{C} = 1992 \ \ \text{ISI Impact Factor (ISI Παράγοντας Αντικτύπου)} \quad (2.7)$$

Αν J είναι το σύνολο των περιοδικών και j είναι ένα συγκεκριμένο περιοδικό, τότε η Εξίσωση (2.7) είναι ισοδύναμη με την Εξίσωση (2.8) στη γενική μορφή:

$$IF_{j,y} = \sum_{z=y-k}^{y-1} \frac{\sum_{i}^{\forall i \in J} N_{i,y \to j,z}}{P_{j,z}} \quad (2.8)$$

Αυτή η μετρική μέθοδος δεν μπορεί να εφαρμοσθεί κατ'ευθείαν σε συνέδρια για αξιολόγηση ανά έτος. Αυτό οφείλεται στο γεγονός ότι όταν υπολογίζουμε τον Παράγοντα Αντικτύπου ISI για ένα συνέδριο c για το έτος y, ουσιαστικά αξιολογούμε τις διοργανώσεις του c που πραγματοποιήθηκαν κατά τη διάρκεια των προηγούμενων k ετών. Για παράδειγμα, για να υπολογίσουμε τον Παράγοντα Αντικτύπου ISI του VLDB'95, ουσιαστικά αξιολογούμε τα VLDB'94 και VLDB'93. Έτσι, δύο ξεχωριστά γεγονότα (που οργανώθηκαν σε διαφορετικές ηπείρους) ενός συγκεκριμένου συνεδρίου ομαδοποιούνται και αξιολογούνται μαζί. Ίσως, θα μπορούσαμε να αξιολογήσουμε το VLDB'95 υπολογίζοντας τον Παράγοντα Αντικτύπου ISI για τα VLDB'96 και VLDB'97. Σε μία τέτοια περίπτωση τα αποτελέσματά μας δεν επηρεάζονται μόνο από την επιτυχία του VLDB'95, αλλά επίσης από την επιτυχία του VLDB'96 και VLDB'97. Επομένως, για την περίπτωση των συνεδρίων, ο Παράγοντας Αντικτύπου ISI δεν μπορεί να χρησιμοποιηθεί για να αξιολογήσουμε ένα συγκεκριμένο συνέδριο c που πραγματοποιήθηκε κατά το έτος y.

Για τους προηγούμενους λόγους, *αντιστρέφουμε* την έννοια του Παράγοντα Αντικτύπου ISI και αντί να μετρούμε τις αναφορές που έγιναν τα προηγούμενα k χρόνια, μετρούμε τις αναφορές που έγιναν κατά τη διάρκεια των επόμενων k χρόνων (Εξίσωση 2.9). Έτσι, αποτιμούμε τον Αντίκτυπο που έχει το συγκεκριμένο συνέδριο κατά τη διάρκεια των επόμενων 2 ετών. Ονομάζουμε αυτόν τον παράγοντα ως *Αντεστραμμένο Παράγοντα Αντικτύπου* ή *Παράγοντα Α-Αντικτύπου* (I-Impact Factor). Η Βαθμολογία Α-Αντικτύπου ανά Έτος (I-Impact Score per Year) ορίζεται ως εξής:

$$IISY_{c,y} = \frac{1}{P_{c,y}} \sum_{i}^{\forall i \in C} \sum_{z=y}^{y+k} N_{i,z \to c,y} \quad (2.9)$$

Οι Εξισώσεις (2.8) και (2.9) μπορεί σημασιολογικά να είναι διαφορετικές αλλά είναι ποιοτικά παρόμοιες καθώς και οι δύο μετρούν το αντίκτυπο μιας συλλογής.

Στην Εξίσωση (2.8) ο αντίκτυπος υπολογίζεται κατά τη διάρκεια ενός συγκεκριμένου έτους (π.χ Ποιός είναι ο αντίκτυπος του VLDB κατά τη διάρκεια του 1998 - στην πραγματικότητα ο αντίκτυπος των VLDB'97 και 96). Στην Εξίσωση (2.9), ο αντίκτυπος υπολογίζεται για ένα συγκεκριμένο έτος του συνεδρίου (π.χ Ποιός είναι ο αντίκτυπος του VLDB'97;). Έτσι μπορούμε να αξιολογήσουμε ατομικά συνέδρια και για παράδειγμα να πάρουμε πληροφορίες όπως: ποιό ήταν το πιο *επιτυχημένο* συνέδριο το 1997. Ίσως με ένα τέτοιο στατιστικό το ίδρυμα VLDB καθιέρωσε το βραβείο της καλύτερης δημοσίευσης μετά την παρέλευση 10ετίας.

Η μετρική μέθοδος της Βαθμολογίας Αντίστροφου Αντικτύπου (Inverted Impact Score) (Εξίσωση 2.9) είναι μία υποπερίπτωση του αλγορίθμου Απλή Βαθμολογία ανά Έτος, αν θέσουμε το $last\_year = y + k$, όπου η συνήθης τιμή για το k που χρησιμοποιείται από το ISI είναι 2 ή 5. Καθώς η έννοια του Παράγοντα Αντικτύπου ISI είναι ευρέως αποδεκτή, χρησιμοποιούμε αυτό το μέτρο στα πειράματά μας ως το βασικό μέτρο σύγκρισης. Δεν μπορούμε να χρησιμοποιήσουμε τον Παράγοντα Αντικτύπου ISI ακριβώς όπως ορίσθηκε από τον Garfield [32], επειδή είναι σημασιολογικά διαφορετικός από τις μετρικές μεθόδους που παρουσιάσθηκαν εδώ.

2.4.3.6 Βαθμολογία A-Αντικτύπου με Βάρη ανά Έτος

Με τον ίδιο τρόπο, αν στη Βαθμολογία με Βάρη ανά Έτος (Εξίσωση 2.6) θέσουμε $last\_year = y + k$, όπου k=2 ή 5, τότε παίρνουμε τη Βαθμολογία A-Αντικτύπου σε ένα ζυγισμένο τρόπο, έστω $WIISY_{c,y}$. Αυτό έχει τα πλεονεκτήματα της μετρικής μεθόδου της Βαθμολογίας A-Αντικτύπου, συν τα πλεονεκτήματα της μετρικής μεθόδου των Βαρών.

2.4.4 Ο Αλγόριθμος Κατάταξης

Ο αλγόριθμος κατάταξης φαίνεται στο Σχήμα 2.3, ο οποίος χρησιμοποιείται και για τους 4 τύπους βαθμολογιών. Ο αλγόριθμος αυτός καλεί την συνάρτηση υπολογισμού βαρών, η οποία παρουσιάζεται στο Σχήμα 2.4. Οι Απλές Βαθμολογίες ανά Έτος είναι τα αποτελέσματα του αλγορίθμου στο επίπεδο 1. Οι Βαθμολογίες με Βάρη είναι τα αποτελέσματα από το τελευταίο επίπεδο που φθάσαμε, και τα αποθηκεύουμε στον πίνακα αποτελεσμάτων με ένα δείκτη επιπέδου που έχει τεθεί στο -1 (αντί του ∞ για πρακτικούς λόγους). Οι Βαθμολογίες A-Αντικτύπου μπορούν να υπολογισθούν με τον ίδιο αλγόριθμο, θέτοντας τη μεταβλητή $last\_year$ σε $y + 2$.

```
last_year=max_year; L=-1;
for(y=max_year;y>=min_year;y--) {
 l=0; ok=0;
 for(c=0;c<n_confs;c++) { W[c,y,l]=1; }
 while(!ok) {
   l++;
   for(c=0;c<n_confs;c++) {
     WSY[c,y,l]=0;
     for(i=0;i<n_confs;i++){
       WSY[c,y,l]+=W[i,y,l-1]*N[i,y,c,y];
       for(z=y+1;z<=last_year;z++) { WSY[c,y,l]+=W[i,y,L]*N[i,z,c,y]; }
     }
     W[0..n_confs,y,l] = compute_weights(WSY[0..n_confs,y,l]);
     ok=check_if_done(l,y);
   }
 }
 for(c=0;c<n_confs;c++) { WSY[c,y,L] = WSY[c,y,l]; }
}
```

Σχήμα 2.3. Γενικός αλγόριθμος κατάταξης.

2.4.5 Το Σύνολο των Βαρών

Για κάθε διαφορετική κατάταξη[2] χρειάζεται να ορίσουμε ένα σύνολο συνόλων:

$$G \;=\; \{G_1, G_2, ..G_n\} \tag{2.10}$$

όπου

$$\begin{aligned} G_i \;&=\; \{\text{ομάδα συνεδρίων}\} && \text{για } 1 \le i \le n \\ G_1 \cup G_2 \cup ... \cup G_n \;&=\; C \\ G_i \cap G_j \;&=\; \emptyset && \text{για } 1 \le i, j \le n, \;\; i \ne j \end{aligned}$$

Πρέπει να αναθέσουμε μία συγκεκριμένη τιμή βάρους σε κάθε σύνολο G_i, για $1 \le i \le n$. Επομένως, πρέπει να ορίσουμε το σύνολο:

$$W \;=\; \left\{W^1, W^2, .., W^n\right\}$$

Σε αυτό το σημείο πρέπει να εισάγουμε δύο σημαντικές παραμέτρους. Τον *αριθμό των ομάδων* ($\equiv n$) και το *εύρος των βαρών*. Για τις δοκιμές μας, έχουμε

[2] Για το επίπεδο l και για το έτος y, ή μόνο για το επίπεδο l αν υπολογίζουμε την κατάταξη για όλα τα χρόνια.

```
function compute_weights( array scores[0..n_confs]) {
 max_groups=5; min_weight=1; max_weight=5;
 groups = new array;
 for(g=0;g<n_confs;g++) {
   groups[g].avg = res[g].score;
   groups[g].members.add(g);
 }
 groups = sort {(a.avg<=>b.avg)} groups;
 while( count(groups)>max_groups) {
   (MinDiff,ToJoin)=find_MinDiff(groups);
   groups[ToJoin] = union(groups[ToJoin], groups[ToJoin-1]);
   delete groups[ToJoin-1];
 }
 while(MinDiff==0) {
   (MinDiff,ToJoin)=find_MinDiff(groups);
   if(MinDiff==0) {
     groups[ToJoin] = union(groups[ToJoin], groups[ToJoin-1]);
     delete groups[ToJoin-1];
   }
 }
 weight=min_weight;
 step = (max_weight-min_weight+1)/max_groups;
 for(g=0;g<count(groups);g++) {
   groups[g].weight=weight; weight+=step;
 }
}
function find_MinDiff(array groups) {
 // Find the minimum difference by
 // comparing the groups[i].avg values
 // and return the index of the second
 // element belonging to the mindiff-pair.
 return(MinDiff,ToJoin);
}
```

Σχήμα 2.4. Αλγόριθμοι ομαδοποίησης και καθορισμού βαρών.

θέσει τον αριθμό των ομάδων ίσο με 5 (το οποίο σημαίνει: πολύ δυνατό, δυνατό, μέσο, αδύναμο, πολύ αδύναμο). Αυτό μας οδηγεί σε 5 διαφορετικά βάρη και 5 ομάδες στην αξιολόγηση των συνεδρίων. Η επιλογή του εύρους βαρών είναι επίσης σημαντική καθώς επηρεάζει τα αποτελέσματα με την έννοια ότι συντονίζει τη σημασία μιας αναφοράς ενός *πολύ δυνατού* συνεδρίου συγκριτικά με τη σημασία μιας

αναφοράς από ένα *πολύ αδύναμο* συνέδριο. Αποφασίσαμε να χρησιμοποιήσουμε το εύρος 1-5 και συγκεκριμένα τα βάρη $W^1 = 1, W^2 = 2, W^3 = 3, W^4 = 4, W^5 = 5$ με σκοπό να τονίσουμε τη διαφορά από την Απλή Βαθμολογία. Για παράδειγμα, διαλέγοντας το εύρος 1-2 (δηλαδή 1, 1.2, 1.4, 1.6, 1.8, 2) θα είχε ελάχιστη διαφορά.

Ουσιαστικά, εφ'όσον οι βαθμολογίες κανονικοποιούνται διαιρώντας με το άθροισμα των βαρών, ο σημαντικός παράγοντας είναι το κλάσμα των βαρών διαιρούμενο με το μικρότερο, και όχι οι απόλυτες τιμές. Επομένως, είναι ασφαλές να δεχθούμε ως ελάχιστο βάρος το 1. Είναι προφανές, ότι δεν έχει νόημα να χρησιμοποιήσουμε αρνητικό ή μηδενικό βάρος[3].

Επίσης, ορίσαμε ότι τα συνέδρια που ανήκουν σε μία διαφορετική επιστημονική περιοχή από αυτήν για την οποία υπολογίζεται η αξιολόγηση, να είναι μέλη του συνόλου G_1. Επιπρόσθετα σε αυτό, συνέδρια με μηδενική βαθμολογία ($\Leftarrow$ 0 αναφορές σε αυτά) επίσης τοποθετούνται στην ομάδα G^1 από τον αλγόριθμο κατάταξης.

2.4.6 Ομαδοποίηση Συνεδρίων με Βάση τις Αναφορές

Ο αλγόριθμος ομαδοποίησης είναι ένας ιεραρχικός αλγόριθμος ομαδοποίησης που εφαρμόζεται σε σημεία μιας διάστασης [46]. Αρχικά, ορίζεται ένας αριθμός ομάδων N, όπου N είναι ο αριθμός των συνεδρίων υπό αξιολόγηση.

$$G_1 = \{S_1\} \quad G_2 = \{S_2\} \quad ... \quad G_N = \{S_N\}$$

Για κάθε ομάδα G_x θέτουμε G_x^A ως τη μέση τιμή των S_i μελών της. Σε κάθε βήμα του αλγορίθμου, βρίσκουμε δύο σύνολα G_i και G_j, για τα οποία η διαφορά των μέσων τιμών τους ($|G_i^A - G_j^A|$) είναι η ελάχιστη οποιουδήποτε άλλου ζεύγους. Ορίζουμε ένα νέο σύνολο $G_k = G_i \cup G_j$ και διαγράφουμε τα σύνολα G_i και G_j. Η διαδικασία επαναλαμβάνεται έως ότου ο αριθμός των ομάδων γίνει n. Αν, όταν φτάσουμε το n, υπάρχει ένα ζεύγος με μηδενική διαφορά των μέσων τιμών τους ($|G_i^A - G_j^A| = 0$)[4], συνεχίζουμε ενώνοντας τις ομάδες μέχρι να πάρουμε $|G_i^A - G_j^A| > 0 \quad \forall G_i, G_j \in G$. Στο Σχήμα 2.4 απεικονίζουμε τον αλγόριθμο ομαδοποίησης.

[3] Ένα βάρος μπορεί να τεθεί μηδέν αν και μόνο αν το αντίστοιχο συνέδριο δεν υπάρχει. Αυτό μπορεί να συμβεί στην περίπτωση που κάνουμε κατάταξη ανά έτος όπου στη λίστα μας έχουμε μεν όλα τα συνέδρια, αλλά μπορεί κάποιο από αυτά να μην πραγματοποιήθηκε τη συγκεκριμένη χρονιά ή απλά να λείπουν από τη βάση μας τα αντίστοιχα στοιχεία.

[4] Αυτό συμβαίνει μόνον όταν ΟΛΑ τα μέλη των δυο ομάδων έχουν ακριβώς το ίδιο σκορ.

2.4.7 Βελτίωση Βαρών

Ο αλγόριθμος Βαθμολογίας με Βάρη, όπως έχει περιγραφεί προηγουμένως, είναι ευάλωτος σε αδιέξοδα. Αυτό οφείλεται στο γεγονός ότι δεν υπάρχει εγγύηση ότι ένα συνέδριο δεν θα μετακινείται συνεχώς από μία ομάδα σε μία άλλη κατά τη διάρκεια εκτέλεσης του αλγορίθμου. Παρουσιάζουμε αυτή την κατάσταση με ένα απλό παράδειγμα δύο συνεδρίων A και B για τα οποία ισχύει:

$$\begin{array}{ll} P_A = P_B = x(=10) & \\ N_{A\to B} = 4 & N_{B\to B} = 0 \\ N_{B\to A} = 3 & N_{A\to A} = 0 \end{array}$$

Σε αυτή την περίπτωση:

$$\begin{array}{lll} level:1 & \left.\begin{array}{l} WS_{A,1} = 0.15 \\ WS_{B,1} = 0.2 \end{array}\right\} \Rightarrow & \left.\begin{array}{l} W_{A,1} = 1 \\ W_{B,1} = 2 \end{array}\right\} \Rightarrow \\ level:2 & \left.\begin{array}{l} WS_{A,2} = 0.2 \\ WS_{B,2} = 0.13 \end{array}\right\} \Rightarrow & \left.\begin{array}{l} W_{A,2} = 2 \\ W_{B,2} = 1 \end{array}\right\} \Rightarrow \\ level:3 & \left.\begin{array}{l} WS_{A,3} = 0.1 \\ WS_{B,3} = 0.26 \end{array}\right\} \Rightarrow & \begin{array}{l} W_{A,3} = 1 \\ W_{B,3} = 2 \end{array} \end{array}$$

Αυτό οδηγεί σε μία επανάληψη επ'άπειρο, αφού στο επίπεδο 4 θα πάρουμε ακριβώς τα ίδια αποτελέσματα με το επίπεδο 2. Για να αποφύγουμε αυτή την περίπτωση, μετά τον υπολογισμό των ομάδων $G_1, G_2, ..., G_n$ στο επίπεδο l, και πριν να αναθέσουμε τα βάρη $(W^1, W^2, ..., W^n)$ για κάθε συνέδριο, ελέγχουμε εάν έχει προκύψει η ίδια κατάσταση σε κάποιο προηγούμενο επίπεδο d. Αν υπάρχει ένα επίπεδο $d < (l-1)$ για το οποίο ισχύει $W_{c,d} = W^k$ $(c \in G_k)$, $\quad \forall c \in C$[5], τότε δεν θέτουμε $W_{c,l} = W^k$ (όπως θα έπρεπε), αλλά αντί για αυτό θέτουμε:

$$W_{c,l} \;=\; avg(W_{c,d}...W_{c,l-1}) \;=\; \frac{\sum_{p=d}^{l-1} W_{c,p}}{l-d}$$

Έτσι ισχύει: $W_{c,l} \overset{l\to\infty}{\longrightarrow} x$, όπου x είναι ένας πραγματικός αριθμός. Στην πραγματικότητα, εφ'όσον έχουμε διαφορετικά βάρη, φτάνουμε στο x πολύ γρήγορα. Στο προηγούμενο παράδειγμα τα επόμενα βήματα θα έπρεπε να είναι:

$$\begin{array}{lll} level:3 \;\Rightarrow & \left.\begin{array}{l} W_{A,3} = 1.5 \\ W_{B,3} = 1.5 \end{array}\right\} \Rightarrow & \\ level:4 & \left.\begin{array}{l} WS_{A,4} = 0.15 \\ WS_{B,4} = 0.2 \end{array}\right\} \Rightarrow termination & \end{array}$$

[5] Αν $d = l-1$ τότε ισχύει η συνθήκη τερματισμού.

2.5 Πειραματικά Αποτελέσματα

Πρώτα, επισημαίνουμε ότι η κατάταξη γίνεται μόνο για μία από τις τέσσερις ομάδες που παρουσιάσθηκαν στην Παράγραφο 2.4.1. Κι αυτό γιατί εστιάζουμε στην ομάδα Βάσεις Δεδομένων, καθώς είναι η πιο ολοκληρωμένη ομάδα στην ψηφιακή βιβλιοθήκη DBLP. Η βάση δεδομένων περιέχει συνέδρια από το 1959 έως το 2003 (αλλά *ολοκληρωμένα* δεδομένα για αυτά τα συνέδρια υπάρχουν μόνο για τα έτη από 1980 και μετά). Επομένως, βρίσκουμε την κατάταξη για κάθε έτος ξεχωριστά χρησιμοποιώντας:

- την Απλή Βαθμολογία ανά Έτος
- τη Βαθμολογία με Βάρη ανά Έτος
- την Απλή Α-Αντικτύπου Βαθμολογία ανά Έτος
- τη Βαθμολογία Α-Αντικτύπου με Βάρη ανά Έτος

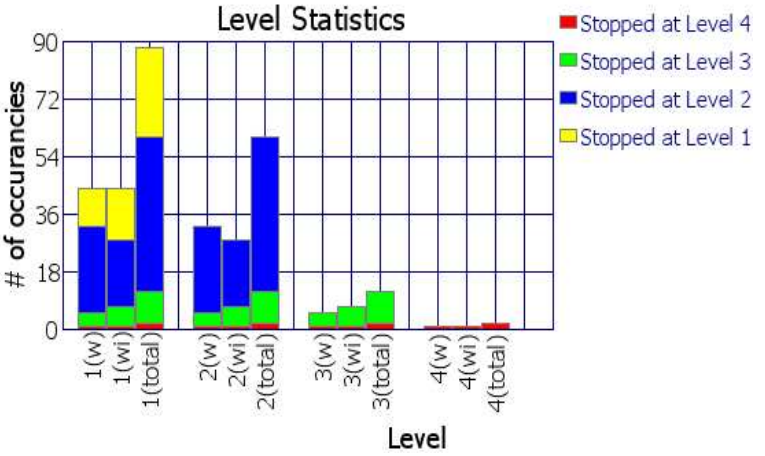
Για όλες τις εκτελέσεις χρησιμοποιείται ο αλγόριθμος του Σχήματος 2.3. Η Απλή Βαθμολογία είναι το αποτέλεσμα του αλγορίθμου στο επίπεδο 1 (όλα τα βάρη στο επίπεδο 0 είναι ίσα με 1). Τα αποτελέσματα [6] μιας κατάταξης με βάρη είναι τα αποτελέσματα του τελευταίου επιπέδου όπου φθάσαμε. Η Βαθμολογία Α-Αντικτύπου είναι μία υποπερίπτωση των προηγουμένων, και επομένως φθάνουμε σε αυτήν αν θέσουμε τη μεταβλητή *last_year* ίση με $y + 2$. Κατά συνέπεια, υπολογίζουμε για 44 έτη τη Βαθμολογία με Βάρη και τη Βαθμολογία Α-Αντικτύπου με Βάρη δηλαδή ένα σύνολο 88 ξεχωριστών κατατάξεων (η απλή κατάταξη είναι ένα επιμέρους αποτέλεσμα της κατάταξης με βάρη).

Μία σημαντική ανησυχία είναι το υπολογιστικό κόστος, π.χ. πόσες φορές πρέπει να επαναλάβουμε τον υπολογισμό, ώστε να πάρουμε τη συνθήκη τερματισμού "καμία αλλαγή" κατά τη διάρκεια της κατάταξης. Όπως φαίνεται στο Σχήμα 2.5, στις περισσότερες περιπτώσεις, δύο επίπεδα ήταν αρκετά για να πάρουμε την τελική κατάταξη και μόνο σε τρεις περιπτώσεις (μία για τη Βαθμολογία με Βάρη και δύο για τη Βαθμολογία Α-Αντικτύπου) έπρεπε να φθάσουμε στο επίπεδο τέσσερα.

2.5.1 Συγκρίσεις Αξιολογήσεων

Για να εικονοποιήσουμε τη σύγκριση των διαφόρων αποτελεσμάτων των κατατάξεων, χρησιμοποιούμε τα διαγράμματα q-q (quantile-quantile plots), τα οποία

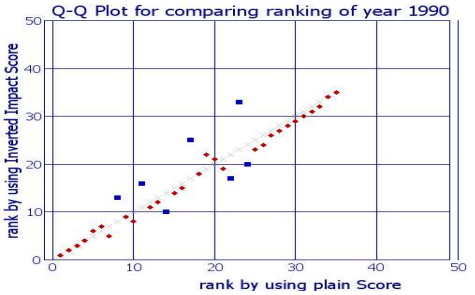
[6] Όλα τα αποτελέσματα είναι προσβάσιμα μέσω Διαδικτύου στη διεύθυνση **http://delab.csd.auth.gr/sceas**



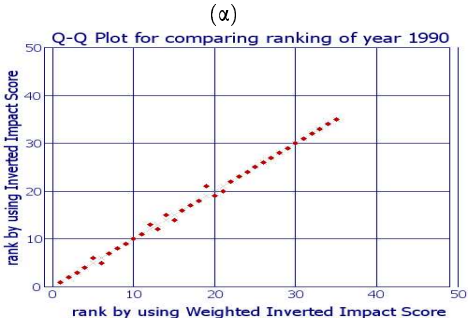
Σχήμα 2.5. Στατιστικά πλήθους επαναλήψεων.

απεικονίζουν τα quantiles μιας univariate κατανομής έναντι των αντιστοίχων quantiles μίας άλλης (στην περίπτωση μας συγκρίνουμε τις κατατάξεις). Επομένως, για να συγκρίνουμε την κατάταξη τύπου A με την κατάταξη τύπου B, για κάθε συνέδριο c στον πίνακα των κατατάξεών μας, τοποθετούμε μία κουκίδα στο διάγραμμα στο σημείο (x, y) όπου x είναι η θέση του c χρησιμοποιώντας τον τύπο A, ενώ y είναι η θέση του c χρησιμοποιώντας τον τύπο B. Επομένως, ο άξονας x αναπαριστά τις θέσεις που υπολογίζονται από τον A και ο άξονας y τις θέσεις που υπολογίζονται από τον B. Οι δύο κατατάξεις θα είναι ισοδύναμες εάν $y = x$ για κάθε σημείο του διαγράμματος. Είναι εύκολο να παρατηρήσει κανείς από τα διαγράμματα q-q (Σχήματα 2.6 και 2.7), ότι τα διάφορα αποτελέσματα δεν διαφέρουν ουσιωδώς για τα *πολύ δυνατά* ή για τα *πολύ αδύναμα* συνέδρια, αλλά κυρίως για τις ομάδες στο μέσο της κλίμακας.

Σε όλα τα διαγράμματα q-q που συγκρίνουν την κατάταξη με Βαθμολογία A-Αντικτύπου (είτε με Βάρη είτε Απλή) και την κατάταξη Βαθμολογίας (είτε με Βάρη, είτε την Απλή) (Σχήματα 2.6(α), 2.7(α) και 2.7(β)), υπάρχουν μερικοί *σκόπελοι* (outliers) (οι οποίοι σημειώνονται ως τετράγωνα (μπλε στην έγχρωμη έκδοση), για τους οποίους ισχύει $x \ll y$. Αυτό σημαίνει ότι κατέχουν πολύ καλύτερη θέση στην κατάταξη που χρησιμοποιεί τη Βαθμολογία παρά αυτήν που χρησιμοποιεί τη Βαθμολογία A-Αντικτύπου. Αυτό οφείλεται στη φύση της έννοιας της Βαθμολογίας Αντικτύπου, όπου μόνο οι αναφορές που έγιναν τα επόμενα k (2 στα δικά μας τεστ) χρόνια λαμβάνονται υπ'όψη. Επομένως, αυτά τα συγκεκριμένα συνέδρια δεν έχουν μεγάλο Αντίκτυπο, που σημαίνει ότι δεν έχουν πολλές



(α)



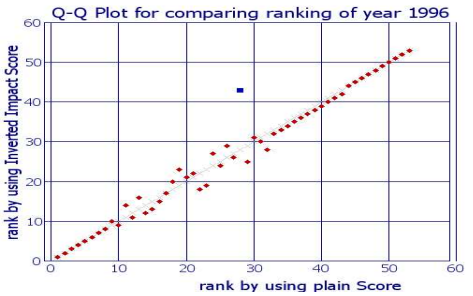
(β)

Σχήμα 2.6. Σύγκριση κατατάξεων συνεδρίων για το έτος 1990.

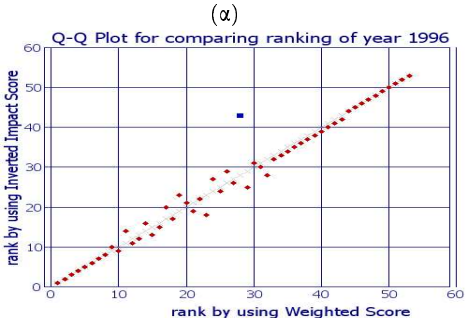
αναφορές κατά τη διάρκεια των επόμενων δύο ετών, αλλά έχουν αναφορές έως *τώρα*.

Συγκεκριμένα, στα Σχήματα 2.7(α) και 2.7(β), ο σκόπελος που βρίσκεται πάνω από τη γραμμή $y = x$ είναι το συνέδριο CPM'96 (Combinatorial Pattern Matching). Το συγκεκριμένο συνέδριο δεν παίρνει καθόλου αναφορές κατά τη διάρκεια των επόμενων 2 ετών, κι έτσι η Βαθμολογία A-Αντικτύπου είναι χαμηλή. Ωστόσο, αν πρέπει να αξιολογήσουμε τη συνολική συνεισφορά του στην ακαδημαϊκή κοινότητα, πρέπει να δούμε την Αξιολόγηση Βαθμολογίας.

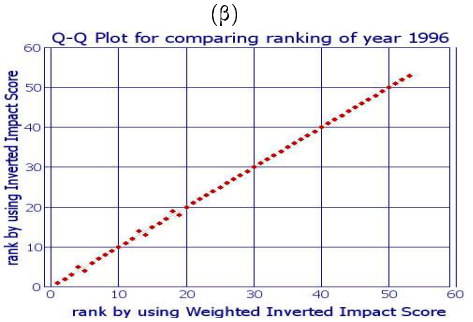
Με ανάλογο τρόπο, οι σκόπελοι για τους οποίους ισχύει $y \ll x$, είναι συνέδρια με μεγάλη *Βαθμολογία A-Αντικτύπου*, αλλά χαμηλή *Βαθμολογία*. Αυτά τα συνέδρια παίρνουν πολλές αναφορές κατά τη διάρκεια των επόμενων k ετών, αλλά οι αναφορές μειώνονται με το χρόνο, πράγμα που σημαίνει ότι δεν περιλαμβάνουν κλασικά αναφερόμενες δημοσιεύσεις (citation classics). Στις περιπτώσεις των Σχημάτων 2.7(β), 2.7(γ) και 2.7(δ), οι σκόπελοι είναι πολύ κοντά στη γραμμή $y = x$ και ποσοτικά λίγοι. Αυτό σημαίνει, ότι δεν υπάρχει δραματική μετατόπιση στην απλή αξιολόγηση προσθέτοντας την έννοια του βάρους, αν και το κλάσμα W^5/W^1 που χρησιμοποιήσαμε είναι υψηλό ($= 5$). Υπάρχουν μερικές ανακατατά-



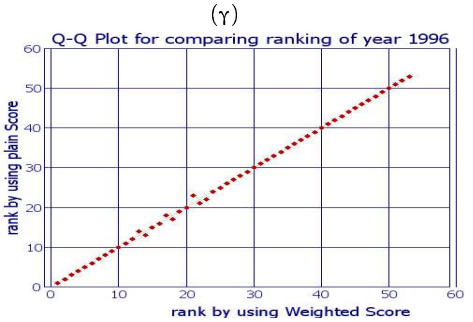
(α)



(β)



(γ)



(δ)

Σχήμα 2.7. Σύγκριση κατατάξεων συνεδρίων για το έτος 1996.

ξεις, οι οποίες βοηθούν στη βελτίωση της αξιολόγησης. Τα συνέδρια του Σχήματος 2.7(γ), για τα οποία ισχύει $x \neq y$, φαίνονται λεπτομερώς στον Πίνακα 2.1. Βλέπουμε ότι το συνέδριο HT (ACM Conference on Hypertext) και το συνέδριο SPIESR (Storage and Retrieval for Image and Video Databases) άλλαξαν θέσεις αφού υπολογίσθηκε η Βαθμολογία με Βάρη. Στην Απλή Βαθμολογία, οι βαθμοί αυτών των δύο συνεδρίων είναι πολύ κοντά. Η Βαθμολογία με Βάρη για το "HT" είναι μεγαλύτερη από αυτήν του "SPIESR", καθώς έχει περισσότερες αναφορές από *πολύ δυνατά* συνέδρια.

Weighted Score		Plain Score		
pos	score	pos	score	conference
13	0.404208	14	0.287565	ACM Conference on Hypertext
14	0.382026	13	0.287772	Storage and Retrieval for Image and Video Databases (SPIE)
17	0.236508	18	0.168072	Database and Expert Systems Applications (DEXA)
18	0.223758	17	0.168552	Digital Libraries
21	0.160647	23	0.110927	Advances in Databases and Information Systems (ADBIS)
22	0.158213	21	0.119178	Australasian Database Conference (ADC)
23	0.154697	22	0.116530	British National Conference on Databases (BNCOD)

Πίνακας 2.1. Σύγκριση του Plain Score με Weighted Score για το έτος 1996.

2.5.2 Σχολιασμός Αποτελεσμάτων

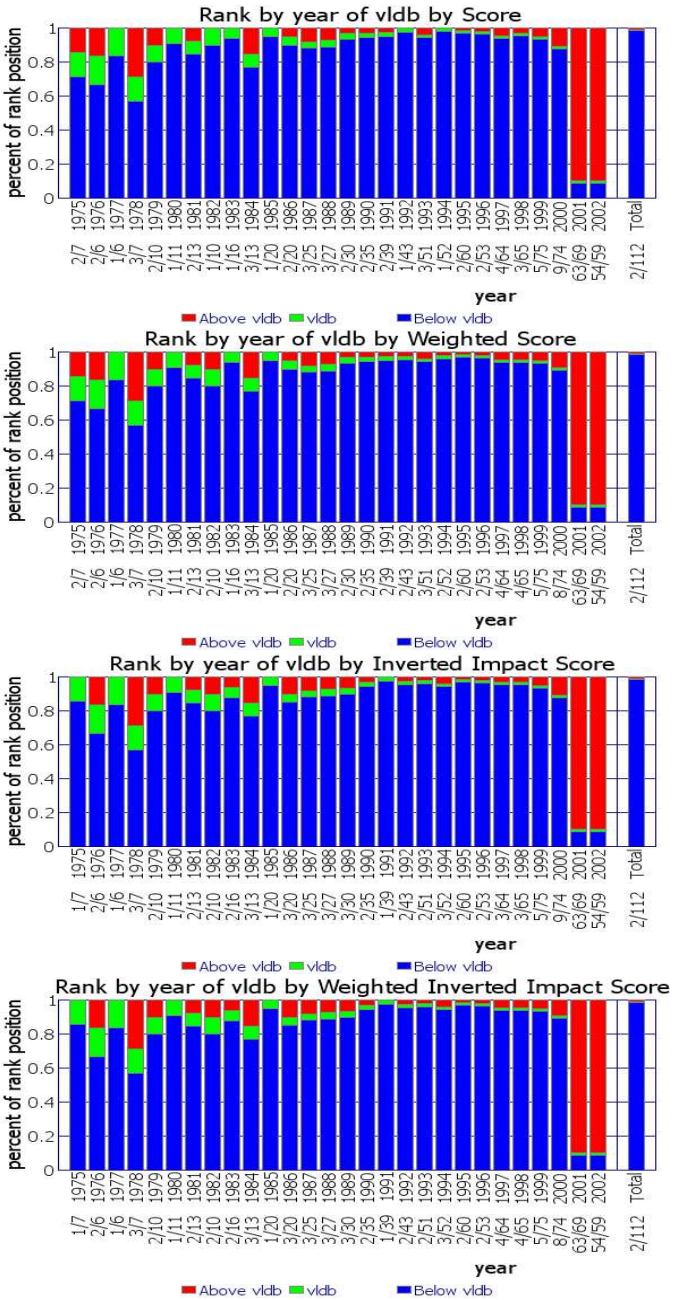
Εκτός από την παρουσίαση των νέων μετρικών μεθόδων για την ανάλυση αναφορών, σε αυτή την εργασία αναφέρουμε κάποιες κατατάξεις για συνέδρια που σχετίζονται με την περιοχή των Βάσεων Δεδομένων[7]. Η παρουσίαση των αποτελεσμάτων της κατάταξης, τα οποία προήλθαν από το σύστημα SCEAS, γίνεται χρησιμοποιώντας δύο τρόπους:

- Με έναν Πίνακα Αξιολόγησης, υποθέτοντας ένα συγκεκριμένο τύπο κατάταξης και ένα επιλεγόμενο έτος. Για παράδειγμα, στον Πίνακα 2.2 παρουσιάζουμε την κατάταξη χρησιμοποιώντας τη Βαθμολογία με Βάρη για το έτος 1995.

- Με ένα Ιστορικό διάγραμμα, όπου μπορούμε να δούμε όλη την ιστορία ενός συνεδρίου για κάθε συγκεκριμένο τύπο κατάταξης.

Στο Σχήμα 2.8, παρουσιάζεται η ιστορία της κατάταξης του συνεδρίου VLDB, σύμφωνα με όλους τους τύπους κατάταξης. Κάθε στήλη αποτελείται από τρία

[7]Πλήρης παρουσίαση των αποτελεσμάτων είναι διαθέσιμη στην ηλεκτρονική διεύθυνση `http://delab.csd.auth.gr/sceas/` . Τα αποτελέσματα που εμφανίζονται εκεί μπορεί να διαφέρουν ελάχιστα με τα παρόντα δεδομένου ότι η βάση ενημερώνεται συχνά.



Σχήμα 2.8. Η ιστορία κατάταξης του συνεδρίου VLDB.

pos	score	#papers	weight	conference
1	10.82818	64	5	ACM SIGMOD Conference
2	7.52698	72	4	Very Large Data Bases (VLDB) Conference
3	6.14474	26	4	Symposium on Principles of Database Systems (PODS)
4	3.71273	27	3	Conf. on Parallel and Distributed Information Systems (PDIS)
5	3.61389	81	3	Int. Conference on Data Engineering (ICDE)
6	2.60681	47	3	Int. Conf. on Extending Database Technology (EDBT)
7	1.65844	17	2	Research Issues in Data Engineering (RIDE)
8	1.19002	74	2	Knowledge Discovery and Data Mining (KDD)
9	0.68100	23	2	Int. Conf. on Cooperative Information Systems (CoopIS)
10	0.59669	28	2	Statistical and Scientific Database Management (SSDBM)

Πίνακας 2.2. Κατάταξη με Weighted Score για το έτος 1996.

μέρη. Το κάτω μέρος (μαύρο, μπλε στην έγχρωμη έκδοση) δίνει το ποσοστό των συνεδρίων που έχουν καταταχθεί χαμηλότερα, το επάνω μέρος (γκρι, κόκκινο στην έγχρωμη έκδοση) δίνει το ποσοστό των συνεδρίων που έχουν καταταχθεί υψηλότερα, ενώ το μεσαίο τμήμα (ανοιχτό γκρι, πράσινο στην έγχρωμη έκδοση) δίνει το ποσοστό των συνεδρίων που έχουν ισοδύναμη κατάταξη. Επιπλέον, η αναλογία κάτω από τον άξονα x δίνει τη σχετική θέση στην κατάταξη για κάθε έτος. Μία διαφορετική τοποθέτηση συμβαίνει για πολλά χρόνια, αλλά όλοι οι γράφοι είναι πολύ παρόμοιοι, αφού το συγκεκριμένο συνέδριο είναι καθαρά ένα *πολύ δυνατό* συνέδριο κατά τη διάρκεια όλων των ετών[8].

Ας σημειώσουμε ότι η κατάταξη των τελευταίων 2 ετών (δηλαδή 2001 και 2002) δεν θα μπορούσε να θεωρηθεί αξιόπιστη, αφού δεν υπάρχουν αναφορές στη βάση μας σε συνέδρια που διοργανώθηκαν κατά τη διάρκεια αυτών των ετών. Οι βαθμολογίες για όλα τα συνέδρια το 2001 και 2002 είναι μηδέν και επομένως, η κατάταξη εξαρτάται από τον αριθμό των δημοσιεύσεων. Συνεπώς, θα πρέπει να αγνοήσουμε τα τελευταία τρία χρόνια από το σκεπτικό της κατάταξης.

2.6 Συμπεράσματα και Μελλοντική Εργασία

Στο κεφάλαιο αυτό αρχικά παρουσιάσαμε μία γενική επισκόπηση των δύο μεγάλων σημερινών συστημάτων που κατατάσουν συνέδρια και περιοδικά, χρησιμοποιώντας την ανάλυση αναφορών, το CiteSeer και το SCI. Ένα αδύναμο σημείο αυτών των συστημάτων είναι ότι βασίζονται στον Παράγοντα Αντικτύπου ISI, κι επομένως, μελετώνται οι αναφορές με έναν επίπεδο τρόπο, π.χ. χωρίς να δίνεται σημασία στην ποιότητα της αντίστοιχης δημοσίευσης. Γι' αυτό παρου-

[8]Τα συνέδρια VLDB και SIGMOD δεν είναι ευθέως συγκρίσιμα με τα υπόλοιπα. Αυτό διότι αυτά περιέχουν *industrial* publications που θα έπρεπε να εκκαθαρισθούν με το χέρι. Άρα στην πραγματικότητα η απόσταση των δυο αυτών συνεδρίων από τα υπόλοιπα είναι ακόμη μεγαλύτερη.

σιάσαμε τέσσερις νέες μετρικές μεθόδους, ώστε να αποκαταστήσουμε αυτή την ανεπάρκεια, οι οποίες είναι κατάλληλες για την κατάταξη των περιοδικών και των συνεδριακών εκδόσεων. Αυτές οι μετρικές μέθοδοι χρησιμοποιούνται από ένα σύστημα που υλοποιήσαμε, το σύστημα SCEAS (Scientific Collection Evaluator by using Advanced Scoring). Το σύστημα είναι αυτόνομο και έχει τα εξής χαρακτηριστικά:

- εισάγει τις βιβλιογραφικές εγγραφές της ψηφιακής βιβλιοθήκης DBLP σε μία τοπική βάση δεδομένων (θα μπορούσε να επεκταθεί ώστε να εισάγει οποιαδήποτε άλλη επιστημονική συλλογή δημοσιεύσεων),
- κατανέμει τη συλλογή που έχει εισαχθεί σε ομάδες, ανάλογα με το θέμα του συνεδρίου και εκτελεί ένα βήμα εκκαθάρισης ώστε να παρέχει αξιόπιστες πληροφορίες,
- υπολογίζει την κατάταξη χρησιμοποιώντας και τις τέσσερις μετρικές μεθόδους για τα συνέδρια που εστιάζουν στις Βάσεις Δεδομένων.

Ο χρήστης του συστήματος SCEAS από τον Παγκόσμιο Ιστό έχει πρόσβαση σε όλα τα αποτελέσματα που παράγονται σε κάθε στάδιο της αξιολογητικής διαδικασίας, μπορεί να συγκρίνει τις διάφορες μετρικές μεθόδους και μπορεί να μελετήσει τα αποτελέσματα της αξιολόγησης προκειμένου να αποκομίσει χρήσιμες πληροφορίες σχετικά με την ποιότητα των συνεδρίων για τις Βάσεις Δεδομένων. Στο μέλλον, το σύστημα μπορεί να επεκταθεί στις εξής κατευθύνσεις:

- Να υπολογίσουμε περισσότερες παραλλαγές των μετρικών μεθόδων με Βάρη όπου οι αυτο-αναφορές της συλλογής θα μπορούν να αποκλείονται ή να λαμβάνονται υπ'όψη πολλαπλασιασμένες με μικρότερο βάρος.
- Να εκτελέσουμε μια λεπτομερή ανάλυση αναφορών κάθε άρθρου και να υπολογίσουμε συγκεντρωτικά αποτελέσματα για κάθε συλλογή.
- Να επεκτείνουμε την κατάταξη/αξιολόγηση σε περισσότερες συλλογές/επιστημονικές περιοχές. Αυτό θα μπορούσε να δώσει τη δυνατότητα, όταν κατατάσουμε μία ομάδα (π.χ. συνέδρια ΒΔ) να παίρνουμε υπ'όψη μας αναφορές "με Βάρος" από άλλου τύπου συλλογές που ανήκουν στην ίδια επιστημονική περιοχή (π.χ. Περιοδικά και Βιβλία ΒΔ).
- Να τροποποιήσουμε την ομαδοποίηση σύμφωνα με όσα αναφέρονται σχετικά στη συγκεκριμένη επιστημονική υποπεριοχή, επιτρέποντας μία οντότητα να είναι μέλος περισσοτέρων του ενός ομάδων.

ΚΕΦΑΛΑΙΟ 3

Κατάταξη Δημοσιεύσεων και Συγγραφέων

Περιεχόμενα

3.1 Εισαγωγή

Η ανάλυση αναφορών βοηθά στον υπολογισμό του αντικτύπου που έχουν οι επιστημονικές συλλογές (δηλαδή, περιοδικά και συνέδρια), οι δημοσιεύσεις και οι συγγραφείς-ερευνητές. Σε αυτό το κεφάλαιο εξετάζουμε γνωστούς αλγορίθμους που χρησιμοποιούνται σήμερα για την Κατάταξη με Ανάλυση Συνδέσμων και παρουσιάζουμε τις αδυναμίες τους σε συγκεκριμένα παραδείγματα. Επίσης, εισάγουμε νέες εναλλακτικές μεθόδους, που έχουν σχεδιασθεί ειδικά για γράφους αναφορών. Για την παρουσίαση αυτών των νέων μεθόδων χρησιμοποιούμε ως βασική πλατφόρμα το σύστημα SCEAS και εκτελούμε μία γενικευμένη σύγκριση όλων των μεθόδων. Επίσης εισάγουμε μία συνάρτηση συνάθροισης (aggregate

function) για τη δημιουργία μίας κατάταξη συγγραφέων βασιζόμενη στη βαθμολογία των δημοσιεύσεων. Τέλος, γίνεται μία προσπάθεια αποτίμησης των αποτελεσμάτων βασιζόμενη στα βραβεία '*VLDB 10 Year Award*', '*SIGMOD Test of Time Award*' και '*SIGMOD E.F.Codd Innovations Award*'.

Οι αλγόριθμοι κατάταξης που χρησιμοποιούνται στη Βιβλιομετρία μπορούν γενικά να χωρισθούν σε δύο κατηγορίες. Στην πρώτη κατηγορία συμπεριλαμβάνονται οι λεγόμενοι *Αλγόριθμοι Κατάταξης επιπέδου Συλλογών*. Σε αυτήν την κατηγορία χρησιμοποιείται ένας γράφος αναφορών με βάρη, όπου οι κόμβοι αναπαριστούν συλλογές, ενώ οι ακμές με βάρη αναπαριστούν το συνολικό αριθμό αναφορών που έγιναν από μία συλλογή σε μία άλλη. Ο Παράγοντας Αντικτύπου ISI ανήκει σε αυτή την κατηγορία [30, 31, 32] και χρησιμοποιεί τα περιοδικά ως συλλογές. Οι συλλογές θα μπορούσαν επίσης να είναι πρακτικά συνεδρίων. Στο [86] καθώς και στο Κεφάλαιο 2 της παρούσης, παρουσιάσαμε μία εναλλακτική μέθοδο ως προς τον Παράγοντα Αντικτύπου, όπου η σημαντικότητα της συλλογής υπολογίζεται από έναν αλγόριθμο ομαδοποίησης (clustering algorithm). Η τελευταία αυτή δουλειά ανήκει σε αυτή την κατηγορία αλγορίθμων. Εναλλακτικά, οι συλλογές μπορούν να είναι τεχνικές αναφορές πανεπιστημίων και ιδρυμάτων, ή οποιοδήποτε άλλο σύνολο δημοσιεύσεων. Υπό αυτή την έννοια, θα μπορούσαν να θεωρηθούν ως συλλογές και οι δημοσιεύσεις ενός συγγραφέα. Επομένως, θα μπορούσαμε να κατατάξουμε συγγραφείς χρησιμοποιώντας αλγορίθμους αυτής της κατηγορίας.

Από την άλλη μεριά, υπάρχει μία κατηγορία Αλγορίθμων *Κατάταξης Επιπέδου Δημοσιεύσεων*. Σύμφωνα με αυτή την προσέγγιση, οι κόμβοι του γράφου αναπαριστούν δημοσιεύσεις, ενώ μία ακμή από τον κόμβο x στον κόμβο y αναπαριστά μία αναφορά από την εργασία x στην εργασία y. Ο υπολογισμός της κατάταξης-βαθμολογίας στο επίπεδο των δημοσιεύσεων έχει το πλεονέκτημα ότι εκτελείται μία μοναδική διαδικασία για να κάνει κατάταξη περισσότερων από μία οντοτήτων: την ίδια την εργασία, τη συλλογή όπου ανήκει και τέλος τους συγγραφείς ταυτοχρόνως. Οι τελευταίοι δύο υπολογισμοί μπορούν να γίνουν από έναν αθροιστικό μέσο όρο του πρώτου υπολογισμού ή χρησιμοποιώντας μία πιο εξελιγμένη συνάρτηση συνάθροισης. Όλοι οι αλγόριθμοι κατάταξης που χρησιμοποιήθηκαν αρχικά για την κατάταξη σελίδων του Παγκόσμιου Ιστού ανήκουν σε αυτή την κατηγορία. Δύο από τους γνωστότερους αλγορίθμους αυτής της κατηγορίας είναι ο `PageRank` των Brin και Page [16] και ο `HITS` του Kleinberg, ο οποίος κατατάσσει τα στοιχεία ως `Hubs` και `Authorities` [51, 52, 53].Ένας άλλος ευρέως αποδεκτός αλγόριθμος είναι ο `SALSA`, που αποτελεί μία παραλλαγή του `HITS` [59].

Σε αυτό το κεφάλαιο θα επικεντρωθούμε στη δεύτερη κατηγορία αλγορίθμων κατάταξης. Πιο συγκεκριμένα, η δομή του κεφαλαίου είναι η εξής. Στην επόμενη παράγραφο παρουσιάζουμε γνωστούς αλγορίθμους που έχουν χρησιμοποιηθεί για την κατάταξη δημοσιεύσεων ή ιστοσελίδων. Επίσης, κάνουμε μία ανακεφαλαίωση των αδύναμων σημείων τους όταν αυτοί χρησιμοποιούνται στη Βιβλιομετρία. Στην Παράγραφο 3.3 παρουσιάζουμε ένα σύνολο παραδειγμάτων κατάταξης συζητώντας τις αδυναμίες τους και τα δυνατά τους σημεία. Πιο συγκεκριμένα, εξετάζουμε διάφορες παραλλαγές του αλγορίθμου `SCEASRank`, που παρουσιάστηκε στο [83], όπως και παραλλαγές των αλγορίθμων `HITS` και `SALSA`. Στην Παράγραφο 3.4 παρουσιάζουμε τα πειράματα που εκτελέσαμε. Συγκεκριμένα, συγκρίνουμε την ταχύτητα υπολογισμού όλων των αλγορίθμων και τα αποτελέσματα των κατατάξεων, τα οποία συγκρίνονται χρησιμοποιώντας διάφορες μεθόδους. Το σύνολο δεδομένων που χρησιμοποιούμε είναι τα περιεχόμενα της βιβλιοθήκης `SCEAS`. Τέλος, στην Παράγραφο 3.5 κάνουμε μία υποθετική εκτίμηση των αποτελεσμάτων με το εξής κριτήριο: αν έπρεπε να αποφασίσουμε για τα βραβεία του 'VLDB 10 Year Award' και 'SIGMOD Test of Time Award' χρησιμοποιώντας έναν από τους προαναφερθέντες αλγορίθμους, θα είμαστε σε θέση να βραβεύσουμε τις σωστές (ίδιες) δημοσιεύσεις και τους σωστούς συγγραφείς; Στο δεύτερο τμήμα της παραγράφου αυτής παρουσιάζεται μία συναθροιστική συνάρτηση βαθμολόγησης συγγραφέων. Η κατάταξη συγγραφέων υπολογίζεται χρησιμοποιώντας ως δεδομένα εισόδου την κατάταξη-βαθμολογία δημοσιεύσεων. Η εκτίμηση των αποτελεσμάτων γίνεται συγκρίνοντάς τα με τη λίστα βραβευθέντων με το 'SIGMOD E.F.Codd Innovations Award'. Η τελευταία παράγραφος είναι συμπερασματική του κεφαλαίου.

3.2 Μέθοδοι Κατάταξης

Σε αυτήν την παράγραφο παρουσιάζουμε γνωστούς αλγορίθμους που χρησιμοποιούνται για την κατάταξη-αξιολόγηση ιστογράφων. Αυτοί οι αλγόριθμοι θα μπορούσαν επίσης να χρησιμοποιηθούν στη Βιβλιομετρία για αξιολόγηση δημοσιεύσεων, η οποία βασίζεται σε γράφους αναφορών. Σε όλο το κεφάλαιο χρησιμοποιούμε τα σύμβολα του Πίνακα 3.1 ώστε να παρουσιάσουμε τους αλγορίθμους με έναν κοινό τρόπο.

I_x : Το σύνολο των δημοσιεύσεων που αναφέρουν το x
$\|I_x\|$: Το πλήθος των δημοσιεύσεων που αναφέρουν το x (Έσω-βαθμός x)
O_x : Το σύνολο των δημοσιεύσεων που αναφέρονται από το x
$\|O_x\|$: Το πλήθος των δημοσιεύσεων που αναφέρονται από το x (Έξω-βαθμός του x)
d : Συντελεστής Τυχαίας Μετάβασης - Damping Factor (0.85 για τον PageRank)
b : Σημαντικότητα της Αναφοράς (συνήθως ίση με 1)
a : Εκθετικός Παράγοντας (> 1, συνήθως ίσος με ϵ)

Πίνακας 3.1. Συμβολισμοί.

3.2.1 Καταμέτρηση Αναφορών

Η κατάταξη των δημοσιεύσεων με τη μέθοδο της καταμέτρησης των εισερχόμενων αναφορών είναι ο απλούστερος και ταχύτερος τρόπος. Αναφερόμαστε σε αυτό τον αλγόριθμο ως Καταμέτρηση Αναφορών (Citation Count) (CC). Επομένως, η βαθμολογία μίας δημοσίευσης x είναι ο έσω-βαθμός (in-degree) του κόμβου x:

$$CC_x = |I_x| \tag{3.1}$$

Αυτή η κατάταξη, χωρίς χρήση βαρών είναι η μέθοδος που έχει χρησιμοποιηθεί για πολλά χρόνια. Ωστόσο, η προσέγγιση αυτή είναι υπό αμφισβήτηση καθώς δεν θα έπρεπε όλες οι αναφορές να προσμετρώνται με τον ίδιο τρόπο. Για παράδειγμα, όταν μία δημοσίευση αναφέρεται από καλές δημοσιεύσεις, τότε θα έπρεπε να παίρνει καλύτερη βαθμολογία. Αυτός είναι ο λόγος που δημιουργήθηκαν πολλοί αλγόριθμοι κατάταξης.

3.2.2 Ισορροπημένη Καταμέτρηση Αναφορών

Μία άλλη απλή μέθοδος κατάταξης είναι η Ισορροπημένη Καταμέτρηση Αναφορών (Balanced Citation Count) (BCC). Σε αυτό το μοντέλο οι αναφορές δεν προσμετρώνται ισοδύναμα, αλλά η σπουδαιότητά τους είναι συνάρτηση του έξω-βαθμού του κόμβου που κάνει τις αναφορές:

$$BCC_x = \sum_{\forall y \in I_x} \frac{1}{|O_y|} \tag{3.2}$$

Εδώ, το βάρος μίας αναφοράς από τον κόμβο y είναι ίσο με $\frac{1}{|O_y|}$ και όχι 1, όπως θα ήταν στη μέθοδο CC (Εξίσωση 3.1). Αυτό σημαίνει ότι η ενίσχυση που δίνει ένας κόμβος y στους κόμβους που δείχνει, αθροίζεται στο 1 παρά στο $|O_y|$, όπως είναι στην CC. Επομένως, όλες οι βαθμολογίες BCC_x αθροίζονται σε $|V|$, που είναι ο αριθμός των κόμβων του γράφου. Ωστόσο, αυτή η μέθοδος έχει τα

ίδια μειονεκτήματα με την CC. Δεν υπάρχουν βάρη για να αντιπροσωπεύσουν τη σπουδαιότητα των δημοσιεύσεων που κάνουν την αναφορά.

3.2.3 PageRank

Ο PageRank λαμβάνει υπ'όψη τη σπουδαιότητα των δημοσιεύσεων που κάνουν τις αναφορές. Αρχικά, η βαθμολογία στον PageRank ορίσθηκε ως εξής [16]:

$$PR(A) \;=\; (1-d) \;+\; d\,(\frac{PR(t1)}{C(t1)} + \ldots + \frac{PR(tn)}{C(tn)})$$

όπου $t1, \ldots, tn$ είναι οι σελίδες που περιέχουν υπερσυνδέσμους προς τη σελίδα A, C είναι ο αριθμός των εξερχόμενων υπερσυνδέσμων από μία σελίδα (έξω-βαθμός/out-degree) και d είναι ένας συντελεστής τυχαίας μετάβασης (damping factor), που συνήθως τίθεται ίσος με 0.85. Χρησιμοποιώντας τα σύμβολα του Πίνακα 3.1, η τελευταία εξίσωση είναι ισοδύναμη με:

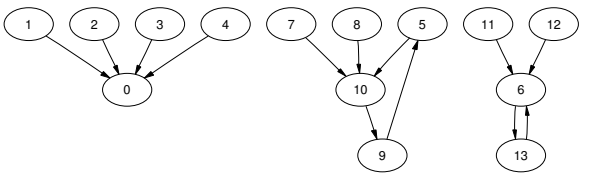
$$PR_x \;=\; (1-d) \;+\; d \sum_{\forall y \in I_x} \frac{PR_y}{|O_y|} \tag{3.3}$$

Με απλά λόγια, ο PageRank αναθέτει υψηλή βαθμολογία σε έναν κόμβο, αν αυτός δείχνεται από κόμβους που έχουν καταταχθεί υψηλά.

Εξ' 0ορισμού, ο PageRank δίνει υψηλή βαθμολογία σε έναν κόμβο x, όταν υπάρχει ένα μεγάλο συνδεδεμένο τμήμα C, που κάποιοι από τους κόμβους του δείχνουν στο x. Όσο περισσότεροι και μεγαλύτεροι κύκλοι περιέχονται στο C, τόσο μεγαλύτερη βαθμολογία θα πάρει ο x. Αυτό συμβαίνει επειδή υπάρχει αυτο-τροφοδότηση από τους κόμβους που ανήκουν σε κύκλους.

Στη Βιβλιομετρία, στην ασυνήθη περίπτωση ύπαρξης κύκλων, κυρίως αυτοί αντιπροσωπεύουν αυτο-αναφορές (self-citations). Κατά συνέπεια δεν είναι λογικό να επιτρέψουμε τις αυτο-αναφορές να επηρεάζουν τη βαθμολογία. Από την άλλη μεριά, η εξάλειψη των κύκλων θα άλλαζε τα αποτελέσματα. Για παράδειγμα, ο Πίνακας 3.2 δείχνει τα αποτελέσματα κατάταξης του γράφου του Σχήματος 3.1 (ο οποίος αποτελείται από τρία συνδεδεμένα μέρη). Παρατηρούμε ότι ο κόμβος 0 λαμβάνει 4 αναφορές, ενώ οι κόμβοι 10 και 6 παίρνουν ο καθένας από 3 αναφορές. Ωστόσο, η βαθμολογία του PageRank για τους κόμβους 10 και 6 είναι περίπου κατά 3 φορές υψηλότερη από τη βαθμολογία του κόμβου 0. Αυτό συμβαίνει επειδή οι κόμβοι 10 και 6 είναι μέλη κύκλων αναφορών.

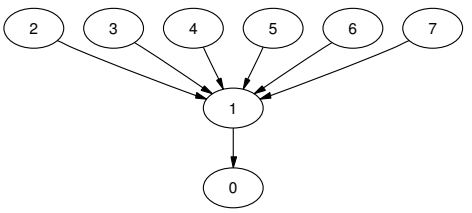
Ένα δεύτερο χαρακτηριστικό του PageRank είναι ότι είναι σχεδιασμένος (ειδικά για τον Παγκόσμιο Ιστό) έτσι ώστε η βαθμολογία μίας σελίδας να επηρεάζεται περισσότερο από τις βαθμολογίες των σελίδων που δείχνουν σε αυτή και



Σχήμα 3.1. Παράδειγμα τριών γράφων.

CC	PR	HA	BHA	SA	BSA	P	PS	BPS	EPS	BEPS	SCEAS
0 4.000	6 1.919	0 1.000	0 1.000	0 0.658	13 0.803	5 0.333	6 2.757	13 39.000	0 1.150	0 1.472	0 1.472
6 3.000	13 1.781	6 0.000	6 0.000	6 0.493	6 0.596	13 0.333	10 2.506	6 37.000	6 1.030	6 1.433	6 1.433
10 3.000	10 1.661	10 0.000	13 0.000	10 0.493	10 0.000	6 0.111	0 2.183	5 33.000	10 0.993	10 1.356	10 1.356
9 1.000	9 1.562	5 0.000	10 0.000	5 0.164	9 0.000	9 0.111	13 2.051	9 31.000	13 0.584	13 0.895	13 0.895
5 1.000	5 1.477	9 0.000	9 0.000	9 0.164	5 0.000	10 0.111	9 1.913	10 31.000	9 0.573	9 0.867	9 0.867
13 1.000	0 0.660	13 0.000	5 0.000	13 0.164	0 0.000	0 0.000	5 1.590	0 4.000	5 0.452	5 0.687	5 0.687
8 0.000	1 0.150	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000	1 0.000
7 0.000	2 0.150	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000
4 0.000	3 0.150	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000
3 0.000	4 0.150	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000
2 0.000	7 0.150	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000
12 0.000	8 0.150	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000
11 0.000	11 0.150	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000
1 0.000	12 0.150	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000

Πίνακας 3.2. Αποτελέσματα κατάταξης του γράφου του Σχήματος 3.1.



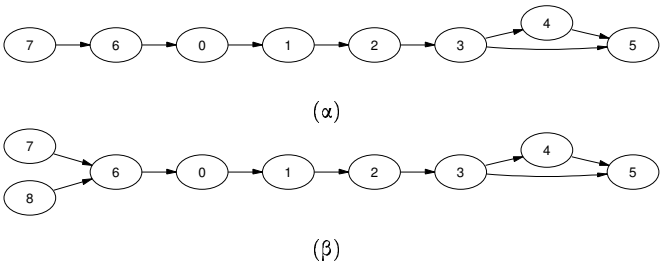
Σχήμα 3.2. Δεύτερο Παράδειγμα γράφων.

CC	PR	HA	BHA	SA	BSA	P	PS	BPS	EPS	BEPS	SCEAS
1 6.000	0 0.928	1 1.000	1 0.973	1 0.986	0 1.000	0 0.000	1 3.865	0 7.000	1 1.763	1 2.207	1 2.207
0 1.000	1 0.915	0 0.000	0 0.233	0 0.164	1 0.002	1 0.000	0 3.135	1 6.000	0 0.812	0 1.180	0 1.180
7 0.000	2 0.150	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000
6 0.000	3 0.150	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000	3 0.000
5 0.000	4 0.150	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000
4 0.000	5 0.150	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000
3 0.000	6 0.150	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000
2 0.000	7 0.150	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000

Πίνακας 3.3. Αποτελέσματα κατάταξης του γράφου του Σχήματος 3.2.

λιγότερο από τον αριθμό των εισερχόμενων συνδέσμων (δηλαδή στην περίπτωση μας από τον αριθμό των αναφορών). Για παράδειγμα, στον Πίνακα 3.3 (σε συνδυασμό με το Σχήμα 3.2), ο κόμβος 0 λαμβάνει μεγαλύτερη βαθμολογία από τον κόμβο 1, αν και ο κόμβος 1 δέχεται 6 αναφορές. Αυτό είναι ένα άλλο αδύναμο σημείο στην περίπτωση της Βιβλιομετρίας.

Ένα τρίτο χαρακτηριστικό είναι ότι μία αλλαγή στη βαθμολογία του κόμβου



Σχήμα 3.3. Τρίτο Παράδειγμα γράφων.

CC	PR	HA	BHA	SA	BSA	P	PS	BPS	EPS	BEPS	SCEAS
5 2.000	5 0.767	5 0.851	5 0.909	5 0.496	5 0.935	0 0.000	5 2.302	5 7.000	5 0.773	5 0.765	5 0.765
6 1.000	3 0.623	4 0.526	4 0.416	0 0.372	4 0.355	1 0.000	4 1.144	3 5.000	4 0.386	3 0.578	3 0.578
4 1.000	2 0.556	0 0.000	3 0.000	1 0.372	3 0.000	2 0.000	3 1.120	2 4.000	3 0.386	2 0.571	2 0.571
3 1.000	1 0.478	1 0.000	2 0.000	2 0.372	2 0.000	3 0.000	2 1.074	1 3.000	2 0.384	1 0.553	1 0.553
2 1.000	4 0.415	2 0.000	1 0.000	3 0.372	1 0.000	4 0.000	1 0.989	4 3.000	1 0.378	0 0.503	0 0.503
1 1.000	0 0.386	3 0.000	0 0.000	6 0.372	0 0.000	5 0.000	0 0.831	0 2.000	0 0.357	6 0.368	6 0.368
0 1.000	6 0.278	6 0.000	6 0.000	4 0.248	6 0.000	6 0.000	6 0.540	6 1.000	6 0.279	4 0.290	4 0.290
7 0.000	7 0.150	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000

Πίνακας 3.4. Αποτελέσματα κατάταξης του γράφου του Σχήματος 3.3(α).

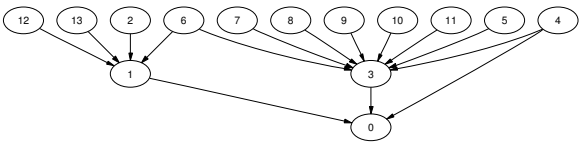
CC	PR	HA	BHA	SA	BSA	P	PS	BPS	EPS	BEPS	SCEAS
6 2.000	5 0.820	5 0.851	5 0.909	6 0.626	5 0.935	0 0.000	5 2.287	5 8.000	5 0.769	5 0.767	5 0.767
5 2.000	3 0.689	4 0.526	4 0.416	5 0.417	4 0.355	1 0.000	4 1.143	3 6.000	6 0.555	6 0.736	6 0.736
4 1.000	2 0.635	6 0.000	6 0.000	0 0.313	3 0.000	2 0.000	3 1.140	2 5.000	0 0.432	0 0.639	0 0.639
3 1.000	1 0.570	0 0.000	3 0.000	1 0.313	2 0.000	3 0.000	2 1.134	1 4.000	1 0.397	1 0.603	1 0.603
2 1.000	0 0.494	1 0.000	0 0.000	2 0.313	1 0.000	4 0.000	1 1.124	4 3.500	2 0.388	2 0.590	2 0.590
1 1.000	4 0.443	2 0.000	1 0.000	3 0.313	0 0.000	5 0.000	0 1.104	0 3.000	3 0.385	3 0.585	3 0.585
0 1.000	6 0.405	3 0.000	2 0.000	4 0.209	6 0.000	6 0.000	6 1.068	6 2.000	4 0.385	4 0.292	4 0.292
8 0.000	7 0.150	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000
7 0.000	8 0.150	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000

Πίνακας 3.5. Αποτελέσματα κατάταξης του γράφου του Σχήματος 3.3(β).

j επηρεάζει τη βαθμολογία του κόμβου i, ακόμα κι αν το συνδετικό μονοπάτι μεταξύ τους είναι πολύ μακρύ. Και αυτή η περίπτωση δεν θα έπρεπε να προκύπτει στη Βιβλιομετρία. Με άλλα λόγια, η πρόσθεση ενός νέου κόμβου (δηλαδή, μιας δημοσίευσης) στο γράφο θα πρέπει να επηρεάζει κυρίως τους γειτονικούς κόμβους (δηλαδή, τις δημοσιεύσεις που αναφέρει αυτή η νέα δημοσίευση). Από την άλλη μεριά, οι βαθμολογίες των υπολοίπων κόμβων (οι γείτονες των γειτόνων κ.τ.λ.) θα πρέπει να επηρεάζονται πολύ λιγότερο από αυτή τη νέα προσθήκη. Για παράδειγμα, ας θεωρήσουμε το γράφο του Σχήματος 3.3(α), όπου ένας νέος κόμβος προστίθεται με έναν επιπλέον σύνδεσμο στον κόμβο 6, έχοντας ως αποτέλεσμα το Σχήμα 3.3(β). Σε αυτή την περίπτωση, η βαθμολογία του κόμβου 5 αυξάνεται κατά 7% και η βαθμολογία του κόμβου 4 αυξάνεται κατά 6.8%, αν και βρίσκονται μακριά ο ένας από τον άλλον κατά 5 κόμβους. Αυτή είναι μία άλλη περίπτωση όπου ο PageRank έχει αποδειχθεί να μην συμπεριφέρεται ομαλά σε ένα

βιβλιογραφικό περιβάλλον.

Στον Πίνακα 3.6 (Σχήμα 3.4) ο PageRank κατατάσσει τον κόμβο 0 πρώτο και τον κόμβο 3 δεύτερο. Για ένα γράφο του Παγκόσμιου Ιστού αυτό είναι ένα λογικό αποτέλεσμα. Ωστόσο, σε ένα γράφο αναφορών ο κόμβος 3 θα έπρεπε να είναι πρώτος στον πίνακα κατάταξης.



Σχήμα 3.4. Τελευταίο παράδειγμα γράφου.

CC	PR	HA	BHA	SA	BSA	P	PS	BPS	EPS	BEPS	SCEAS
3 8.000	0 1.607	3 0.960	3 0.886	3 0.848	0 0.797	0 0.000	0 7.130	0 13.000	3 2.378	3 2.575	3 2.575
1 4.000	3 1.043	1 0.218	0 0.424	1 0.424	3 0.587	1 0.000	3 5.247	3 7.000	0 1.952	0 2.341	0 2.341
0 3.000	1 0.596	0 0.177	1 0.189	0 0.318	1 0.143	2 0.000	1 2.623	1 3.500	1 1.189	1 1.288	1 1.288
9 0.000	2 0.150	2 0.000	2 0.000	2 0.000	2 0.000	3 0.000	2 0.000	2 0.000	2 0.000	2 0.000	2 0.000
8 0.000	4 0.150	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000	4 0.000
7 0.000	5 0.150	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000	5 0.000
6 0.000	6 0.150	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000	6 0.000
5 0.000	7 0.150	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000	7 0.000
4 0.000	8 0.150	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000	8 0.000
2 0.000	9 0.150	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000	9 0.000
13 0.000	10 0.150	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000	10 0.000
12 0.000	11 0.150	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000	11 0.000
11 0.000	12 0.150	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000	12 0.000
10 0.000	13 0.150	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000	13 0.000

Πίνακας 3.6. Αποτελέσματα κατάταξης του γράφου του Σχήματος 3.4.

3.2.4 HITS

Ο HITS προτάθηκε για την κατάταξη σελίδων του Παγκόσμιου Ιστού, οι οποίες ανακτώνται κατά την αναζήτηση μέσω ενός φυλλομετρητή ιστοσελίδων (web browser). Η ιδέα πίσω από τον HITS είναι η διάκριση μεταξύ *εστιακών σημείων* (hubs) και *αυθεντιών* (authorities). Τα εστιακά σημεία είναι σελίδες με καλούς συνδέσμους, ενώ οι αυθεντίες είναι σελίδες με καλό περιεχόμενο. Ένας κόμβος μπορεί να είναι εστιακό σημείο ή αυθεντία. Επομένως, ο HITS υπολογίζει δύο διανύσματα βαθμολογιών. Αρχικά, οι βαθμολογίες για τα εστιακά σημεία και τις αυθεντίες είχαν ορισθεί ως [52]:

$$\begin{aligned} \vec{a'} &= A^T * \vec{h} \\ \vec{h'} &= A * \vec{a} \end{aligned}$$

όπου A είναι ο πίνακας γειτνίασης (adjacency matrix) του γράφου αναφορών, με $A_{i,j} = 1$ εάν η δημοσίευση i κάνει αναφορά στη δημοσίευση j, και μηδέν διαφο-

ρετικά. Το $\vec{a}$ είναι ένα διάνυσμα όπου το i-οστό στοιχείο του αντιπροσωπεύει τη βαθμολογία μίας δημοσίευσης ως αυθεντίας, ενώ το διάνυσμα $\vec{h}$ περιέχει τις βαθμολογίες των κόμβων που είναι εστιακά σημεία. Χρησιμοποιώντας την ορολογία του Πίνακα 3.1, οι βαθμολογίες των `HITS Authority` (`HA`) και `HITS Hub` (`HH`) μπορούν να υπολογισθούν ως:

$$\begin{aligned} HA_x &= \sum_{\forall y \in I_x} HH_y \\ HH_x &= \sum_{\forall y \in O_x} HA_y \end{aligned} \tag{3.4}$$

Τυπικά, ο `HITS` υπολογίζεται επάνω σε ένα υποσύνολο ενός γράφου, το οποίο είναι το σύνολο που προέκυψε από την αναζήτηση ενός χρήστη. Στην περίπτωση της Βιβλιομετρίας, θα μπορούσαμε να χρησιμοποιήσουμε ως σύνολο για εφαρμογή στον `HITS` ολόκληρο το γράφο αναφορών και να κατατάξουμε όλες τις δημοσιεύσεις. Πρακτικά, η κατάταξη-αξιολόγηση του `HITS` δεν είναι κατάλληλη για τον τομέα της Βιβλιομετρίας. Ο λόγος είναι ότι οι δημοσιεύσεις λαμβάνουν υψηλές βαθμολογίες αυθεντικότητας, εάν υπάρχουν εστιακά σημεία που δείχνουν σε αυτές. Επιπρόσθετα, όπως απέδειξε ο Borodin et al. στο [12], ένα εστιακό σημείο τιμωρείται όταν δείχνει κάποια *φτωχή* αυθεντία. Αυτό επίσης παρουσιάζεται στο Σχήμα 3.1, όπου η βαθμολογία αυθεντικότητας του κόμβου 0 συγκλίνει στο 1, ενώ άλλες αυθεντίες παίρνουν βαθμολογία 0. Η ίδια συμπεριφορά του `HITS` εμφανίζεται στον Πίνακα 3.3 (Σχήμα 3.2), όπου η βαθμολογία του κόμβου 1 συγκλίνει στο 1.

3.2.5 Prestige

Ειδικά για τη Βιβλιομετρία, ο Kleinberg στο [52] προτείνει ένα μοντέλο, όπου οι αυθεντίες *τροφοδοτούν* απ'ευθείας άλλες αυθεντίες. Αυτό είναι ό,τι ο Chakrabarti ονομάζει `Prestige` [17]. Ο τρόπος υπολογισμού της βαθμολογίας με αυτή τη μέθοδο είναι:

$$\vec{a'} = A^T * \vec{a}$$

ενώ σύμφωνα με τους δικούς μας συμβολισμούς, η προηγούμενη έκφραση είναι ισοδύναμη με:

$$P_x = \sum_{\forall y \in I_x} P_y \tag{3.5}$$

Από την Εξίσωση 3.5 είναι προφανές ότι στην ιδεατή περίπτωση που δεν υπάρχουν κύκλοι στο γράφο αναφορών:

- Οι τιμές του $\vec{P}$ συγκλίνουν στο $\vec{0}$.

- Ακόμα κι αν βρούμε έναν τρόπο ώστε το $\vec{P}$ να μην συγκλίνει στο μηδέν, ο κόμβος x ποτέ δεν θα πάρει μεγαλύτερη βαθμολογία από τον κόμβο y, αν ο x δείχνει στον y. Η απόδειξη αυτής της προτάσεως είναι απλή, αφού δεν υπάρχουν αρνητικές τιμές στο διάνυσμα βαθμολογιών. Επομένως, στην καλύτερη περίπτωση θα είναι $P_x = P_y$, αλλά στη γενική και πλέον κοινή περίπτωση θα ισχύει: $P_x \leq P_y$.

Από όλα τα παραδείγματα που παρουσιάσθηκαν έως τώρα, βλέπουμε ότι ο `Prestige` δίνει βαθμό κατάταξης μόνο σε κόμβους που ανήκουν σε κύκλους ή σε κόμβους που δείχνονται από μέλη κύκλων. Η βαθμολογία όλων των υπόλοιπων κόμβων συγκλίνει στο 0.

3.2.6 SALSA

Ο αλγόριθμος `SALSA`, που προτάθηκε από τους Lempel και Moran [59], είναι μία παραλλαγή του `HITS` καθώς χρησιμοποιεί μία ισορροπία με βάση τους έσω- και έξω-βαθμούς. Η διατύπωση της βαθμολογίας του `SALSA` χρησιμοποιώντας τους δικούς μας συμβολισμούς είναι:

$$\begin{aligned} SA_x &= \sum_{\forall y \in I_x} \frac{SH_y}{|O_y|} \\ SH_x &= \sum_{\forall y \in O_x} \frac{SA_y}{|I_y|} \end{aligned} \tag{3.6}$$

Ο `SALSA` συμπεριφέρεται πολύ καλύτερα από τον `HITS` στις περιπτώσεις των Σχημάτων 3.1, 3.2 και 3.4. Από την άλλη πλευρά, όπως μπορούμε να δούμε στο παράδειγμα του Σχήματος 3.3, ο κόμβος 6 έρχεται πρώτος στον πίνακα κατάταξης μετά την πρόσθεση μίας αναφοράς.

3.3 Οι Νέες Μέθοδοι Κατάταξης

Κάνοντας μία περίληψη των αδυναμιών των προηγούμενων αλγορίθμων παρατηρούμε ότι:

- Ο `CC` και ο `BCC` δεν λαμβάνουν υπ'όψη τη σημασία των αναφερόμενων δημοσιεύσεων

- Ο `Prestige` δεν έχει καλή συμπεριφορά όταν υπάρχουν κύκλοι αναφορών, οπότε η βαθμολογία αυτών που δεν ανήκουν σε κύκλους συγκλίνει στο μηδέν.

- Ο PageRank έχει το ίδιο πρόβλημα με τον Prestige. Τα μέλη κύκλων επιτυγχάνουν υψηλότερες βαθμολογίες.
- Ο HITS και ο SALSA βασίζονται στην έννοια των εστιακών σημείων και αυθεντιών, η οποία δεν είναι κατάλληλη για την κατάταξη-βαθμολόγηση δημοσιεύσεων.

Έχοντας υπ'όψη τις παρατηρήσεις αυτές, εν συνεχεία θα ορίσουμε κάποιες νέες μεθόδους κατάταξης, οι οποίες είναι κατάλληλες για την κατάταξη δημοσιεύσεων.

3.3.1 B-HITS

Για να προσπεράσουμε την αδυναμία του HITS, προτείνουμε τον αλγόριθμο B-HITS (Balanced HITS). Σύμφωνα με αυτή την εκδοχή, ένας κόμβος δεν ενισχύεται μόνο από συνδέσμους που προέρχονται από εστιακά σημεία, αλλά και από συνδέσμους από άλλες αυθεντίες. Έτσι ουσιαστικά αλλάζουμε την αρχική έννοια της *αυθεντίας* σε *αξιοσημείωτα στοιχεία ή στοιχεία με αξία* και επομένως, η φόρμουλα υπολογισμού των βαθμολογιών γίνεται:

$$\begin{aligned} \vec{a'} &= (1-p) * A^T \vec{h} + p * A^T \vec{a} \\ \vec{h'} &= A * \vec{a} \end{aligned}$$

όπου το p είναι το ποσοστό της ενίσχυσης των αυθεντιών σε άλλες αυθεντίες. Χρησιμοποιώντας την ορολογία μας, οι εξισώσεις αυτές είναι ισοδύναμες προς:

$$\begin{aligned} BHA_x &= (1-p) * \sum_{\forall y \in I_x} BHH_y + p * \sum_{\forall y \in I_x} BHA_y \\ BHH_x &= \sum_{\forall y \in O_x} BHA_y \end{aligned} \tag{3.7}$$

Σε αυτό το σημείο, για λόγους συντομίας αποφεύγουμε την εύρεση μίας βέλτιστης τιμής για το p, αν και χρειάζεται επιπλέον διερεύνηση (πιθανόν να βασίζεται στα χαρακτηριστικά του γράφου). Υποθέτοντας μία δίκαιη εξισορρόπηση, χρησιμοποιήσαμε στα πειράματα μας $p = 0.5$.

3.3.2 B-SALSA

Η εξισορρόπηση της ενίσχυσης στις αυθεντίες από τις αυθεντίες και από τα εστιακά σημεία του B-HITS μπορεί επίσης να χρησιμοποιηθεί και στον SALSA. Η εξίσωση που προκύπτει είναι:

$$\begin{aligned} BSA_x &= (1-p) * \sum_{\forall y \in I_x} \frac{BSH_y}{|O_y|} + p * \sum_{\forall y \in I_x} \frac{BSA_y}{|O_y|} \\ BSH_x &= \sum_{\forall y \in O_x} \frac{BSA_y}{|I_y|} \end{aligned} \tag{3.8}$$

Ουσιαστικά, αυτό θα προκαλέσει τον SALSA πρακτικά να συμπεριφερθεί σαν τον PageRank όπως θα δούμε και στη συνέχεια.

3.3.3 SCEAS PS: Απλή Βαθμολογία Δημοσίευσης

Στο σύστημά μας SCEAS [86], η Απλή Βαθμολογία Δημοσίευσης (Plain Score - PS) ενός κόμβου x είναι ίση με το άθροισμα των βαθμολογιών ($PS+b$) όλων των κόμβων που δείχνουν απ'ευθείας στο x. Επομένως, κάθε αναφορά στο x από μία δημοσίευση y δίνει ένα σταθερό παράγοντα b συν τη βαθμολογία PS_y προκειμένου να αναπαραστήσει τη σπουδαιότητά της. Έτσι, η εξίσωση που προκύπτει είναι:

$$PS_x = \sum_{\forall y \in I_x} (b + PS_y) \tag{3.9}$$

Αυτή η προσέγγιση, που ονομάζεται PS, είναι υβριδική μεταξύ του CC και του Prestige, όπου ο CC πολλαπλασιάζεται κατά τον παράγοντα b (δηλαδή $\vec{PS} = b * \vec{CC} + \vec{P}$). Ο Prestige έχει το μειονέκτημα ότι εν απουσία κύκλων συγκλίνει στο μηδέν. Χρησιμοποιώντας την PS επιλύουμε το πρόβλημα του Prestige.

Αν δεν υπάρχουν κύκλοι στο γράφο, τότε η PS θα μπορούσε να υπολογισθεί αναδρομικά. Σε αυτή την περίπτωση, αν ο κόμβος x δείχνει στον y, τότε ο κόμβος x ποτέ δεν θα λάβει βαθμολογία μεγαλύτερη από του y. Για παράδειγμα, για $b = 1$ στο Σχήμα 3.2 το διάνυσμα που προκύπτει θα έπρεπε να είναι $\vec{PS} = (7, 6, 0, 0, 0, 0, 0, 0)$, αφού $PS_1 = 6 * b = 6$ και $PS_0 = PS_1 + b = 7$.

Συνήθως όμως υπάρχουν κύκλοι. Σε μία τέτοια περίπτωση, η PS, όπως ο HITS και ο Prestige, χρειάζεται ένα βήμα κανονικοποίησης ώστε να συγκλίνει. Χωρίς κανονικοποίηση, όλοι αυτοί οι αλγόριθμοι θα οδηγούσαν το διάνυσμα της βαθμολογίας να συγκλίνει στο ∞. Η κανονικοποίηση διατηρεί την *ενέργεια* σταθερή κατά τη διάρκεια των επαναλήψεων. Συνήθως, για την κανονικοποίηση χρησιμοποιείται η νόρμα-1 και στις περισσότερες περιπτώσεις κανονικοποιούμε έτσι ώστε $||.||_1 = 1$.

Για την περίπτωση της PS, ο παράγοντας b καθιστά περισσότερο πολύπλοκη την πράξη της κανονικοποίησης. Αν κανονικοποιήσουμε έτσι ώστε $||\vec{PS}||_1 = 1$, τότε για μεγάλους γράφους θα έχουμε $\vec{PS} \approx \vec{CC}$. Το αρχικό διάνυσμα για το $\vec{PS}$ είναι $\vec{0}$, πράγμα που σημαίνει ότι μετά την πρώτη επανάληψη το $||\vec{PS}||_1$ θα πρέπει να είναι ίσο με $|E| * b$. Επομένως, κρατούμε το $||\vec{PS}||_1$ σταθερό κανονικοποιώντας το με:

$$||\vec{PS}||_1 = |E| * b$$

Στο παράδειγμα του Σχήματος 3.2, το διάνυσμα που προκύπτει είναι $\vec{PS} =$

$(3.13, 3.86, 0, 0, 0, 0, 0, 0)$, το οποίο σημαίνει ότι τελικά η $PS_1 > PS_0$. Επομένως, με την κανονικοποίηση παρακάμπτουμε το γεγονός ότι ο κόμβος x δεν θα μπορούσε ποτέ να λάβει υψηλότερη βαθμολογία από τον κόμβο y, αν ο x δείχνει στον y.

3.3.4 SCEAS BPS: Ισορροπημένη Βαθμολογία Δημοσίευσης

Μπορούμε να επεκτείνουμε την Εξίσωση 3.9 υιοθετώντας έναν παράγοντα εξισορρόπησης ως μία συνάρτηση του αριθμού των εξερχόμενων αναφορών. Αυτό καταλήγει στην Ισορροπημένη Βαθμολογία Δημοσίευσης (BPS - Balanced Publication Score) ως ακολούθως:

$$BPS_x = \sum_{\forall y \in I_x} \frac{BPS_y + b}{|O_y|} \tag{3.10}$$

Η BPS ενσωματώνει τη λογική των PageRank και SALSA στον PS. Από τους πίνακες των παραδειγμάτων που δίνουμε, σημειώνουμε ότι ο BPS καταλήγει στην ίδια περίπου κατάταξη με τον PageRank. Αυτό επίσης θα φανεί στην επόμενη παράγραφο.

3.3.5 SCEAS EPS: Βαθμολογία Δημοσίευσης με Εκθετικά Βάρη

Μία άλλη βελτιωμένη έκδοση της PS (Εξίσωση 3.9) είναι η Βαθμολογία Δημοσίευσης με Εκθετικά Βάρη (EPS - Exponentially Weighted Publication Score). Η Βαθμολογία EPS ενός κόμβου x είναι το άθροισμα με εκθετικά βάρη των EPS όλων των κόμβων που δείχνουν απ'ευθείας το x:

$$EPS_x = \sum_{\forall y \in I_x} (EPS_y + b) * a^{-1} \tag{3.11}$$

Αυτή η μέθοδος μέτρησης ουσιαστικά παίρνει υπ'όψη το μέγεθος του δέντρου που σχηματοποιείται από τις άμεσες ή έμμεσες αναφορές στο x. Αν υπάρχει ένας έμμεσος σύνδεσμος από τον κόμβο x στον κόμβο y, τότε η βαθμολογία του y είναι μία συνάρτηση της βαθμολογίας του x πολλαπλασιασμένο με a^{-d}, όπου d είναι η απόσταση μεταξύ των δύο κόμβων.

Εδώ, όπως και στην περίπτωση της PS, δεν είναι απαραίτητη η κανονικοποίηση εφ'όσον δεν υπάρχουν κύκλοι. Για $a = e$, ο υπολογισμός της κατάταξης για το Σχήμα 3.2 (χωρίς κανονικοποίηση) θα κατέληγε στην τιμή του διανύσματος

$\vec{EPS} = (1.179, 2.207, 0, 0, 0, 0, 0, 0)$, αφού $EPS_1 = 6 * b * a^{-1} = 6 * e^{-1} = 2.207$ και $EPS_0 = (EPS_1 + b) * a^{-1} = (2.207 + 1) * e^{-1} = 1.179$.

Γενικά χρειαζόμαστε ένα βήμα κανονικοποίησης όταν υπάρχουν κύκλοι. Για τους ίδιους λόγους με την περίπτωση της PS, το $||\vec{EPS}||_1$ δεν θα πρέπει να κανονικοποιηθεί σε 1. Επιπλέον, δεν θα πρέπει να χρησιμοποιήσουμε την ίδια τιμή κανονικοποίησης με την PS, καθώς αυτό θα μπορούσε να οδηγήσει την EPS να είναι πανομοιότυπη με την PS. Για παράδειγμα, έστω ότι για την περίπτωση PS καταλήγουμε σε έναν παράγοντα κανονικοποίησης n, τέτοιον ώστε $||n * \vec{PS}||_1 = |E| * b$. Αν κατά τη διάρκεια της ίδιας επανάληψης αναζητούσαμε έναν παράγοντα κανονικοποίησης k για την EPS, τέτοιον ώστε $||k * \vec{EPS}||_1 = |E| * b$, τότε θα έπρεπε να ισχύει $k = a * n$. Αυτό θα οδηγούσε τους δύο αλγορίθμους να είναι πανομοιότυποι, αφού θα παρήγαγε $\vec{EPS} = \vec{PS} * (k/n) * a^{-1} = \vec{PS} * (a * n/n) * a^{-1} = \vec{PS}$, πράγμα που σημαίνει ότι η κατάταξη θα ήταν ακριβώς ίδια. Για να το αποφύγουμε αυτό, θα έπρεπε να συμπεριλάβουμε στη διαδικασία κανονικοποίησης τον παράγοντα a, έτσι ώστε $||\vec{EPS}||_1 = |E| * b * a^{-1}$. Επομένως, ο παράγοντας κανονικοποίησης είναι εξίσου σημαντικός στην περίπτωση της EPS.

3.3.6 SCEAS BEPS: Ισορροπημένη Βαθμολογία Δημοσίευσης με Εκθετικά Βάρη

Μία υβριδική μέθοδος που βασίζεται στις Εξισώσεις 3.10 και 3.11 είναι η Ισορροπημένη Βαθμολογία Δημοσίευσης με Εκθετικά Βάρη (BEPS - Balanced Exponentially Weighted Publication Score). Η βαθμολογία $BEPS$ του κόμβου x είναι το άθροισμα με εκθετικά βάρη του $BEPS_y$ διαιρούμενου με το πλήθος των αναφορών που έγιναν από τη δημοσίευση y, $\forall y \in I_x$:

$$BEPS_x = \sum_{\forall y \in I_x} \frac{BEPS_y + b}{|O_y|} * a^{-1} \tag{3.12}$$

3.3.7 SCEAS General

Ένας συντελεστής τυχαίας μετάβασης d μπορεί να προστεθεί στην Εξίσωση 3.12, πράγμα θα οδηγούσε στην ακόλουθη εξίσωση, ως μία γενικευμένη έκφραση του BEPS και του PageRank:

$$S_x = (1 - d) + d * \sum_{\forall y \in I_x} \frac{S_y + b}{|O_y|} * a^{-1} \tag{3.13}$$

Από εδώ και στο εξής θα αναφερόμαστε σε αυτή τη μέθοδο με το όνομα SCEASRank. Για $d = 1$, ο SCEASRank είναι ισοδύναμος με τον BEPS. Για $b = 0$ και $a = 1$ ο

SCEAS είναι ισοδύναμος με τον PageRank. Ο PageRank χρησιμοποιεί την τιμή $d = 0.85$ για να εξισορροπήσει την ακρίβεια και την ταχύτητα σύγκλισης. Μία τιμή για το d πλησιέστερη στο 1 έχει ως αποτέλεσμα μεγαλύτερη ακρίβεια στις βαθμολογίες. Επίσης, η τιμή $d = 1$ θα έπρεπε να οδηγήσει τον PageRank να συγκλίνει στο μηδέν. Επομένως, για τον PageRank είναι απαραίτητη μία τιμή $d < 1$. Επιπλέον όσο πιο κοντά το d στο 1, τόσο πιο ευαίσθητος είναι ο PageRank στις κυκλικές αναφορές. Για τον SCEAS είναι ασφαλές να χρησιμοποιήσουμε οποιαδήποτε τιμή για το d (όπου $0 < d \leq 1$) αν $b > 0$. Επίσης, η ταχύτητα σύγκλισης κυρίως επηρεάζεται από τον παράγοντα a παρά από τον παράγοντα d. Στην περίπτωση μας είναι ασφαλές να χρησιμοποιήσουμε οποιαδήποτε συντελεστή d μεγαλύτερο του 0.85, όπως για παράδειγμα 0.99.

3.4 Πειραματικά Αποτελέσματα

3.4.1 Σύνολο Δεδομένων

Το σύστημα SCEAS χρησιμοποιεί τα δεδομένα της ψηφιακής βιβλιοθήκης DBLP. Τη στιγμή των πειραμάτων του παρόντος κεφαλαίου τα δεδομένα είχαν χρονική σφραγίδα 19/5/2005. Ο Πίνακας 3.7 περιγράφει λεπτομερώς τα ποιοτικά χαρακτηριστικά του γράφου αναφορών της DBLP. Παρατηρούμε ότι μόνο το 1.31% των δημοσιεύσεων έχουν αποθηκευμένες τις αναφορές τους (V_O), ενώ μόνο το 2.92% αυτών έχουν έσω-βαθμό (V_I). Επομένως, πρακτικά μόνο αυτές οι δημο-

Σύμβολο	Ορισμός	Ιδιότητες
$G_\infty = (V_\infty, E_\infty)$	Ο παγκόσμιος γράφος αναφορών	
$G_{DBLP} = (V_{DBLP}, E_{DBLP})$	Η δική μας βάση δεδομένων (αντίγραφο DBLP)	$G_{DBLP} \subset G_\infty$
V_{DBLP}	Κόμβοι στο γράφο αναφορών	$\lvert V_{DBLP}\rvert = 624893$
V_{INP}	Το σύνολο των πρακτικών συνεδρίων	$\lvert V_{INP}\rvert = 391543$
V_{ART}	Το σύνολο των άρθρων στα περιοδικά	$\lvert V_{ART}\rvert = 233350$
E_{DBLP}	Αναφορές μες το γράφο	$\lvert E_{DBLP}\rvert = 100210$
$E_{DBLP-} = \{(i \to j) : i \in V_{DBLP} \wedge j \notin V_{DBLP}\}$	Αναφορές σε οντότητες που δεν ανήκουν στο DBLP	$\lvert E_{DBLP-}\rvert = 67971$
$V_I = \{x \in V_{DBLP} : \lvert I_x\rvert > 0\}$		$\lvert V_I\rvert = 18273$
$V_O = \{x \in V_{DBLP} : \lvert O_x\rvert > 0\}$		$\lvert V_O\rvert = 8183$
		$\lvert V_O \cup V_I\rvert = 20831$

Πίνακας 3.7. Ιδιότητες της βάσης δεδομένων και του γράφου αναφορών.

σιεύσεις μπορούν να βαθμολογηθούν. Η κατανομή αυτών των αναφορών ανά έτος και πηγή περιγράφεται στο [83]. Εδώ παρατηρούμε, ότι οι διαθέσιμες αναφορές (π.χ. το σύνολο V_O) είναι σχετικές με δημοσιεύσεις που εκδίδονται μέσα σε ένα επιλεγμένο σύνολο των σημαντικότερων πρακτικών συνεδρίων και περιοδικών, σύμφωνα με τους διαχειριστές της DBLP. Με άλλα λόγια, το σύνολο των δημοσιεύσεων που κάνουν αναφορές έχει ήδη φιλτραρισθεί σύμφωνα με το κριτήριο της σπουδαιότητας του συνεδρίου/περιοδικού. Πρακτικά, στο σύνολο δεδομένων μας, όλες οι δημοσιεύσεις που κάνουν αναφορές σε άλλες έχουν κάποιο ελάχιστο μέτρο ποιότητας και επομένως είναι αναμενόμενο, ότι οι διαφορές στις σειρές κατάταξης που παράγονται κατά τα πειράματά μας από τις διάφορες μεθόδους, θα είναι μάλλον μικρές.

3.4.2 Ταχύτητα Υπολογισμού

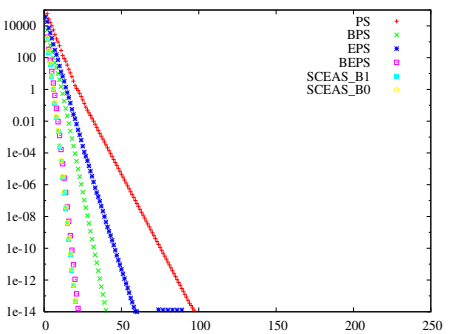
Σύμφωνα με τον ορισμό του `SCEASRank` είναι προφανές ότι καταλήγουμε σε μία πολύ ταχεία σύγκλιση για $a = e$. Στο Σχήμα 3.5, ο άξονας-x αποικονίζει το πλήθος των επαναλήψεων που απαιτούνται από κάθε αλγόριθμο για τον υπολογισμό των κατατάξεων για το σύνολο δεδομένων της DBLP. Ο άξονας-y παρουσιάζει σε λογαριθμική κλίμακα την τιμή του:

$$\delta \;=\; ||\vec{x}_l - \vec{x}_{l-1}||_1 \tag{3.14}$$

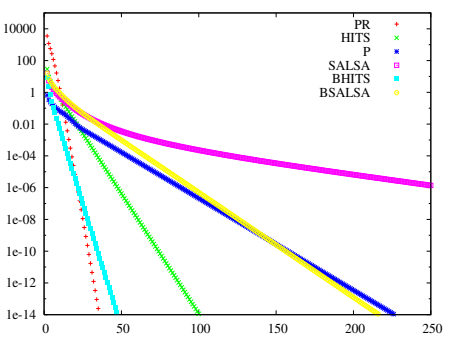
όπου $\vec{x}_l$ είναι το διάνυσμα με τις βαθμολογίες $\{S_1, S_2, \ldots, S_V\}$ μετά από l επαναλήψεις. Στις περιπτώσεις του `HITS` και του `SALSA`, το δ εκφράζεται ως το άθροισμα της νόρμας-1 του διανύσματος αυθεντιών συν τη νόρμα-1 του διανύσματος κομβικών σημείων:

$$\delta \;=\; ||\vec{a}_l - \vec{a}_{l-1}||_1 + ||\vec{h}_l - \vec{h}_{l-1}||_1 \tag{3.15}$$

Η συνθήκη τερματισμού κάθε αλγορίθμου είναι $\delta < \epsilon$, όπου ϵ είναι ένας πολύ μικρός αριθμός. Στην πραγματικότητα, όπως περιγράφεται στο [47], αυτός ο αριθμός δεν μπορεί να ορισθεί εκ των προτέρων για τον `PageRank`, αφού εξαρτάται από το γράφο αναφορών. Είναι προφανές ότι κάθε αλγόριθμος χρειάζεται μία διαφορετική τιμή για το ϵ ως συνθήκη τερματισμού. Ανεξαρτήτως όμως από αυτό, στο διάγραμμα είναι εμφανές ότι οι καμπύλες του `SCEAS` είναι πολύ πιο απότομες από τις καμπύλες των άλλων αλγορίθμων. Αυτό σημαίνει ότι ο `SCEAS` συγκλίνει γρηγορότερα από τις άλλες μεθόδους, ανεξάρτητα από το ποιές είναι οι πραγματικές τιμές του δ και του ϵ. Για καλύτερη κατανόηση, ο Πίνακας 3.8 δείχνει την εφαπτόμενη της γωνίας $((x2 - x1)/(y1 - y2))$ κάθε καμπύλης. Από



(α) Ταχύτητα σύγκλησης των παραλλαγών SCEAS πάνω στη βάση DBLP.



(β) Ταχύτητα σύγκλησης των PageRank, HITS, SALSA, Prestige πάνω στη βάση DBLP.

Σχήμα 3.5. Ταχύτητα σύγκλησης/υπολογισμών

τον τελευταίο πίνακα είναι φανερό ότι ο SCEAS είναι ο ταχύτερος αλγόριθμος, ακολουθεί ο BEPS με μικρή διαφορά, ο PageRank και ο BPS έρχονται αργότερα με σχεδόν μισή ταχύτητα σύγκλισης σε σχέση με τον SCEAS, ενώ οι υπόλοιποι αλγόριθμοι έχουν πολύ αργή ταχύτητα σύγκλισης.

Σε αυτό το σημείο αξίζει να αναφερθεί ότι για όλους τους αλγορίθμους εκτός από τον HITS, τον SALSA και τις παραλλαγές τους, ο υπολογισμός θα μπορούσε να γίνει σε μόνο ένα βήμα, αν και μόνο αν:

1. Ο γράφος δεν είχε κύκλους, και
2. Ο υπολογισμός γινόταν αναδρομικά ξεκινώντας από τους κόμβους που δεν δέχονται αναφορές (dangling nodes).

Ωστόσο, ο γράφος περιέχεις κάποιους κύκλους και έτσι ο υπολογισμός σε ένα βήμα είναι ανέφικτος.

Μέθοδος	Εφαπτομένη
SCEAS_B0	1.029
SCEAS_B1	1.030
BEPS	1.106
PR	1.804
BPS	2.057
EPS	3.176
BHITS	3.193
PS	5.428
HITS	6.605
BSALSA	15.184
P	17.270
SALSA	71.621

Πίνακας 3.8. Γωνία εφαπτομένης των γραμμών του Σχήματος 3.5.

3.4.3 Συγκρίσεις Αξιολογήσεων

Σε αυτήν την παράγραφο συγκρίνουμε τα αποτελέσματα όλων των αλγορίθμων που αναφέρθηκαν προηγουμένως. Το έργο αυτό δεν είναι εύκολο, αφού όλοι οι αλγόριθμοι είναι ευρέως αποδεκτοί αλλά ο καθένας στο δικό του τομέα εφαρμογής. Πιο συγκεκριμένα, εκτελούμε μία στατιστική σύγκριση, η οποία βασίζεται στα αποτελέσματα κατάταξης για τα δεδομένα της ψηφιακής βιβλιοθήκης DBLP:

- Μετρώντας τον αριθμό των κοινών στοιχείων μεταξύ των x κορυφαίων στοιχείων του πίνακα κατάταξης για κάθε ζεύγος αλγορίθμων.
- Καταρτίζοντας διάγραμμα της συνάρτησης $Top(x)$, όπου x είναι ο αριθμός των κορυφαίων στοιχείων στους πίνακες κατάταξης.
- Υπολογίζοντας την απόσταση μεταξύ όλων των ζευγών των πινάκων κατάταξης με βάση το Kendall's tau [50], που επίσης χρησιμοποιείται στο [12].
- Υπολογίζοντας την απλή απόσταση μεταξύ όλων των ζευγών των πινάκων κατάταξης.
- Υπολογίζοντας την απόσταση με βάρη μεταξύ όλων των ζευγών των πινάκων κατάταξης.
- Χρησιμοποιώντας διαγράμματα q-q.

3.4.3.1 Σύγκριση με Βάση τον Αριθμό των Κοινών Στοιχείων

Ας θεωρήσουμε ότι όλες οι μέθοδοι εκτελούνται στο σύνολο δεδομένων της ψηφιακής βιβλιοθήκης DBLP, ενώ για κάθε μέθοδο εξάγονται τα 20 κορυφαία στοι-

χεία. Ο Πίνακας 3.9 απεικονίζει τον αριθμό των κοινών στοιχείων μεταξύ των 20 κορυφαίων στοιχείων για κάθε ζεύγος μεθόδων κατάταξης. Όπως φαίνεται στον πίνακα, ο SCEAS_B1 (για $b = 1, d = 0.85, a = e$), ο SCEAS_B0 (για $b = 0, d = 0.85, a = e$) και ο BEPS (για $b = 1, d = 1, a = e$) έχουν περίπου την ίδια συμπεριφορά. Ο PageRank είναι πολύ κοντά στον BPS, αφού ο BPS είναι ισοδύναμος με τον PageRank χωρίς το συντελεστή τυχαίας μετάβασης. Ο SALSA authority δίνει τα κορυφαία 20 στοιχεία βασιζόμενος στον αριθμό αναφορών (CC). Οι παραλλαγές PS και EPS είναι επίσης πολύ κοντά μεταξύ τους.

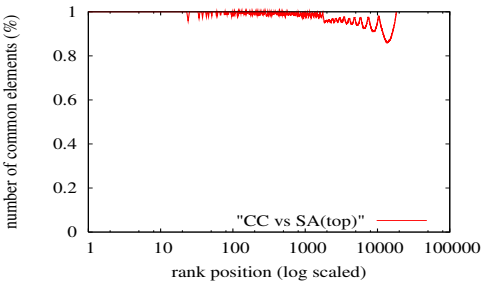
	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1	SCEAS_B0
CC	-	18	12	13	5	20	7	13	10	11	11	15	16	16
BCC	18	-	12	11	5	18	7	13	10	11	11	14	15	15
PR	12	12	-	11	12	12	8	16	17	19	18	17	16	16
HA	13	11	11	-	6	13	10	12	10	11	11	12	11	11
P	5	5	12	6	-	5	8	10	13	13	12	9	9	9
SA	20	18	12	13	5	-	7	13	10	11	11	15	16	16
BHA	7	7	8	10	8	7	-	10	8	8	9	8	7	7
BSA	13	13	16	12	10	13	10	-	14	16	15	17	16	16
PS	10	10	17	10	13	10	8	14	-	18	19	14	13	13
BPS	11	11	19	11	13	11	8	16	18	-	18	16	15	15
EPS	11	11	18	11	12	11	9	15	19	18	-	15	14	14
BEPS	15	14	17	12	9	15	8	17	14	16	15	-	19	19
SCEAS_B1	16	15	16	11	9	16	7	16	13	15	14	19	-	20
SCEAS_B0	16	15	16	11	9	16	7	16	13	15	14	19	20	-

Πίνακας 3.9. Πλήθος κοινών στοιχείων στις 20 πρώτες δημοσιεύσεις για κάθε ζεύγος μεθόδων.

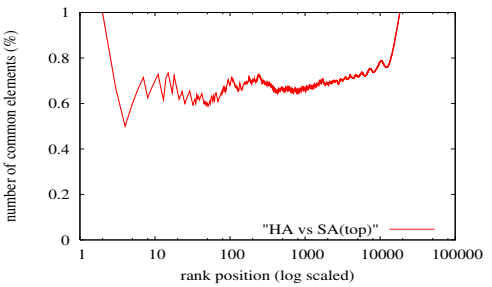
Στον Πίνακα 3.9, όπως επίσης και στους επόμενους πίνακες που παρουσιάζουν τις αποστάσεις των αλγορίθμων κατάταξης, τα κελιά με πλαίσιο δηλώνουν ότι οι δύο συγκρινόμενοι αλγόριθμοι είναι ανόμοιοι (δηλαδή, λίγα κοινά στοιχεία ή σε μεγάλη απόσταση). Τα κελιά που σημειώνονται με γκρι φόντο σημαίνουν ότι οι δύο αλγόριθμοι είναι παρόμοιοι. Το ανοιχτό γκρι υποδηλώνει μεγάλη ομοιότητα, ενώ το σκούρο γκρι απλή ομοιότητα.

3.4.3.2 Σύγκριση με Βάση τη Συνάρτηση $Top(x)$

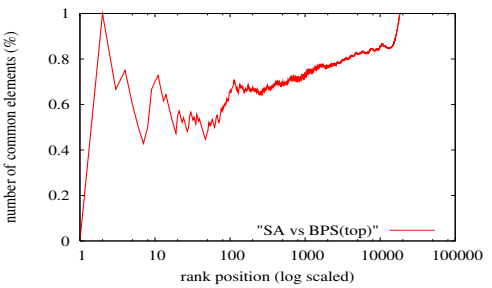
Ίσως κάποιος θα μπορούσε διαφωνήσει με την επιλογή για τη μέτρηση του αριθμού των κοινών στοιχείων μεταξύ των 20 κορυφαίων στοιχείων από δύο λίστες κατάταξης με σκοπό τη διερεύνηση της ομοιότητας αυτών των λιστών. Για αυτό το λόγο, ορίζουμε τη συνάρτηση $Top(a_1, a_2, x)$, έτσι ώστε για δύο αλγόριθμους κατάταξης a_1 και a_2, η $Top(a_1, a_2, x)$ δίνει τον αριθμό των κοινών στοιχείων μεταξύ των x κορυφαίων στοιχείων από κάθε λίστα κατάταξης, διαιρεμένο με τον



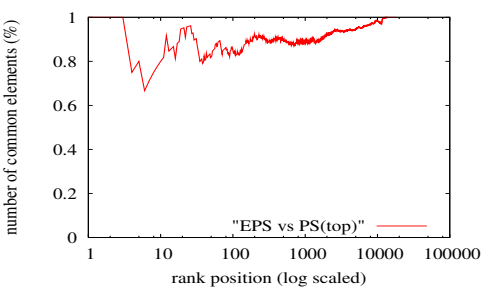
(α) CC vs. SA



(β) HA vs. SA



(γ) SA vs. BPS



(δ) EPS vs. PS

Σχήμα 3.6. Σύγκριση αποτελεσμάτων κατάταξης με τη συνάρτηση $Top(a_1, a_2, x)$.

αριθμό των κόμβων (το x δηλαδή):

$$R(a,x) \;=\; \{i \in V : P_a(i) \leq x\}$$

$$Top(a_1,a_2,x) \;=\; |R(a_1,x) \cap R(a_2,x)|/x \tag{3.16}$$

όπου $P_a(i)$ είναι η θέση του κόμβου i στον πίνακα κατάταξης που κατασκευάσθηκε από τον αλγόριθμο a και $R(a,x)$ είναι το σύνολο των κορυφαίων x στοιχείων του πίνακα κατάταξης. Είναι προφανές ότι $Top(a_1,a_2,x) \underset{x\to|V|}{\longrightarrow} 1$. Κάποια διαγράμματα της συνάρτησης αυτής φαίνονται στο Σχήμα 3.6. Για παράδειγμα, το Σχήμα 3.6(α) επιβεβαιώνει ότι ο SALSA είναι σχεδόν ισοδύναμος με τον CC όπως φαίνεται στον Πίνακα 3.9. Επίσης, ο EPS είναι πολύ κοντά στον PS (βλέπε Σχήμα 3.6(δ)). Από την άλλη μεριά, ο SALSA απέχει από τον BPS (βλέπε Σχήμα 3.6(γ)), ενώ τέλος ο HITS Authorities είναι σταθερά στο 50% σε σχέση με τον SALSA Authorities.

3.4.3.3 Σύγκριση με Βάση τη Μέθοδο Kendall's tau

Η απόσταση μεταξύ οποιουδήποτε ζεύγους πινάκων κατάταξης μπορεί να υπολογισθεί με το Kendall's tau με *πέναλτυ* p [50, 12]. Για να εκτελέσουμε αυτόν τον υπολογισμό, το *σύνολο διαφορών κατάταξης* (violating set) $\mathcal{V}(a_1,a_2)$ και το *σύνολο ασθενών διαφορών κατάταξης* (weakly violating set) $\mathcal{W}(a_1,a_2)$ πρέπει να ορισθούν ως εξής:

$$\mathcal{V}(a_1,a_2) = \{(i,j) \in V : (a_1(i) < a_2(j) \wedge a_2(i) > a_1(j)) \vee (a_1(i) > a_2(j) \wedge a_2(i) < a_1(j))\}$$
$$\mathcal{W}(a_1,a_2) = \{(i,j) \in V : (a_1(i) = a_1(j) \wedge a_2(i) \neq a_2(j)) \vee (a_1(i) \neq a_1(j) \wedge a_2(i) = a_2(j))\}$$

όπου $a_1(i)$ είναι η υπολογισμένη βαθμολογία για το στοιχείο i από τη μέθοδο κατάταξης a_1. Το σύνολο διαφορών κατάταξης είναι το σύνολο όλων των ζευγών κόμβων για τα οποία η κατάταξή τους είναι διαφορετική στους δύο πίνακες κατάταξης (a_1 και a_2). Το σύνολο ασθενών διαφορών κατάταξης είναι το σύνολο όλων των ζευγών κόμβων, τα οποία βαθμολογούνται με τον ίδιο βαθμό από έναν αλγόριθμο, αλλά με διαφορετικό βαθμό από έναν άλλον. Το Kendall's tau τιμωρεί με 1 κάθε μέλος του συνόλου διαφορών κατάταξης, και με p κάθε μέλος του συνόλου ασθενών διαφορών κατάταξης:

	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1	SCEAS_B0
CC	-	7.6%	7.6%	12.8%	15.0%	0.8%	15.5%	5.0%	7.6%	7.8%	5.8%	7.3%	7.3%	7.3%
BCC	7.6%	-	6.7%	34.3%	25.2%	11.1%	34.5%	23.0%	24.1%	7.8%	22.7%	3.0%	2.6%	2.6%
PR	7.6%	6.7%	-	31.5%	20.2%	11.7%	30.6%	19.1%	18.5%	1.0%	17.3%	3.6%	4.1%	4.1%
HA	12.8%	34.3%	31.5%	-	15.9%	21.3%	12.1%	16.0%	17.6%	31.1%	17.3%	33.1%	33.3%	33.3%
P	15.0%	25.2%	20.2%	15.9%	-	18.9%	12.1%	13.4%	9.2%	19.4%	10.9%	22.9%	23.2%	23.2%
SA	0.8%	11.1%	11.7%	21.3%	18.9%	-	24.0%	13.8%	13.5%	12.0%	11.8%	11.1%	11.1%	11.1%
BHA	15.5%	34.5%	30.6%	12.1%	12.1%	24.0%	-	15.3%	16.0%	30.0%	16.6%	32.8%	33.0%	33.0%
BSA	5.0%	23.0%	19.1%	16.0%	13.4%	13.8%	15.3%	-	10.2%	18.7%	9.2%	21.0%	21.3%	21.3%
PS	7.6%	24.1%	18.5%	17.6%	9.2%	13.5%	16.0%	10.2%	-	17.6%	1.9%	21.5%	21.9%	21.9%
BPS	7.8%	7.8%	1.0%	31.1%	19.4%	12.0%	30.0%	18.7%	17.6%	-	16.6%	4.7%	5.2%	5.2%
EPS	5.8%	22.7%	17.3%	17.3%	10.9%	11.8%	16.6%	9.2%	1.9%	16.6%	-	20.2%	20.6%	20.6%
BEPS	7.3%	3.0%	3.6%	33.1%	22.9%	11.1%	32.8%	21.0%	21.5%	4.7%	20.2%	-	0.4%	0.4%
SCEAS_B1	7.3%	2.6%	4.1%	33.3%	23.2%	11.1%	33.0%	21.3%	21.9%	5.2%	20.6%	0.4%	-	0.0%
SCEAS_B0	7.3%	2.6%	4.1%	33.3%	23.2%	11.1%	33.0%	21.3%	21.9%	5.2%	20.6%	0.4%	0.0%	-

Πίνακας 3.10. Διαφορές στις κατατάξεις δημοσιεύσεων του DBLP υπολογισμένες με τη συνάρτηση $d^{(0)}(a_1, a_2)$.

	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1	SCEAS_B0
CC	-	30.2%	30.5%	35.8%	37.9%	11.5%	38.6%	28.1%	25.6%	30.7%	23.9%	30.2%	30.2%	30.2%
BCC	30.2%	-	7.1%	35.0%	46.4%	23.6%	35.2%	23.6%	29.4%	8.2%	28.0%	3.5%	3.0%	3.0%
PR	30.5%	7.1%	-	31.7%	41.3%	24.2%	30.8%	19.3%	23.4%	1.0%	22.2%	3.6%	4.1%	4.1%
HA	35.8%	35.0%	31.7%	-	37.2%	33.7%	12.1%	16.1%	22.6%	31.3%	22.3%	33.3%	33.5%	33.5%
P	37.9%	46.4%	41.3%	37.2%	-	43.0%	33.4%	34.7%	25.5%	40.5%	27.1%	44.0%	44.3%	44.3%
SA	11.5%	23.6%	24.2%	33.7%	43.0%	-	36.4%	26.2%	26.8%	24.5%	25.1%	23.6%	23.5%	23.5%
BHA	38.6%	35.2%	30.8%	12.1%	33.4%	36.4%	-	15.3%	21.1%	30.1%	21.7%	33.0%	33.2%	33.2%
BSA	28.1%	23.6%	19.3%	16.1%	34.7%	26.2%	15.3%	-	15.2%	18.8%	14.3%	21.2%	21.4%	21.4%
PS	25.6%	29.4%	23.4%	22.6%	25.5%	26.8%	21.1%	15.2%	-	22.5%	1.9%	26.4%	26.8%	26.8%
BPS	30.7%	8.2%	1.0%	31.3%	40.5%	24.5%	30.1%	18.8%	22.5%	-	21.4%	4.7%	5.2%	5.2%
EPS	23.9%	28.0%	22.2%	22.3%	27.1%	25.1%	21.7%	14.3%	1.9%	21.4%	-	25.1%	25.5%	25.5%
BEPS	30.2%	3.5%	3.6%	33.3%	44.0%	23.6%	33.0%	21.2%	26.4%	4.7%	25.1%	-	0.4%	0.4%
SCEAS_B1	30.2%	3.0%	4.1%	33.5%	44.3%	23.5%	33.2%	21.4%	26.8%	5.2%	25.5%	0.4%	-	0.0%
SCEAS_B0	30.2%	3.0%	4.1%	33.5%	44.3%	23.5%	33.2%	21.4%	26.8%	5.2%	25.5%	0.4%	0.0%	-

Πίνακας 3.11. Διαφορές στις κατατάξεις δημοσιεύσεων του DBLP υπολογισμένες με τη συνάρτηση $d^{(1)}(a_1, a_2)$.

$$K^{(p)}(a_1, a_2) = \sum_{\forall i,j \in V} \mathcal{I}^{(p)}_{a_1 a_2}(i,j)$$

$$\mathcal{I}^{(p)}_{a_1 a_2}(i,j) = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{V}(a_1, a_2) \\ p & \text{if } (i,j) \in \mathcal{W}(a_1, a_2) \\ 0 & \text{otherwise} \end{cases}$$

Ουσιαστικά, το Kendall's tau με πέναλτυ $p = 0$ δίνει τον αριθμό των ανταλλαγών που θα εκτελεσθούν από την ταξινόμηση φυσαλίδας (bubble sort) ώστε να κατασκευασθεί ο δεύτερος πίνακας κατάταξης από τον πρώτο. Τέλος, η ασθενής απόσταση κατάταξης, όπως ορίζεται στο [12] είναι:

$$d^{(0)}(a_1, a_2) \;=\; \frac{1}{|V|(|V|-1)/2} K^{(0)}(a_1, a_2) \tag{3.17}$$

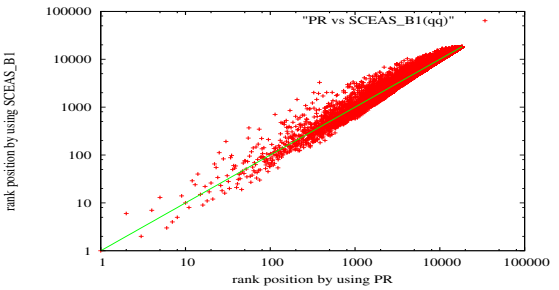
ενώ τα αποτελέσματα φαίνονται στον Πίνακα 3.10. Η αυστηρή απόσταση κατάταξης (βλέπε Πίνακα 3.11) είναι:

$$d^{(1)}(a_1, a_2) \;=\; \frac{1}{|V|(|V|-1)/2} K^{(1)}(a_1, a_2) \tag{3.18}$$

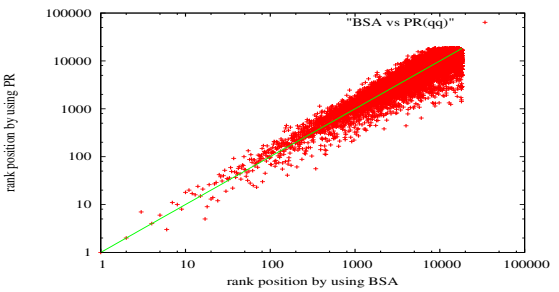
όπου $|V|$ είναι ο αριθμός των κορυφών του γράφου. Επομένως, το $d^{(0)}(a_1, a_2)$ και το $d^{(1)}(a_1, a_2)$ είναι κανονικοποιημένα στην κλίμακα [0..1].

3.4.3.4 Σύγκριση με Βάση τα Διαγράμματα q-q

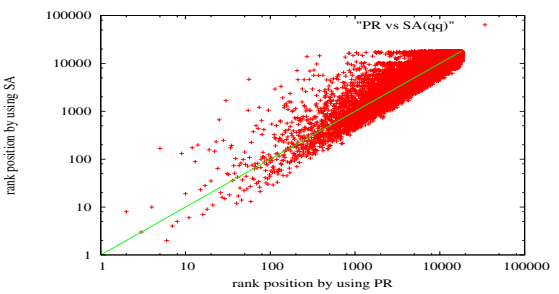
Ένα άλλο εργαλείο σύγκρισης των κατατάξεων είναι με τη χρήση διαγραμμάτων q-q. Κάθε σημείο $x(i,j)$ ενός διαγράμματος q-q αντιστοιχεί σε κάποιο κόμβο x. Οι συντεταγμένες του σημείου (i,j) σημαίνουν ότι ο κόμβος έχει καταταχθεί στην i-οστή θέση από τον πρώτο αλγόριθμο (άξονας-x) και στην j-οστή θέση από το δεύτερο αλγόριθμο (άξονας-y). Στο Σχήμα 3.7(α) ο `PageRank` συγκρίνεται με τον `SCEAS`. Βλέπουμε ότι όλα τα σημεία είναι αρκετά κοντά στη γραμμή $y = x$. Τα Σχήματα 3.7(β) και 3.7(γ) δείχνουν τη σύγκριση του `PageRank` με τον `SALSA` και τον `B-SALSA`. Είναι φανερό ότι ο `B-SALSA` είναι πλησιέστερα στον `PageRank` απ' ότι στον `SALSA`. Τέλος, το Σχήμα 3.7(δ) δείχνει την ομοιότητα του `SCEAS` (με $b = 1$) με τον `BEPS`. Μπορούμε να δούμε ότι όλα τα σημεία είναι κατανεμημένα πολύ κοντά στη γραμμή $y = x$.



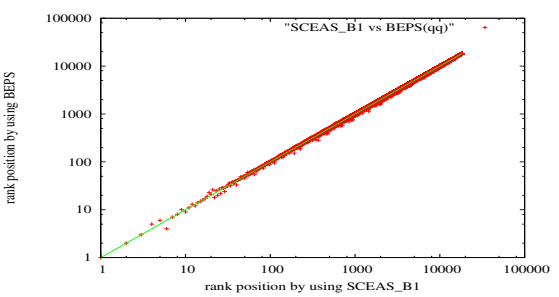
(α) PR vs. SCEAS_B1



(β) BSA vs. PR



(γ) PR vs. SA



(δ) SCEAS_B1 vs. BEPS

Σχήμα 3.7. Συγκρίσεις των αποτελεσμάτων κατάταξης του DBLP με q-q plots.

	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1	SCEAS_B0
CC	-	13.3%	13.3%	17.1%	23.0%	7.9%	18.8%	11.5%	13.9%	13.4%	12.6%	13.1%	13.1%	13.1%
BCC	13.3%	-	5.3%	23.0%	24.3%	12.3%	23.3%	15.7%	17.8%	6.1%	16.9%	2.5%	2.2%	2.2%
PR	13.3%	5.3%	-	21.2%	21.2%	12.8%	20.7%	13.1%	14.1%	0.9%	13.3%	2.8%	3.2%	3.2%
HA	17.1%	23.0%	21.2%	-	18.3%	18.4%	8.6%	11.2%	14.0%	20.9%	13.8%	22.1%	22.2%	22.2%
P	23.0%	24.3%	21.2%	18.3%	-	22.7%	15.9%	17.0%	13.1%	20.7%	14.3%	22.9%	23.1%	23.1%
SA	7.9%	12.3%	12.8%	18.4%	22.7%	-	20.5%	13.2%	14.7%	13.1%	13.3%	12.3%	12.3%	12.3%
BHA	18.8%	23.3%	20.7%	8.6%	15.9%	20.5%	-	10.9%	13.1%	20.3%	13.6%	22.1%	22.2%	22.2%
BSA	11.5%	15.7%	13.1%	11.2%	17.0%	13.2%	10.9%	-	8.9%	12.8%	8.2%	14.3%	14.5%	14.5%
PS	13.9%	17.8%	14.1%	14.0%	13.1%	14.7%	13.1%	8.9%	-	13.6%	1.5%	16.1%	16.3%	16.3%
BPS	13.4%	6.1%	0.9%	20.9%	20.7%	13.1%	20.3%	12.8%	13.6%	-	12.8%	3.6%	4.0%	4.0%
EPS	12.6%	16.9%	13.3%	13.8%	14.3%	13.3%	13.6%	8.2%	1.5%	12.8%	-	15.2%	15.4%	15.4%
BEPS	13.1%	2.5%	2.8%	22.1%	22.9%	12.3%	22.1%	14.3%	16.1%	3.6%	15.2%	-	0.4%	0.4%
SCEAS_B1	13.1%	2.2%	3.2%	22.2%	23.1%	12.3%	22.2%	14.5%	16.3%	4.0%	15.4%	0.4%	-	0.0%
SCEAS_B0	13.1%	2.2%	3.2%	22.2%	23.1%	12.3%	22.2%	14.5%	16.3%	4.0%	15.4%	0.4%	0.0%	-

Πίνακας 3.12. Διαφορές στις κατατάξεις δημοσιεύσεων του DBLP υπολογισμένες με τη συνάρτηση $D(a_1, a_2)$.

3.4.3.5 Σύγκριση με Βάση την Απλή Απόσταση

Για λόγους συντομίας αποφεύγουμε την παρουσίαση περισσότερων συγκρίσεων με τη χρήση των διαγραμμάτων q-q. Για να συνοψίσουμε ένα διάγραμμα q-q σε έναν αριθμό μπορούμε να υπολογίσουμε τις αποστάσεις όλων των σημείων (i, j) από τη γραμμή $y = x$ και κατόπιν να τις προσθέσουμε. Η απόσταση του σημείου (i, j) από τη γραμμή $y = x$ είναι ίση με $|j - i|/\sqrt{2}$. Επομένως, ένα εναλλακτικό μέτρο *Απλής Απόστασης* των αλγορίθμων a_1 και a_2 είναι:

$$D(a_1, a_2) = \frac{1}{|V|} \sum_{\forall i} \frac{|P_{a_1}(i) - P_{a_2}(i)|}{|V|} \tag{3.19}$$

όπου $|V|$ είναι ο αριθμός των κόμβων και $P_{a_1}(i)$ είναι η θέση του κόμβου i στον πίνακα κατάταξης του αλγορίθμου a_1. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.12. Η συνάρτηση $D(a_1, a_2)$ κανονικοποιείται επίσης στην κλίμακα $[0..0.5]$, εφ'όσον το 0.5 είναι η τιμή της απόστασης D για την αντίστροφη διάταξη. Αυτή η μέθοδος σύγκρισης είναι γνωστή ως το Spearman's footrule [19].

3.4.3.6 Σύγκριση με Βάση την Απόσταση με Βάρη

Κατά τη σύγκριση κατατάξεων οποιαδήποτε διαφορά σε μία υψηλή θέση είναι πρακτικά σημαντικότερη απ'ότι μια διαφορά σε μία χαμηλή θέση. Ούτε το Kendall's tau, ούτε το footrule του Spearman εφαρμόζουν μία τέτοιου είδους διάκριση. Για αυτό το λόγο ορίζουμε ένα εναλλακτικό μέτρο απόστασης κατάταξης. Θέτουμε την *Απόσταση με Βάρη* D_w δύο μεθόδων κατάταξης a_1 και a_2 ως:

$$\begin{aligned} w(a_1, a_2, i) &= \frac{1}{min(P_{a_1}(i), P_{a_2}(i))} \\ d_w(a_1, a_2, i) &= |P_{a_1}(i) - P_{a_2}(i)| * w(a_1, a_2, i) \\ D_w(a_1, a_2) &= \frac{1}{|V| * \sum_{\forall i \in V} w(a_1, a_2, i)} \sum_{\forall i \in V} d_w(a_1, a_2, i) \end{aligned} \tag{3.20}$$

Η συνάρτηση με βάρη $w(a_1, a_2, i)$ είναι γραμμική και αντιστρόφως ανάλογη στην καλύτερη θέση κατάταξης του κόμβου i. Αυτό σημαίνει ότι εάν ένα στοιχείο καταταχθεί πρώτο από τον a_1 και δεύτερο από τον a_2, τότε το βάρος θα είναι 1 και η Απόσταση με Βάρη $d_w = 1$. Στον Πίνακα 3.13 φαίνονται περισσότερες ενδεικτικές περιπτώσεις.

Στον Πίνακα 3.14 μπορούμε να δούμε ποιοί αλγόριθμοι βρίσκονται κοντά μεταξύ τους. Ο `PageRank` είναι πολύ κοντά στους `SCEAS, BEPS, BPS` και τέλος στον `BCC`. Ο `PageRank` υπολογίζει μια ισορροπημένη βαθμολογία βασισμένη στους έξω-βαθμούς και επομένως, είναι μάλλον πλησιέστερα στον `BCC` παρά στον `CC`. Από την άλλη μεριά, ο `HITS Authorities (HA)` είναι πλησιέστερα στον `CC` παρά

$P_{a_1}(i)$	$P_{a_2}(i)$	$w(a_1, a_2, i)$	$d_w(a_1, a_2, i)$
1	2	1	1
10	11	0.1	0.1
50	52	0.02	0.04
1	11	1	10
100	110	0.01	0.1
1010	1000	0.001	0.01

Πίνακας 3.13. Παραδείγματα περιπτώσεων με μέτρο την Απόσταση με Βάρη.

στον BCC. Ο Prestige (P) είναι αρκετά κοντά στον PS, αλλά μακριά από τους υπόλοιπους αλγορίθμους. Ο SALSA (SA) είναι αρκετά κοντά στον CC και σε άλλους αλγορίθμους "αυθεντιών μόνο" (πιο κοντά από τον HA). Ο B-HITS (BHA) δείχνει να αποκλίνει από τον HA. Ο B-SALSA (BSA) κρατά μία απόσταση σε σχέση με όλους τους υπόλοιπους αλγορίθμους.

3.4.3.7 Συζήτηση Σχετικά με τις Συγκρίσεις

Ξεκινώντας με την πρώτη μέθοδο σύγκρισης στον Πίνακα 3.9, μπορούμε να ανιχνεύσουμε 3 ομάδες αλγορίθμων με παρόμοια αποτελέσματα:

1. BEPS, SCEAS_B1, SCEAS_B0
2. CC, BCC, SA
3. BPS, PR, EPS, PS

Από την άλλη μεριά, ο Prestige (P) δείχνει να είναι ο πιο ανόμοιος σε σχέση με τους περισσότερους αλγορίθμους και κυρίως σε σχέση με τους CC, BCC, SA και HA. Είναι εμφανές ότι ο Πίνακας 3.9, δεν δίνει εσωτερικές πληροφορίες, αφού βασίζεται σε μία πολύ απλή μετρική μέθοδο.

Αναφερόμενοι στη συνάρτηση *Top(x)* μπορούμε να έχουμε καλύτερη κατανόηση της (αν-)ομοιότητας των αλγορίθμων. Δυστυχώς, είναι πολύ δύσκολο να μελετήσει κανείς 91 διαγράμματα (που είναι ο συνολικός αριθμός των συνδυασμών). Από τα διαγράμματα του Σχήματος 3.6, βλέπουμε ότι στα 100 πρώτα στοιχεία, το ποσοστό των κοινών στοιχείων είναι σχεδόν σταθερό και πολύ κοντά στην τιμή που απεικονίζεται στον Πίνακα 3.9.

	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1	SCEAS_B0
CC	-	3.6%	3.6%	5.1%	10.4%	1.6%	7.1%	2.7%	5.0%	3.8%	3.9%	3.3%	3.3%	3.3%
BCC	3.6%	-	1.8%	8.0%	12.0%	3.2%	9.3%	4.4%	6.8%	2.2%	5.8%	0.7%	0.6%	0.6%
PR	3.6%	1.8%	-	6.5%	8.6%	3.5%	6.9%	3.3%	4.4%	0.2%	3.9%	0.9%	1.0%	1.0%
HA	5.1%	8.0%	6.5%	-	6.6%	5.6%	4.3%	3.4%	4.2%	6.4%	3.8%	7.0%	7.1%	7.1%
P	10.4%	12.0%	8.6%	6.6%	-	10.8%	5.2%	6.2%	3.7%	8.1%	4.5%	10.4%	10.6%	10.6%
SA	1.6%	3.2%	3.5%	5.6%	10.8%	-	7.9%	3.3%	5.5%	3.7%	4.4%	3.0%	3.0%	3.0%
BHA	7.1%	9.3%	6.9%	4.3%	5.2%	7.9%	-	3.9%	3.9%	6.6%	4.0%	8.0%	8.2%	8.2%
BSA	2.7%	4.4%	3.3%	3.4%	6.2%	3.3%	3.9%	-	2.7%	3.2%	2.2%	3.9%	3.9%	3.9%
PS	5.0%	6.8%	4.4%	4.2%	3.7%	5.5%	3.9%	2.7%	-	4.1%	0.6%	5.5%	5.7%	5.7%
BPS	3.8%	2.2%	0.2%	6.4%	8.1%	3.7%	6.6%	3.2%	4.1%	-	3.6%	1.2%	1.3%	1.3%
EPS	3.9%	5.8%	3.9%	3.8%	4.5%	4.4%	4.0%	2.2%	0.6%	3.6%	-	4.8%	4.9%	4.9%
BEPS	3.3%	0.7%	0.9%	7.0%	10.4%	3.0%	8.0%	3.9%	5.5%	1.2%	4.8%	-	0.1%	0.1%
SCEAS_B1	3.3%	0.6%	1.0%	7.1%	10.6%	3.0%	8.2%	3.9%	5.7%	1.3%	4.9%	0.1%	-	0.0%
SCEAS_B0	3.3%	0.6%	1.0%	7.1%	10.6%	3.0%	8.2%	3.9%	5.7%	1.3%	4.9%	0.1%	0.0%	-

Πίνακας 3.14. Διαφορές στις κατατάξεις δημοσιεύσεων του DBLP υπολογισμένες με τη συνάρτηση $D_w(a_1, a_2)$

Η επόμενη μετρική μέθοδος, το Kendall's tau, μπορεί να δώσει περισσότερες πληροφορίες. Στον πίνακα ασθενούς απόστασης (βλέπε Πίνακα 3.10), παρατηρούμε ότι η ομάδα αλγορίθμων BEPS - SCEAS_B0 - SCEAS_B1 είναι πολύ δυνατή και παραμένει επίσης δυνατή με την αυστηρή μέτρηση απόστασης. Η επόμενη ομάδα που βρήκαμε με τη μετρική μέθοδο *Top20* (CC-BCC-SA) φαίνεται να παραβιάζεται στον πίνακα ασθενούς απόστασης, αφού μόνο ο CC και ο SA παραμένουν στην ομάδα. Από τον Πίνακα 3.11 παρατηρούμε ακόμα και οι CC και SA αποκλίνουν περισσότερο ο ένας από τον άλλον και επομένως η ομάδα εξαφανίζεται. Και στους δύο Πίνακες (3.10 και 3.11), φαίνεται ότι η τελευταία ομάδα των (BPS-PR-EPS-PS) χωρίζεται σε δύο ομάδες: την BPS-PR και την EPS-PS. Επιπλέον, η ομάδα BPS-PR είναι πολύ κοντά στην ομάδα BEPS-SCEAS_B1-SCEAS_B0, ενώ ο BCC είναι κοντά στην BEPS-SCEAS_B1-SCEAS_B0, ενώ ο BCC είναι κοντά στην BEPS-SCEAS_B1-SCEAS_B0, αλλά όχι τόσο κοντά στην BPS-PR.

Ο Πίνακας 3.12 (Spearman's footrule) δείχνει τις ίδιες ομάδες όπως και οι Πίνακες 3.10 και 3.11. Επιπλέον, στον πίνακα αυτό είναι σαφέστερο ότι οι HA, P και BHA απέχουν σχεδόν από όλους τους υπόλοιπους αλγορίθμους.

3.5 Σχολιασμός Αποτελεσμάτων

Αυτή η παράγραφος αποτελείται από δύο τμήματα. Πρώτα, θεωρούμε τις δημοσιεύσεις των συνεδρίων VLDB και SIGMOD, εκτελούμε μία κατάταξη και επιχειρούμε μια *αξιολόγηση της κατάταξης* συγκρίνοντας τα αποτελέσματα με τις βραβευμένες δημοσιεύσεις με το 'VLDB 10 Year Award' και το 'SIGMOD Test of Time Award'. Στο δεύτερο τμήμα, χρησιμοποιούμε τα αποτελέσματα της κατάταξης-βαθμολογίας των δημοσιεύσεων για τη δημιουργία ενός πίνακα κατάταξης-βαθμολογίας συγγραφέων. Αξιολογούμε τα αποτελέσματα συγκρίνοντάς τα με τη λίστα βραβευμένων συγγραφέων-ερευνητών με το 'Edgar F. Codd Innovations Award'.

3.5.1 Κατάταξη Δημοσιεύσεων

Η αξιολόγηση των αλγορίθμων κατάταξης είναι μία μάλλον δύσκολη εργασία για επιστημονικές δημοσιεύσεις, εφ'όσον είναι υποκειμενική η απόφαση για το ποιά είναι καλύτερη. Ένα κριτήριο επιβεβαίωσης της ορθότητας των αποτελεσμάτων της κατάταξης είναι να βασισθούμε σε δύο πολύ γνωστά βραβεία δημοσιεύσεων: το 'VLDB 10 Year Award' και το 'SIGMOD Test of Time Award'. Δεχόμαστε ότι εφ'όσον ένας αλγόριθμος δίνει υψηλές θέσεις κατάταξης στις βραβευμένες

δημοσιεύσεις, τότε αυτός ο αλγόριθμος μπορεί να χρησιμοποιηθεί με ασφάλεια για την αξιολόγηση-κατάταξη δημοσιεύσεων.

Συγκρίνουμε όλες τις μεθόδους κατάταξης: τους `PageRank`, `SCEAS` και τις παραλλαγές τους, τον `HITS` και τους `B-HITS (Authorities)`, `SALSA` και τους `B-SALSA (Authorities)`, `Prestige`, `CC` και `BCC`. Ο `SCEAS` με $b = 0$ δεν παρουσιάζεται εδώ, επειδή παράγει παρόμοια αποτελέσματα με τον `SCEAS` με $b = 1$ για το σύνολο δεδομένων της ψηφιακής βιβλιοθήκης DBLP. Το σενάριο της δοκιμής έχει ως εξής:

1. Εκτέλεση όλων των αλγορίθμων κατάταξης στο γράφο αναφορών της DBLP.
2. Διαχωρισμός των εργασιών που έχουν δημοσιευθεί στα πρακτικά των συνεδρίων VLDB και SIGMOD (προβολή του V_{DBLP} στο V_{vldb} και V_{sigmod}, αντίστοιχα). Πέραν αυτού του σημείου αγνοούνται όλες οι υπόλοιπες.
3. Οργάνωση των πινάκων κατάταξης κάνοντας ομαδοποιήσεις σύμφωνα με το συνέδριο και το έτος. Κάποιοι από αυτούς τους πίνακες παρουσιάζονται στο Παράρτημα Β.
4. Εξέταση της θέσης των βραβευμένων δημοσιεύσεων στους προηγούμενους πίνακες κατάταξης.

Στους Πίνακες 3.15 και 3.16 παρουσιάζουμε τις βραβευμένες δημοσιεύσεις και τη θέση κατάταξης τους που προκύπτει από όλες τις μεθόδους κατάταξης. Για παράδειγμα, η εργασία με τίτλο 'Fast Algorithms for Mining Association Rules in Large DBs, 1994' (βλέπε Πίνακα 3.16) κατατάσσεται 1η από τον `PageRank`, 1η από τον `SCEAS`, 4η από τον `HITS` (ως αυθεντία), 16η από τον `Prestige` και 1η από τους `CC`, `BCC`, `PR`, `SA`, `BSA`. Σε αυτό το σημείο, ας σημειώσουμε πάλι, ότι έτσι δεν κάνουμε αξιολόγηση-έλεγχο των επιτροπών βραβείων, αλλά των μεθόδων κατάταξης. Σε αυτούς τους λεπτομερείς πίνακες παρατηρούμε ότι οι βραβευμένες δημοσιεύσεις έχουν γενικά καταταχθεί υψηλά. Φυσικά, κάποιες παρεκκλίσεις και εξαιρέσεις υπάρχουν. Αυτές οι εξαιρέσεις μπορεί να υπάρχουν για διάφορους λόγους:

- Το δείγμα των αναφορών μας μπορεί να μην είναι αρκετά μεγάλο: π.χ. μία βραβευμένη δημοσίευση μπορεί να δέχεται πολλές αναφορές από επιστημονικές περιοχές που δεν περιλαμβάνονται στο σύνολο δεδομένων της ψηφιακής βιβλιοθήκης DBLP.
- Εξ ορισμού τα βραβεία είναι υποκειμενικά: π.χ. η απόφαση μίας επιτροπής βραβείων μπορεί να βασίζεται σε αντικειμενικούς παράγοντες (όπως ο

Έτος	Τίτλος	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1
1988	A Case for Redundant Arrays of Inexpensive Disks (RAID). D.A. Patterson, G.A. Gibson, R.H. Katz	2	1	1	8	7	2	10	3	4	1	4	1	1
1989	F-Logic: a Higher-Order language for Reasoning about Objects, Inheritance, and Scheme. M. Kifer, G. Lausen	5	4	6	5	4	5	5	5	8	6	7	4	4
1990	Encapsulation of Parallelism in the Volcano Query Processing System. G. Graefe	10	9	10	5	9	10	7	10	7	10	7	9	9
1990	Set-Oriented Production Rules in Relational Database Systems. J. Widom, S.J. Finkelstein	3	3	3	4	15	3	4	3	3	3	3	3	3
1992	Extensible/Rule Based Query Rewrite Optimization in Starburst. H. Pirahesh, J.M. Hellerstein, W. Hasan	3	2	1	2	5	3	5	2	3	1	3	1	1
1992	Querying Object-Oriented Databases. M. Kifer, W. Kim, Y. Sagiv	1	5	2	1	8	1	1	3	2	2	1	3	3
1993	Mining Association Rules between Sets of Items in Large Databases. R. Agrawal, T. Imielinski, A.N. Swami	1	1	1	6	26	1	5	1	1	1	1	1	1
1994	From Structured Documents to Novel Query Facilities. V. Christophides, S. Abiteboul, S. Cluet, M. Scholl	2	2	2	7	8	2	3	2	1	2	2	2	2
1994	Shoring Up Persistent Applications. M.J. Carey, D.J. DeWitt, M.J. Franklin, N.E. Hall, M.L. McAuliffe, J.F. Naughton, D.T. Schuh, M.H. Solomon, C.K. Tan, O.G. Tsatalos, S.J. White, M.J. Zwilling	1	1	1	1	1	1	2	1	2	1	1	1	1
	Άθροισμα Θέσεων	28	28	27	39	83	28	42	30	31	27	29	25	25

Πίνακας 3.15. Βραβευμένες δημοσιεύσεις του συνεδρίου SIGMOD.

Έτος	Τίτλος	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1
1986	Object and File Management in the EXODUS Extensible Database System. M.J. Carey, D.J. DeWitt, J.E. Richardson, E.J. Shekita	2	3	2	1	2	2	2	2	1	2	2	3	3
1987	The R^+-Tree: a Dynamic Index for Multi-Dimensional Objects. T.K. Sellis, N. Roussopoulos, C. Faloutsos	1	1	1	1	19	1	1	1	1	1	1	1	1
1988	Disk Shadowing. D. Bitton, J. Gray	2	1	1	5	8	2	11	3	2	1	2	1	1
1989	ARIES/NT: a Recovery Method Based on Write-Ahead Logging for Nested Transactions. K. Rothermel, C. Mohan	6	9	6	14	1	6	13	7	1	6	2	7	7
1990	Deriving Production Rules for Constraint Maintainance. S. Ceri, J. Widom	1	1	1	3	17	1	6	1	1	1	1	1	1
1991	A Transactional Model for Long-Running Activities. U. Dayal, M. Hsu, R. Ladin	4	2	2	24	22	4	17	7	12	2	9	2	2
1992	Querying in Highly Mobile Distributed Environments. T. Imielinski, B.R. Badrinath	12	7	3	43	10	12	30	15	11	3	12	4	4
1993	Universality of Serial Histograms. Y.E. Ioannidis	5	9	6	8	11	5	8	4	4	6	4	7	7
1994	Fast Algorithms for Mining Association Rules in Large Databases. R. Agrawal, R. Srikant	1	1	1	4	16	1	8	1	1	1	1	1	1
1995	W3QS: a Query System for the World-Wide Web. D. Konopnicki, O. Shmueli	3	2	4	7	3	3	1	4	2	4	2	3	3
	Άθροισμα Θέσεων	37	36	27	110	109	37	97	45	36	27	36	30	30

Πίνακας 3.16. Βραβευμένες δημοσιεύσεις του συνεδρίου VLDB.

αριθμός των αναφορών που έχει δεχθεί η δημοσίευση) αλλά επίσης μπορεί να συνδυάζει και άλλα μέτρα και ενδείξεις του αντικτύπου που έχει.

- Είναι μάλλον ασυνήθιστο για ένα συγγραφέα να βραβευθεί δύο φορές από τον ίδιο οργανισμό.

Για να έχουμε μια γενική εικόνα της απόδοσης των μεθόδων κατάταξης, στην τελευταία γραμμή προσθέτουμε τις θέσεις των βραβευμένων δημοσιεύσεων. Όσο μικρότερο είναι το άθροισμα, τόσο καλύτερη είναι η μέθοδος κατάταξης. Παρατηρούμε ότι και στους δύο πίνακες ο `HITS (Authorities)` και ο `Prestige`

είναι οι με διαφορά χειρότερες μέθοδοι. Αυτό επιβεβαιώνει τις παρατηρήσεις μας, που εξηγούνται στην Παράγραφο 3.2. Ο BHITS είναι ελαφρώς καλύτερος από τον HITS για την περίπτωση του Πίνακα 3.16. Μεταξύ των άλλων δέκα μεθόδων, CC, BCC, SA, BSA, PS και EPS φαίνονται να είναι στην ίδια μεσαία κατηγορία, αλλά ωστόσο παραμένουν αρκετά καλές μέθοδοι κατάταξης. Τέλος, βλέπουμε ότι οι PageRank, BPS, BEPS και SCEAS πλησιάζουν μεταξύ τους και εναλάσσονται στη νικήτρια θέση στους δύο Πίνακες 3.15 και 3.16. Στη συνέχεια, θα αγνοήσουμε τους HITS (Authorities) και Prestige, αφού δεν είναι κατάλληλοι για τους δικούς μας σκοπούς κατάταξης.

Εφ'όσον όλοι οι αλγόριθμοι (εκτός από τον HITS και τον Prestige) κατέληξαν με μία παρόμοια συμπεριφορά, θα προσπαθήσουμε να παράγουμε ένα μοναδικό πίνακα κατάταξης λαμβάνοντας το μέσο όρο των αποτελεσμάτων τους σύμφωνα με τη λογική του [79]. Με απλά λόγια, θα υπολογίσουμε και τις δέκα κατατάξεις και θα αναθέσουμε σε κάθε δημοσίευση έναν αριθμό πόντων (5 έως 1) ανάλογα με τη θέση της στο συγκεκριμένο πίνακα κατάταξης. Για παράδειγμα, σε κάθε πίνακα η πρώτη δημοσίευση παίρνει 5 πόντους, η δεύτερη παίρνει 4 πόντους κ.ο.κ. Έτσι, αν μία δημοσίευση καταταχθεί ως πρώτη και στις 10 κατατάξεις, τότε θα λάβει 50 πόντους. Επομένως, για να παράγουμε τους νέους πίνακες κατάταξης επαναλαμβάνουμε τα βήματα 3 και 4 του προηγούμενου σεναρίου.

Αυτός ο τελευταίος υπολογισμός αποικονίζεται στους Πίνακες 3.17 και 3.18, όπου η στήλη 'Pos' αντιπροσωπεύει τη θέση της αντίστοιχης δημοσίευσης. Είναι εύκολο να παρατηρήσουμε ότι η πλειοψηφία των βραβευμένων δημοσιεύσεων κατατάσσονται στις κορυφαίες 3 θέσεις αυτών των νέων πινάκων κατάταξης. Έπειτα από εξαντλητικά πειράματα καταλήξαμε επίσης στο συμπέρασμα, ότι το άθροισμα των θέσεων είναι μικρότερο χρησιμοποιώντας αυτή την προσέγγιση των μέσων όρων, σε σύγκριση με οποιαδήποτε ανεξάρτητη μέθοδο κατάταξης. Πιο συγκεκριμένα, παρατηρούμε ότι στην περίπτωση του SIGMOD (βλέπε Πίνακα 3.17) το άθροισμα των θέσεων (δηλαδή, 26) είναι καλύτερο από το μέσο όρο αθροίσματος του Πίνακα 3.15 και ελαφρώς μεγαλύτερο από τους καλύτερους αλγορίθμους αυτής της περίπτωσης, (δηλαδή, 25 για τον BEPS και για τον SCEAS General) κυρίως λόγω της δημοσίευσης του 1990/2000 η οποία αποτελεί μια εξαίρεση. Στην περίπτωση του VLDB (Πίνακας 3.18) το άθροισμα (δηλαδή, 28) είναι και πάλι μικρότερο από το μέσο άθροισμα του Πίνακα 3.16 και ελαφρώς μεγαλύτερο από τους PageRank (27) και BPS (27). Στο Παράρτημα Β παρουσιάζουμε μία λεπτομερή κατάταξη των δημοσιεύσεων του SIGMOD'95 για κάποιες από τις μεθόδους, όπως και τις 3 κορυφαίες δημοσιεύσεις για κάθε έτος από το 1995 έως το 1998.

Έτος	Τίτλος	Θέση	Βαθμοί
1988	A Case for Redundant Arrays of Inexpensive Disks (RAID). D.A. Patterson, G.A. Gibson, R.H. Katz	1	37
1989	F-Logic: a Higher-Order language for Reasoning about Objects, Inheritance, and Scheme. M. Kifer, G. Lausen	5	9
1990	Encapsulation of Parallelism in the Volcano Query Processing System. G. Graefe	8	0
1990	Set-Oriented Production Rules in Relational Database Systems. J. Widom, S. J. Finkelstein	3	27
1992	Extensible/Rule Based Query Rewrite Optimization in Starburst. H. Pirahesh, J.M. Hellerstein, W. Hasan	2	37
1992	Querying Object-Oriented Databases. M. Kifer, W. Kim, Y. Sagiv	3	31
1993	Mining Association Rules between Sets of Items in Large Databases. R. Agrawal, T. Imielinski, A.N. Swami	1	45
1994	From Structured Documents to Novel Query Facilities. V. Christophides, S. Abiteboul, S. Cluet, M. Scholl	2	37
1994	Shoring Up Persistent Applications. M.J. Carey, D.J. DeWitt, M.J. Franklin, N.E. Hall, M.L. McAuliffe, J.F. Naughton, D.T. Schuh, M.H. Solomon, C.K. Tan, O.G. Tsatalos, S.J. White, M.J. Zwilling	1	44
	Άθροισμα θέσεων	26	

Πίνακας 3.17. Αθροιστικές θέσεις των βραβευμένων δημοσιεύσεων με το SIGMOD Test of Time.

Έτος	Τίτλος	Θέση	Βαθμοί
1986	Object and File Management in the EXODUS Extensible Database System. M.J. Carey, D.J. DeWitt, J.E. Richardson, E.J. Shekita	2	34
1987	The R^+-Tree: a Dynamic Index for Multi-Dimensional Objects. T.K. Sellis, N. Roussopoulos, C. Faloutsos	1	42
1988	Disk Shadowing. D. Bitton, J. Gray	1	38
1989	ARIES/NT: a Recovery Method Based on Write-Ahead Logging for Nested Transactions. K. Rothermel, C. Mohan	6	5
1990	Deriving Production Rules for Constraint Maintenance. S. Ceri, J. Widom	1	43
1991	A Transactional Model for Long-Running Activities. U. Dayal, M. Hsu, R. Ladin	3	24
1992	Querying in Highly Mobile Distributed Environments. T. Imielinski, B.R. Badrinath	4	10
1993	Universality of Serial Histograms. Y.E. Ioannidis	6	6
1994	Fast Algorithms for Mining Association Rules in Large Databases. R. Agrawal, R. Srikant	1	45
1995	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli	3	30
	Άθροισμα θέσεων	28	

Πίνακας 3.18. Αθροιστικές θέσεις των βραβευμένων δημοσιεύσεων με το VLDB 10 Year Award.

3.5.2 Κατάταξη Συγγραφέων

Μπορούμε να βασισθούμε στη μέθοδο για τον υπολογισμό βαθμολογιών για δημοσιεύσεις, όπως επίσης και για τον υπολογισμό βαθμολογιών για συγγραφείς. Μία προσέγγιση θα ήταν να υπολογίσουμε το μέσο όρο βαθμολογίας όλων των δημοσιεύσεών τους. Και πάλι, η διαδικασία αυτή δεν είναι εύκολη. Για παρά-

δειγμα, ο συγγραφέας A έχει 200 δημοσιεύσεις εκ των οποίων μόνο οι 40 είναι *πρώτης κατηγορίας*. Ας υποθέσουμε ότι αυτές οι υψηλής ποιότητας δημοσιεύσεις έχουν η καθεμία βαθμολογία 10 βαθμών, ενώ οι υπόλοιπες έχουν 1 βαθμό. Ο συγγραφέας B έχει συνολικά 20 δημοσιεύσεις, εκ των οποίων οι 10 ανήκουν στην πρώτη κατηγορία. Είναι λογικό να θεωρήσουμε ότι ο συγγραφέας A θα έπρεπε να καταταχθεί υψηλότερα από το συγγραφέα B λόγω της επιστημονικής του συνεισφοράς, αφού ο συγγραφέας A έχει 4 φορές περισσότερες πρώτης κατηγορίας δημοσιεύσεις σε σχέση με το συγγραφέα B. Ωστόσο, αν απλά υπολογίσουμε το μέσο όρο όλων των βαθμολογιών των δημοσιεύσεων, τότε οι συγγραφείς θα είχαν 2.8 και 5.5 βαθμούς αντίστοιχα. Επομένως, δεν είναι δίκαιο να λάβουμε υπ'όψη μας όλες τις δημοσιεύσεις που έχει συγγράψει ένα άτομο.

Επιπλέον, δεν είναι δίκαιο να λάβουμε υπ'όψη μας διαφορετικό αριθμό δημοσιεύσεων για κάθε συγγραφέα (π.χ. 40 δημοσιεύσεις για τον A και 10 για τον B). Στην προσέγγιση μας λαμβάνουμε υπ'όψη μας τον ίδιο αριθμό δημοσιεύσεων για όλους τους συγγραφείς, έτσι ώστε τα αποτελέσματά μας να είναι συγκρίσιμα. Επομένως, τώρα το πρόβλημα έγκειται στην επιλογή του αριθμού των δημοσιεύσεων κάθε συγγραφέα που θα συμμετέχει στην αξιολόγηση. Εκτελέσαμε το ακόλουθο πείραμα, ώστε να επιλέξουμε αυτόν τον αριθμό. Υπολογίσαμε το μέσο όρο βαθμολογιών για κάθε συγγραφέα, χρησιμοποιώντας τις x καλύτερες δημοσιεύσεις του, $\forall x \in \{1, 3, 5, 8, 10, 15, 20, 25, 30, 40\}$. Έτσι, δημιουργήσαμε 10 κατατάξεις για κάθε μέθοδο κατάταξης. Ως δεδομένα ελέγχου χρησιμοποιήσαμε τους συγγραφείς που βραβεύθηκαν στο 'SIGMOD Edgar F. Codd Innovations Award'. Όσο υψηλότερα κατατάχθηκαν αυτοί οι συγγραφείς, τόσο καλύτερη θεωρήθηκε η εκτίμηση που έγινε. Στην Παράγραφο 3.5.2.1 παρουσιάζουμε τα αποτελέσματα που παρήχθησαν από αυτό το πείραμα. Βασιζόμενοι σε αυτό το πείραμα, συμπεράναμε ότι ένας μέσος όρος των 25 καλύτερων δημοσιεύσεων είναι το καταλληλότερο μέτρο (και εναλλακτικά ο μέσος όρος των 30 καλύτερων δημοσιεύσεων κάθε συγγραφέα).

Στους Πίνακες 3.19 και 3.20 συγκρίνουμε τις διάφορες μεθόδους κατάταξης. Στους πίνακες αυτούς παρουσιάζουμε τις θέσεις κατάταξης των βραβευμένων συγγραφέων για κάθε μέθοδο κατάταξης, παίρνοντας υπ'όψη μας τη μέση βαθμολογία των καλύτερων 25 και 30 δημοσιεύσεων κάθε συγγραφέα, αντίστοιχα. Είναι προφανές ότι η στήλη `SCEAS_B1` από τον Πίνακα 3.19 είναι πανομοιότυπη με τη στήλη 'best25' του Πίνακα 3.23, και η στήλη `SCEAS_B1` του Πίνακα 3.20 είναι πανομοιότυπη με τη στήλη 'best30' του Πίνακα 3.23.

Οι δύο τελευταίες γραμμές των Πινάκων 3.19-3.20 δείχνουν τη θέση κατάταξης του βραβευμένου συγγραφέα που κατατάχθηκε ως τελευταίος *το χαμηλότερο*

Όνομα	Θέση κατάταξης με:												
	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1
Michael Stonebraker	2	3	4	2	4	2	3	2	2	5	2	4	4
Jim Gray	3	1	2	9	9	3	10	4	6	2	5	1	1
Philip Bernstein	6	7	6	6	2	6	1	5	5	6	6	8	8
David DeWitt	1	4	8	4	55	1	21	8	11	8	8	5	5
C. Mohan	29	32	40	62	176	29	95	40	41	43	34	34	33
David Maier	9	10	13	8	21	9	23	12	14	14	13	11	11
Serge Abiteboul	13	19	23	16	15	13	62	21	26	24	22	22	21
Hector Garcia-Molina	21	16	20	115	291	21	201	36	69	21	50	19	19
Rakesh Agrawal	15	12	16	82	407	15	182	28	70	17	45	12	12
Rudolf Bayer	73	51	24	117	32	74	19	30	17	20	24	36	40
Patricia Selinger	38	39	28	24	73	38	41	25	19	26	18	32	34
Donald Chamberlin	14	9	5	11	3	14	9	7	3	4	4	7	7
Ronald Fagin	17	18	10	17	5	17	7	9	9	10	11	14	14
Χαμηλότερη Θέση	73	51	40	117	407	74	201	40	70	43	50	36	40
Άθροισμα Θέσεων	241	221	199	473	1093	242	674	227	292	200	242	205	209

Πίνακας 3.19. Θέσεις των βραβευμένων συγγραφέων με το M.O. των 25 καλύτερων δημοσιεύσεών τους.

Όνομα	Θέση κατάταξης με:												
	CC	BCC	PR	HA	P	SA	BHA	BSA	PS	BPS	EPS	BEPS	SCEAS_B1
Michael Stonebraker	1	2	4	1	4	1	3	2	2	4	1	3	3
Jim Gray	3	1	2	10	9	3	10	4	6	2	5	1	1
Philip A. Bernstein	6	7	6	6	2	6	1	5	5	6	6	7	7
David DeWitt	2	3	8	3	50	2	21	8	11	8	8	5	5
C. Mohan	29	30	39	57	162	29	90	41	39	42	34	32	32
David Maier	8	12	14	9	20	8	23	13	14	14	13	11	12
Serge Abiteboul	12	19	22	15	14	12	58	20	23	23	21	21	21
Hector Garcia-Molina	21	14	20	101	270	20	184	35	64	20	46	18	17
Rakesh Agrawal	14	11	16	68	371	14	168	26	61	17	40	12	11
Rudolf Bayer	72	53	24	111	31	73	19	32	17	22	25	41	41
Patricia Selinger	42	43	31	25	68	42	39	28	19	27	19	34	35
Donald Chamberlin	16	9	5	11	3	16	9	7	3	5	4	8	8
Ronald Fagin	19	18	10	17	5	19	7	9	9	10	11	15	16
Χαμηλότερη Θέση	72	53	39	111	371	73	184	41	64	42	46	41	41
Άθροισμα Θέσεων	245	222	201	434	1009	245	632	230	273	200	233	208	209

Πίνακας 3.20. Θέσεις των βραβευμένων συγγραφέων με το M.O. των 30 καλύτερων δημοσιεύσεών τους.

σημείο κατάταξης και το *άθροισμα των θέσεων κατάταξης* όλων των βραβευμένων συγγραφέων. Αυτοί οι δύο αριθμοί εξυπηρετούν ως μέτρο για τη σύγκριση των κατατάξεων. Όσο μικρότεροι είναι αυτοί οι αριθμοί, τόσο καλύτερη κατάταξη

έχει επιτευχθεί. Σε αυτό τον πίνακα μπορούμε να δούμε ότι ο HITS authorities και ο Prestige είναι και πάλι οι κατά πολύ χειρότερες μέθοδοι, αφού το άθροισμα των θέσεων κατάταξης και το χαμηλότερο σημείο κατάταξης είναι περίπου 2-3 φορές μεγαλύτερα από τους αντίστοιχους αριθμούς που υπολογίζονται για τις άλλες μεθόδους. Η Απλή Καταμέτρηση Αναφορών (Plain Citation Count - CC) και η Ισορροπημένη Καταμέτρηση Αναφορών (Balanced Citation Count - BCC) δίνουν αποδεκτά αποτελέσματα. Με διαφορά οι καλύτερες μέθοδοι είναι οι PageRank, B-SALSA, BPS, BEPS και SCEAS. Επίσης, ο B-SALSA βελτιώνει τον SALSA κατά περισσότερο από 50%. Στο Παράρτημα Β παρουσιάζουμε τα πλήρη αποτελέσματα κατάταξης συγγραφέων.

3.5.2.1 Το Πείραμα της Κατάταξης Συγγραφέων

Σε αυτήν την παράγραφο παρουσιάζουμε τα αποτελέσματα που παρήχθησαν από το πείραμα που αναφέρεται στην Παράγραφο 3.5.2, βοηθώντας στην εύρεση του αριθμού των δημοσιεύσεων που θα έπρεπε να λαμβάνουμε υπ'όψη μας για κάθε συγγραφέα. Παρουσιάζουμε τα αποτελέσματα κατάταξης των μεθόδων CC (Πίνακας 3.21), BEPS (Πίνακας 3.22) και SCEAS (Πίνακας 3.23). Χάριν συντομίας, παρουσιάζουμε μόνο τους βραβευμένους συγγραφείς και τη θέση τους για κάθε επιλεγμένο αριθμό *κορυφαίων* δημοσιεύσεων. Για παράδειγμα, ο Hector-Garcia Molina κατατάσσεται $81^{ος}$ στην κατάταξη του SCEAS (Πίνακας 3.23) όταν η βαθμολογία παράγεται από τη βαθμολογία μόνο της μιας κορυφαίας δημοσίευσης κάθε συγγραφέα. Κατατάσσεται $41^{ος}$, εάν η βαθμολογία του παράγεται από το μέσο όρο 3 κορυφαίων δημοσιεύσεων, κ.ο.κ.

Οι δύο τελευταίες γραμμές των Πινάκων 3.21-3.23 δείχνουν τη θέση κατάταξης του βραβευμένου συγγραφέα που κατατάχθηκε τελευταίος (Χαμηλότερη Θέση) και το άθροισμα των θέσεων κατάταξης για όλους τους βραβευμένους συγγραφείς. Αυτοί οι δύο αριθμοί είναι το μέτρο σύγκρισης των κατατάξεων (Άθροισμα Θέσεων). Όσο μικρότεροι είναι αυτοί οι αριθμοί, τόσο καλύτερη κατάταξη επιτυγχάνεται. Το χαμηλότερο σημείο κατάταξης είναι σημαντικά υψηλότερο, όταν το μέσο όρο των κορυφαίων 1-3 δημοσιεύσεων κάθε συγγραφέα. Αυτό οφείλεται στο γεγονός ότι οι συν-συγγραφείς μιας κορυφαίας δημοσίευσης πλεονεκτούν και *ανεβαίνουν* στην κατάταξη. Επομένως, αυξάνοντας τον αριθμό των επιλεγμένων κορυφαίων δημοσιεύσεων, οι βραβευμένοι συγγραφείς κινούνται προς την κορυφή του πίνακα κατάταξης. Αυτή η τάση διαρκεί έως ότου ο αριθμός των επιλεγμένων δημοσιεύσεων γίνει 25. Η ίδια παρατήρηση ισχύει εάν θεωρήσουμε την έννοια του αθροίσματος των θέσεων κατάταξης. Επίσης παρατηρούμε

ότι στη στήλη 40 χάνουμε δύο βραβευμένους συγγραφείς, επειδή έχουν λιγότερες από 40 δημοσιεύσεις στην ψηφιακή βιβλιοθήκη DBLP. Για λόγους συντομίας, δεν συμπεριλάβαμε τους αντίστοιχους πίνακες για τις άλλες μεθόδους κατάταξης, αφού τα αποτελέσματα είναι παρόμοια και η χαμηλότερη θέση κατάταξης είναι η ίδια. Έτσι, έπειτα από αυτό το πείραμα αποφασίσαμε να κατατάξουμε τους συγγραφείς παίρνοντας το μέσο όρο των 25 καλύτερων δημοσιεύσεών τους. Ως δεύτερη επιλογή, διαλέξαμε το μέσο όρο των 30 καλύτερων δημοσιεύσεων.

Μία τελευταία παρατήρηση σε σχέση με αυτό τον πίνακα είναι η ακόλουθη. Αρκετά ενδιαφέρον και εύκολα εξηγήσιμο είναι το γεγονός ότι υπάρχουν πολ-

Όνομα	Θέση κατάταξης με:										
	best1	best3	best5	best8	best10	best15	best20	best25	best30	best40	BCAvg
C. Mohan	102	92	79	54	48	36	34	29	29	26	52
David J. DeWitt	34	16	12	7	5	4	1	1	2	2	5
David Maier	44	19	13	11	9	9	8	9	8	7	12
Donald D. Chamberlin	5	4	4	6	7	10	11	14	16		6
Hector Garcia-Molina	146	58	42	35	31	27	26	21	21	15	43
Jim Gray	8	3	2	2	2	2	2	3	3	4	2
Michael Stonebraker	23	9	5	4	4	3	3	2	1	1	4
Patricia G. Selinger	3	13	17	21	23	31	36	38	42		21
Philip A. Bernstein	52	21	15	13	10	7	6	6	6	5	13
Rakesh Agrawal	101	44	35	27	25	20	15	15	14	12	31
Ronald Fagin	86	33	23	20	20	19	16	17	19	18	26
Rudolf Bayer	45	42	48	62	65	73	73	73	72	69	66
Serge Abiteboul	83	31	25	19	19	16	14	13	12	8	20
Χαμηλότερη Θέση	146	92	79	62	65	73	73	73	72	69	66
Άθροισμα Θέσεων	732	385	320	281	268	257	245	241	245	167	301

Πίνακας 3.21. Θέσεις των βραβευμένων συγγραφέων με τη μέθοδο CC.

Όνομα	Θέση κατάταξης με:										
	best1	best3	best5	best8	best10	best15	best20	best25	best30	best40	BCAvg
C. Mohan	104	105	84	69	61	48	42	34	32	27	65
David J. DeWitt	30	20	15	13	12	7	7	5	5	3	11
David Maier	39	28	21	17	16	13	12	11	11	8	17
Donald D. Chamberlin	9	4	4	4	4	4	5	7	8		4
Hector Garcia-Molina	81	44	38	30	27	22	22	19	18	13	34
Jim Gray	3	3	2	2	2	2	2	1	1	1	2
Michael Stonebraker	21	10	9	7	5	5	4	4	3	2	6
Patricia G. Selinger	7	22	24	26	26	32	33	32	34		27
Philip A. Bernstein	57	27	18	14	11	8	8	8	7	5	15
Rakesh Agrawal	48	33	27	22	20	17	15	12	12	7	24
Ronald Fagin	41	29	22	18	18	15	17	14	15	14	23
Rudolf Bayer	22	25	29	33	38	35	37	36	41	35	36
Serge Abiteboul	148	64	48	41	37	26	25	22	21	16	46
Χαμηλότερη Θέση	148	105	84	69	61	48	42	36	41	35	65
Άθροισμα Θέσεων	610	414	341	296	277	234	229	205	208	131	310

Πίνακας 3.22. Θέσεις των βραβευμένων συγγραφέων με τη μέθοδο BEPS.

Όνομα	Θέση κατάταξης με:										
	best1	best3	best5	best8	best10	best15	best20	best25	best30	best40	BCAvg
C. Mohan	104	103	84	69	61	48	41	33	32	26	65
David J. DeWitt	30	18	14	12	11	7	5	5	5	3	10
David Maier	41	26	20	17	16	12	12	11	12	8	16
Donald D. Chamberlin	9	5	4	4	4	5	7	7	8		5
Hector Garcia-Molina	81	41	36	28	26	22	22	19	17	12	30
Jim Gray	3	3	2	2	2	2	1	1	1	1	2
Michael Stonebraker	21	9	9	7	5	4	4	4	3	2	7
Patricia G. Selinger	7	21	23	25	27	32	33	34	35		27
Philip A. Bernstein	67	28	17	13	12	9	8	8	7	5	15
Rakesh Agrawal	47	33	25	20	20	15	14	12	11	7	21
Ronald Fagin	49	29	22	18	17	16	17	14	16	14	23
Rudolf Bayer	23	24	29	33	40	35	40	40	41	39	39
Serge Abiteboul	146	64	46	41	35	26	25	21	21	16	46
Χαμηλότερη Θέση	146	103	84	69	61	48	41	40	41	39	65
Άθροισμα Θέσεων	628	404	331	289	276	233	229	209	209	133	306

Πίνακας 3.23. Θέσεις των βραβευμένων συγγραφέων με τη μέθοδο SCEAS_B1.

λοί συγγραφείς, των οποίων η θέση κατάταξης γίνεται υψηλότερη αυξάνοντας τον αριθμό των κορυφαίων δημοσιεύσεων (με τον S. Abiteboul να πλεονεκτεί περισσότερο), ενώ το ανάποδο ισχύει για λίγους άλλους συγγραφείς (π.χ. η P. Selinger λόγω της συγκεκριμένης συνεισφοράς της στο System R και ο R. Bayer λόγω της συνεισφοράς του στα B-δένδρα και τις σχεσιακές δομές). Ο Jim Gray σταθερά κρατά κορυφαίες θέσεις.

Στην Παράγραφο 3.5.1 χρησιμοποιήσαμε μία μέθοδο σύνοψης των λιστών κατάταξης για κάθε μέθοδο. Μία άλλη κατεύθυνση είναι να ελέγξουμε εάν μία μέθοδος data fusion, όπως η Καταμέτρηση Borda, θα μπορούσε να συνοψίσει τις λίστες κατάταξης που παρουσιάζονται στους Πίνακες 3.21-3.23. Δεδομένων πινάκων κατάταξης των N στοιχείων ο καθένας, η Καταμέτρηση Borda δίνει N πόντους σε κάθε πρώτο στοιχείο των λιστών κατάταξης, N-1 πόντους σε κάθε δεύτερο στοιχείο κ.ο.κ. Ουσιαστικά, στην Παράγραφο 3.5.1 λάβαμε υπ'όψη μας μόνο τα πρώτα 5 στοιχεία κάθε πίνακα κατάταξης. Εδώ αποφασίσαμε να πάρουμε από κάθε λίστα κατάταξης τους κορυφαίους 10.000 (το οποίο είναι αρκετό). Έτσι, σε κάθε πρώτο συγγραφέα δίνονται 10.000 πόντοι, σε κάθε δεύτερο 9.999 κ.ο.κ. Η απουσία κάποιων συγγραφέων από τον πίνακα κατάταξης (π.χ στήλη best40) δεν οφείλεται στο γεγονός ότι οι συγγραφείς κατατάχθηκαν πολύ χαμηλά, αλλά στο γεγονός ότι δεν έχουν αρκετές δημοσιεύσεις ώστε να υπολογισθεί η βαθμολογία τους. Επομένως, διαιρούμε το άθροισμα της Καταμέτρησης Borda για κάθε συγγραφέα με τον αριθμό των εμφανίσεών του στη λίστα κατάταξης (π.χ. ένας συγγραφέας που κατατάχθηκε πρώτος σε όλους τους πίνακες κατάταξης θα

λάβει τη βαθμολογία 10.000 παρά 100.000). Στους Πίνακες 3.21-3.23, η τελευταία στήλη που ονομάζεται BCAgv δείχνει τη θέση των συγγραφέων στη λίστα κατάταξης της Καταμέτρησης Borda.

Μία εναλλακτική μέθοδος data fusion είναι η μέθοδος Condorcet. Συγκεκριμένα, υλοποιήσαμε την εκδοχή `Black` της μεθόδου Condorcet και καταλήξαμε σε παρόμοια αποτελέσματα με την Καταμέτρηση Borda.

3.6 Συμπεράσματα και Μελλοντική Εργασία

Σε αυτό το κεφάλαιο προτείναμε και εξετάσαμε πειραματικά μία οικογένεια νέων εναλλακτικών μεθόδων για την εκτίμηση επιστημονικών δημοσιεύσεων, πέρα από τους γνωστούς αλγορίθμους `PageRank` και `HITS`. Λεπτομερής περιγραφή των αλγορίθμων και της βελτίωσης στην απόδοση δίνεται στο [85]. Επίσης, αξιολογούμε τις προηγούμενες μεθόδους χρησιμοποιώντας το σύνολο δεδομένων της ψηφιακής βιβλιοθήκης DBLP ως σύνολο εκπαίδευσης (training set) και τις βραβευμένες δημοσιεύσεις των 'VLDB 10 Year Award' και 'SIGMOD Test of Time Award' ως σύνολο εκτίμησης (evaluation set) για τη μέθοδο αξιολόγησης δημοσιεύσεων. Επιπλέον, παρουσιάζουμε κατάταξη συγγραφέων βασισμένη στα αποτελέσματα της κατάταξης δημοσιεύσεων και χρησιμοποιώντας το 'SIGMOD Edgar F. Codd Innovations Award' ως σύνολο αξιολόγησης (evaluation set). Προφανώς, δεν είναι σκοπός μας να προτείνουμε ποιος από αυτούς θα έπρεπε να βραβευθεί τα επόμενα χρόνια. Ωστόσο, πιστεύουμε ότι αυτή η αντικειμενική μας προσέγγιση είναι κατάλληλη προς βοήθεια των αντίστοιχων επιτροπών. Σε όλες τις προηγούμενες περιπτώσεις, η απόδοση της μεθόδου κατάταξής μας `SCEAS` ήταν γενικά καλύτερη από άλλες μεθόδους.

Ώς μελλοντική εργασία μπορούμε να αναφέρουμε την εφαρμογή των παραπάνω μετρικών και στο χώρο της Ιστομετρίας. Η εφαρμογή αυτή παρουσιάζεται στο επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 4

Κατάταξη στον Παγκόσμιο Ιστό

Περιεχόμενα

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα μεταφέρουμε στο χώρο της Ιστομετρίας τους αλγορίθμους που παρουσιάσθηκαν για το χώρο της Βιβλιομετρίας. Συνεπώς στο παρόν κεφάλαιο δεν παρουσιάζουμε κάποια νέα μέθοδο, αλλά εφαρμόζουμε αυτές που ορίσαμε στο Κεφάλαιο 3. Στην συνέχεια θα φανεί ότι αυτή η εφαρμογή μπορεί να γίνει με πολύ καλά αποτελέσματα. [1]

[1] Για λόγους σαφήνειας στην παρουσίαση, προτιμήθηκε το υλικό του κεφαλαίου αυτού να παρουσιαστεί αυτόνομα και όχι από κοινού με το υλικό του προηγούμενου κεφαλαίου.

4.2 Προετοιμασία Πειραμάτων

4.2.1 Σύνολο Δεδομένων

Προκειμένου να εφαρμόσουμε τους αλγορίθμους του προηγούμενου κεφαλαίου και σε δεδομένα του Παγκόσμιου Ιστού, πρώτο μας μέλημα ήταν η συλλογή των δεδομένων αυτών. Το σενάριο που ακολουθήθηκε είναι το εξής:

1. Επιλογή κάποιων λέξεων-κλειδιά με βάση τις οποίες θα συλλεχθεί το σύνολο δεδομένων. Οι λέξεις-κλειδιά που επελέγησαν φαίνονται στον Πίνακα 4.1.

2. Για κάθε λέξη-κλειδί:

   (a) Αποστολή ερωτήματος στην μηχανή αναζήτησης Google[2] για τη λέξη-κλειδί και εντοπισμός των 200 πρώτων σελίδων που επιστρέφονται. Εφ'εξής θεωρούμε ότι (α) η σειρά κατάταξης που επιστρέφεται από τη Google είναι η βέλτιστη και (β) ότι οι σελίδες αυτές περιέχουν τις λέξεις-κλειδιά. Το σύνολο αυτό των σελίδων έστω ότι αποτελεί το σύνολο C_1.

   (b) Για κάθε στοιχείο του συνόλου C_1 αποστολή στη Google του ερωτήματος σχετικά με το ποιές σελίδες περιέχουν υπερσύνδεσμο προς αυτό. Έτσι δημιουργούμε το σύνολο C_2. Με τον ίδιο τρόπο προχωρούμε μέχρι το επίπεδο 4 και δημιουργούμε το αντίστοιχο σύνολο. Ο Πίνακας 4.1 περιέχει το πλήθος των στοιχείων του κάθε συνόλου C_i.

   (c) Για κάθε στοιχείο του συνόλου $C_{tot} = \{C_1 \cup C_2 \cup ...\}$ με τη χρήση του προγράμματος webbot[3] *φέρνουμε* τη σελίδα, καθώς και όλες όσες αναφέρονται από αυτήν μέσα σε μια απόσταση 4 υπερσυνδέσμων. Στον Πίνακα 4.1 στις στήλες F_i φαίνεται πόσα από τα αρχικά URLS καταφέραμε να φέρουμε. Η επόμενη στήλη αντιστοιχεί στο ποσοστό C_i/F_i.

[2] `http://www.google.com`

[3] Το webbot είναι ένα πρόγραμμα που παρέχεται από το W3C, όπου δίνοντάς του ένα αρχικό URL, αποθηκεύει μετα-δεδομένα για αυτό καθώς και για όλες τις σελίδες προς τις οποίες υπάρχει υπερσύνδεσμος από την αρχική μέχρι το επίπεδο που θα ορίσει ο χρήστης. Επίσης ο χρήστης μπορεί να ορίσει και διάφορους περιορισμούς όπως να μην ακολουθώνται υπερσύνδεσμοι που δείχνουν σε συγκεκριμένους τύπους αρχείων (πχ. εικόνες, βίντεο κ.τ.λ.).

Query	C_1	F_1		C_2	F_2		C_3	F_3		C_{tot}	F_{tot}	
"antonis sidiropoulos"	92	92	100%	302	302	100%	1059	1059	100%	1453	1453	100%
"computational complexity"	199	106	53%	2577	2548	99%	3890	3836	99%	6666	6490	97%
"computational geometry"	197	197	100%	2875	2844	99%	5529	5479	99%	8601	8520	99%
"star wars"	201	173	86%	19451	6382	33%	53414	14689	28%	73066	21244	29%
basketball	198	177	89%	22830	6189	27%	47052	15462	33%	70080	21828	31%
complexity	197	197	100%	6808	6719	99%	11995	11650	97%	19000	18566	98%
jaguar	201	160	80%	127537	43631	34%	2606182	1927193	74%	2733920	1970984	72%
jordan	198	197	99%	11664	11196	96%	33456	29474	88%	45318	40867	90%
movies	199	2	1%	30161	4223	14%	257063	174872	68%	287423	179097	62%
sidiropoulos	201	198	99%	441	439	100%	900	900	100%	1542	1537	100%
snowstorm	199	86	43%	2020	2012	100%	1984	1944	98%	4203	4042	96%
tsunami indian ocean	199	58	29%	3686	3669	100%	7356	7329	100%	11241	11056	98%
twister	199	51	26%	3919	3818	97%	13277	12611	95%	17395	16480	95%
twister weather	198	91	46%	1936	1921	99%	4201	4155	99%	6335	6167	97%
weather	201	2	1%	24723	3793	15%	198618	133426	67%	223542	137221	61%

Πίνακας 4.1. Στατιστικά στοιχεία για τα δεδομένα συλλογής.

Η διαδικασία αυτή ξεκίνησε τον Ιανουάριο του 2005 και σταμάτησε τον Απρίλιο του 2006 και μας οδήγησε σε ένα σύνολο δεδομένων, το οποίο αποτελείται από σχεδόν 80 εκατομμύρια σελίδες (79.029.279) και περίπου 410 εκατομμύρια υπερσυνδέσμους (410.257.262) μεταξύ των σελίδων αυτών. Δεδομένου του τεράστιου μεγέθους του συνόλου αυτού (αν και όχι πλήρες), δεν μπορούμε με τους διαθέσιμους πόρους να επιτύχουμε αξιολόγηση-κατάταξη των σελίδων για τα δεδομένα ολόκληρης της βάσης μας. Για το λόγο αυτό, για κάθε λέξη-κλειδί απομονώνουμε ένα υποσύνολο όπου θα εφαρμόσουμε τις μεθόδους μας. Για κάθε λέξη κλειδί δημιουργούμε το σύνολο $K_{tot} = \{K_1 \cup K_2 \cup ...\}$. Το σύνολο K_1 είναι ίσο με το σύνολο C_1. Το K_2 αποτελείται από όλες τις σελίδες της βάσης, οι οποίες δεν ανήκουν στο K_1 και είτε δείχνουν σε στοιχεία του K_1 είτε δείχνονται από κάποιο στοιχείο του K_1. Αντίστοιχα, προκύπτουν και τα υπόλοιπα σύνολα K_i. Στα πειράματά μας θα χρησιμοποιήσουμε μόνο το $K_{1,2,3} = K_1 \cup K_2 \cup K_3$. Σταματούμε στην απόσταση 3, διότι σε απόσταση 4 το σύνολο μας πάλι αποτελείται από τόσο μεγάλο πλήθος σελίδων, το οποίο δεν μπορούμε να διαχειρισθούμε με τους διαθέσιμους πόρους.

Query	K_1	K_2	K_3	K_4	K_{tot}	$K_{1,2,3}$
"antonis sidiropoulos"	46	7007	536600	10098323	10641976	543653
"computational complexity"	182	8187	1306281	23653600	24968250	1314650
"computational geometry"	176	5776	792244	14200068	14998264	798196
"star wars"	194	46882	3345110	40631472	44023658	3392186
basketball	197	136140	3358161	34554537	38049035	3494498
complexity	197	31859	2601198	37924977	40558231	2633254
jaguar	181	24341	2260527	31555624	33840673	2285049
jordan	197	46159	3507458	38889260	42443074	3553814
movies	182	85339	5061493	32857759	38004773	5147014
sidiropoulos	198	6372	1158201	23114686	24279457	1164771
snowstorm	162	7991	1712913	28284471	30005537	1721066
tsunami indian ocean	175	14879	2064281	30004441	32083776	2079335
twister	164	7682	1846134	28326803	30180783	1853980
twister weather	172	6763	1516035	25368665	26891635	1522970
weather	195	39859	2751588	30710074	33501716	2791642

Πίνακας 4.2. Στατιστικά στοιχεία για τα σύνολα εισόδου K_i.

4.2.2 Μεθοδολογία Πειραματισμού

Για κάθε μια λέξη-κλειδί έχουμε δημιουργήσει τον αντίστοιχο γράφο $K_{1,2,3}$. Σε κάθε γράφο εφαρμόζουμε τους αλγορίθμους `PageRank`, `Prestige`, `HITS` και `SALSA` που έχουν παρουσιασθεί στο προηγούμενο κεφάλαιο. Δεν χρησιμοποιούμε

τους CC και BCC διότι αυτοί δεν χρησιμοποιούνται στην περίπτωση της Ιστομετρίας. Από την οικογένεια SCEAS δοκιμάσαμε τις εξής παραλλαγές της βασικής Εξίσωσης 3.13:

- SCEAS: $d = 0.85, b = 1, a = \epsilon$
- SCEAS_B0: $d = 0.85, b = 0, a = \epsilon$
- SCEAS_D99: $d = 0.99, b = 1, a = \epsilon$

Τέλος, προφανώς δεν περιλαμβάνεται σε αυτό το κεφάλαιο εφαρμογή των B-HITS και B-SALSA, εφ'όσον ορίσαμε αυτές τις παραλλαγές ειδικά για τον τομέα της Βιβλιομετρίας.

Συνεπώς σε κάθε γράφο εφαρμόζουμε επτά αλγορίθμους ταξινόμησης. Επίσης, έχουμε και την ταξινόμηση από τη Google, η οποία είναι η σειρά με την οποία μας έδωσε τα αποτελέσματα στις αναζητήσεις μας το google.com. Δεδομένου ότι η Google, εκτός από εφαρμογή του αλγορίθμου PageRank, λαμβάνει και άλλες μετρικές υπ'όψη του (και ελλείψει κάτι καλύτερου), θεωρούμε ότι η βέλτιστη κατάταξη είναι αυτή που δίνεται από τη Google. Άρα όλους τους αλγορίθμους μας θα τους συγκρίνουμε με την Google για να επιβεβαιώσουμε την ποιότητά τους. Θα αναφερόμαστε στη βέλτιστη κατάταξη με όνομα αλγορίθμου GOOGLE.

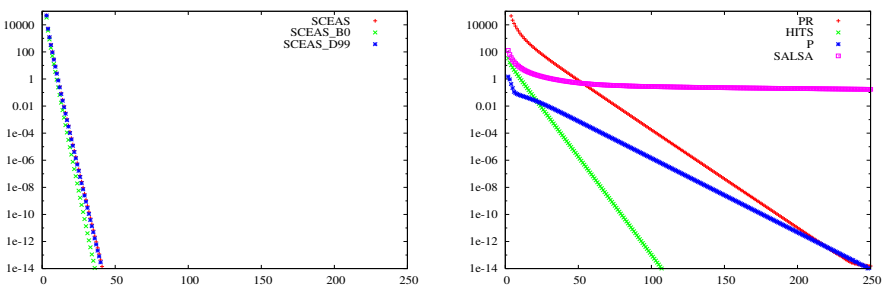
Επίσης, εδώ θα πρέπει να αναφέρουμε ότι η κατάταξη που δίνεται από τη Google, ίσχυε το καλοκαίρι του 2005. Οι γράφοι που έχουμε στη διάθεσή μας συλλέχθηκαν από τον Ιανουάριο του 2005 έως τον Απρίλιο του 2006. Άρα προφανώς κατά την αναζήτηση στη Google, η βάση της Google περιέχει διαφορετικό ιστογράφο από αυτόν που έχουμε συλλέξει. Κρατούμε όμως την κατάταξη που μας δόθηκε από τη Google το καλοκαίρι του 2005 δεδομένου ότι αυτή η χρονική στιγμή βρίσκεται περίπου στη μέση της περιόδου συλλογής του ιστογράφου. Άλλωστε όλοι οι υπόλοιποι αλγόριθμοι που θα αξιολογήσουμε εφαρμόσθηκαν επάνω ακριβώς στον ίδιον ιστογράφο.

Τέλος, επισημαίνεται το εξής: Η βαθμολογία από κάθε μέθοδο υπολογίζεται για κάθε λέξη κλειδί επάνω στο σύνολο $K_{1,2,3}$. Έτσι για κάθε σελίδα προκύπτει μια βαθμολογία (για κάθε αλγόριθμο). Θεωρούμε ότι από το σύνολο $K_{1,2,3}$, μόνο οι σελίδες που ανήκουν στο K_1 πρέπει να είναι στο σύνολο αποτελεσμάτων. Άρα, η τελική μας κατάταξη είναι τα μέλη του γράφου K_1, ταξινομημένα με βάση τη βαθμολογία που προέκυψε από τον ιστογράφο $K_{1,2,3}$.

4.3 Πειραματικά Αποτελέσματα

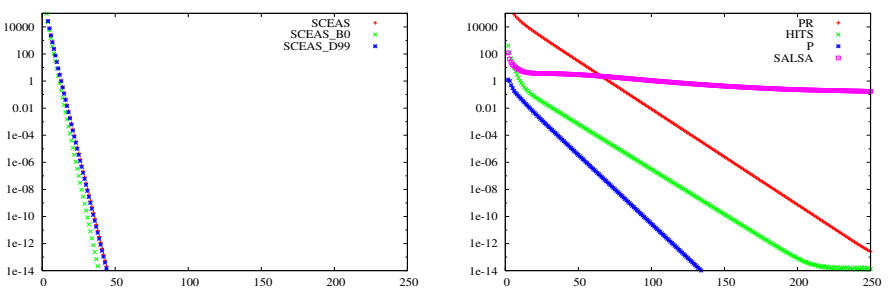
4.3.1 Ταχύτητα Υπολογισμού

Στα Σχήματα 4.1, 4.2 και 4.3 βλέπουμε την ταχύτητα σύγκλισης των αλγορίθμων μας. Όπως είναι φανερό, όλες οι παραλλαγές του `SCEASRank` είναι ταχύτερες από όλους τους υπόλοιπους αλγορίθμους. Μπορούμε να διαπιστώσουμε ότι ο `SCEASRank` τρέχει περίπου στο 20% του χρόνου του `PageRank`. Ο `SALSA` είναι ο βραδύτερος απ'όλους. Οι ανωτέρω αλγόριθμοι παρουσιάζουν σχετικά σταθερή ταχύτητα σύγκλισης και στα τρία παραδείγματα που περιλαμβάνουμε. Από την άλλη ο `HITS` και ο `Prestige` (P στο διάγραμμα), φαίνεται να βελτιώνονται όσο μεγαλύτερος είναι ο γράφος. Αυτή η βελτίωση είναι εικονική, και οφείλεται στο γεγονός ότι γίνεται κανονικοποίηση του διανύσματος $||.||_1$ στη μονάδα. Άρα, όσο μεγαλύτερος ο γράφος (το διάνυσμα) τόσο μικρότερα νούμερα, και άρα υπάρχει απώλεια ακρίβειας. Έτσι ο Prestige, ενώ στο Σχήμα 4.1 συγκλίνει μετά από 250 επαναλήψεις, στα Σχήματα 4.2 και 4.3 η σύγκλιση επιτυγχάνεται μετά από τις 130-150 επαναλήψεις.



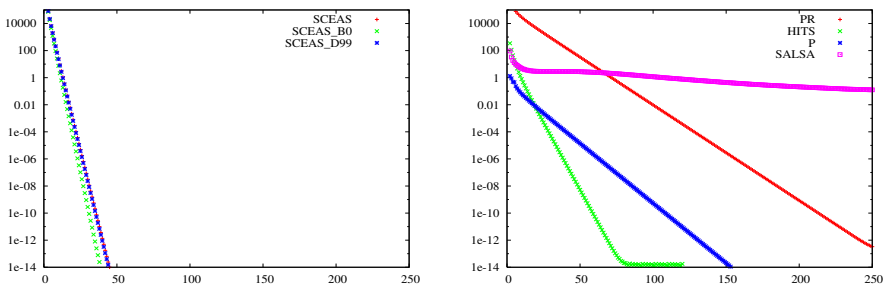
(α) (β)

Σχήμα 4.1. Ταχύτητα σύγκλισης για το "antonis sidiropoulos".



(α) (β)

Σχήμα 4.2. Ταχύτητα σύγκλισης για το "jordan".



(α) (β)

Σχήμα 4.3. Ταχύτητα σύγκλισης για το "movies".

4.3.2 Συγκρίσεις Αξιολογήσεων

Για οικονομία χώρου σε αυτό το κεφάλαιο, στους επόμενους πίνακες παρουσιάζουμε συγκρίσεις των αλγορίθμων μόνο με βάση το Kendal tau ($d^{(p)}(a_1, a_2)$), απλή απόσταση ($D(a_1, a_2)$) και απόσταση με βάρος ($D_w(a_1, a_2)$). Στον Πίνακα 4.3 συκρίνουμε τα αποτελέσματα κατάταξης του συνόλου K_1 για το ερώτημα "antonis sidiropoulos". Στον Πίνακα 4.3(α) βλέπουμε τις αποστάσεις με βάση το Kendal tau με πέναλτυ 0. Η βέλτιστη κατάταξη θεωρούμε ότι δίνεται από τη μηχανή αναζήτησης Google, που αν και χρησιμοποιεί τον αλγόριθμο PageRank, λαμβάνει υπ'όψη του και άλλα στοιχεία. Βλέπουμε λοιπόν ότι σε αυτήν την περίπτωση όλοι οι αλγόριθμοι ισαπέχουν από το βέλτιστο.

Στον Πίνακα 4.3(β) όπου έχουμε αποστάσεις με βάση το Kendal tau με πέναλτυ 1, βλέπουμε ότι οι παραλλαγές SCEAS καθώς και ο PageRank είναι περισσότερο κοντά στο βέλτιστο σε σχέση με τους υπόλοιπους. Στα ίδια συμπεράσματα καταλήγουμε και από τους Πίνακες 4.3(γ) και 4.3(δ).

Στον Πίνακα 4.4 παρουσιάζονται οι αποστάσεις που προέκυψαν από το γράφο που αντιστοιχεί στο ερώτημα "basketball". Παρατηρούμε ότι τα αποτελέσματα είναι αντίστοιχα.

	Google	PR	HA	P	SA	SCEAS	SCEAS B0	SCEAS D99
Google	-	18.07%	23.29%	18.07%	19.81%	18.65%	18.74%	18.65%
PR	18.07%	-	16.04%	11.88%	4.83%	1.55%	1.64%	1.55%
HA	23.29%	16.04%	-	6.28%	13.62%	16.04%	15.94%	16.04%
P	18.07%	11.88%	6.28%	-	11.11%	13.04%	12.95%	13.04%
SA	19.81%	4.83%	13.62%	11.11%	-	5.99%	5.89%	5.99%
SCEAS	18.65%	1.55%	16.04%	13.04%	5.99%	-	0.10%	0.00%
SCEAS B0	18.74%	1.64%	15.94%	12.95%	5.89%	0.10%	-	0.10%
SCEAS D99	18.65%	1.55%	16.04%	13.04%	5.99%	0.00%	0.10%	-

(α) $d^{(0)}(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS B0	SCEAS D99
Google	-	41.06%	52.66%	58.07%	42.80%	41.64%	41.74%	41.64%
PR	41.06%	-	22.42%	28.89%	4.83%	1.55%	1.64%	1.55%
HA	52.66%	22.42%	-	25.60%	20.00%	22.42%	22.32%	22.42%
P	58.07%	28.89%	25.60%	-	28.12%	30.05%	29.95%	30.05%
SA	42.80%	4.83%	20.00%	28.12%	-	5.99%	5.89%	5.99%
SCEAS	41.64%	1.55%	22.42%	30.05%	5.99%	-	0.10%	0.00%
SCEAS B0	41.74%	1.64%	22.32%	29.95%	5.89%	0.10%	-	0.10%
SCEAS D99	41.64%	1.55%	22.42%	30.05%	5.99%	0.00%	0.10%	-

(β) $d^{(1)}(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS B0	SCEAS D99
Google	-	14.27%	17.30%	17.39%	16.26%	15.03%	15.03%	15.03%
PR	14.27%	-	12.57%	10.78%	3.69%	1.42%	1.51%	1.42%
HA	17.30%	12.57%	-	6.33%	10.59%	13.04%	13.04%	13.04%
P	17.39%	10.78%	6.33%	-	9.17%	11.06%	11.06%	11.06%
SA	16.26%	3.69%	10.59%	9.17%	-	4.06%	3.97%	4.06%
SCEAS	15.03%	1.42%	13.04%	11.06%	4.06%	-	0.09%	0.00%
SCEAS B0	15.03%	1.51%	13.04%	11.06%	3.97%	0.09%	-	0.09%
SCEAS D99	15.03%	1.42%	13.04%	11.06%	4.06%	0.00%	0.09%	-

(γ) $D(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS B0	SCEAS D99
Google	-	17.94%	18.38%	20.34%	21.71%	18.87%	19.41%	18.87%
PR	17.94%	-	18.34%	15.14%	3.49%	1.21%	1.54%	1.21%
HA	18.38%	18.34%	-	8.62%	16.16%	19.15%	18.94%	19.15%
P	20.34%	15.14%	8.62%	-	12.92%	17.08%	16.76%	17.08%
SA	21.71%	3.49%	16.16%	12.92%	-	4.02%	3.70%	4.02%
SCEAS	18.87%	1.21%	19.15%	17.08%	4.02%	-	0.32%	0.00%
SCEAS B0	19.41%	1.54%	18.94%	16.76%	3.70%	0.32%	-	0.32%
SCEAS D99	18.87%	1.21%	19.15%	17.08%	4.02%	0.00%	0.32%	-

(δ) $D_w(a_1, a_2)$

Πίνακας 4.3. Αποστάσεις κατάταξης των αλγορίθμων στον ιστογράφο "antonis sidiropoulos"

	Google	PR	HA	P	SA	SCEAS	SCEAS_B0	SCEAS_D99
Google	-	39.86%	37.38%	40.43%	37.24%	40.15%	40.21%	40.15%
PR	39.86%	-	25.74%	29.29%	13.53%	2.86%	3.18%	2.89%
HA	37.38%	25.74%	-	21.52%	20.09%	25.97%	25.95%	25.96%
P	40.43%	29.29%	21.52%	-	25.95%	29.62%	29.70%	29.63%
SA	37.24%	13.53%	20.09%	25.95%	-	13.79%	13.82%	13.80%
SCEAS	40.15%	2.86%	25.97%	29.62%	13.79%	-	0.32%	0.03%
SCEAS_B0	40.21%	3.18%	25.95%	29.70%	13.82%	0.32%	-	0.28%
SCEAS_D99	40.15%	2.89%	25.96%	29.63%	13.80%	0.03%	0.28%	-

(α) $d^{(0)}(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS_B0	SCEAS_D99
Google	-	39.88%	37.41%	40.53%	37.27%	40.17%	40.23%	40.17%
PR	39.88%	-	25.74%	29.36%	13.53%	2.86%	3.18%	2.89%
HA	37.41%	25.74%	-	21.59%	20.09%	25.97%	25.95%	25.96%
P	40.53%	29.36%	21.59%	-	26.02%	29.69%	29.77%	29.70%
SA	37.27%	13.53%	20.09%	26.02%	-	13.79%	13.82%	13.80%
SCEAS	40.17%	2.86%	25.97%	29.69%	13.79%	-	0.32%	0.03%
SCEAS_B0	40.23%	3.18%	25.95%	29.77%	13.82%	0.32%	-	0.28%
SCEAS_D99	40.17%	2.89%	25.96%	29.70%	13.80%	0.03%	0.28%	-

(β) $d^{(1)}(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS_B0	SCEAS_D99
Google	-	26.74%	25.80%	27.62%	25.20%	27.02%	27.05%	27.02%
PR	26.74%	-	17.76%	19.68%	9.50%	2.10%	2.34%	2.12%
HA	25.80%	17.76%	-	14.87%	13.93%	17.91%	17.90%	17.90%
P	27.62%	19.68%	14.87%	-	17.36%	19.78%	19.80%	19.79%
SA	25.20%	9.50%	13.93%	17.36%	-	9.55%	9.56%	9.55%
SCEAS	27.02%	2.10%	17.91%	19.78%	9.55%	-	0.28%	0.03%
SCEAS_B0	27.05%	2.34%	17.90%	19.80%	9.56%	0.28%	-	0.25%
SCEAS_D99	27.02%	2.12%	17.90%	19.79%	9.55%	0.03%	0.25%	-

(γ) $D(a_1, a_2)$

	Google	PR	HA	P	SA	SCEAS	SCEAS_B0	SCEAS_D99
Google	-	26.03%	31.42%	36.03%	24.86%	26.59%	26.63%	26.60%
PR	26.03%	-	16.30%	19.63%	5.48%	1.02%	1.15%	1.03%
HA	31.42%	16.30%	-	9.20%	11.28%	16.44%	16.44%	16.44%
P	36.03%	19.63%	9.20%	-	17.34%	20.01%	20.06%	20.01%
SA	24.86%	5.48%	11.28%	17.34%	-	5.44%	5.43%	5.45%
SCEAS	26.59%	1.02%	16.44%	20.01%	5.44%	-	0.13%	0.01%
SCEAS_B0	26.63%	1.15%	16.44%	20.06%	5.43%	0.13%	-	0.12%
SCEAS_D99	26.60%	1.03%	16.44%	20.01%	5.45%	0.01%	0.12%	-

(δ) $D_w(a_1, a_2)$

Πίνακας 4.4. Αποστάσεις κατάταξης των αλγορίθμων στον ιστογράφο "basketball"

4.3.3 Σχολιασμός Αποτελεσμάτων

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://skyblue.csd.auth.gr/members/asidirop.html	1	3	11	7	11	12	12	12
http://delab.csd.auth.gr/~dimitris/dkatsaro.htm	2	22	9	24	23	9	9	9
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html	3	1	1	2	1	1	1	1
http://aetos.it.teithe.gr/~asidirop/	4	4	17	8	15	16	16	16
http://users.auth.gr/~asidirop/	5	23	3	16	6	4	3	3
http://aetos.it.teithe.gr/~asidirop/submission/	6	24	12	17	10	10	10	10
http://deslab.mit.edu/DesignLab/new_deslab/new-person.html	7	15	4	19	4	5	5	5
http://www.robotstxt.org/wc/active/html/dienstspider.html	8	2	10	1	8	14	14	14
http://ii.pmf.ukim.edu.mk/boi2000/teams.html	9	5	18	9	16	17	17	17
http://delab.csd.auth.gr/~asidirop/	10	26	26	25	26	26	26	26
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html	11	17	16	3	12	24	24	24
http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos	12	20	15	6	25	13	13	13
http://www.hostsun.com/gr/bots_index3.php	13	12	14	18	13	15	15	15
http://www.cs.kuleuven.ac.be/~dirk/ada-belgium/events/03/030612-wdas.html	14	27	27	26	27	27	27	27
http://www.iei.pi.cnr.it/DELOS/TOM/list.html	15	21	13	27	24	11	11	11

Πίνακας 4.5. Συγκεντρωτικά αποτελέσματα κατάταξης του "antonis sidiropoulos".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.euroleague.net/	1	70	45	144	43	47	47	47
http://www.fiba.com/	2	5	5	9	3	6	6	6
http://www.nba.com/	3	2	2	4	2	2	2	2
http://www.usabasketball.com/	4	14	7	68	4	7	7	7
http://www.wnba.com/	5	12	4	2	5	4	4	4
http://www.infoplease.com/ipsa/A0003691.html	6	152	133	55	117	131	131	131
http://dmoz.org/Sports/Basketball/	7	40	20	3	53	32	30	30
http://sports.espn.go.com/ncb/index	8	34	19	22	10	18	18	18
http://sports.espn.go.com/nba/index	9	13	15	25	9	14	15	15
http://www.forbes.com/2005/04/01/ cx_da_0401bookreview_miracle.html	10	115	122	42	61	117	117	117
http://probasketball.about.com/	11	30	59	31	28	57	57	57
http://www.nba.com/lakers/	12	6	12	15	13	15	14	14
http://www.basketball.com/	13	54	22	63	27	20	20	20
http://www.usabasketball.com/women/	14	44	163	186	153	162	162	162
http://www.bbhighway.com/	15	41	61	67	54	62	62	62

Πίνακας 4.6. Συγκεντρωτικά αποτελέσματα κατάταξης του "basketball".

Στους Πίνακες 4.5-4.19 παρουσιάζουμε τα αποτελέσματα της κατάταξης με τους διάφορους αλγορίθμους, καθώς επίσης και την κατάταξη που δόθηκε από την μηχανή αναζήτησης Google. Σε όλους τους αλγορίθμους και για όλους τους ιστογράφους μας παρατηρούμε ότι βρίσκουν λιγότερες από τις μισές σελίδες που θα περιμέναμε στην πρώτη δεκαπεντάδα με βάση τη Google. Αυτό πιθανότατα οφείλεται σε δυο κυρίως λόγους. Ο πρώτος λόγος είναι η χρονική διάρκεια στη συλλογή σελίδων - έχουμε στη διάθεσή μας διαφορετικό γράφο από αυτόν που διατηρεί η Google. Ο δεύτερος λόγος είναι η έλλειψη αρκετών δεδομένων. Αν και προσπαθήσαμε να συλλέξουμε όσο το δυνατόν περισσότερες σελίδες, προφανώς έχουμε στην διάθεσή μας πολύ μικρότερο γράφο από αυτόν που έχει κατασκευάσει

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.ipam.ucla.edu/programs/ptac2002/	1	164	137	177	158	131	133	133
http://journals.wiley.com/1076-2787/	2	47	149	82	92	153	152	152
http://www.comdig.org/	3	13	1	13	1	1	1	1
http://journal-ci.csse.monash.edu.au/	4	7	26	7	30	26	26	26
http://www.calresco.org/	5	14	16	54	12	12	12	12
http://www.calresco.org/themes.htm	6	29	22	87	15	20	20	20
http://eccc.uni-trier.de/eccc/	7	85	7	4	24	6	6	6
http://eccc.uni-trier.de/eccc/info/people.html	8	109	86	18	122	84	84	84
http://life.csu.edu.au/discontinued.html	9	107	174	132	159	172	172	172
http://pespmc1.vub.ac.be/COMPLEXI.html	10	110	65	33	62	88	86	86
http://pespmc1.vub.ac.be/COMSELLI.html	11	25	41	34	28	43	43	43
http://www.geocities.com/ResearchTriangle/1402/	12	97	66	81	63	55	57	57
http://www.geocities.com/ResearchTriangle/1402/ ComplexityAndEvolution-Links.htm	13	53	158	120	133	156	156	156
http://www.complexityzoo.com/	14	23	11	3	7	7	7	7
http://www.complexityclan.com/	15	91	53	117	21	54	56	56

Πίνακας 4.7. Συγκεντρωτικά αποτελέσματα κατάταξης του "complexity".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://eccc.uni-trier.de/eccc/	1	28	2	29	8	2	2	2
http://eccc.uni-trier.de/eccc/info/people.html	2	40	37	49	48	38	38	38
http://weblog.fortnow.com/	3	6	1	3	2	1	1	1
http://www.springerlink.com/openurl.asp? genre=journal&issn=1016-3328	4	1	54	2	3	43	45	45
http://www.springerlink.com/link.asp?id=101499	5	3	33	65	45	26	26	26
http://www.cs.rochester.edu/u/www/courses/486/	6	71	39	43	74	33	34	34
http://www.uncg.edu/mat/avg/avgcomp/avgcomp.html	7	137	21	63	24	18	19	19
http://facweb.cs.depaul.edu/jrogers/complexity/	8	15	4	20	22	6	6	6
http://facweb.cs.depaul.edu/jrogers/complexity/cfp.htm	9	19	40	33	43	47	47	47
http://www.math.cas.cz/~ccc06/	10	33	24	8	32	25	25	25
http://www.almaden.ibm.com/software/theory/CCC05/	11	30	49	23	31	56	55	55
http://www.brics.dk/Complexity2003/	12	17	22	30	39	15	15	15
http://www.cs.utep.edu/longpre/complexity.html	13	16	66	40	53	89	88	88
http://link.springer-ny.com/link/service/journals/00037/	14	143	143	135	143	143	143	143
http://www.iscid.org/boards/ubb-get_topic-f-1-t-000247.html	15	84	56	11	104	93	91	91

Πίνακας 4.8. Συγκεντρωτικά αποτελέσματα κατάταξης του "computational complexity".

η Google. Επίσης, υπενθυμίζουμε ότι για τεχνικούς λόγους[4] χρησιμοποιήσαμε για τον υπολογισμό των βαθμολογιών το σύνολο-γράφο $K_{1,2,3}$ και όχι το K_{tot}[5].

Η δεύτερη παρατήρηση που ισχύει για όλους τους πίνακες, είναι ότι όλοι οι αλγόριθμοι κατατάσσουν τις επιλεγμένες σελίδες περίπου στις ίδιες θέσεις. Η οι-

[4] Όλα τα πειράματα έχουν πραγματοποιηθεί σε Intel Pentium 4 @ 3GHz με μέγεθος κύριας μνήμης (RAM) 4GB.

[5] Το πρόγραμμα που αναπτύξαμε για τους υπολογισμούς των βαθμολογιών με βάση τους διάφορους αλγορίθμους που παρουσιάζονται, αναπτύχθηκε έτσι ώστε να κάνει καλή χρήση της μνήμης και να είναι αποδοτικό στην ταχύτητα εκτέλεσης. Διατηρεί λοιπόν το γράφο καθώς και το διάνυσμα τον βαθμολογιών στην κύρια μνήμη του υπολογιστή. Για να φορτωθεί ο γράφος K_{tot} από το πρόγραμμά μας απαιτείται (σύμφωνα με τους υπολογισμούς μας) πάνω από 8GB στην κύρια μνήμη. Κάτι τέτοιο ήταν αδύνατο να το επιτευχθεί σε υπολογιστή με μέγεθος διευθυνσιοδότησης μνήμης 32bit (Intel Architecture 32). Θα απαιτούνταν εξειδικευμένες τεχνικές με χρήση μνήμης στο δίσκο - κάτι που είναι εκτός των σκοπών της παρούσας εργασίας.

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://cccg.cs.uwindsor.ca/	1	25	41	40	39	51	51	51
http://www.cs.virginia.edu/~robins/	2	31	70	62	60	72	71	71
http://www.cgal.org/	3	13	3	16	7	5	5	5
http://www.geometryfactory.com/	4	84	27	37	57	46	44	43
http://elib.cs.sfu.ca/Collections/CMPT/cs-journals/P-Elsevier/J-Elsevier-CG.html	5	133	68	44	89	101	96	96
http://elib.cs.sfu.ca/Collections/CMPT/cs-journals/P-Springer/J-Springer-DCG.html	6	132	63	43	80	93	91	91
http://www.cccg.ca/	7	11	31	49	37	32	32	32
http://www-cgrl.cs.mcgill.ca/~godfried/teaching.html	8	93	13	51	24	9	11	11
http://www-cgrl.cs.mcgill.ca/~godfried/teaching/cg-projects.html	9	95	67	20	72	68	67	67
http://www.socg05.org/	10	34	17	55	18	18	18	18
http://cgw2004.csail.mit.edu/	11	17	74	56	38	69	69	69
http://pages.cpsc.ucalgary.ca/~marina/ Newweb/session.htm	12	36	72	45	76	80	80	80
http://www.ried.tokai.ac.jp/JCDCG/	13	19	39	69	45	30	30	30
http://netlib.bell-labs.com/netlib/compgeom/readme.html	14	21	12	63	12	8	8	8
http://www-gdz.sub.uni-goettingen.de/cgi-bin/digbib.cgi?PPN362609810	15	3	61	2	28	56	56	56

Πίνακας 4.9. Συγκεντρωτικά αποτελέσματα κατάταξης του "computational geometry".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.jaguarvehicles.com/	1	29	73	6	70	88	86	86
http://www.jaguar.com/	2	3	9	15	5	6	6	6
http://www.jaguar-racing.com/	3	6	20	17	20	19	19	19
http://www.jaguarcars.com/nl/	4	53	145	111	122	141	141	141
http://www.jaguar.com/uk/	5	75	82	35	82	86	87	87
http://www.jaguarcars.com/	6	21	26	31	22	23	24	24
http://www.schrodinger.com/SiteMap.php? mID=3&sID=0&cID=0	7	105	13	75	13	16	16	16
http://www.jag-lovers.org/	8	5	10	4	8	9	9	9
http://www.jaguar.com.au/	9	57	43	63	71	37	37	37
http://www.apple.com/macosx/	10	2	1	11	1	1	1	1
http://www.apple.com/pr/library/2002/may/06jaguar.html	11	68	129	76	95	116	116	116
http://www.savethejaguar.com/	12	74	48	87	43	41	41	41
http://www.kidsplanet.org/factsheets/jaguar.html	13	112	41	92	93	43	40	40
http://www.jcna.com/	14	7	39	48	39	38	38	38
http://www.jagweb.com/	15	33	11	12	9	12	12	12

Πίνακας 4.10. Συγκεντρωτικά αποτελέσματα κατάταξης του "jaguar".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.sheilajordanjazz.com/	1	70	92	87	103	87	87	87
http://jordan.sportsline.com/	2	23	35	45	17	32	31	32
http://www.jordanfanclub.co.uk/	3	73	82	77	62	71	71	71
http://www.cia.gov/cia/publications/factbook/geos/jo.html	4	4	19	8	13	21	21	21
http://www.jumpman23.com/	5	96	32	64	64	51	51	51
http://www.f1jordan.com/	6	28	34	52	42	50	50	50
http://www.jordanrudess.com/	7	14	26	20	33	19	19	19
http://www.nic.gov.jo/	8	54	44	59	32	38	39	39
http://www.ju.edu.jo/	9	40	24	24	36	14	14	14
http://www.jordantimes.com/	10	7	10	6	3	6	7	7
http://lcweb2.loc.gov/frd/cs/jotoc.html	11	3	3	14	15	2	2	2
http://www.cs.berkeley.edu/~jordan/	12	32	29	120	74	27	27	27
http://www.see-jordan.com/	13	37	27	43	26	18	18	18
http://www.jordanfashions.com/	14	117	69	68	92	77	78	78
http://www.jrsoftware.org/	15	72	18	131	35	16	16	16

Πίνακας 4.11. Συγκεντρωτικά αποτελέσματα κατάταξης του "jordan".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://movies.yahoo.com/	1	6	15	11	8	10	10	10
http://www.reel.com/	2	64	82	77	47	81	81	81
http://www.imdb.com/	3	2	1	1	1	1	1	1
http://www.lordoftherings.net/	4	7	19	17	9	12	12	12
http://www.brainpop.com/	5	37	40	102	31	37	36	36
http://www.hollywood.com/	6	42	26	20	13	18	19	19
http://www.romanm.ch/	7	59	156	143	109	150	151	151
http://www.mgm.com/	8	13	16	33	28	16	16	16
http://www.rottentomatoes.com/	9	4	2	2	2	2	2	2
http://dmoz.org/Arts/Movies/	10	8	8	42	60	19	18	18
http://movies.aol.com/	11	43	11	18	21	11	11	11
http://www.onwisconsin.com/movies/	12	123	57	84	32	58	59	59
http://www.apple.com/trailers/	13	3	3	4	5	4	3	3
http://www.hd.net/	14	19	39	3	29	32	33	33
http://www.foxmovies.com/index1.html	15	76	60	96	48	61	60	60

Πίνακας 4.12. Συγκεντρωτικά αποτελέσματα κατάταξης του "movies".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://skyblue.csd.auth.gr/members/asidirop.html	1	16	46	15	59	53	53	53
http://www.fhw.gr/fhw/en/news/present/991004_baloukli.html	2	73	68	40	51	67	68	68
http://www.ilsp.gr/homepages/sidiropoulos.html	3	62	69	18	52	69	69	69
http://www.ilsp.gr/homepages/sidiropoulos_eng.html	4	14	66	4	49	63	63	63
http://cgi.di.uoa.gr/~bitsikas/Pavlos.html	5	3	8	32	15	6	6	6
http://www.studio52.gr/english/ SIDIROPOULOS_PAYLOS.htm	6	39	23	35	58	45	44	44
http://www.studio52.gr/english/ SIDIROPOULOS_GIORGOS.htm	7	63	77	39	73	93	93	93
http://www.greeka.com/greece-music.htm	8	4	11	19	5	12	12	12
http://www.indianjgastro.com/article.asp?issn=0254-8860;year=2004;volume=23;issue=4;spage=131;epage=134;aulast=Hatzitolios	9	61	49	29	37	47	47	47
http://aetos.it.teithe.gr/~asidirop/	10	17	94	16	81	90	90	90
http://delab.csd.auth.gr/~asidirop/	11	106	106	96	106	106	106	106
http://www-vlsi.stanford.edu/~sidirop/	12	107	107	97	107	107	107	107
http://www.telecom.tuc.gr/~nikos/	13	58	40	34	53	36	36	36
http://www.telecom.tuc.gr/~nikos/TUCwebCV.pdf	14	92	92	38	98	89	89	89
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Nikolaos.html	15	11	22	1	11	28	28	28

Πίνακας 4.13. Συγκεντρωτικά αποτελέσματα κατάταξης του "sidiropoulos".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.jackhodgson.com/weblog/archives/000023.html	1	59	50	39	80	49	49	49
http://en.wikipedia.org/wiki/Snowstorm	2	1	22	2	17	37	33	33
http://www.cnn.com/2005/WEATHER/01/22/winter.storm/	3	12	79	12	51	91	90	90
http://www.flashinsider.com/2005/08/25/webcam-snowstorm/	4	4	94	1	19	93	94	94
http://antwrp.gsfc.nasa.gov/apod/ap000131.html	5	31	71	9	60	73	73	73
http://politicalgraveyard.com/death/weather.html	6	6	44	8	20	57	56	56
http://nslog.com/archives/2005/04/04/april_snowstorm.php	7	45	70	18	98	94	92	92
http://www.snowstormrecords.com/	8	20	10	26	22	10	10	10
http://earthobservatory.nasa.gov/Newsroom/NewImages/images.php3?img_id=16745	9	29	111	10	95	111	111	111
http://dict.aiedu.com/word/snowstorm	10	2	8	57	1	3	3	3
http://www.chicagotribune.com/news/local/chi-weatherblog-link,1,7316166,framedurl?coll=chi-news-hed	11	13	78	6	53	85	85	85
http://www.schillmania.com/content/entries/2003/12/06/	12	42	6	42	10	7	7	7
http://www.ae-pro.com/happy2003.html	13	3	49	63	38	44	44	44
http://www.gskinner.com/blog/archives/2005/08/ flash_8_webcam.html	14	14	90	5	50	90	91	91
http://www.helsinginsanomat.fi/english/ article/1101978076599	15	55	106	101	76	100	101	101

Πίνακας 4.14. Συγκεντρωτικά αποτελέσματα κατάταξης του "snowstorm".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.starwars.com/	1	3	1	3	1	1	1	1
http://www.theforce.net/	2	6	2	19	3	2	2	2
http://www.jedimaster.net/	3	14	44	4	32	42	43	43
http://www.starwars.com/episode-iii/	4	12	8	8	4	8	8	8
http://www.imdb.com/title/tt0121766/	5	8	20	1	10	19	19	19
http://www.imdb.com/title/tt0076759/	6	21	21	2	11	18	18	18
http://www.asciimation.co.nz/	7	24	76	14	37	71	71	71
http://www.lucasarts.com/	8	22	6	16	2	5	5	5
http://www.lego.com/starwars/	9	95	42	60	53	61	59	59
http://www.lego.com/starwars/videogame/default.asp	10	185	166	143	182	171	171	171
http://www.lucasarts.com/games/swbattlefront/	11	35	37	47	28	37	37	37
http://starwarsgalaxies.station.sony.com/	12	69	18	15	14	17	17	17
http://www.theforce.net/swtc/	13	46	5	51	19	4	4	4
http://www.hasbro.com/starwars/	14	94	53	118	57	47	47	47
http://www.starwarsspoofs.com/	15	96	130	66	107	133	133	133

Πίνακας 4.15. Συγκεντρωτικά αποτελέσματα κατάταξης του "star wars".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://europa.eu.int/comm/press_room/presspacks/tsunami_asia/index_en.htm	1	120	86	72	113	92	92	92
http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake	2	1	1	5	1	1	1	1
http://www.noaa.gov/tsunamis.html	3	38	16	19	32	17	17	17
http://www.ngdc.noaa.gov/spotlight/tsunami/tsunami.html	4	55	36	25	48	33	34	34
http://www.pmel.noaa.gov/tsunami/sumatra20041226.html	5	8	29	11	29	30	30	30
http://www.usaid.gov/locations/asia_near_east/tsunami/	6	26	2	1	4	2	2	2
http://walrus.wr.usgs.gov/tsunami/srilanka05/	7	64	48	21	57	43	43	43
http://serc.carleton.edu/NAGTWorkshops/visualization/collections/tsunami.html	8	31	23	26	35	21	21	21
http://www.ngdc.noaa.gov/seg/hazard/tsuintro.shtml	9	71	63	53	55	58	59	59
http://wcatwc.arh.noaa.gov/IndianOSite/IndianO12-26-04.htm	10	9	75	10	36	84	81	81
http://www.prh.noaa.gov/ptwc/bulletins.htm	11	22	12	20	7	11	11	10
http://staff.aist.go.jp/kenji.satake/Sumatra-E.html	12	7	6	15	24	14	14	14
http://www.alertnet.org/thenews/emergency/SA_TID.htm	13	25	24	40	19	23	23	23
http://www.waveofdestruction.org/	14	4	9	12	2	8	8	8
http://www.firstgov.gov/Citizen/Topics/Asia_Tsunamis.shtml	15	43	27	2	34	25	26	26

Πίνακας 4.16. Συγκεντρωτικά αποτελέσματα κατάταξης του "tsunami indian ocean".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.movie-list.com/trailers.php?id=twister	1	92	89	27	80	88	88	88
http://socialtwister.com/	2	6	17	3	3	14	15	15
http://www.imdb.com/title/tt0117998/	3	2	18	6	1	11	11	11
http://www.boxofficemojo.com/movies/?id=twister.htm	4	76	77	17	76	99	99	99
http://www.turdtwister.com/index.php	5	30	99	23	87	95	96	96
http://www.logotwister.com/	6	48	11	15	11	10	10	10
http://www.twistermedia.com/	7	35	12	31	10	7	7	7
http://whyfiles.org/013tornado/	8	42	45	26	46	40	40	40
http://www.nationalgeographic.com/xpeditions/lessons/01/g912/fontwister.html	9	115	139	63	139	140	140	140
http://www.nationalgeographic.com/xpeditions/lessons/15/g35/tornadosafety.html	10	116	140	64	140	141	141	141
http://books.guardian.co.uk/departments/generalfiction/story/0,6000,1558403,00.html	11	73	144	34	136	143	143	143
http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html	12	21	13	12	19	8	8	8
http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/mt.html	13	71	10	2	27	6	6	6
http://www.onlineseats.com/auto-racing-tickets/tropicana-twister-300/index.asp	14	80	76	22	106	110	105	105
http://www.cdw.com/techtwister	15	7	143	7	96	144	144	144

Πίνακας 4.17. Συγκεντρωτικά αποτελέσματα κατάταξης του "twister".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://abcnews.go.com/Technology/Weather/	1	18	38	5	27	46	46	46
http://en.wikipedia.org/wiki/Tornado	2	1	1	1	1	1	1	1
http://www.wunderground.com/tornadoFAQ.asp	3	12	78	7	62	90	89	89
http://www.nationalgeographic.com/xpeditions/lessons/01/g912/fontwister.html	4	90	115	42	113	115	115	115
http://www.guardian.co.uk/weather/0,2759,179784,00.html	5	10	20	3	9	22	22	22
http://www.computerworld.com/securitytopics/security/recovery/story/0,10801,84579,00.html	6	14	40	74	42	35	36	36
http://www.greatweather.co.uk/	7	4	8	31	5	5	5	5
http://www.amazon.com/exec/obidos/tg/detail/-/0790729636?v=glance	8	29	15	40	23	10	10	10
http://www.kansascity.com/mld/kansascity/news/special_packages/star_weather/	9	20	46	16	16	42	44	44
http://www.disastercenter.com/tornado.htm	10	36	76	52	51	82	82	82
http://www.hprcc.unl.edu/nebraska/ summer2004photos2.html	11	5	34	8	73	47	47	47
http://en.wikipedia.org/wiki/Severe_weather	12	2	29	2	8	32	32	32
http://www.enchantedlearning.com/rhymes/Twisters.shtml	13	31	28	15	12	36	35	35
http://twister.sbs.ohio-state.edu/	14	9	3	10	4	3	3	3
http://www.967thetwister.com/	15	3	39	4	26	40	40	39

Πίνακας 4.18. Συγκεντρωτικά αποτελέσματα κατάταξης του "twister_weather".

URL	GOOGLE	HITS_A	PageRank	Prestige	SALSA_A	SCEAS_B0	SCEAS_D99	SCEASRank
http://www.ntua.gr/weather/	1	144	29	146	100	39	37	37
http://www.wunderground.com/	2	2	11	14	5	11	11	11
http://www.ananova.com/	3	4	31	7	10	19	19	19
http://www.ecmwf.int/	4	47	73	67	54	54	56	56
http://www.noaa.gov/	5	6	3	13	6	2	2	2
http://www.accuweather.com/	6	38	22	25	22	22	22	22
http://www.bbc.co.uk/weather/	7	10	2	5	3	3	3	3
http://www.cnn.com/WEATHER/	8	19	7	2	7	7	7	7
http://www.earthwatch.com/	9	40	93	103	55	93	93	93
http://www.fastweatherforecast.com/	10	116	121	117	105	142	141	141
http://www.findlocalweather.com/	11	118	119	137	97	139	138	138
http://www.intellicast.com/	12	13	17	8	19	18	18	18
http://www.metoffice.com/	13	16	89	71	58	88	88	88
http://www.my-cast.com/	14	111	90	130	86	94	94	94
http://www.tvweather.com/	15	43	103	23	74	108	107	107

Πίνακας 4.19. Συγκεντρωτικά αποτελέσματα κατάταξης του "weather".

κογένεια SCEAS κατατάσσει παρόμοια θα μπορούσαμε να πούμε με τον PageRank. Από την άλλη μεριά, HITS, SALSA και Prestige διαφέρουν σε αρκετές περιπτώσεις - κάτι που εντοπίσαμε και στην Παράγραφο 4.3.2.

Στο Παράρτημα Γ παρουσιάζεται η πλήρης λίστα αποτελεσμάτων κατάταξης των ερωτημάτων για τις λέξεις-κλειδιά "antonis sidiropoulos", "movies" και "tsunami indian ocean".

4.4 Συμπεράσματα και Μελλοντική Εργασία

Στο παρόν κεφάλαιο εφαρμόσαμε τους αλγορίθμους της οικογένειας SCEAS στο χώρο της Ιστομετρίας. Διαπιστώσαμε ότι οι αλγόριθμοι SCEAS δίνουν ποιοτικά αντίστοιχα αποτελέσματα με τους γνωστούς αλγορίθμους HITS, SALSA και

PageRank. Η ταχύτητα υπολογισμού όμως της οικογένειας SCEAS υπερτερεί κατά πολύ των ανταγωνιστών της. Από την άλλη, ίσως θα μπορούσε ο αλγόριθμος SCEAS να δώσει και ποιοτικά καλύτερα αποτελέσματα σε σχέση με τους υπολοίπους. Για κάτι τέτοιο όμως απαιτείται ραφινάρισμα των παραμέτρων d, b και a σε τιμές καταλληλότερες για τον τομέα της Ιστομετρίας. Η διαδικασία αυτή αποτελεί μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 5

Ομαδοποίηση σε Γράφους Αναφορών

Περιεχόμενα

5.1 Εισαγωγή

Ο Παγκόσμιος Ιστός είναι ένα τυπικό παράδειγμα ενός κοινωνικού δικτύου. Αντιστοίχως, ένας γράφος αναφορών είναι επίσης ένα κοινωνικό δίκτυο, που αποτελείται από τις λεγόμενες *κοινότητες* (communities). Η αναγνώριση των κοινοτήτων στους γράφους αναφορών μπορεί να βοηθήσει τη βιβλιογραφική έρευνα, όπως και την εξόρυξη δεδομένων. Στο κεφάλαιο αυτό θα παρουσιάσουμε ένα γρήγορο αλγόριθμο, που μπορεί να αναγνωρίσει τις κοινότητες σε ένα μη κατευθυνόμενο χωρίς βάρη γράφο. Αυτός ο γράφος μπορεί να αναπαριστά έναν ιστογράφο ή ένα γράφο αναφορών.

Κατά τη διάρκεια της περασμένης δεκαετίας ο Παγκόσμιος Ιστός έγινε το δημοφιλέστερο δίκτυο στον κόσμο. Ο Παγκόσμιος Ιστός μεγαλώνει με πολύ υψηλή

ταχύτητα κι επομένως ο όγκος της πληροφορίας που μπορεί να βρεθεί μέσω αυτού είναι τεράστιος. Δύο είναι τα κύρια προβλήματα του Παγκόσμιου Ιστού: (α) πως να βρει κανείς την πληροφορία, και (β) πως να την ανακτήσει γρήγορα. Για το πρώτο πρόβλημα έχουν παρουσιασθεί πολλές λύσεις τα τελευταία χρόνια. Από την ταξινόμηση σε βάση με τη λέξη-κλειδί έχουμε φθάσει στους Αλγορίθμους Ανάλυσης Αξιολόγησης (Link Analysis Ranking Algorithms - LAR). Οι αλγόριθμοι LAR έδωσαν μία αποδεκτή λύση για το πρόβλημα της αναζήτησης με το οποίο και ασχοληθήκαμε στο Κεφάλαιο 4. Για το δεύτερο πρόβλημα, οι εξυπηρετητές proxy και τα Δίκτυα Κατανομής Περιεχομένων (Content Distribution Networks) έδωσαν μία ανάσα στο πρόβλημα της ταχύτητας. Ωστόσο, ο Παγκόσμιος Ιστός ακόμη αυξάνεται με γρήγορους ρυθμούς, ενώ οι ιστοχώροι είναι τεράστιοι και συνήθως έχουν δημιουργηθεί ημι-αυτοματοποιημένα. Έτσι, αυτές οι περιοχές έρευνας πρέπει να επεκτείνουν τις προτάσεις τους.

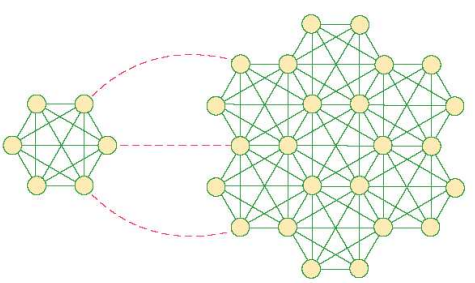
Από την άλλη μεριά, ο Παγκόσμιος Ιστός είναι ένα χαρακτηριστικό παράδειγμα ενός κοινωνικού δικτύου [93, 70]. Άλλα παραδείγματα κοινωνικών δικτύων περιλαμβάνουν το δίκτυο ιστού τροφίμων, τα δίκτυα επιστημονικών συνεργασιών, τα δίκτυα σεξουαλικών σχέσεων, τα μεταβολικά δίκτυα, τα δίκτυα εναέριας μεταφοράς κ.ο.κ. Τα κοινωνικά δίκτυα θεωρούνται συνήθως γράφοι που απαρτίζονται από κορυφές και ακμές (κατευθυνόμενες ή μη), όπου συχνά έχουν βάρη στις ακμές τους. Η θεωρία των κοινωνικών δικτύων ασχολείται με τις ιδιότητες που σχετίζονται με τη συνδεσιμότητα (όπως, βαθμός, δομή, κεντρικότητα), τις αποστάσεις (όπως, διάμετρος, κοντινότερα μονοπάτια), την προσαρμοστικότητα (όπως, γεωδαιτικές (geodesic) ακμές ή κορυφές, αρθρωτικές (articulation) κορυφές) αυτών των γράφων-μοντέλων αύξησης του δικτύου. Τα κοινωνικά δίκτυα μελετήθηκαν πολύ πριν τη σύλληψη του Παγκόσμιου Ιστού. Πρωτοποριακές εργασίες για το χαρακτηρισμό του Παγκόσμιου Ιστού ως κοινωνικό δίκτυο και για τη μελέτη των βασικών ιδιοτήτων οφείλονται στην εργασία του Barabasi και των συνεργατών του [2]. Αργότερα, πολλές μελέτες ερεύνησαν άλλες πλευρές του, όπως η ελεύθερης-κλίμακας φύση του [4, 1], η ανάπτυξή του [94, 10, 76, 66], κ.τ.λ.

Ένα από τα πλέον ενδιαφέροντα χαρακτηριστικά του Παγκόσμιου Ιστού, όπως επίσης και άλλων κοινωνικών δικτύων, είναι η αυτο-οργανωτική συμπεριφορά του, που συνήθως φαίνεται από την ύπαρξη κοινοτήτων. Ομάδες κορυφών που έχουν μεγάλη πυκνότητα ακμών μεταξύ τους και μικρότερη πυκνότητα ακμών προς άλλες ομάδες είναι ένας συχνός πληροφοριακός ορισμός μίας κοινότητας. Η έννοια της κοινότητας είναι πολύ χρήσιμη από πρακτική άποψη, επειδή μπορεί να χρησιμοποιηθεί για τη βελτίωση της αποτελεσματικότητας των μηχανών αναζήτησης

[26], για σκοπούς προτοποθέτησης (prefetching) [87], για την αξιολόγηση βιβλιογραφικών αναφορών [86], την ανίχνευση ανεπιθύμητων ηλεκτρονικών μηνυμάτων (spam) [34], τη δημιουργία χαρτών και γράφων ιστοχώρων, κ.τ.λ. Επιπλέον, η ανακάλυψη των κοινοτήτων στους γράφους αναφορών μπορεί να διευκολύνει και να επεκτείνει τη βιβλιογραφική αναζήτηση. Για παράδειγμα, θα ήταν δυνατή η εύρεση σχετικών κειμένων ακόμα κι αν δεν υπήρχαν κοινές λέξεις-κλειδιά και απ'ευθείας σύνδεση (αναφορά) μεταξύ τους. Επίσης θα ήταν δυνατόν να βρούμε συγγραφείς με τα ίδια ενδιαφέροντα, όπως επίσης και κοινότητες συγγραφέων που εργάζονται στην ίδια επιστημονική περιοχή.

5.2 Επισκόπηση Βιβλιογραφίας

Η έννοια μίας κοινότητας του Παγκόσμιου Ιστού δεν είναι πολύ αυστηρή. Γενικά περιγράφεται ως ένα υποσύνολο δομής (υποσύνολο κορυφών) ενός γράφου με πυκνή σύνδεση μεταξύ των μελών της κοινότητας και αραιή πυκνότητα εκτός της κοινότητας. Με εργαλείο μία τέτοια περιγραφή είναι πολύ εύκολο να αναγνωρίσουμε δύο κοινότητες στον γράφο που απεικονίζεται στο Σχήμα 5.1. Η ύπαρξη κοινοτήτων στον Παγκόσμιο Ιστό αναφέρθηκε για πρώτη φορά στο [33]. Ωστόσο, ο προαναφερόμενος ποιοτικός ορισμός δεν είναι αρκετός όταν προσπαθούμε επινοήσουμε αλγορίθμους για τον προσδιορισμό των κοινοτήτων σε γράφους του Παγκόσμιου Ιστού. Επομένως, χρειαζόμαστε ακριβέστερους και με ποσοτική ανάλυση ορισμούς για τις κοινότητες.



Σχήμα 5.1. Παράδειγμα ενός ιστογράφου.

Προκειμένου να παρέχουμε έναν τέτοιον ποσοτικό ορισμό, χρειάζεται να εισαγάγουμε κάποιες ποσότητες. Η βασική ποσότητα που πρέπει να θεωρήσουμε είναι το d_i, ο βαθμός ενός κόμβου i του μη κατευθυνόμενου γράφου G (που αντιπροσωπεύει το υπό εξέταση δίκτυο). Το d_i σύμφωνα με τον πίνακα γειτνίασης

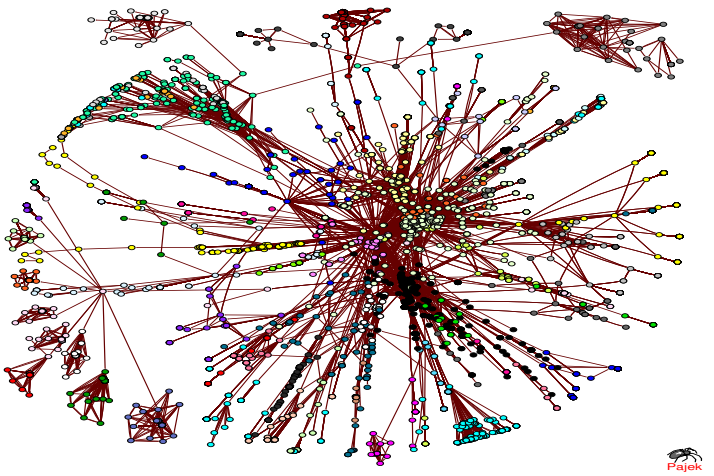
$A[i,j]$, είναι $d_i = \sum_j A[i,j]$. Εάν θεωρήσουμε έναν υπο-γράφο $V \subset G$, όπου ανήκει ο κόμβος i, μπορούμε να χωρίσουμε το συνολικό βαθμό d σε δύο ποσότητες: $d_i(V) = d_i^{in}(V) + d_i^{out}(V)$. Ο πρώτος όρος του αθροίσματος δηλώνει το πλήθος των ακμών που συνδέουν τον κόμβο i με τους άλλους κόμβους που ανήκουν στον V, δηλ, $d_i^{in} = \sum_{j \in V} A[i,j]$. Ο δεύτερος όρος τύπου του αθροίσματος δηλώνει το πλήθος των συνδέσεων προς κόμβους του υπόλοιπου γράφου, δηλ, $d_i^{out} = \sum_{j \notin V} A[i,j]$. Ο πρώτος ορισμός των κοινοτήτων οφείλεται στον Flake [26, 27], ο οποίος όρισε μία κοινότητα ως ένα σύνολο κόμβων C (όπου $C \subset G$), τέτοιο ώστε

$$d_i^{in}(C) > d_i^{out}(C) \quad \forall i \in C \tag{5.1}$$

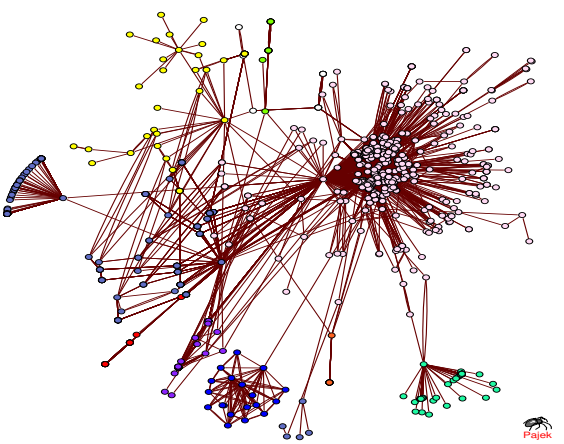
Γενικά, μπορούμε να δώσουμε διάφορους ποσοτικούς ορισμούς σε μία κοινότητα, οι οποίοι εξαρτώνται και από το γενικότερο πλαίσιο της εφαρμογής που αναπτύσσεται. Η δομή μίας κοινότητας μπορεί να συναντηθεί σε διάφορες κλίμακες στον Παγκόσμιο Ιστό. Αυτές που έχουν ερευνηθεί λεπτομερέστερα είναι οι κοινότητες που εκτείνονται σε πολλούς ιστοχώρους (inter-site communities) και συνήθως ορίζουν ευρείες θεματικές περιοχές, που προσδιορίζονται από ένα σύνολο λέξεων-κλειδιών, π.χ. κοινότητα της 9/11 [28]. Οι έννοιες των σύνθετων κειμένων [24, 21] και των λογικών πληροφοριακών μονάδων [92, 61] συνδέονται στενά με τις κοινότητες του Παγκόσμιου Ιστού, αλλά σε πολύ μικρότερη κλίμακα, αποτελούμενες από ένα μικρό αριθμό αντικειμένων του Παγκόσμιου Ιστού σε ένα μόνο ιστοχώρο και επομένως είναι ενδο-τοπικές κοινότητες. Ένα σύνθετο κείμενο είναι ένα *λογικό κείμενο* (logical document) που έχει γραφεί (συνήθως) από ένα συγγραφέα και παρουσιάζεται σε πολλές ιστοσελίδες που συνδέονται μεταξύ τους με υπερ-συνδέσμους. Παρομοίως, μία λογική πληροφοριακή μονάδα δεν είναι μία μοναδική ιστοσελίδα, αλλά είναι ένας συνδεδεμένος υπό-γράφος που αντιστοιχεί σε ένα λογικό κείμενο, οργανωμένο σε ένα σύνολο ιστοσελίδων μέσω συνδέσμων που παρέχονται από την ιστοσελίδα που ο συγγραφέας θεωρεί ως *συνηθισμένες διαδρομές πλοήγησης* (standard navigation routes).

Επεκτείνουμε την έννοια των κοινοτήτων που είναι εσωτερικοί ενός ιστοχώρου (intra-site) και προτείνουμε κοινότητες με θέμα πολύ ευρύτερο από αυτό των λογικών κειμένων. Η ύπαρξή τους προσδιορίζεται από την πυκνότητα των συνδέσμων μεταξύ των ιστοσελίδων από τις οποίες αποτελούνται οι κοινότητες. Για να υποστηρίξουμε αυτό τον ισχυρισμό, εξετάσαμε πολλούς τόπους (domains) του Παγκόσμιου Ιστού με ένα ρομπότ[1]. Ως ένα διαισθητικό βήμα, επιβεβαιώσαμε την ύπαρξη τέτοιων κοινοτήτων χρησιμοποιώντας οπτική απεικό-

[1] Χρησιμοποιήθηκε η εφαρμογή webbot από το W3C - `http://www.w3c.org/Robot/`.



(α) www.hollins.edu



(β) noc.auth.gr

Σχήμα 5.2. Οπτικοποίηση εσωτερικών κοινοτήτων.

νιση των γράφων[2]. Ως δείγμα παρουσιάζουμε την απεικόνιση του ιστοχώρου `http://www.hollins.edu` (δες Σχήμα 5.2(α) όπως ήταν διαθέσιμο στον Παγκόσμιο Ιστό τον Ιανουάριο του 2004). Στον ιστογράφο, μπορούμε εύκολα να δούμε τη συνύπαρξη σύνθετων κειμένων (στην κάτω δεξιά γωνία), με συμπαγείς ομάδες κόμβων (επάνω κέντρο), και λιγότερο φανερών ομάδων (επάνω δεξιά στην εικόνα). Επίσης η απεικόνιση του ιστοχώρου `http://noc.auth.gr/` (με

[2]Η οπτική απεικόνιση όλων αυτών των δικτύων πραγματοποιήθηκε με το πακέτο απεικόνισης `Pajek`.

δεδομένα του Φεβρουαρίου 2006) που φαίνεται στο Σχήμα 5.2(β), έχει τα ίδια χαρακτηριστικά.

Περιληπτικά, σε μεγάλο επίπεδο μπορούμε να έχουμε κοινότητες που κατανέμονται σε διάφορους ιστοχώρους [26, 27, 52, 39], σε μικρό επίπεδο έχουμε λογικά κείμενα [61] και σε μεσαίο επίπεδο (πχ. σε επίπεδο ενός ιστοχώρου) μπορούμε να έχουμε επίσης κοινότητες [87].

Με ανάλογο τρόπο, διαφορετικοί τύποι κοινοτήτων υπάρχουν σε ένα γράφο αναφορών (ή συνεργασιών) ενός συγγραφέα. Κάποιος συγγραφέας μπορεί να έχει εργασθεί σε δύο ινστιτούτα (ομάδες εργασίας) κι έτσι θα πρέπει να ανήκει σε δύο κοινότητες που ορίζονται από τις ομάδες εργασίας. Κάποια ομάδα εργασίας μπορεί να συνεργάζεται με μία άλλη, και έτσι και οι δύο ομάδες εργασίας ανήκουν σε μία ομάδα υψηλότερου επιπέδου (ιεραρχική ομαδοποίηση). Ταυτοχρόνως, μόνο ένα άτομο μίας ομάδας εργασίας μπορεί να συνεργάζεται με ένα άλλο, έτσι αυτό το άτομο θα έπρεπε να ανήκει σε μία ομάδα που ορίζεται από την ομάδα εργασίας του και σε υψηλότερο επίπεδο στην ομάδα που ορίζεται από τη δεύτερη ομάδα εργασίας συν τον εαυτό του. Έτσι, στους γράφους αναφορών, υπάρχουν ταυτόχρονα διαφορετικοί τύποι κοινοτήτων.

Για να καλύψουμε τις περισσότερες από τις περιπτώσεις κοινοτήτων που μπορούν να βρεθούν με βάση τον ορισμό του Flake, δίνουμε έναν ασθενέστερο ορισμό για τις κοινότητες. Ορίζουμε μία κοινότητα ως C (όπου $C \subset G$) τέτοια ώστε [87]:

$$\frac{d^{out}(C)}{d^{in}(C)} < s \tag{5.2}$$

όπου $d^{in}(C)$ είναι ο αριθμός των συνδέσμων μέσα στην κοινότητα και $d^{out}(C)$ είναι ο αριθμός των συνδέσμων από μέλη σε μη μέλη. Θέτουμε τον παράγοντα s ίσο με 1, κι έτσι έχουμε μία *βασική κοινότητα*, αλλά επίσης μπορούμε να δώσουμε στο s τιμή ίση με οποιοδήποτε αριθμό μικρότερο από 1, ώστε να βρούμε ισχυρότερες κοινότητες.

Η αναγνώριση κοινοτήτων είναι ουσιαστικά μία διαδικασία ομαδοποίησης ενός γράφου, η οποία σκοπό έχει την αναγνώριση ισχυρά συνδεδεμένων υποσυνόλων κορυφών (δηλαδή, τις κοινότητες). Η ανακάλυψη των *βέλτιστων* κοινοτήτων του Παγκόσμιου Ιστού, όπως και κάθε ομαδοποίηση γράφου είναι ένα πρόβλημα NP-hard. Επομένως, όλες οι μέθοδοι που προτάθηκαν βασίζονται σε κάποιες ιδιότητες των γράφων ώστε να βρεθεί (ή να προσεγγισθεί) η βέλτιστη ομαδοποίηση.

Κάποιες μέθοδοι αξιολογούν μόνο την τοπική γειτονιά μιας κορυφής, ώστε να αποφασίσουν αν ανήκει σε μία συγκεκριμένη κοινότητα, ενώ κάποιες άλλες μέθοδοι απαιτούν εξέταση όλου του ιστογράφου, ώστε να ανακαλύψουν τέτοιες

κοινότητες. Όποια και αν είναι η μέθοδος που επιλέγεται για να αναγνωρίσει τις κοινότητες, πάντα υπάρχει κάποιο αντιστάθμισμα που συνδέεται με αυτή τη διαδικασία. Το αντιστάθμισμα αυτό σχετίζεται με την *ποιότητα* των ανακαλυφθέντων κοινοτήτων, δηλαδή την πυκνότητα της συνδεσιμότητας μέσα σε αυτές, με το αντίστοιχο υπολογιστικό κόστος χρόνου και μνήμης. Στο υπόλοιπο αυτού του κεφαλαίου, υπογραμμίζουμε τις σημαντικότερες μεθόδους εύρεσης κοινοτήτων οι οποίες ομαδοποιούνται σε: (α) βιβλιομετρικές, (β) φασματικές, (γ) μέγιστης ροής και (δ) γραφοθεωρητικές. Η πρώτη οικογένεια μεθόδων εκμεταλλεύεται μόνο τοπικές πληροφορίες, η δεύτερη οικογένεια βασίζεται σε πληροφορίες από όλο το γράφο, ενώ οι άλλες δύο οικογένειες μπορούν να προσαρμοσθούν ώστε να χρησιμοποιούν είτε τοπικές είτε γενικές πληροφορίες ή συνδυασμό αυτών.

5.2.1 Βιβλιομετρικές Μέθοδοι

Οι βιβλιομετρικές μέθοδοι (bibliometric methods) επιχειρούν να αναγνωρίσουν τις κοινότητες αναζητώντας για ομοιότητα μεταξύ ζευγών κορυφών. Επομένως, πρέπει να απαντήσουν στην ερώτηση "Είναι αυτές οι δύο σελίδες παρόμοιες;". Για να απαντηθεί αυτή η ερώτηση, πρέπει να ορισθεί ένα μέτρο ομοιότητας για τις κορυφές. Υπάρχουν δύο τέτοια μέτρα που χρησιμοποιούνται ευρέως. Το πρώτο είναι το *συναναφερόμενο ζευγάρωμα* (*co-citation coupling*) και το δεύτερο είναι το *βιβλιογραφικό ζευγάρωμα* (*bibliographic coupling*). Οι βιβλιογραφικές τεχνικές είναι σχετικά παλιές. Περισσότερες πληροφορίες για την εφαρμογή τους μπορούν να βρεθούν στα [17, 29].

5.2.2 Φασματικές Μέθοδοι

Οι δημοφιλέστερες φασματικές μέθοδοι (spectral method) για την αναγνώριση κοινοτήτων είναι οι αλγόριθμοι `HITS` [52] και `PageRank` [16, 75, 55, 9, 6]. Λέγονται έτσι διότι χρησιμοποιούν ιδιοδιανύσματα (eigenvectors) και ιδιοτιμές (eigenvalues), τα οποία αποτελούν το φάσμα ενός πίνακα.

Η δημοφιλέστερη φασματική μέθοδος για την αναγνώριση κοινοτήτων είναι ο `HITS` [52]. Ο αλγόριθμος `HITS` παίρνει ένα υποσύνολο του γράφου του Παγκόσμιου Ιστού και δημιουργεί δύο βάρη για κάθε σελίδα στο υποσύνολο. Τα βάρη συνήθως αναφέρονται ως η βαθμολογία του εστιακού σημείου και της αυθεντίας, αντίστοιχα. Ο `HITS` εκτελείται σε δύο μέρη: το πρώτο είναι ένα βήμα προ–επεξεργασίας που χρησιμοποιείται για να επιλέξει το υποσύνολο του ιστογράφου που πρόκειται να χρησιμοποιηθεί, ενώ το δεύτερο μέρος είναι μία επαναληπτική αριθμητική διαδικασία. Το πρώτο μέρος δουλεύει ως εξής:

1. Στείλε την ερώτηση που ενδιαφέρει στη μηχανή αναζήτησης.
2. Πάρε τα περίπου 400 κορυφαία αποτελέσματα.
3. Αναγνώρισε όλες τις ιστοσελίδες που είναι έναν ή δύο συνδέσμους μακριά (προς κάθε κατεύθυνση) από τα αποτελέσματα που συγκεντρώθηκαν στο προηγούμενο βήμα.

Η αναφερόμενη διαδικασία δημιουργεί ένα βασικό σύνολο ιστοσελίδων που είτε περιέχουν τις λέξεις-κλειδιά του ερωτήματος που μας ενδιαφέρει, είτε βρίσκονται σε απόσταση δύο συνδέσμων από κάποια σελίδα που περιέχει τις λέξεις-κλειδιά. Επειδή και τα δύο γινόμενα των πινάκων A^TA και AA^T είναι συμμετρικά και θετικά, καθένα θα έχει την ιδιότητα ότι το αριστερό και το δεξιό ιδιοδιάνυσμα θα είναι πανομοιότυπο (λόγω συμμετρίας), ενώ το πρώτο ιδιοδιάνυσμα θα έχει όλα τα στοιχεία του θετικά (με θετική ιδιοτιμή). Όλα τα άλλα ιδιοδιανύσματα για αυτούς τους πίνακες μπορεί να είναι ετερογενή κατά το ότι τα στοιχεία τους μπορούν να έχουν μικτά πρόσημα. Αυτά τα ακόλουθα ιδιοδιανύσματα μπορούν να χρησιμοποιηθούν στο διαχωρισμό σελίδων σε διαφορετικές κοινότητες με ένα τρόπο που σχετίζεται με το φασματικό καταμερισμό γράφων. Χρησιμοποιώντας μία τέτοια μέθοδο, βρέθηκε ότι μη-μέγιστα ιδιοδιανύσματα μπορούν να χρησιμοποιηθούν για να μοιράσουμε σελίδες από ένα βασικό σύνολο σε πολλαπλές κοινότητες που περιέχουν παρόμοιο κείμενο, αλλά είναι ανόμοιες στις έννοιες.

Αντίστοιχο σκεπτικό ακολουθείται και από τον `PageRank`. Η αντίστοιχη μέθοδος περιγράφεται στο [41]. Άλλες φασματικές μέθοδοι για την εύρεση κοινοτήτων περιγράφονται στα [8, 73, 95]

5.2.3 Μέθοδοι Μέγιστης Ροής

Ο Flake σε μία σειρά δημοσιεύσεων χρησιμοποίησε την έννοια της *Μέγιστης Ροής/Ελάχιστης Τομής* (max-flow/min-cut) για να ανακαλύψει κοινότητες στους ιστογράφους. Δεδομένων ενός γράφου G και δύο κορυφών του, s (πηγή) και t (προορισμός), το $s-t$ πρόβλημα μέγιστης ροής είναι η εύρεση της μέγιστης ροής που μπορεί να δρομολογηθεί από την κορυφή s προς στην κορυφή t. Η μέγιστη ροή από κάθε ακμή ορίζεται από τη συνάρτηση χωρητικότητας $c(\cdot)$ σε σχέση με το γράφο G. Το θεώρημα των Ford και Fulkerson για τη Μέγιστη Ροή/Ελάχιστη Τομή αποδεικνύει ότι η μέγιστη ροή από την κορυφή s στην κορυφή t ενός γράφου ταυτίζεται με την ελάχιστη τομή που χωρίζει τις s και t.

Η διαίσθηση πίσω από το θεώρημα είναι ότι οι ροές περιορίζονται από ανασχετικούς παράγοντες (bottlenecks), οπότε απομακρύνοντας αυτούς τους ανα-

σχετικούς παράγοντες μπορούμε να διαχωρίσουμε δύο σημεία σε ένα δίκτυο. Ο αλγόριθμος που προτάθηκε [26, 28] δουλεύει ως ακολούθως: Η είσοδος του είναι ένας γράφος G, ένα σύνολο αρχικών ιστοσελίδων S, και μία μοναδική (δηλωμένη από το χρήστη) παράμετρος α. Η διαδικασία δημιουργεί έναν καινούργιο γράφο, G_α, που έχει μία τεχνητή κορυφή t. Η κορυφή προορισμού t συνδέεται με όλες τις αρχικές κορυφές με ακμές μικρής χωρητικότητας ίσης με α. Αφού κατασκευασθεί ο γράφος G_α, η διαδικασία υπολογίζει την ελάχιστη τομή για μια τυχαία επιλεγμένη κορυφή πηγής s προς την κορυφή t. Το τμήμα R που παραμένει συνδεδεμένο στην κορυφή s είναι εγγυημένα μία κοινότητα για την οποία ισχύει η Εξίσωση 5.1. Όλα τα μέλη της κοινότητας ομαδοποιούνται σε *έναν* κόμβο. Έτσι προκύπτει ένας νέος γράφος και η διαδικασία επαναλαμβάνεται με νέο (τυχαίο) s.

5.2.4 Γραφο-Θεωρητικές Μέθοδοι

Είδαμε προηγουμένως ότι οι βιβλιομετρικές μέθοδοι προσπαθούν να αναγνωρίσουν τις *ισχυρότερες* κορυφές ώστε να εισάγουν τις γειτονικές κορυφές σε μία κοινότητα. Οι Girvan και Newman ([35]) υιοθέτησαν την αντίθετη προσέγγιση, ακολουθώντας ένα γραφο-θεωρητικό τρόπο. Αντί να προσπαθούν να κατασκευάσουν ένα μέτρο που να υπαγορεύει ποιες κορυφές είναι οι κεντρικότερες στις κοινότητες, αντιθέτως, εστίασαν στις ακμές που είναι λιγότερο κεντρικές, τις ακμές που είναι *ανάμεσα* στις κοινότητες. Αντί να κατασκευάζουν κοινότητες προσθέτοντας τις ισχυρότερες κορυφές σε ένα αρχικά κενό σύνολο, τις κατασκευάζουν αφαιρώντας προοδευτικά ακμές από τον αρχικό γράφο. Αξιοποίησαν την έννοια της Ενδιάμεσης Κεντρικότητας (betweenness centrality) των ακμών, η οποία είχε μελετηθεί στο παρελθόν ως μέτρο για την κεντρικότητα και επιρροή των κόμβων στα δίκτυα. Η Ενδιάμεση Κεντρικότητα (BC[3]) μιας κορυφής i ορίζεται ως ο αριθμός των κοντινότερων μονοπατιών μεταξύ ζευγών άλλων κορυφών που περνούν διαμέσου του i[4]. Η έννοια αυτή είναι ένα μέτρο της επιρροής ενός κόμβου στη ροή πληροφοριών ανάμεσα σε άλλους κόμβους, ειδικά σε περιπτώσεις όπου η ροή πληροφορίας επάνω από ένα δίκτυο κατά κύριο λόγο ακολουθεί το κοντινότερο διαθέσιμο μονοπάτι. Πρότειναν έναν απλό αλγόριθμο, του οποίου τα βήματα είναι τα ακόλουθα:

1. Υπολόγισε την BC για όλες τις ακμές στο δίκτυο.
2. Αφαίρεσε την ακμή με τη μεγαλύτερη BC.

[3]Εφεξής θα αναφερόμαστε στην Ενδιάμεση Κεντρικότητα με την συντομογραφία BC.
[4]Αντίστοιχα ορίζεται και η Ενδιάμεση Κεντρικότητα ακμής.

3. Ξαναϋπολόγισε την BC για όλες τις ακμές που επηρεάστηκαν από την αφαίρεση.

4. Επανάλαβε από το βήμα 2 έως ότου δεν απομείνουν ακμές.

Με αυτή τη διαδικασία ο γράφος σταδιακά αποσυνδέεται αποκαλύπτοντας τις κοινότητες που υπάρχουν.

5.3 Κίνητρο και Συνεισφορά

Οι τεχνικές που περιγράφηκαν προηγουμένως έχουν κάποια δυνατά και κάποια αδύναμα σημεία. Ξεκινώντας από τις βιβλιογραφικές μεθόδους τονίζουμε ότι μπορούν να εφαρμοσθούν μόνο σε ένα γράφο αναφορών και όχι σε έναν ιστογράφο, επειδή συνήθως χρειάζονται επιπλέον πληροφορίες για τη σχέση των κορυφών (π.χ. τους συν-συγγραφείς).

Οι φασματικές μέθοδοι μπορούν να εφαρμοσθούν μόνο όταν η πληροφορία για τη λέξη-κλειδί είναι διαθέσιμη επάνω στο γράφο. Επίσης, δεν είναι ικανές να βρουν όλες τις κοινότητες του γράφου, παρά μόνο μία κοινότητα που σχετίζεται με την έρευνα μιας λέξης-κλειδιού. Αυτό σημαίνει ότι δεν μπορούν να εφαρμοσθούν όταν η πληροφορία της λέξης-κλειδί δεν είναι διαθέσιμη ή δεν γνωρίζουμε από πριν τις λέξεις-κλειδιά.

Οι μέθοδοι μέγιστης ροής έχουν κάποια μεγάλα μειονεκτήματα. Το πρώτο μειονέκτημα είναι ότι βασίζονται σε έναν πολύ *αυστηρό* ορισμό των κοινοτήτων. Αν ο γράφος μας δεν έχει τόσο αυστηρές κοινότητες, η μέθοδος δεν είναι ικανή να βρει κάποια. Το δεύτερο είναι ότι κυρίως χρησιμοποιούνται για να βρουν κοινότητες εξωτερικές ενός ιστοχώρου, με το να αφαιρούν τους συνδέσμους που είναι εσωτερικοί στους ιστοχώρους. Η ύπαρξη εσωτερικών συνδέσμων στον ιστοχώρο πρακτικά ακυρώνει τα αποτελέσματα. Από την άλλη μεριά, ο χρόνος υπολογισμού είναι μεγάλος και βασίζεται σε πολλές αποφάσεις που πρέπει να ληφθούν κατά τη διάρκεια της υλοποίησης του αλγορίθμου. Έτσι υπάρχουν πολλές παραλλαγές του ίδιου αλγορίθμου. Καθώς οι παραλλαγές αποκτούν καλύτερη απόδοση, ο χρόνος υπολογισμού αυξάνεται δραματικά. Όμως το σημαντικότερο μειονέκτημα αυτών των μεθόδων είναι η ύπαρξη του παράγοντα α, που πρέπει να εισαχθεί από το χρήστη και δεν υπάρχει κάποιος κανόνας για να τον θέσουμε σε σχέση με τα χαρακτηριστικά του γράφου. Η μόνη μέθοδος είναι να εκτελέσουμε δυαδική αναζήτηση για να βρεθεί μια τιμή του α με βάση την οποία να προκύπτει ομαδοποίηση. Αυτό πρακτικά σημαίνει πολλές αποτυχημένες προσπάθειες για να πάρουμε μία ομαδοποίηση.

Τέλος, οι γραφο-θεωρητικές μέθοδοι, όπως θα περιγράψουμε και στη συνέχεια, απαιτούν μεγάλη χρήση μνήμης όπως και μεγάλο υπολογιστικό χρόνο. Επιπλέον, στον πραγματικό κόσμο, μία ιστοσελίδα μπορεί να μην ανήκει αυστηρά μόνο σε μία αλλά σε περισσότερες κοινότητες. Ομοίως, μπορεί να μην ανήκει σε κάποια κοινότητα. Επομένως, για το σύνολο των κοινοτήτων $C_1, C_2, \ldots, C_k$ μπορεί να ισχύει $\bigcup_i C_i \subset G$ και όχι πάντα $\bigcup_i C_i = G$. Μπορεί επίσης να υπάρχουν τομές μεταξύ των κοινοτήτων. Έτσι, μπορεί να υπάρχουν i, j (όπου $i \neq j$) τέτοια ώστε $C_i \cap C_j \neq \emptyset$. Αυτό είναι η κύρια θεωρητική διαφορά με όλες τις υπόλοιπες μεθόδους που δεν καλύπτουν αυτήν την περίπτωση.

Έτσι, η μέθοδος μας ψάχνει για ένα σύνολο ομάδων $C = \{C_1, C_2, ...\}$ τέτοιο ώστε $\forall C_i \in C$ να ισχύει $\frac{d^{out}(C_i)}{d^{in}(C_i)} < s$, όπου s θα μπορούσε να ορίζεται από το χρήστη, αλλά κανονικά τίθεται στο 1. Μπορεί να υπάρχουν άπειρες ομαδοποιήσεις που ικανοποιούν τη συνθήκη αυτή. Εμείς επικεντρωνόμαστε στις ομαδοποιήσεις που ελαχιστοποιούν την έκφραση:

$$Q_C = \frac{1}{|C|} \sum_{\forall C_i \in C} \frac{d^{out}(C_i)}{d^{in}(C_i)} \tag{5.3}$$

5.4 Προτεινόμενη Μέθοδος

Στόχος μας είναι να βρούμε ομάδες τέτοιες ώστε να ισχύει η Εξίσωση 5.2. Μπορούμε να κατασκευάσουμε αυτές τις ομάδες ξεκινώντας από κάποιους αντιπροσωπευτικούς κόμβους για κάθε μία από αυτές, εφ'όσον μπορούμε να τους βρούμε. Αναφερόμαστε σε αυτούς τους κόμβους ως *πυρήνες*. Το πρώτο μας μέλημα θα πρέπει να είναι η εύρεση κάποιων πυρήνων. Από την άλλη μεριά, η BC είναι ένας τρόπος να βρούμε πόσο κεντρικός είναι κάθε κόμβος του γράφου G. Έχουν παρουσιασθεί πολλοί αλγόριθμοι στη βιβλιογραφία για τον υπολογισμό της BC. Ο ευφυέστερος από αυτούς είναι ο αλγόριθμος που παρουσιάζεται στο [15][5], ο οποίος έχει πολυπλοκότητα $\mathcal{O}(nm)$ και απαίτηση μνήμης $\mathcal{O}(m+n)$, όπου n είναι το πλήθος των κόμβων και m το πλήθος των ακμών. Έτσι, μπορούμε να χρησιμοποιήσουμε τους κόμβους με χαμηλή BC ως *πυρήνες*.

5.4.1 Ομαδοποίηση Χρησιμοποιώντας την έννοια της BC

Η έννοια της BC χρησιμοποιείται στην εργασία [71] για την ομαδοποίηση ενός γράφου. Σε αυτή την εργασία, η BC υπολογίζεται για κάθε ακμή του γράφου. Σε κάθε βήμα, η ακμή e με τη μεγαλύτερη BC αφαιρείται από το γράφο και η BC

[5]Ευχαριστώ τον Ulrik Brandes που μου παρείχε την υλοποίηση του υπολογισμού της BC [15].

ξανα-υπολογίζεται για κάποιες από τις ακμές. Με άλλα λόγια, όλα τα κοντινότερα μονοπάτια που περιείχαν τις διεγραμμένες ακμές e ξανα-υπολογίζονται και η BC υπολογίζεται εκ νέου για όλες τις ακμές. Η διαδικασία επαναλαμβάνεται έως ότου πάρουμε ομάδες (συνδεδεμένα τμήματα - connected components) που δεν είναι συνδεδεμένες μεταξύ τους.

Η πολυπλοκότητα αυτού του αλγορίθμου είναι υψηλή $\mathcal{O}(n^3)$, αφού η BC ξανα-υπολογίζεται σε κάθε βήμα του αλγορίθμου. Επιπλέον, έχει υψηλές απαιτήσεις μνήμης $\mathcal{O}(n^2)$, αφού πρέπει να αποθηκεύουμε όλα τα κοντινότερα μονοπάτια για όλα τα ζεύγη κόμβων. Αυτό απαγορεύει τη χρήση του συγκεκριμένου αλγορίθμου για μεγάλους γράφους και ειδικά για πυκνούς γράφους, επειδή η απαίτηση μνήμης καθίσταται τεράστια. Οι συγγραφείς του [71] αναφέρουν ότι χρησιμοποιώντας υπολογιστές της εποχής μας (2003), η μέθοδος που περιγράφεται μπορεί να εφαρμοσθεί σε γράφους με περίπου 10.000 κόμβους.

Ανεξάρτητα από το ότι πρακτικά υπάρχουν δυσκολίες στην εφαρμογή του αλγορίθμου, το συμπέρασμα της ανωτέρω εργασίας είναι ότι οι κορυφές με μεγάλη BC είναι κοντά στα σύνορα των ομάδων, όπως επίσης και ακμές με υψηλή BC είναι ακμές εσωτερικές της ομάδας. Από την άλλη μεριά, κορυφές και ακμές με χαμηλή BC βρίσκονται στο κέντρο των ομάδων ή απλά δεν είναι συνδεδεμένες με άλλες ομάδες.

Ο ισχυρισμός αυτός δεν αληθεύει όταν μέρος του γράφου έχει δομή δένδρου. Τμήμα του γράφου με μορφή δένδρου σημαίνει ότι σε αυτόν τον υπο-γράφο δεν υπάρχουν κύκλοι και ότι το πλήθος των ακμών είναι ίσο με το πλήθος των κορυφών. Όλοι οι κόμβοι σε έναν τέτοιον υπο-γράφο έχουν μεγάλη BC, οπότε εμφανώς αποτελούν μία ομάδα. Υποθέτουμε ότι σε έναν ιστογράφο τα τμήματα αυτά δεν καλύπτουν μεγάλο ποσοστό του γράφου. Αναφερόμαστε σε αυτά τα μέρη του γράφου που έχουν δενδροειδή μορφή ως *ουρές γράφων*. Πρακτικά, οι ουρές με τη μορφή δένδρου σε έναν ιστογράφο αναπαριστούν εικονικά κείμενα, που δεν έχουν συνδέσμους που δείχνουν έξω από το κείμενο. Έτσι, μπορεί να αποτελούν ανεξάρτητες ομάδες του γράφου μας ή μπορεί να είναι μέλη ομάδων.

5.4.2 Η Μέθοδος Ομαδοποίησης CBC

Ο αλγόριθμος CBC (Ομαδοποίηση με Ενδιάμεση Κεντρικότητα - Clustering by Betweenness Centrality) ξεκινά με τη γνώση ότι οι κόμβοι με τη χαμηλότερη BC είναι μέλη ομάδων και δεν συνδέονται απ' ευθείας με άλλες ομάδες. Αρχικά αφαιρούμε τις *ουρές του γράφου* από το γράφο G όπως περιγράψαμε νωρίτερα, καταλήγοντας σε ένα νέο γράφο G'. Υπολογίζουμε την BC βασιζόμενοι στο G'.

Το υπόλοιπο της διαδικασίας απεικονίζεται στο Σχήμα 5.3, όπου C είναι αρχικά ένα κενό σύνολο ομάδων.

```
function InitClustering(graph G,G', clustering C, int max_clique_size) {
   InitiateCliques(G',C,max_clique_size);
   ExtentTailedClusters(G,C); // Add the tails of size 1
                              // to the cliques that they are connected
   MakeTailedClusters(G,C);   // Make the tree-like tails independent Cliques
   MergeTailedClusters(G,C);  // Merge tail-clusters until reach the max-cluster-size
}
```

Σχήμα 5.3. Αρχικοποίηση ομαδοποίησης.

5.4.2.1 Διαμόρφωση Κλίκας

Οι κόμβοι με τη χαμηλότερη BC είναι οι πυρήνες των ομάδων. Έτσι, μπορούμε να κατασκευάσουμε κάποιες αρχικές μικρές ομάδες γύρω από αυτούς. Αυτό απεικονίζεται ως ψευδοκώδικας στο Σχήμα 5.4. Αυτές οι μικρές ομάδες είναι οι *κλίκες* του γράφου. Πρέπει όμως εδώ να προσεχθεί ότι με τον όρο *κλίκα* εννοούμε περιοχές γύρω από τους πυρήνες, ενώ στη βιβλιογραφία ο όρος αυτός σημαίνει έναν πλήρως συνδεδεμένο υπο-γράφο. Το μέγεθος της κλίκας μπορεί να ποικίλει και είναι συνάρτηση του πόσο πυκνός ή αραιός είναι ο γράφος. Για έναν πυκνό γράφο, οι κλίκες αποτελούνται από όλους τους κόμβους που συνδέονται κατ'ευθείαν με τον πυρήνα. Για αραιούς γράφους οι κλίκες μπορεί να είναι μεγαλύτερες. Το βέλτιστο μέγεθος κλίκας δεν μπορεί να είναι γνωστό εκ των προτέρων, αφού ο

```
function InitiateCliques(graph G,clustering C) {
  static max_clique_size=sqrt(G.n_nodes);
  for all nodes n in G ordered by BC desc {
    cluster c;
    if(n belongs to any c in C) {
        next;
    }
    c = {n};
    for all p neighbors of n {
        if(not p belongs to any c in C) {
           c.add(p);
        }
    }
    C.add(c)
    while(ExtentClique(G,c,C,max_clique_size)) {;}
  }
  max_clique_size=next value;
}

function ExtentClique(graph G,
     cluster c, clustering C,
     int max_size) {
     if(c.n_nodes>max_size) {
        return false
     }
     Extent c using BFS until the
     size of 2*max_size
     return true
}
```

Σχήμα 5.4. Αρχικοποίηση κλικών.

γράφος μπορεί να είναι πυκνός σε κάποιες περιοχές, αλλά αραιός σε κάποιες άλλες. Την πρώτη φορά που θα κληθεί η *InitiateCliques*, η παράμετρος *Maximum Clique Size* τίθεται στο μηδέν. Επομένως, οι Κλίκες που θα κατασκευασθούν κατά την πρώτη επανάληψη του αλγορίθμου έχουν διάμετρο δύο. Εδώ αξίζει να σημειωθεί ότι αν μία αρχική κλίκα (Initial Clique) έχει μέγεθος ενός ή δύο κόμβων, σβήνεται και αγνοείται. Αυτό σημαίνει ότι μετά από το πρώτο βήμα μπορεί να έχουμε κάποιους *ορφανούς* κόμβους. Αυτοί θα είναι κόμβοι με μεγάλη BC και συνήθως βρίσκονται ανάμεσα στις ομάδες. Το υπολογιστικό κόστος αυτού του βήματος είναι γραμμική συνάρτηση του μεγέθους του γράφου: $\mathcal{O}(n)$.

5.4.2.2 Συγχώνευση Κλικών

Έχοντας ένα σύνολο κλικών, το επόμενο βήμα είναι να τις συγχωνεύσουμε, έτσι ώστε να κατασκευάσουμε συσχετιζόμενες ομάδες (correlated clusters). Οι κλίκες από μόνες τους μπορεί να μην αποτελούν συσχετιζόμενες ομάδες. Ο ψευδοκώδικας παρουσιάζεται στο Σχήμα 5.5. Υποθέτοντας ότι στο προηγούμενο βήμα βρήκαμε ένα πλήθος ομάδων l (κλικών κατά την πρώτη επανάληψη), στη συνάρτηση *Merge* κατασκευάζουμε έναν $l \times l$ πίνακα B. Κάθε στοιχείο $B[i,j]$ αντιστοιχεί στο πλήθος των ακμών από την ομάδα (ή κλίκα) C_i στην ομάδα (ή κλίκα) C_j. Τα διαγώνια στοιχεία $B[i,i]$ αντιστοιχούν στο πλήθος των εσωτερικών ακμών της κλίκας. Ο πίνακας B είναι προφανώς συμμετρικός. Σημειώνεται επίσης ότι ο κόμβος x ανήκει στις ομάδες C_i και C_j και υπάρχει μία ακμή $x \rightarrow y$ ($y \in C_i$), τότε αυτή η ακμή μετρά μία φορά για το $B[i,i]$, αφού είναι εσωτερική ακμή, αλλά επίσης μετρά και για το $B[i,j]$, αφού το y ανήκει και στο C_j. Έτσι, το άθροισμα μιας γραμμής του πίνακα B δεν είναι ίσο με το συνολικό πλήθος των ακμών που ανήκουν ή συνδέονται με μια ομάδα, αλλά συνήθως μεγαλύτερο.

```
function ClusterMerge(graph G,clustering C...) {
   while(! ok) {
      MakeCliquesStronger(C,G);
      Merge(G,C);
      if(c->best_quality==0) {manage_subsets(c);}
      delete_the_worst(C);  // Delete a cluster if does not fulfil parameters
      add_orphans_to_cliques(G,C);
      ok=check(C);
      if(!ok) InitClustering(G,C);
   }
}
```

Σχήμα 5.5. Συνένωση ομάδων.

Αυτό το βήμα συγχώνευσης του αλγορίθμου συνίσταται από πολλές επαναλήψεις. Σε κάθε επανάληψη εκτελείται μία συγχώνευση. Οι επαναλήψεις σταματούν όταν δεν υπάρχει άλλο ζεύγος προς συγχώνευση. Το ζεύγος που θα συγχωνευθεί είναι εκείνο με το μέγιστο $B[i,j]/B[i,i]$. Οι συνθήκες για μία συγχώνευση μπορεί να ποικίλουν και εξαρτώνται από τις παραμέτρους του χρήστη, αν υπάρχουν[6]. Έτσι, σε αυτό το βήμα ελέγχουμε κάθε ζεύγος ομάδων (κλικών) και επιλέγουμε το καλύτερο προς συγχώνευση. Η συγχώνευση δεν μπορεί να συμβεί, εάν οι δύο ομάδες είναι ήδη συσχετισμένες ή η ένωσή τους είναι μεγαλύτερη από το *μέγιστο μέγεθος ομάδας* που μπορεί να θέσει ο χρήστης. Η συγχώνευση των ομάδων C_i, C_j μπορεί να γίνει εάν $B[i,j] > \sum_k B[i,k]/2$ ή εάν τουλάχιστον μία από τις C_i, C_j δεν είναι συσχετιζόμενη ή αν $B[i,j]/B[i,i] \geq s$ και δεν μπορούμε να βρούμε καλύτερο ζεύγος.

Αν το αρχικό πλήθος των κλικών είναι l, τότε ο μέγιστος αριθμός επαναλήψεων που θα εκτελεσθούν είναι l. Κάθε επανάληψη ελέγχει κάθε ζεύγος κλικών, οπότε ο χρόνος που απαιτείται είναι $l^2 + (l-1)^2 + ... + 1$, οπότε η πολυπλοκότητα είναι $\mathcal{O}(l^3)$. Η τιμή του l εξαρτάται από τα χαρακτηριστικά του γράφου, αλλά είναι συνάρτηση του $\sqrt{n}$, όπου n είναι το πλήθος των κόμβων του G[7]. Έτσι, η χρονική πολυπλοκότητα σε σχέση με το πλήθος των κόμβων, στη χειρότερη περίπτωση είναι $\mathcal{O}(n\sqrt{n})$. Οι απαιτήσεις σε μνήμη είναι ένας πίνακας $l \times l$, που σημαίνει $\mathcal{O}(l^2)$, δηλαδή $\mathcal{O}(n)$.

Στο Σχήμα 5.5 υπάρχουν διάφορα βήματα προκειμένου να βελτιώσουμε την ποιότητα και/ή την ταχύτητα, ανάλογα με τις ανάγκες μας. Για παράδειγμα, η διαδικασία που λέγεται *MakeCliquesStronger* μπορεί να χρησιμοποιηθεί για τη μετακίνηση κόμβων μεταξύ των ομάδων. Η δεδομένη συμπεριφορά της είναι να ελέγχει για όλους τους κόμβους (σε χρόνο $\mathcal{O}(n)$), τον αριθμό των συνδέσμων που έχουν προς κάθε ομάδα. Έτσι, ένας κόμβος μπορεί να αλλάξει ομάδα, αν υπάρχει κάποια άλλη ομάδα προς την οποία αυτός ο κόμβος έχει περισσότερους συνδέσμους. Η συνάρτηση *ManageSubsets* χρησιμοποιείται για να διαγράψει τυχόν ομάδες που είναι υποσύνολα άλλων ομάδων και θα κληθεί μόνο για βελτιστοποίηση της ταχύτητας.

Τέλος, μπορεί να προσθέσουμε τους ορφανούς κόμβους στις ομάδες, ακόμα κι αν η προκύπτουσα ομαδοποίηση δεν είναι καλύτερη από την υπάρχουσα[8]. Αυτό θα

[6] Μία παράμετρος μπορεί να είναι ο παράγοντας s που αναφέρεται στον ορισμό της κοινότητας.

[7] Δεδομένου ότι η παράμετρος *μέγιστο μέγεθος κλίκας* (Maximum Clique Size) αρχικώς τίθεται σε $\sqrt{n}$.

[8] Αυτό επιλέγεται από το χρήστη με την παράμετρο *Minimize Orphans*.

οδηγήσει τον αλγόριθμο να ελαχιστοποιήσει τον αριθμό των ορφανών κόμβων, αλλά ο παράγοντας ποιότητας Q_C από την Εξίσωσης 5.3 που θα προκύψει, θα είναι χειρότερος.

Τέλος, η ομαδοποίηση ελέγχεται εάν εκπληρώνει όλους τους τυχόν περιορισμούς. Εάν όχι, διαγράφονται οι ομάδες που δεν εκπληρούν τους περιορισμούς και αρχικοποιούμε εκ νέου σε κλίκες τους κόμβους που παρέμειναν ορφανοί. Κάθε φορά που καλείται η *InitClustering* χρησιμοποιούμε μία νέα τιμή για τον παράγοντα *μέγιστο μέγεθος κλίκας*. Στην υλοποίησή μας, η ακολουθία των τιμών είναι: $0, \sqrt{n}, \sqrt{n}/2, 2\sqrt{n}, \sqrt{n}/3, 3\sqrt{n}, ...$ Στις περισσότερες περιπτώσεις που αντιμετωπίσαμε κατά τη διάρκεια των πειραμάτων, η ομαδοποίηση υπολογίζεται σε δύο επαναλήψεις της συνάρτησης *InitClustering* και σε λίγες περιπτώσεις σε τέσσερις. Φυσικά, εάν τα μεγέθη των ομάδων ποικίλουν, περιμένουμε ότι θα χρειασθούν περισσότερες επαναλήψεις, πρακτικά όμως κατά τα πειράματά μας δεν συναντήσαμε τέτοια περίπτωση.

5.5 Πειραματικά Αποτελέσματα

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, ανάλογη ιδέα είναι αυτή που περιγράφεται στο [71]. Το μεγαλύτερο μειονέκτημα αυτού του αλγορίθμου είναι η τεράστια πολυπλοκότητα, αφού κάθε βήμα του αλγορίθμου απαιτεί εκ νέου υπολογισμό της BC. Αν και ο υπολογισμός είναι αυξητικός (incremental), η χρονική και χωρική πολυπλοκότητα είναι πολύ υψηλή. Στην παρούσα παράγραφο δεν θα συγκρίνουμε σε σχέση με αυτό τον αλγόριθμο, αφού η διαφορά στην πολυπλοκότητα και στις απαιτήσεις μνήμης είναι προφανής.

Από την άλλη μεριά, μία σύγκριση με τον αλγόριθμο του *Flake* δεν θα ήταν εφικτή για τρεις λόγους. Ο πρώτος είναι ότι έχουμε διαφορετικούς ορισμούς του τι είναι ομάδα, κι έτσι οι μέθοδοι του *Flake* δίνουν διαφορετικούς τύπους ομάδων. Ο δεύτερος είναι ότι ο *Flake* επιδιώκει μια ιεραρχική ομαδοποίηση. Στην περίπτωσή μας δεν αναζητούμε για ιεραρχίες στις ομάδες αλλά για ένα σύνολο ομάδων που να ικανοποιεί τους περιορισμούς μας ανεξάρτητα από το βάθος της ιεραρχίας που βρίσκονται οι ομάδες. Τέλος, όπως δηλώνεται και στο [28], η μέθοδός τους εφαρμόζεται για την εύρεση δια-τοπικών (intra-site) κοινοτήτων με την απομάκρυνση των εσωτερικών συνδέσμων των ιστοχώρων (inter-site links). Στην περίπτωσή μας αναζητούμε κοινότητες μέσα στους ιστοχώρους.

5.5.1 Μέθοδος και Σύνολο Δεδομένων Αξιολόγησης

Το σύνολο δεδομένων που χρησιμοποιείται για την αξιολόγηση της μεθόδου αποτελείται από πραγματικούς και συνθετικούς ιστογράφους. Το πραγματικό σύνολο δεδομένων περιλαμβάνει τους ιστοχώρους: `noc.auth.gr` (όπως ήταν το Φεβρουάριο του 2006), `www.hollings.edu` (Ιανουάριος 2004) και `www.unicef.org` (Ιανουάριος 2006).

Οι συνθετικοί ιστογράφοι δημιουργήθηκαν με το εργαλείο FWgen [49]. Οι παράμετροι που πρέπει να δοθούν στο FWgen προκειμένου να δημιουργήσει γράφους, είναι πέντε: (α) το πλήθος των κόμβων, (β) το πλήθος των ακμών ή η πυκνότητα σε σχέση με τον αντίστοιχο πλήρως συνδεδεμένο γράφο[9], (γ) ο αριθμός των ομάδων που πρέπει να δημιουργηθούν, (δ) η ανομοιομορφία (skew) που αντανακλά τα σχετικά μεγέθη των παραγόμενων ομάδων[10], και τέλος (ε) ο συντελεστής *assortativity* που δίνει το ποσοστό των ακμών που πρόκειται να είναι εσωτερικές ακμές στις κοινότητες. Οι τιμές που έχουν νόημα για την *assortativity* είναι μεγαλύτερες του 50%. Όσο μεγαλύτερη είναι η *assortativity*, τόσο ισχυρότερες είναι οι ομάδες που δημιουργούνται. Αν η *assortativity* είναι 100%, τότε οι παραγόμενες ομάδες θα είναι ασύνδετες μεταξύ τους.

Το εργαλείο FWgen δημιουργεί δύο αρχεία. Το πρώτο είναι ο γράφος και το δεύτερο καταγράφει τις παραγόμενες ομάδες, έτσι ώστε μπορούμε να συγκρίνουμε την ομαδοποίηση της CBC με τη *βέλτιστη παραγόμενη*. Αφού η παραγωγή ακμών ακολουθεί τυχαίες αποφάσεις, η βέλτιστη παραγόμενη ομαδοποίηση μπορεί να μην είναι πανομοιότυπη με την *απόλυτα βέλτιστη*. Με τον τελευταίο όρο εννοούμε την ομαδοποίηση που ελαχιστοποιεί τον παράγοντα Q_C, όπως θα φανεί αργότερα στο κεφάλαιο αυτό.

5.5.1.1 Μέθοδος Αξιολόγησης

Η μέθοδος αξιολογείται σε δύο διαστάσεις: ποιότητα και ταχύτητα. Η εκτίμηση της ποιότητας θα μπορούσε να γίνει μόνο με τους συνθετικούς γράφους, για τους οποίους γνωρίζουμε εκ των προτέρων τις ομάδες που κατασκευάζονται. Έτσι, μετρούμε την απόσταση της μεθόδου μας από τη βέλτιστη παραγομένη ομαδοποίηση, χρησιμοποιώντας ένα μέτρο απόστασης που εξηγείται στην Παράγραφο 5.5.1.2. Μετρούμε επίσης την τιμή Q_C της Εξίσωσης 5.3. Όσο μικρότερη είναι

[9] Η πυκνότητα d ενός γράφου είναι $d = \frac{2m}{n(n-1)}$ όπου m το πλήθος των ακμών και n το πλήθος των κορυφών.

[10] $Skew = 0.3$ σημαίνει ότι δυο τυχαίες ομάδες μπορεί να έχουν διαφορά στο μέγεθός τους μέχρι 30%.

αυτή η τιμή, τόσο καλύτερη είναι η ομαδοποίηση που παράγεται. Από την άλλη μεριά, η αξιολόγηση της ταχύτητας ομαδοποίησης είναι τετριμμένη. Μετρούμε τον πραγματικό χρόνο της εκτέλεσης του αλγορίθμου (ο χρόνος που απασχολήθηκε η ΚΜΕ από τη διαδικασία). Για το πραγματικό σύνολο δεδομένων παρουσιάζουμε τα στατιστικά αποτελέσματα της ομαδοποίησης.

5.5.1.2 Σύγκριση Ομαδοποιήσεων

Σε αυτό το κεφάλαιο θα ορίσουμε μια συνάρτηση σύγκρισης δυο ομαδοποιήσεων[11] $\mathcal{C}_A$ και $\mathcal{C}_B$. Η κάθε ομαδοποίηση ορίζει ένα σύνολο ομάδων όπου: $\mathcal{C}_A = \{c_{A1}, c_{A2}, ...c_{Ax}\}$. Η κάθε ομάδα c_{Ax} περιέχει ένα σύνολο κόμβων του γράφου $G = (V, E)$. Στη γενική περίπτωση δυο ομάδες μπορεί να περιέχουν και κοινούς κόμβους, δηλαδή να ισχύει: $c_{Ai} \cap c_{Aj} \neq \emptyset$. Επίσης, θα μπορεί ένας κόμβος να μην ανήκει σε κάποια ομάδα, δηλαδή: $n \in V \nRightarrow \exists c_{Ai} \in \mathcal{C}_A : n \in c_{Ai}$.

Πριν προχωρήσουμε, ας δούμε κάποιους ορισμούς που θα μας βοηθήσουν να κατασκευάσουμε τη ζητούμενη συνάρτηση:

$$\mathcal{N}(n, c) = \begin{cases} 1 & if\ n \in c \\ 0 & if\ n \notin c \end{cases}$$

Ορίζουμε τη συνάρτηση $\mathcal{N}(n, c)$ να είναι ίση με 1, εάν ο κόμβος n ανήκει στην ομάδα c και ίση με 0 διαφορετικά.

$$\mathcal{K}(n, \mathcal{C}) = \{c \in \mathcal{C} : n \in c\}$$

Επίσης, η συνάρτηση $\mathcal{K}(n, \mathcal{C})$ μας δίνει το σύνολο των ομάδων όπου ανήκει ο κόμβος n. Το πλήθος των ομάδων όπου μπορεί να ανήκει ένας κόμβος ενδέχεται να είναι:

- μηδέν. Τότε ο κόμβος είναι ορφανός.
- ένα ή περισσότερες. Προφανώς, το μέγιστο πλήθος ομάδων όπου μπορεί να ανήκει ένας κόμβος είναι όλες οι ομάδες της ομαδοποίησης $\mathcal{C}$.

Έχουμε τελικά $0 \leq \mathcal{K}(n, \mathcal{C}) \leq |\mathcal{C}|$

[11] Στην ελληνική βιβλιογραφία χρησιμοποιείται και ο όρος Κατάτμηση αντί της ομαδοποίησης. Κατάτμηση (partitioning) όμως δηλώνει το χωρισμό ενός συνόλου σε ομάδες χωρίς αυτές να έχουν κοινά στοιχεία και χωρίς να μένουν ορφανοί κόμβοι. Για το λόγο αυτό προτιμάται ο όρος ομαδοποίηση.

Η ομοιότητα ή η συγγένεια S δυο κόμβων n_1 και n_2 σε μια ομαδοποίηση $\mathcal{C}$ ορίζεται ως το ποσοστό των εμφανίσεων του κόμβου n_2 στο σύνολο των ομάδων όπου ανήκει ο n_1 ($\mathcal{K}(n_1, \mathcal{C})$):

$$S(n_1, n_2, \mathcal{C}) = \frac{\sum\limits_{\forall c \in \mathcal{K}(n_1, \mathcal{C})} \mathcal{N}(n_2, c)}{|\mathcal{K}(n_1, \mathcal{C})|}$$

Αν η ομαδοποίηση επιβάλλει οι κόμβοι να ανήκουν μόνο με μια ομάδα ο καθένας, τότε η συνάρτηση S μπορεί να πάρει μόνο τις τιμές 0 και 1, ανάλογα αν οι δυο κόμβοι ανήκουν στην ίδια ομάδα ή όχι. Στη γενική περίπτωση όπου ένας κόμβος μπορεί να ανήκει σε περισσότερες από μια ομάδες, τότε η συνάρτηση S μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα [0..1]. Η τιμή θα είναι 0 αν οι δυο κόμβοι δεν ανήκουν σε κάποια κοινή ομάδα και 1 αν ο n_2 ανήκει σε όλες τις ομάδες που ανήκει ο n_1. Ενδιάμεσες τιμές θα έχουμε αν ο n_2 ανήκει σε κάποιες από τις ομάδες που ανήκει ο n_1.

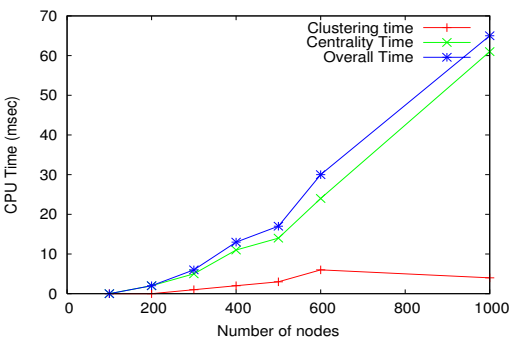
Ας σημειωθεί εδώ ότι εφ'όσον η ομαδοποίηση επιτρέπει ένας κόμβος να ανήκει σε περισσότερες από μια ομάδες, τότε η ισότητα: $S(n_1, n_2, \mathcal{C}) = S(n_2, n_1, \mathcal{C})$ *δεν ισχύει πάντα*. Η ανωτέρω ισότητα ισχύει μόνο αν ο κάθε κόμβος επιτρέπεται να ανήκει *μόνο* σε μια ομάδα (δηλαδή αν ισχύει: $\forall n \in V \quad |\mathcal{K}(n, \mathcal{C})| = 1$).

$$\mathcal{D}(\mathcal{C}_A, \mathcal{C}_B, G) = \frac{\sum\limits_{\forall n_1 \in V} \sum\limits_{\forall n_2 \in V} |S(n_1, n_2, \mathcal{C}_A) - S(n_1, n_2, \mathcal{C}_B)|}{|V|(|V| - 1|)} \tag{5.4}$$

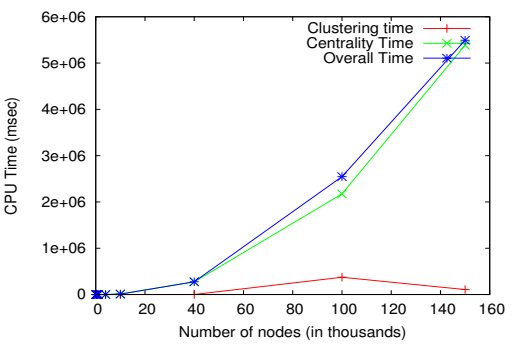
Τέλος, η συνάρτηση $\mathcal{D}(\mathcal{C}_A, \mathcal{C}_B, G)$ (5.4) μας επιτρέπει να συγκρίνουμε δυο ομαδοποιήσεις $\mathcal{C}_A$ και $\mathcal{C}_B$. Ουσιαστικά, η συνάρτηση $\mathcal{D}$ μετρά τη μέση διαφορά ομοιότητας για όλα τα δυνατά ζεύγη κόμβων του γράφου G. Η διαφορά $|S(n_1, n_2, \mathcal{C}_A) - S(n_1, n_2, \mathcal{C}_B)|$ μπορεί να είναι από 0 έως 1. Θα είναι 0 όταν και στις δυο ομαδοποιήσεις οι κόμβοι έχουν τον ίδιο συντελεστή ομοιότητας, και 1 όταν έχουν διαφορετικό (δηλαδή $S(n_1, n_2, \mathcal{C}_A) = 1$ και $S(n_1, n_2, \mathcal{C}_B) = 0$ ή το αντίστροφο). Συνεπώς, η συνάρτηση $\mathcal{D}$ μπορεί να πάρει τιμές στο διάστημα [0..1]. Αν οι δυο ομαδοποιήσεις είναι όμοιες, τότε $\mathcal{D}(\mathcal{C}_A, \mathcal{C}_B, G) = 0$. Στη χειρότερη περίπτωση, η ανωτέρω συνάρτηση μπορεί να είναι ίση με 1. Αυτό συμβαίνει αν στην πρώτη ομαδοποίηση έχουμε μόνο μια ομάδα όπου ανήκουν όλοι οι κόμβοι του γράφου, ενώ στη δεύτερη ομαδοποίηση έχουμε ομάδες με έναν και μόνο κόμβο η κάθε μια.

5.5.2 Σχολιασμός Αποτελεσμάτων

Στα γραφήματα (α) και (β) του Σχήματος 5.6 παρουσιάζουμε την ταχύτητα εκτέλεσης του αλγορίθμου μας σε σχέση με το πλήθος των κορυφών του γράφου. Ο *Χρόνος KME* (CPU time) που παρουσιάζουμε υπολογίζεται μέσω του πυρήνα του unix χρησιμοποιώντας την κλήση συστήματος *times()* και αντιπροσωπεύει το χρόνο που παρέμεινε η διεργασία στην KME. Η γραμμή με τους διαγώνιους σταυρούς αντιπροσωπεύει τον απαιτούμενο χρόνο για τον υπολογισμό της BC. Η γραμμή με τα σημεία-σταυρούς αντιπροσωπεύει τον απαιτούμενο χρόνο για τον υπολογισμό της ομαδοποίησης. Τέλος, η γραμμή με τα σημεία-αστερίσκους αντιπροσωπεύει το άθροισμα όλων των προηγουμένων. Παρατηρούμε ότι ο αλγόριθμός μας απαιτεί πολύ λιγότερο χρόνο από τον υπολογισμό της BC, ο οποίος αποδείχθηκε ότι έχει πολυπλοκότητα $\mathcal{O}(mn)$. Στα Σχήματα 5.7(α) και 5.7(β), επιβεβαιώνεται ότι ο υπολογισμός της BC είναι γραμμικά ανάλογος του $n * m$,

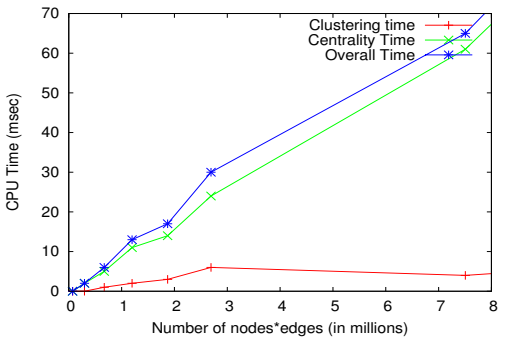


(α)

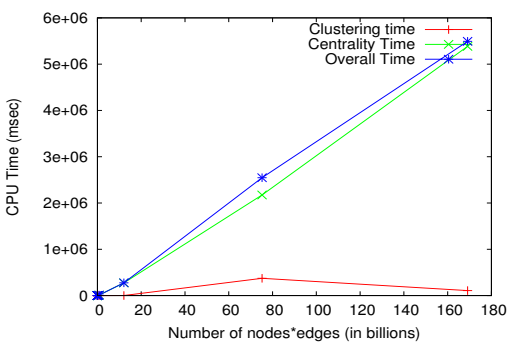


(β)

Σχήμα 5.6. Χαρακτηριστικά ιστογράφων: κορυφές: n, ακμές: $15 * n$, ομάδες: 5 $(n < 1000)$, 10 $(1000 \leq n < 10000)$, 100 $(10000 \leq n)$, assortativity: 0.85, skew:0.1.



(α)

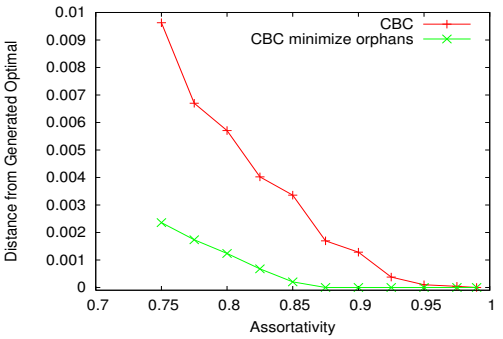


(β)

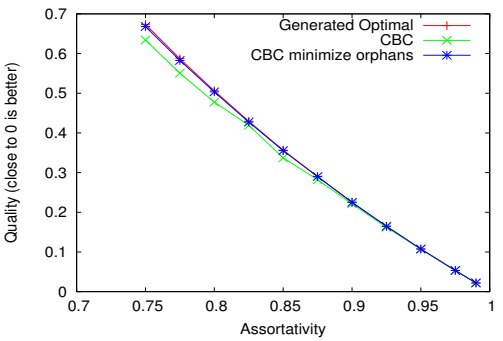
Σχήμα 5.7. Χαρακτηριστικά ιστογράφων: κορυφές: n, ακμές: $15 * n$, ομάδες: 5 $(n < 1000)$, 10 $(1000 \leq n < 10000)$, 100 $(10000 \leq n)$, assortativity: 0.85, skew:0.1.

οπότε οι μετρήσεις μας για την ταχύτητα είναι σωστές. Επομένως, αν χρησιμοποιήσουμε έναν αλγόριθμο προσεγγιστικό της BC με πολυπλοκότητα καλύτερη του $\mathcal{O}(mn)$, θα είμαστε σε θέση να ομαδοποιήσουμε πραγματικά τεράστιους γράφους που αποτελούνται από πολύ περισσότερους από 200.000 κόμβους.

Στο Σχήμα 5.8 παρουσιάζουμε τα αποτελέσματα της ομαδοποίησης που σχετίζονται με την παράμετρο assortativity. Το Σχήμα 5.8(α) δείχνει την απόσταση της ομαδοποίησής μας από τη βέλτιστη παραγόμενη. Η απόσταση υπολογίζεται με τη συνάρτηση $\mathcal{D}(\mathcal{C}_A, \mathcal{C}_B, G)$. Η κόκκινη γραμμή (με τα σημεία-σταυρούς) αντιπροσωπεύει τον αλγόριθμό μας CBC, ενώ η πράσινη γραμμή (με τους διαγώνιους σταυρούς) αναπαριστά τον αλγόριθμό μας με χρήση της επιλογής για ελαχιστοποίηση των ορφανών κόμβων. Παρατηρούμε ότι η έκδοση με την ελαχιστοποίηση των ορφανών κόμβων παράγει ομαδοποίηση που πλησιάζει περισσότερο στη βέλτιστη παραγόμενη. Αυτό συμβαίνει επειδή στη βέλτιστη παραγόμενη δεν υπάρχουν



(α)



(β)

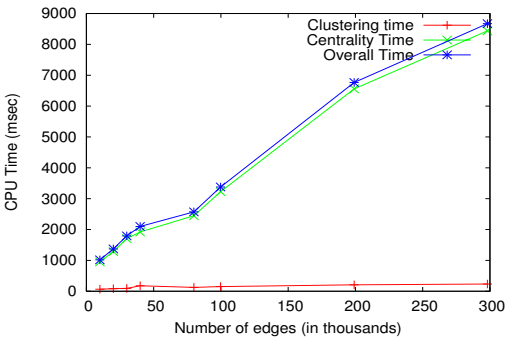
Σχήμα 5.8. Χαρακτηριστικά ιστογράφων: κορυφές: 4000, ακμές: 30000, ομάδες: 10, skew:0.10.

ορφανοί κόμβοι.

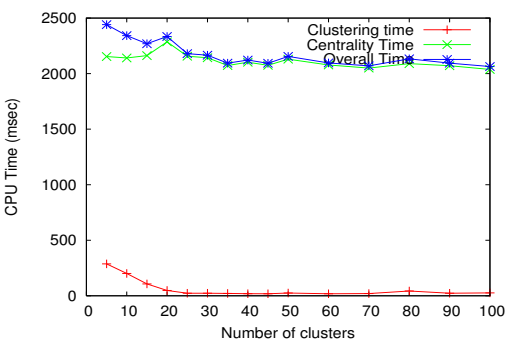
Η απόσταση από τη βέλτιστη παραγόμενη είναι 1% στη χειρότερη περίπτωση και συγκλίνει στο μηδέν καθώς οι ομάδες γίνονται ισχυρότερες. Από την άλλη μεριά, στο Σχήμα 5.8(β) παρουσιάζουμε την ποιότητα της ομαδοποίησης που εκφράζεται με τον παράγοντα Q_C της Εξίσωσης 5.3. Είναι προφανές ότι όταν οι ομάδες είναι ισχυρές, η ποιότητα της ομαδοποίησης είναι καλύτερη.

Εδώ σημειώνεται ότι και οι δύο εκδόσεις μας του CBC κρατούν την ποιότητα πολύ κοντά στη βέλτιστη παραγόμενη ομαδοποίηση και είναι πάντα καλύτερες από αυτήν. Αυτό οφείλεται στο γεγονός ότι το εργαλείο FWGen δεν κατασκευάζει τις βέλτιστες ομάδες, όμως αν υπάρχουν στο γράφο, ο αλγόριθμός μας μπορεί να τις εντοπίσει. Αυτό εξηγεί το γεγονός ότι στο Σχήμα 5.8(α), η απόσταση από τη βέλτιστη παραγόμενη ομαδοποίηση δεν είναι μηδέν.

Στο Σχήμα 5.9 κρατούμε σταθερά τα χαρακτηριστικά του γράφου και αλλά-



Σχήμα 5.9. Χαρακτηριστικά ιστογράφων: κορυφές: 4000, ομάδες: 10, assortativity: 0.90, skew:0.10.



Σχήμα 5.10. Χαρακτηριστικά ιστογράφων: κορυφές: 5000, ακμές: 37500-39000, assortativity: 0.90, skew:0.10.

ζουμε το πλήθος των ακμών, και κατ'επέκταση την πυκνότητα. Παρατηρούμε ότι ο απαιτούμενος χρόνος για την ομαδοποίηση παραμένει μικρός και σχεδόν σταθερός. Τέλος, στο Σχήμα 5.10, μεταβάλλουμε το πλήθος των ομάδων (στον άξονα x). Φαίνεται ότι όταν οι ομάδες είναι λίγες, ο απαιτούμενος χρόνος είναι μεγαλύτερος από αυτόν που απαιτείται για περισσότερες ομάδες. Αυτό οφείλεται στο γεγονός ότι πρέπει να εκτελεσθούν περισσότερες λειτουργίες συγχώνευσης.

Στον Πίνακα 5.1 παρουσιάζουμε μία σύνοψη των αποτελεσμάτων για τους πραγματικούς ιστογράφους. Οι τρεις πρώτες στήλες ορίζουν το γράφο. Οι στήλες MC και MS αντιπροσωπεύουν τις παραμέτρους χρήστη μέγιστο μέγεθος ομάδας και ελάχιστο μέγεθος ομάδας. Η επόμενη στήλη *Ομάδες* περιέχει το πλήθος των ομάδων που βρέθηκαν κατά την κάθε εκτέλεση. Το Q_C είναι η Ποιότητα της ομαδοποίησης που έχει προκύψει ως αποτέλεσμα (Εξίσωση 5.3). Η στήλη *Or*

Ιστοχώρος	Κόμβοι	Ακμές	MC	MS	Ομάδες	Q_C	$Drate$	Or
noc.auth	955	5620	50%	5	3	0.994	1.00	486
noc.auth	955	5620	80%	5	10	0.332	1.04	3
hollins	4487	16373	50%	10	64	0.204	1.05	170
unicef	56852	749666	30%	10	142	0.544	1.14	67
unicef	56852	749666	30%	1000	12	0.206	1.12	91

Πίνακας 5.1. Αποτελέσματα ομαδοποίησης πραγματικών ιστογράφων.

δηλώνει τον αριθμό των ορφανών κόμβων που παρέμειναν στο γράφο. Τέλος, η στήλη $Drate$ δείχνει το ποσοστό των κόμβων που ανήκουν σε περισσότερες από μία ομάδες και υπολογίζεται ως:

$$Drate = \frac{\sum_{\forall i \in G} N(i)}{N_C}$$

όπου $N(i)$ είναι ο αριθμός των ομάδων που ανήκει ο κόμβος i και N_C είναι ο αριθμός των κόμβων που ανήκουν τουλάχιστον σε μία ομάδα. Η τιμή 1 για το $Drate$ σημαίνει ότι όλοι οι κόμβοι ανήκουν σε μία ακριβώς ομάδα.

Εφ'όσον οι πιθανές ομάδες που ενδέχεται να έχουν $d^{out}/d^{in} < s$ μπορεί να είναι άπειρες, πρέπει με κάποιο τρόπο να εστιάσουμε σε μερικές από αυτές. Για αυτό το λόγο, η υλοποίησή μας θεωρεί δύο παραμέτρους. Η πρώτη είναι το *ελάχιστο μέγεθος ομάδας* (MS στον Πίνακα 5.1) και η δεύτερη είναι το *μέγιστο μέγεθος ομάδας* αναλογικά με το μέγεθος του γράφου (MC στον Πίνακα 5.1). Είναι προφανές ότι αυτές οι δύο παράμετροι επηρεάζουν τα αποτελέσματα.

Για παράδειγμα, στην πρώτη προσπάθεια να ομαδοποιήσουμε τον ιστοχώρο `http://noc.auth.gr` χρησιμοποιήσαμε ως MC την προκαθορισμένη τιμή 50%. Αυτό προκάλεσε τον αλγόριθμο να αφήσει πολλούς ορφανούς κόμβους. Η δεύτερη εκτέλεση χρησιμοποίησε $MC = 80\%$. Αυτό παρήγαμε μία μεγάλη ομάδα 496 κόμβων (Πίνακας 5.2), η οποία είναι μεγαλύτερη από το μισό του γράφου και δεν θα μπορούσε να χωρισθεί σε μικρότερες ομάδες. Τα αποτελέσματα της ομαδοποίησης απεικονίζονται και στο Σχήμα 5.2(β), όπου κάθε ομάδα παρουσιάζεται με διαφορετικό χρώμα. Δυστυχώς, είναι αδύνατον να απεικονίσουμε τους κόμβους που ανήκουν σε περισσότερες από μία ομάδες, αφού κάθε κόμβος μπορεί να έχει μόνο ένα χρώμα. Αυτοί οι κόμβοι παίρνουν χρώμα από μία από τις ομάδες όπου ανήκουν με τυχαία επιλογή. Συνολικά αποτελέσματα για όλα τα πειράματα είναι διαθέσιμα στο `http://delab.csd.auth.gr/~asidirop/clustering`.

Τέλος, μπορούμε να συμπεράνουμε ότι ο αλγόριθμος CBC απαιτεί χρόνο $\mathcal{O}(nm)$, που ουσιαστικά είναι ο απαιτούμενος χρόνος για τον υπολογισμό της BC.

Κόμβοι	C^{in}	C^{out}	C^{out}/C^{in}
496	4846	143	0.0295089
272	322	154	0.478261
62	61	2	0.0327869
46	83	46	0.554217
44	47	33	0.702128
19	59	1	0.0169492
16	15	2	0.133333
16	15	2	0.133333
13	11	6	0.545455
12	10	7	0.7

Πίνακας 5.2. Αποτελέσματα ομαδοποίησης του `http://noc.auth.gr`.

Οι απαιτήσεις μνήμης είναι της τάξης του $\mathcal{O}(n)$ για τη διαδικασία της συγχώνευσης και περίπου $\mathcal{O}(n)$ για να αποθηκεύει σε ποιές ομάδες ανήκει ο κάθε κόμβος. Σε συνθετικούς γράφους, η ποιότητα ομαδοποίησης που έχει μετρηθεί είναι πάντοτε καλύτερη από την ποιότητα της βέλτιστης παραγόμενης ομαδοποίησης. Η απόσταση αυτών των δύο ομαδοποιήσεων είναι το πολύ 1% και εξαρτάται από το τι αναζητούμε. Είναι προφανές ότι σε πραγματικούς ιστογράφους η ποιότητα των ομάδων εξαρτάται από τον ίδιο τον ιστογράφο. Όταν ψάχνουμε για μεγάλες ομάδες, είναι προφανές ότι ο παράγοντας ποιότητας Q_C θα είναι καλύτερος. Τα μεγέθη των ομάδων μπορεί να ποικίλουν ανάλογα με την εφαρμογή όπου προβλέπεται να χρησιμοποιηθούν τα αποτελέσματα της ομαδοποίησης. Ο αλγόριθμός μας είναι ικανός να βρίσκει ομάδες οποιουδήποτε επιθυμητού μεγέθους, εφ'όσον υπάρχουν στον ιστογράφο.

5.6 Συμπεράσματα και Μελλοντική Εργασία

Στο κεφάλαιο αυτό επιχειρήσαμε μία επισκόπηση των μεθόδων ομαδοποίησης. Παρουσιάσαμε επίσης τη δική μας μέθοδο, που ονομάζεται CBC και έχει απαίτηση χρόνου $\mathcal{O}(mn)$ και μνήμης $\mathcal{O}(n)$. Επίσης παρουσιάσαμε πειράματα σε συνθετικά και πραγματικά σύνολα δεδομένων. Όπως αναμενόταν, τα πειράματα δείχνουν ότι αυτή η μέθοδος είναι πολύ γρήγορη και μπορεί να χρησιμοποιηθεί στην ομαδοποίηση τεράστιων ιστογράφων. Διαπιστώνουμε ότι το αργό τμήμα της μεθόδου είναι ο υπολογισμός της Ενδιάμεσης Κεντρικότητας. Ως μελλοντική εργασία σχεδιάζουμε να χρησιμοποιήσουμε κάποιον προσεγγιστικό αλγόριθμο της Εν-

διαμέσης Κεντρικότητας για να ελέγξουμε την ταχύτητα ομαδοποίησης και την απόδοση της ποιότητας. Αυτή η μέθοδος χρησιμοποιείται επίσης για τις μεθόδους προτοποθέτησης (prefetching) σε ένα CDN [87].

ΚΕΦΑΛΑΙΟ 6

H-index για Συγγραφείς, Συνέδρια και Περιοδικά

Περιεχόμενα

6.1 Εισαγωγή

Εκτός από τους αλγορίθμους που έχουν παρουσιασθεί μέχρι τώρα για αξιολόγηση συγγραφέων και συλλογών (συνεδρίων/περιοδικών), περί το τέλος του 2005 εμφανίσθηκε μια νέα μέθοδος για την αξιολόγηση συγγραφέων [44]. Ο ορισμός της μετρικής *h-index* είναι ο εξής:

ΟΡΙΣΜΟΣ 6.1. Ένας ερευνητής έχει δείκτη h (*h-index*) εάν h από τις N_p δημοσιεύσεις του έχουν συγκεντρώσει τουλάχιστον h αναφορές η κάθε μια και οι υπόλοιπες $(N_p - h)$ δημοσιεύσεις έχουν συγκεντρώσει λιγότερο από h αναφορές.

Αυτή η μετρική υπολογίζει πόσο ευρεία είναι η ερευνητική συνεισφορά του κάθε συγγραφέα. Για να έχει κάποιος συγγραφέας μεγάλο *h-index* δεν αρκεί απλώς να έχει μερικές καλές εργασίες, αλλά χρειάζονται πολλές καλές εργασίες. Προφανώς νέοι ερευνητές είναι εκτός συναγωνισμού σε αυτή τη μετρική διότι (α) δεν έχουν τον απαιτούμενο χρόνο να δημοσιεύσουν πολλές καλές εργασίες και (β) οι εργασίες τους δεν έχουν προλάβει να γίνουν γνωστές και άρα να συγκεντρώσουν αρκετές αναφορές.

Με βάση τον προηγούμενο ορισμό, το h θα είναι πάντα μικρότερο ή ίσο από το πλήθος N_p των δημοσιεύσεων ενός ερευνητή. Επίσης θα ισχύει πάντα η σχέση: $h^2 \leq N_{c,tot}$, όπου $N_{c,tot}$ είναι το συνολικό πλήθος των αναφορών που έχουν γίνει στο συγκεκριμένο ερευνητή. Προφανώς η ισότητα θα ισχύει μόνον όταν όλες οι δημοσιεύσεις που συνεισφέρουν στο δείκτη *h-index* έχουν ακριβώς από h αναφορές η κάθε μια (κάτι απίθανο). Άρα, στην περισσότερο αναμενόμενη περίπτωση ισχύει: $h^2 < N_{c,tot}$. Από την προηγουμένη σχέση, μπορεί να ορισθεί και ο δείκτης a ως εξής:

$$N_{c,tot} = ah^2 \tag{6.1}$$

Ο δείκτης a μπορεί να λειτουργήσει ως ένας δεύτερος δείκτης-κριτήριο για την αξιολόγηση και κατάταξη των συγγραφέων. Πιο συγκεκριμένα, δείχνει το μέγεθος των *επιτυχιών* του συγγραφέα. Το a θα είναι μεγάλο όταν κάποια ή κάποιες από τις δημοσιεύσεις έχουν δεχθεί πολλές αναφορές, αρκετά περισσότερες από αυτό που μας δείχνει το h. Εμπειρικά τυπικές τιμές για τον δείκτη a κυμαίνονται μεταξύ 3 και 5.

Σύμφωνα με τον Hirsch, τον εμπνευστή του *h-index*, οι υπόλοιπες παραδοσιακές μετρικές έχουν τα εξής χαρακτηριστικά:

1. Το πλήθος δημοσιεύσεων μετρά παραγωγικότητα αλλά όχι αντίκτυπο ή σημαντικότητα της έρευνας.

2. Το συνολικό πλήθος αναφορών έχει τα μειονεκτήματα ότι είναι δύσκολο να βρεθεί και επιπλέον επηρεάζεται από τις επιτυχίες ενός ερευνητή, οι οποίες δεν είναι πάντα αντιπροσωπευτικές της συνολικής συνεισφοράς του στην ακαδημαϊκή κοινότητα. Αν χρησιμοποιούμε τον *h-index* τέτοιες περιπτώσεις απεικονίζονται στο δείκτη a, ο οποίος και θα έχει υπερβολικά μεγάλες τιμές.

3. Το μέσο πλήθος αναφορών ανά δημοσίευση έχει το πλεονέκτημα ότι είναι ένας κανονικοποιημένος δείκτης και άρα μπορούν να γίνουν συγκρίσεις ερευνητών διαφορετικής ηλικίας. Από την άλλη πριμοδοτεί τη μικρή παραγωγικότητα και *τιμωρεί* την υψηλή παραγωγικότητα.

4. Το πλήθος των *σημαντικών* δημοσιεύσεων που ορίζεται ως το πλήθος των δημοσιεύσεων με περισσότερες από y αναφορές είναι ένας αρκετά καλός δείκτης, αλλά απαιτεί τον ορισμό του y. Ανάλογα με την τιμή που θα δοθεί στο y μπορεί να πριμοδοτούνται ή το αντίστροφο κάποιοι συγγραφείς. Επιπλέον, πρέπει να χρησιμοποιείται διαφορετικό y ανάλογα με την *αρχαιότητα* του εκάστοτε ερευνητή.

5. Τέλος, το πλήθος των αναφορών στις q καλύτερες εργασίες καθενός συγγραφέα έχει και αυτό το μειονέκτημα καθορισμού του συντελεστή q το οποίο επηρεάζει τα αποτελέσματα.

6.2 H-index για Συνέδρια και Περιοδικά

Ορμώμενοι από την προηγούμενη ιδέα μπορούμε να ορίσουμε αντίστοιχα το *h-index* για συνέδρια και περιοδικά. Δηλαδή, το *h-index* ενός περιοδικού ή συνεδρίου να είναι ο αριθμός h εάν h από τις N_p δημοσιεύσεις που έχουν γίνει στο περιοδικό ή συνέδριο έχουν συγκεντρώσει τουλάχιστον h αναφορές η κάθε μια, ενώ οι υπόλοιπες $(N_p - h)$ δημοσιεύσεις έχουν συγκεντρώσει λιγότερο από h αναφορές. Πάλι όμως δεν έχουμε ισορροπία στη σύγκριση συνεδρίων/περιοδικών με μικρή ιστορία σε σχέση με αντίστοιχα που έχουν μεγάλο ιστορικό.

Η λύση στο πρόβλημα αυτό είναι να υπολογίζουμε τον *h-index* ανά έτος.

ΟΡΙΣΜΟΣ 6.2. Ένα συνέδριο ή περιοδικό έχει δείκτη h_y για το έτος y (*yearly h-index*) εάν h_y από τις $N_{p,y}$ δημοσιεύσεις του έτους y έχουν συγκεντρώσει περισσότερες από h_y αναφορές η κάθε μια, ενώ οι υπόλοιπες $(N_{p,y} - h_y)$ δημοσιεύσεις έχουν συγκεντρώσει λιγότερες από h_y αναφορές.

Για παράδειγμα, το h_{1992} του συνεδρίου VLDB είναι το πλήθος των δημοσιεύσεων του VLDB'92 που έχουν συγκεντρώσει περισσότερο από h_{1992} αναφορές. Τα μειονεκτήματα τις μέτρησης αυτής είναι τα εξής δυο:

1. Τα συνέδρια/περιοδικά δεν δημοσιεύουν το ίδιο πλήθος άρθρων. Έτσι ένα συνέδριο A που δημοσιεύει περίπου 50 άρθρα ανά έτος, έχει προφανώς άνω φράγμα στο δείκτη h_y το 50. Ένα άλλο συνέδριο B που δημοσιεύει 150 άρθρα έχει προφανώς ως άνω φράγμα για το δείκτη h_y το 150 και αρκετές πιθανότητες να ξεπεράσει το σκορ των 50. Το πλήθος βέβαια των δημοσιεύσεων που γίνονται σε ένα συνέδριο/περιοδικό ανά έτος αντικατοπτρίζει και την προτίμηση που έχουν οι ερευνητές σε αυτό. Αν λοιπόν θεωρήσουμε ότι το A δημοσίευσε 50 άρθρα επειδή δεν μπόρεσε να συλλέξει περισσότερα

αρκετά αξιόλογα, τότε ορθώς έχει ως άνω φράγμα το δείκτη 50 και ορθώς ας μην μπορέσει να ξεπεράσει το B. Από την άλλη μεριά όμως ίσως μας ενδιαφέρει και ο μέσος αντίκτυπος των άρθρων ενός συνεδρίου/περιοδικού ανεξαρτήτως του όγκου.

2. Το h_y συνεχώς αλλάζει. Ακόμη κι αν μιλούμε για ένα συνέδριο του 1970 ο δείκτης που μπορούμε να υπολογίσουμε σήμερα είναι πιθανό να αλλάξει μετά από μερικά χρόνια. Το μειονέκτημα, λοιπόν, της μεθόδου είναι ότι ποτέ δεν μπορούμε να έχουμε τελική αξιολόγηση για τα συνέδρια/περιοδικά μιας χρονολογίας, όσο παλιά κι αν είναι αυτή.

Για το πρώτο μειονέκτημα, μπορούμε να ορίσουμε έναν παράλληλο δείκτη που να είναι κανονικοποιημένος σε σχέση με το πλήθος των δημοσιεύσεων που υπάρχουν. Ορίζουμε λοιπόν:

ΟΡΙΣΜΟΣ 6.3. Ένα συνέδριο ή περιοδικό έχει κανονικοποιημένο δείκτη $h_y^n = h_y/N_{p,y}$ για το έτος y (*normalized yearly h-index*) εάν h_y από τις $N_{p,y}$ εργασίες του έτους y έχουν συγκεντρώσει περισσότερες από h_y αναφορές η κάθε μια, ενώ οι υπόλοιπες $(N_{p,y} - h_y)$ δημοσιεύσεις έχουν συγκεντρώσει λιγότερες από h_y αναφορές.

6.3 H-index για Συγγραφείς

Όπως στην περίπτωση των συνεδρίων/περιοδικών, το ίδιο ακριβώς μπορεί να υπολογισθεί και για τους συγγραφείς, δηλαδή όχι μόνο να μετρούμε την ευρύτητα των εργασιών τους, αλλά και το ποσοστό των επιτυχημένων εργασιών σε σχέση με το συνολικό αριθμό εργασιών:

ΟΡΙΣΜΟΣ 6.4. Ένας ερευνητής έχει δείκτη $h^n = h/N_p$ (*normalized h-index*) εάν h από τις N_p δημοσιεύσεις του έχουν συγκεντρώσει τουλάχιστον h αναφορές η κάθε μια, ενώ οι υπόλοιπες $(N_p - h)$ δημοσιεύσεις έχουν συγκεντρώσει λιγότερες από h αναφορές.

Όπως θα δούμε και πειραματικά, η συγκεκριμένη μετρική προφανώς πριμοδοτεί τους συγγραφείς με μικρή παραγωγικότητα και άρα πρέπει να εφαρμόζεται για συγκεκριμένους σκοπούς αξιολόγησης ή ως δευτερεύων δείκτης.

Το μειονέκτημα όλων των προηγούμενων περιπτώσεων είναι ότι δεν λαμβάνουμε υπ'όψη μας το χρόνο κατά τον οποίο έγινε μια δημοσίευση ή μια αναφορά σε αυτήν. Για παράδειγμα, έστω ένας ερευνητής που έχει κάνει καλές εργασίες

τη δεκαετία του '60 και οι οποίες δέχονται πολλές αναφορές. Ο ερευνητής αυτός θα έχει μεγάλο *h-index* εξαιτίας των εργασιών που έχει κάνει στο παρελθόν, και εφ'όσον οι εργασίες του συνεχίζουν να δέχονται αναφορές, το *h-index* του θα αυξάνεται. Αυτό όμως δεν αντικατοπτρίζει την ερευνητική συνεισφορά του *σήμερα*. Σήμερα μπορεί αυτός ο ερευνητής να είναι ανενεργός. Άρα υπάρχει η ανάγκη για ακόμη ένα δείκτη που θα πριμοδοτεί τους ερευνητές που συνεισφέρουν ακόμη και σήμερα ή μπορεί να μην συνεισέφεραν πολύ στο παρελθόν αλλά πλέον έχουν μια δυναμική που μας κάνει να περιμένουμε ενδιαφέρουσες εργασίες από αυτούς στο μέλλον.

Μπορούμε να ορίσουμε λοιπόν ένα νέο δείκτη που βασίζεται στη φιλοσοφία του *h-index* για τους συγγραφείς. Ορίζουμε τη βαθμολογία $S^c(i)$ (Contemporary Score) της δημοσίευσης i ως:

$$S^c(i) = \gamma * (Y(now) - Y(i) + 1)^{-\delta} * |C(i)|$$

Όπου $Y(i)$ και $C(i)$ είναι:

$Y(i)$=Το έτος της δημοσίευσης i
$C(i)$=Οι δημοσιεύσεις που κάνουν αναφορά στη δημοσίευση i

Θέτοντας δ=1, το $S^c(i)$ είναι το πλήθος των αναφορών που έχει λάβει η δημοσίευση i, διαιρεμένο με την ηλικία της. Εφ'όσον διαιρούμε το πλήθος των αναφορών με τη χρονική απόσταση, θα προκύψουν πολύ μικρά $S^c(i)$ και δεν θα μπορέσουν να δημιουργήσουν το *h-index* που θέλουμε. Γι' αυτό χρησιμοποιούμε το συντελεστή γ. Στα πειράματά μας τον έχουμε θέσει ίσο με 4. Έτσι, για μια δημοσίευση που έγινε φέτος οι αναφορές μετρούν πολλαπλασιασμένες επί 4. Για μια δημοσίευση που έγινε πριν από 4 χρόνια, οι αναφορές μετρούν επί 1. Για μια δημοσίευση που έγινε πριν από 6 χρόνια, οι αναφορές μετρούν επί 4/6 κ.ο.κ.

Έτσι, μια παλιά δημοσίευση, ακόμη και αν συνεχίζει να δέχεται αναφορές, χάνει σιγά–σιγά την αξία της, και ουσιαστικά στον υπολογισμό του *h-index* λαμβάνουμε υπ'όψη κυρίως τις καινούργιες εργασίες[1]. Έτσι το *contemporary h-index* ενός ερευνητή ορίζεται:

ΟΡΙΣΜΟΣ 6.5. Ένας ερευνητής έχει δείκτη h^c (*contemporary h-index*) εάν h^c από τις N_p δημοσιεύσεις του έχουν βαθμολογία $S^c(i) \geq h^c$ η κάθε μια, ενώ οι υπόλοιπες $(N_p - h^c)$ δημοσιεύσεις έχουν βαθμολογία $S^c(i) \leq h^c$.

[1] Προφανώς, αν το δ είναι στην περιοχή του μηδενός, μειώνεται ο αντίκτυπος που έχει ο *χρονικός* έλεγχος και προφανώς για $\delta = 0$ και $\gamma = 1$ δεν διαφέρει από τον απλό *h-index*.

Μια άλλη οπτική γωνία είναι να μετρήσουμε πόσο αντίκτυπο έχει η δουλειά ενός ερευνητή δεδομένης μιας χρονικής στιγμής. Έτσι, μπορούμε να πούμε ότι δεν μας ενδιαφέρει πόσο παλιές είναι οι εργασίες κάποιου, αλλά αν συνεχίζουν να δέχονται αναφορές από άλλους. Έτσι αξιολογούμε κατά κάποιον τρόπο και τη διαχρονικότητα των εργασιών ενός ερευνητή.

Με τη βοήθεια των ανωτέρω μαθηματικών ορισμών, ορίζουμε τη βαθμολογία $S^t(i)$ (trend Score) για την δημοσίευση i ως:

$$S^t(i) = \gamma * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta}$$

όπου τα δ και γ ορίζονται όπως και πριν. Συνεπώς:

ΟΡΙΣΜΟΣ 6.6. Ένας ερευνητής έχει δείκτη h^t (*trend h-index*) εάν h^t από τις N_p δημοσιεύσεις του έχουν $S^t(i) \geq h^t$ η κάθε μια, ενώ οι υπόλοιπες $(N_p - h^t)$ δημοσιεύσεις έχουν λιγότερο από h^t.

Στη συνέχεια του κεφαλαίου θα υπολογίσουμε τις διάφορες παραλλαγές του *h-index* που ορίσαμε για συγγραφείς, συνέδρια και περιοδικά με βάση τα στοιχεία που έχουμε συγκεντρώσει από τη συλλογή DBLP.

6.4 Πειραματικά Αποτελέσματα για Συγγραφείς

Σε αυτήν την παράγραφο θα παρουσιάσουμε τα αποτελέσματα της αξιολόγησης συγγραφέων χρησιμοποιώντας όλες τις προηγούμενες μεθόδους. Να υπενθυμίσουμε ότι το σύνολο δεδομένων μας είναι τα δεδομένα αναφορών από τη συλλογή DBLP. Αν και το σύνολο αυτό είναι σχετικά μικρό, όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, εντούτοις είναι ένα πολύ ικανοποιητικό δείγμα. Τη στιγμή που εκτελέσθηκαν τα επόμενα πειράματα (DBLP timestamp: 10/3/2006) η βάση μας περιέχει 100205 αναφορές. Παρ'όλα αυτά, ο κύριος όγκος των αναφορών προέρχεται από δημοσιεύσεις προγενέστερες του 2001. Συνεπώς τα αριθμητικά αποτελέσματα που παρουσιάζουμε μπορούμε να υποθέσουμε ότι ίσχυαν το 2001. Από εδώ και στο εξής με τον όρο *τώρα* εννοούμε κάπου στο 2001.

6.4.1 Αποτελέσματα Κατάταξης Συγγραφέων και Συγκρίσεις

Στους Πίνακες 6.1, 6.2, 6.3 και 6.4 εμφανίζουμε τα αποτελέσματα της κατάταξης των συγγραφέων με βάση τη μέθοδο *h-index* και τις παραλλαγές που ορίσαμε. Με την πρώτη ματιά μπορούμε να δούμε πως οι τιμές που υπολογίζουμε στο *h-index*

(Πίνακας 6.1) απέχουν κατά πολύ από τις τιμές που αναφέρονται στο [44] για την περίπτωση ερευνητών φυσικής. Αυτό συμβαίνει διότι η συλλογή DBLP έχει καταχωρημένο μικρό ποσοστό των συνολικών πραγματικών αναφορών. Από την άλλη όμως η κατάταξη που παίρνουμε είναι ενδεικτική εφ'όσον το υποσύνολο των αναφορών που έχουμε διαθέσιμο προέρχεται από τα πλέον σημαντικά συνέδρια και περιοδικά.

Όνομα	h	a	$N_{c,tot}$	N_p
1.Michael Stonebraker	24	3.78	2180	193
2.Jeffrey D. Ullman	23	3.37	1783	227
3.David J. DeWitt	22	3.91	1896	150
4.Philip A. Bernstein	20	3.39	1359	124
5.Won Kim	19	2.96	1071	143
6.Catriel Beeri	18	3.16	1024	93
7.Rakesh Agrawal	18	3.06	994	154
8.Umeshwar Dayal	18	2.81	913	130
9.Hector Garcia-Molina	17	3.60	1041	314
10.Yehoshua Sagiv	17	3.52	1020	121
11.Ronald Fagin	17	2.83	818	121
12.Jim Gray	16	6.13	1571	118
13.Serge Abiteboul	16	4.33	1111	172
14.Michael J. Carey	16	4.25	1090	151
15.Nathan Goodman	16	3.37	865	68
16.Christos Faloutsos	16	2.89	742	175
17.Raymond A. Lorie	15	6.23	1403	35
18.Jeffrey F. Naughton	15	2.90	653	123
19.Bruce G. Lindsay	15	2.76	623	60
20.David Maier	14	5.56	1090	158

Πίνακας 6.1. Κατάταξη συγγραφέων με βάση το *h-index*.

Μελετώντας και τους υπόλοιπους Πίνακες 6.2, 6.3 και 6.4, η μόνη κατάταξη που διαφέρει εμφανώς από τις υπόλοιπες είναι το *normalized h-index*. Αυτό συμβαίνει διότι πλεονεκτούν οι συγγραφείς με λίγες και καλές δημοσιεύσεις. Πρακτικά όμως δεν μπορούμε να κάνουμε αξιολόγηση της ερευνητικής δουλειάς κάποιου χρησιμοποιώντας μόνο τον *normalized h-index*, αλλά ίσως παράλληλα με τον *h-index*. Μπορούμε εύκολα να υποθέσουμε ότι η πλειοψηφία των ερευνητών που φαίνονται στον Πίνακα 6.2 είναι μάλλον υποψήφιοι διδάκτορες ή μεταπτυχιακοί φοιτητές που έκαναν μερικές πολύ καλές εργασίες με έναν πολύ καλό και γνωστό καθηγητή, και μετά μάλλον έκαναν παύση στο ερευνητικό τους έργο. Η άλλη περίπτωση είναι ο κύριος όγκος των δημοσιεύσεών τους να μην περιλαμβάνεται στη συλλογή DBLP, πιθανόν διότι πλέον έχουν αλλάξει ερευνητικό τομέα.

Όνομα	h_n	h	a	N_{cit}	N_p
1.Rajiv Jauhari	1	5	3.72	93	5
2.Jie-Bing Yu	1	5	2.36	59	5
3.L. Edwin McKenzie	1	5	2.04	51	5
4.Upen S. Chakravarthy	0.88	8	2.60	167	9
5.James B. Rothnie Jr.	0.85	6	6.55	236	7
6.M. Muralikrishna	0.85	6	5.47	197	7
7.Stephen Fox	0.83	5	4.12	103	6
8.Antonin Guttman	0.8	4	20.43	327	5
9.Marc G. Smith	0.8	4	4.81	77	5
10.Gail M. Shaw	0.8	4	4.37	70	5
11.Glenn R. Thompson	0.8	4	4.37	70	5
12.David W. Shipman	0.75	6	11.16	402	8
13.Dennis R. McCarthy	0.75	6	5.30	191	8
14.Spyros Potamianos	0.66	4	10.43	167	6
15.Robert K. Abbott	0.66	4	4.68	75	6
16.Edward B. Altman	0.66	4	3.06	49	6
17.Brian M. Oki	0.66	4	2.56	41	6
18.Gene T.J. Wuu	0.66	6	2.25	81	9
19.Marguerite C. Murphy	0.66	4	1.62	26	6
20.Gerald Held	0.62	5	9.84	246	8

Πίνακας 6.2. Κατάταξη συγγραφέων με βάση το *normalized h-index*.

Τέλος, πάντα είναι πιθανόν να εντοπίσουμε ανάμεσα σε αυτούς και *ανερχόμενους* ερευνητές που θα συνεχίσουν την καλή ερευνητική παρουσία τους.

Βλέποντας τους Πίνακες 6.3 και 6.4, με την πρώτη ματιά δεν διαπιστώνουμε ουσιαστικές αποκλίσεις από τον Πίνακα 6.1 στη σειρά κατάταξης. Εδώ όμως μας ενδιαφέρει ακόμη και η μικρότερη διαφορά στην κατάταξη γιατί μπορεί να μας δώσει αρκετή πληροφορία ανά ερευνητή. Για παράδειγμα, ο Χρήστος Φαλούτσος εμφανίζεται στην $16^η$ θέση στον πίνακα *h-index* ενώ στον πίνακα *contemporary h-index* ανεβαίνει στην $14^η$ θέση. Αυτό σημαίνει ότι ο καλύτερος όγκος των καλών εργασιών του είναι σε πρόσφατα έτη (σε σχέση με τους υπόλοιπους). Επίσης στον *trend h-index* βλέπουμε ότι ανεβαίνει στην $8^η$ θέση δηλαδή οι εργασίες του δέχονται πρόσφατες αναφορές, το οποίο συνεπάγεται ότι οι εργασίες του Φαλούτσου είναι *της μόδας*. Με τον όρο αυτό προφανώς εννοούμε ότι τη συγκεκριμένη χρονική στιγμή υπάρχει γενικότερο ενδιαφέρον για το συγκεκριμένο ερευνητικό τομέα από την υπόλοιπη επιστημονική κοινότητα.

Παρακινούμενοι από τις διαφορές στους πίνακες αυτούς, παρουσιάζουμε τα Σχήματα 6.1 και 6.2. Σε αυτά βλέπουμε το ιστορικό του *h-index* για κάθε ερευνητή, επιλέγοντας αυτούς για τους οποίους έχουμε διαφορές στις παραλλαγές του

Όνομα	h_c	a_c	h	$N_{c,tot}$	N_p
1.David J. DeWitt	14	3.10	22	1896	150
2.Jeffrey D. Ullman	13	3.44	23	1783	227
3.Michael Stonebraker	12	3.98	24	2180	193
4.Rakesh Agrawal	12	3.24	18	994	154
5.Serge Abiteboul	11	4.08	16	1111	172
6.Jennifer Widom	11	3.23	14	709	136
7.Jim Gray	10	3.93	16	1571	118
8.Michael J. Carey	10	3.79	16	1090	151
9.Won Kim	10	3.00	19	1071	143
10.David Maier	10	2.93	14	1090	158
11.Hector Garcia-Molina	9	5.30	17	1041	314
12.Jeffrey F. Naughton	9	3.85	15	653	123
13.Yehoshua Sagiv	9	3.76	17	1020	121
14.Christos Faloutsos	9	3.68	16	742	175
15.Catriel Beeri	9	3.59	18	1024	93
16.Philip A. Bernstein	9	3.49	20	1359	124
17.Umeshwar Dayal	9	3.39	18	913	130
18.Hamid Pirahesh	9	3.34	14	622	67
19.H.V. Jagadish	9	2.88	12	503	151
20.Raghu Ramakrishnan	8	5.05	14	818	147

Πίνακας 6.3. Κατάταξη συγγραφέων με βάση το *contemporary h-index*.

h-index και αυτούς με απότομη ανοδική κλήση στις καμπύλες των γραφημάτων. Εδώ πρέπει να σημειώσουμε ότι το σύνολο δεδομένων μας μάλλον είναι ελλιπές για τα έτη 1999-2000, και έτσι σε όλους τους ερευνητές παρουσιάζεται μια κλήση προς τα κάτω σε αυτά τα δυο έτη. Παρατηρώντας λοιπόν τα Σχήματα ας έχουμε στο μυαλό μας ότι τα στοιχεία για τα δυο τελευταία έτη είναι ενδεικτικά αλλά όχι πραγματικά.

Συγκρίνοντας λοιπόν τα Σχήματα 6.1(α) και 6.1(β) βλέπουμε ότι οι 2 ερευνητές Jim Gray και Χρήστος Φαλούτσος έχουν τώρα τον ίδιο *h-index*. Όμως ο Χρήστος Φαλούτσος έχει μεγαλύτερη ανοδική κλήση από τον Jim Gray δεδομένου ότι άρχισε να δέχεται αναφορές το 1984, ενώ ο Jim Gray το 1976. Επίσης, η καμπύλη του *trend h-index* (h_t) του Χρήστου Φαλούτσου παραμένει σταθερά επάνω από την αντίστοιχη του *h-index* (h). Αυτό σημαίνει ότι ο Χρήστος Φαλούτσος συνεχώς δέχεται νέες αναφορές, άρα περιμένουμε το *h-index* να ανέβει περισσότερο από του Jim Gray. Τέλος, το *contemporary h-index* (h_c) του Jim Gray είναι από το 1985 σταθερά κάτω από το h και με το πέρασμα των χρόνων αποκλίνει. Αυτό μας δείχνει ότι από το 1985 και μετά δεν έχει παρουσιάσει υπερβολικά καλές εργασίες (σε σχέση πάντα με τις προηγούμενες του ιδίου), οπότε

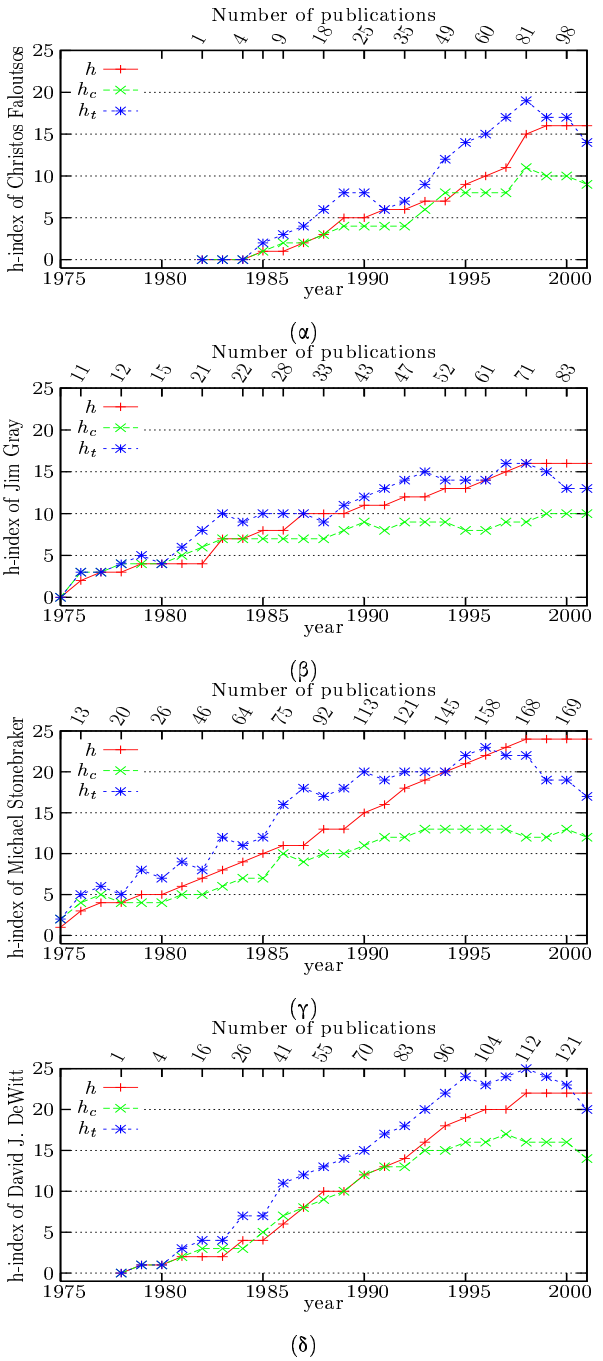
Όνομα	h_t	a_t	h	$N_{c,tot}$	N_p
1.David J. DeWitt	20	2.73	22	1896	150
2.Michael Stonebraker	17	3.61	24	2180	193
3.Jeffrey D. Ullman	17	3.45	23	1783	227
4.Rakesh Agrawal	17	3.06	18	994	154
5.Jennifer Widom	16	2.81	14	709	136
6.Serge Abiteboul	14	4.07	16	1111	172
7.Hector Garcia-Molina	14	4.03	17	1041	314
8.Christos Faloutsos	14	3.15	16	742	175
9.Jim Gray	13	4.46	16	1571	118
10.Jeffrey F. Naughton	13	3.36	15	653	123
11.Won Kim	13	3.23	19	1071	143
12.Michael J. Carey	12	4.79	16	1090	151
13.Yehoshua Sagiv	12	3.96	17	1020	121
14.Umeshwar Dayal	12	3.41	18	913	130
15.Catriel Beeri	12	3.12	18	1024	93
16.Raghu Ramakrishnan	11	4.41	14	818	147
17.Philip A. Bernstein	11	4.03	20	1359	124
18.David Maier	11	3.94	14	1090	158
19.Hamid Pirahesh	11	3.87	14	622	67
20.H.V. Jagadish	11	3.58	12	503	151

Πίνακας 6.4. Κατάταξη συγγραφέων με βάση το *trend h-index*.

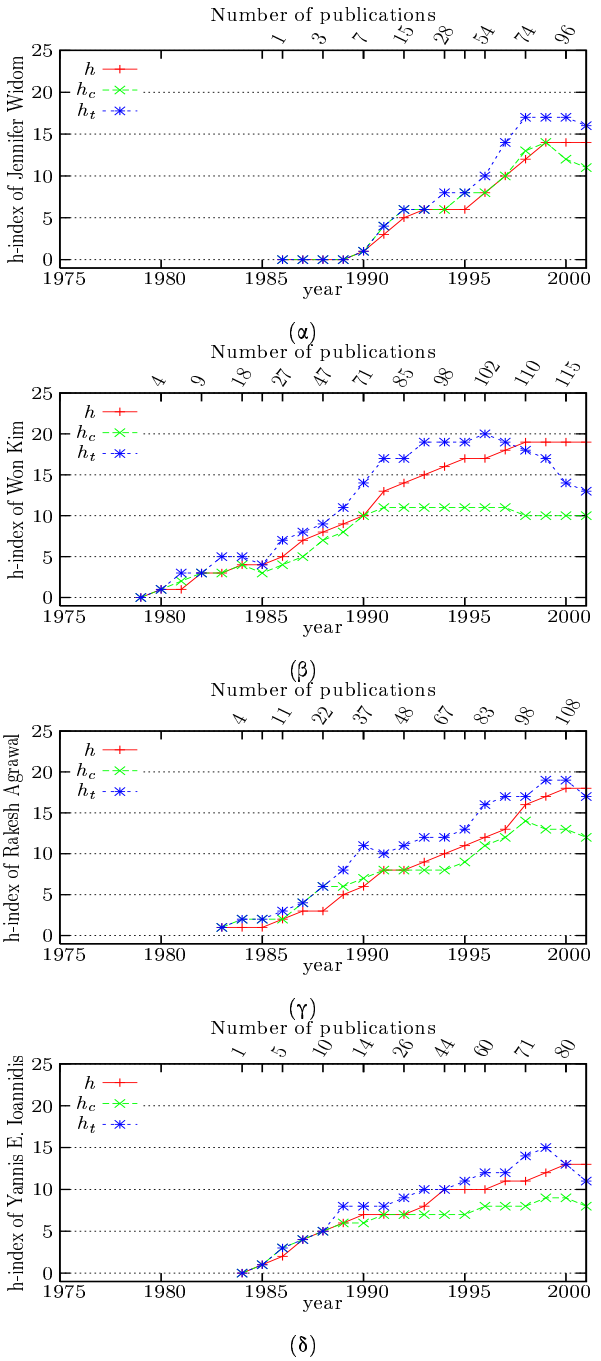
και από εκεί και μετά αρχίζει φθίνουσα πορεία.

Τα Σχήματα 6.1(γ) και 6.1(δ) αντιστοιχούν στους Michael Stonebraker και David J. DeWitt. Και οι δυο αυτοί ερευνητές βρίσκονται στην κορυφή της λίστας μας. Παρατηρούμε ότι το *contemporary h-index* του David J. DeWitt βρίσκεται πολύ κοντά στο *h-index*, πράγμα που σημαίνει ότι σταθερά δημοσιεύει καλές εργασίες. Αντιθέτως, του Michael Stonebraker, από το 1985 και μετά έχει αρχίσει να παρουσιάζει ουσιαστική απόκλιση Αυτό μας προετοιμάζει να καταλάβουμε ότι μετά από λίγα χρόνια θα πέσει και το *trend h-index* του Michael Stonebraker, όπως και φαίνεται στο ίδιο γράφημα. Έτσι, ενώ στην κατάταξη *h-index* ο Michael Stonebraker είναι υψηλότερα από τον David J. DeWitt, στις 2 άλλες παραλλαγές ο David J. DeWitt προηγείται. Αυτό σημαίνει, ότι αν η παραγωγικότητα των δυο ερευνητών παραμείνει στους ίδιους ρυθμούς, τότε σύντομα ο δεύτερος θα ξεπεράσει τον πρώτο και στον *h-index*.

Στο Σχήμα 6.2(α) βλέπουμε την πορεία των δεικτών για την Jennifer Widom. Ενώ η Jennifer Widom δεν εμφανίζεται καν στους πρώτους 20 με βάση το *h-index*, με τους δείκτες *contemporary h-index* και *trend h-index* ανεβαίνει στην $6^η$ και $7^η$ θέση, αντίστοιχα. Η Windom είναι η μοναδική περίπτωση από τη λίστα μας



Σχήμα 6.1. *h-index* ερευτητών της περιοχής των Βάσεων Δεδομένων.



Σχήμα 6.2. *h-index* ερευνητών της περιοχής των Βάσεων Δεδομένων (συνέχεια).

που παρουσιάζει τόση μεγάλη διαφορά στους χρονικούς δείκτες με το βασικό *h-index*. Απ'ότι βλέπουμε όμως και στο σχήμα, η διαφορά αυτή είναι αιτιολογημένη, εφ'όσον η ταχύτητα ανόδου του βασικού *h-index* είναι μεγάλη. Επίσης, είναι η μόνη ερευνήτρια για την οποία το *contemporary h-index* είναι σταθερά δίπλα στο *h-index* και δεν πέφτει χαμηλότερα. Τέλος, αν και ο *trend h-index* σε όλους τους ερευνητές που παρουσιάζουμε πέφτει κάτω από τον *h-index* για το έτος 2000, στην περίπτωση της Windom παραμένει ψηλά. Αυτό σημαίνει ότι υπάρχει ιδιαίτερη δυναμική της Jennifer Widom και στα επόμενα έτη αναμένουμε να κάνει την παρουσία της ψηλά στην κατάταξη του *h-index*.

Στην περίπτωση του Σχήματος 6.2(δ) για τον Γιάννη Ιωαννίδη βλέπουμε μια ανοδική τάση περίπου αντίστοιχη με αυτήν του Jim Gray. Το *trend h-index* παραμένει σταθερά επάνω από τον *h-index*, πράγμα που σημαίνει ότι υπάρχει ιδιαίτερη δυναμική. Από την άλλη, το *contemporary h-index* από το 1993 και μετά παρουσιάζει μια μικρή απόκλιση από τον *h-index*, η οποία είναι ακριβώς αντίστοιχη με αυτήν του Jim Gray από το 1985. Με βάση λοιπόν τα δεδομένα στοιχεία, ο Γιάννης Ιωαννίδης ακολουθεί την πορεία του Jim Gray με χρονική απόκλιση περίπου 10 ετών. Μάλιστα, παρατηρώντας καλύτερα τα σχήματα, μπορούμε επιπλέον να παρατηρήσουμε ότι σε καμιά χρονική στιγμή οι δείκτες για τον Γιάννη Ιωαννίδη δεν παρουσιάζουν πτώση, ενώ ο Jim Gray είχε μικρές πτώσεις στο δείκτη *trend h-index* για τα έτη 1980, 1984 και 1988.

Στο Σχήμα 6.2(β) ο Won Kim παρουσιάζει αντίστοιχη πορεία με αυτήν του Stonebraker από την άποψη ότι υπάρχει μεγάλη ανοδική καμπύλη για το *trend h-index*, ενώ το *contemporary h-index* παραμένει χαμηλά μετά το 1990. Τελικώς είναι εμφανές ότι και το *trend h-index* θα ακολουθήσει φθίνουσα πορεία. Συνεπώς αναμένουμε ότι το *h-index* δεν θα παρουσιάσει μεγάλη άνοδο.

6.4.2 Αξιολόγηση Αποτελεσμάτων με Βάση Βραβευμένους Ερευνητές

Στο Κεφάλαιο 3 χρησιμοποιήσαμε τη λίστα των βραβευμένων συγγραφέων με το βραβείο 'SIGMOD E.F.Codd Innovations Award' για να αξιολογήσουμε την κατάταξη που προκύπτει από τους αλγορίθμους μας. Το ίδιο μπορούμε να κάνουμε και στην προκειμένη περίπτωση, ώστε να διαπιστώσουμε αν η βαθμολογία με βάση το *h-index* αντικατροπτρίζεται και σε βραβεία που έχουν απονεμηθεί για την περιοχή των Βάσεων Δεδομένων. Στον Πίνακα 6.5 παρουσιάζουμε τη λίστα με τους ερευνητές που έχουν πάρει το βραβείο 'SIGMOD E.F.Codd Innovations Award'. Στην πρώτη τριάδα στηλών (h, h_c, h_t) βλέπουμε τη θέση που κατέχει αυτή τη

Όνομα	h	h_c	h_t	h	h_c	h_t	*Έτος*	h	h_c	h_t
Michael Stonebraker	1	3	2	3	2	1	1992	1	2	1
Jim Gray	12	7	9	12	11	10	1993	15	11	8
Philip A. Bernstein	4	16	17	2	6	4	1994	2	9	5
David J. DeWitt	3	1	1	3	1	1	1995	2	1	1
C. Mohan	28	37	31	44	36	35	1996	49	23	19
David Maier	20	10	18	11	9	15	1997	15	10	17
Serge Abiteboul	13	5	6	17	4	11	1998	16	6	11
Hector Garcia-Molina	9	11	7	10	8	4	1999	14	7	5
Rakesh Agrawal	7	4	4	9	4	4	2000	7	4	4
Rudolf Bayer	145	196	183	142	218	222	2001	145	196	183
Patricia G. Selinger	143	144	119	143	144	119	2002	143	133	132
Donald D. Chamberlin	44	87	69	44	72	86	2003	44	68	62
Ronald Fagin	11	39	32	11	28	27	2004	11	24	21
Χαμηλότερη Θέση	145	196	183	143	218	222		145	196	183
Άθροισμα Θέσεων	440	560	498	451	543	539		464	494	469

Πίνακας 6.5. Θέσεις των βραβευμένων ερευνητών.

στιγμή ο κάθε βραβευμένος ερευνητής στην κατάταξη με κριτήριο τον αντίστοιχο δείκτη. Η επόμενη ομάδα στηλών αντιστοιχεί στους δείκτες όπως ήταν διαμορφωμένοι την εποχή κοντά στη βράβευσή τους. Δηλαδή, η μεσαία τριάδα δεικτών αντιστοιχεί στους δείκτες όπως ήταν διαμορφωμένοι την προηγούμενη χρονιά της βράβευσης, ενώ η τελευταία τριάδα στηλών δείχνει τη θέση που κατείχε ο αντίστοιχος ερευνητής τη χρονιά κατά την οποία βραβεύθηκε. Τέλος, η στήλη *Έτος* δείχνει το έτος της βράβευσης.

Εδώ υπενθυμίζεται ότι η βάση δεδομένων μας δεν έχει αρκετά στοιχεία για την περίοδο μετά το 2000, και άρα προφανώς οι κατατάξεις για τα έτη 2000 και μετά είναι περίπου ισοδύναμες. Για τις βραβεύσεις όμως που έγιναν πριν από το 2000 μπορούμε να κάνουμε ενδιαφέρουσες παρατηρήσεις:

- **C. Mohan**: Αν και στην κατάταξη αυτή τη στιγμή εμφανίζεται σχετικά χαμηλά με βάση τα *trend h-index* και *contemporary h-index*, εντούτοις κατά το έτος 1996 βρισκόταν σε πολύ υψηλότερη θέση ιδίως με βάση το δείκτη *trend h-index*. Αυτό απεικονίσθηκε στον *h-index* αρκετά αργότερα, και ενώ κατά το 1996 βρισκόταν στην $49^{η}$ θέση, πλέον, με βάση τον *h-index* βρίσκεται στην $28^{η}$.

- Αντίστοιχες περιπτώσεις με εμφανή τη διαφορά στην κατάταξη είναι αυτές των **Hector Garcia-Molina** και **Philip A. Bernstein**.

- **Serge Abiteboul**: Το έτος της βράβευσης ο *trend h-index* βρίσκεται σχετικά χαμηλά (σε σχέση με τον *contemporary h-index*). Με βάση όμως τον *contemporary h-index* βρίσκονταν σε υψηλή θέση. Αυτό μας δείχνει ότι κατά την εποχή της βράβευσης είχε παρουσιάσει ενδιαφέρουσες εργασίες, και βραβεύθηκε πριν ακόμη αυτό απεικονισθεί στον *trend h-index* ή στον *h-index*. Άρα σε κάποιες περιπτώσεις ο *contemporary h-index* μας δείχνει πληροφορία που δεν μπορεί να απεικονισθεί στους υπόλοιπους δείκτες - την οποία πληροφορία προφανώς έλαβε υπ'όψη της η επιτροπή βράβευσης.

- Για τις περιπτώσεις **Michael Stonebraker** και **David J. DeWitt** βλέπουμε ότι υπάρχει σταθερότητα στην κορυφή ενώ για τις περιπτώσεις μετά το 2000 δεν μπορούμε να βγάλουμε αντίστοιχα συμπεράσματα λόγω έλλειψης δεδομένων.

6.5 Πειραματικά Αποτελέσματα για Συνέδρια

Το σύνολο δεδομένων που έχουμε περιγράφεται στα προηγούμενα κεφάλαια. Από το σύνολο αυτό ξεχωρίζουμε μόνο τα συνέδρια της περιοχής των Βάσεων Δεδομένων σύμφωνα με το [25], και κάνουμε αξιολόγηση μόνο των συγκεκριμένων συνεδρίων. Σε πρώτη φάση θα πειραματισθούμε με του ίδιους δείκτες που ορίσαμε και για τους συγγραφείς. Στον Πίνακα 6.6 παρουσιάζουμε τα 15 πρώτα συνέδρια με βάση το *h-index*. Η κατάταξη αλλάζει δραματικά στον Πίνακα 6.7, αλλά οφείλεται κυρίως στη μη πληρότητα των δεδομένων μας για μερικά συνέδρια.

Όνομα	h	a	$N_{c,tot}$	N_p
1.sigmod	45	6.05	12261	2059
2.vldb	37	7.10	9729	2192
3.pods	26	5.74	3883	776
4.icde	22	6.83	3307	1970
5.er	16	5.80	1486	1338
6.edbt	13	3.89	658	434
7.eds	12	3.65	527	101
8.adbt	12	2.86	412	42
9.icdt	11	4.79	580	313
10.oodbs	11	3.96	480	122

Πίνακας 6.6. Κατάταξη συνεδρίων με βάση το *h-index*.

Όνομα	h_n	h	a	$N_{c,tot}$	N_p
1.adbt	0.28	12	2.86	412	42
2.dpds	0.17	7	2.97	146	39
3.eds	0.11	12	3.65	527	101
4.icod	0.11	6	3	108	52
5.jcdkb	0.11	8	3.32	213	70
6.ddb	0.09	4	6.87	110	44
7.oodbs	0.09	11	3.96	480	122
8.tdb	0.08	3	6.44	58	36
9.berkeley	0.07	10	3.52	352	142

Πίνακας 6.7. Κατάταξη συνεδρίων με βάση το *normalized h-index*.

Όνομα	h_c	a_c	h	$N_{c,tot}$	N_p
1.sigmod	21	9.49	45	12261	2059
2.vldb	17	11.34	37	9729	2192
3.pods	12	9.73	26	3883	776
4.icde	11	11.88	22	3307	1970
5.icdt	8	5.04	11	580	313
6.edbt	7	6.16	13	658	434
7.oodbs	6	3.63	11	480	122
8.er	5	16.21	16	1486	1338
9.kdd	5	6.89	6	243	1074
10.dood	5	6.57	8	440	171

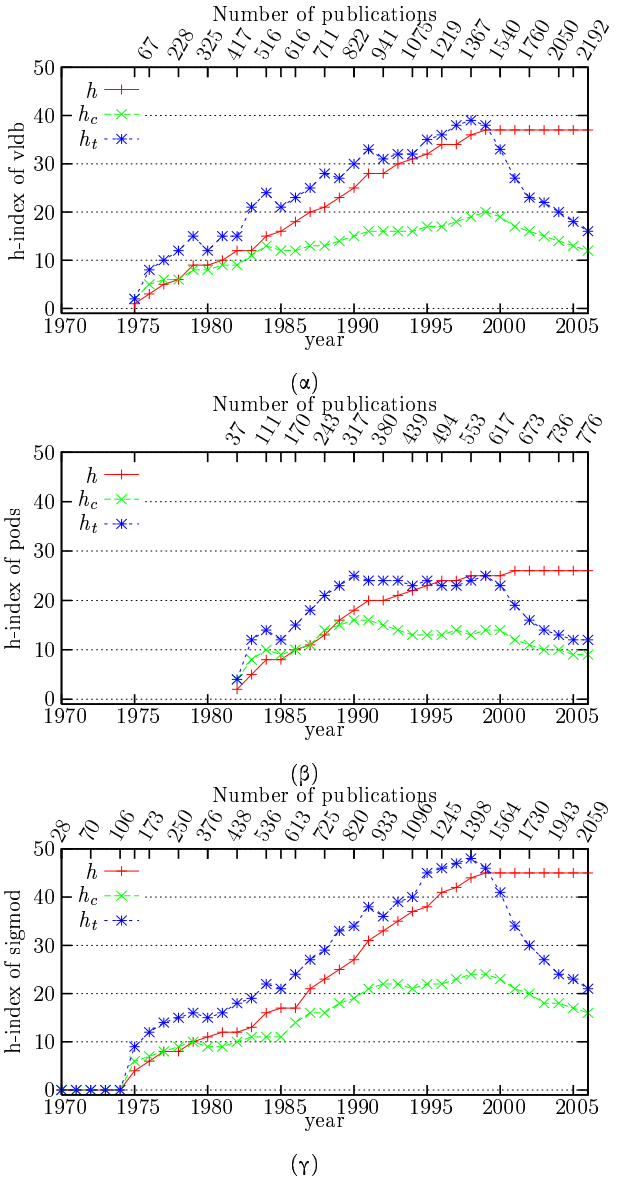
Πίνακας 6.8. Κατάταξη συνεδρίων με βάση το *contemporary h-index*.

Όνομα	h_t	a_t	h	$N_{c,tot}$	N_p
1.sigmod	34	6.67	45	12261	2059
2.vldb	27	8.00	37	9729	2192
3.pods	19	6.53	26	3883	776
4.icde	16	9.52	22	3307	1970
5.icdt	12	3.67	11	580	313
6.edbt	9	6.02	13	658	434
7.er	8	10.35	16	1486	1338
8.dood	8	4.43	8	440	171
9.kdd	7	6.42	6	243	1074
10.dbpl	7	5.11	8	410	228

Πίνακας 6.9. Κατάταξη συνεδρίων με βάση το *trend h-index*.

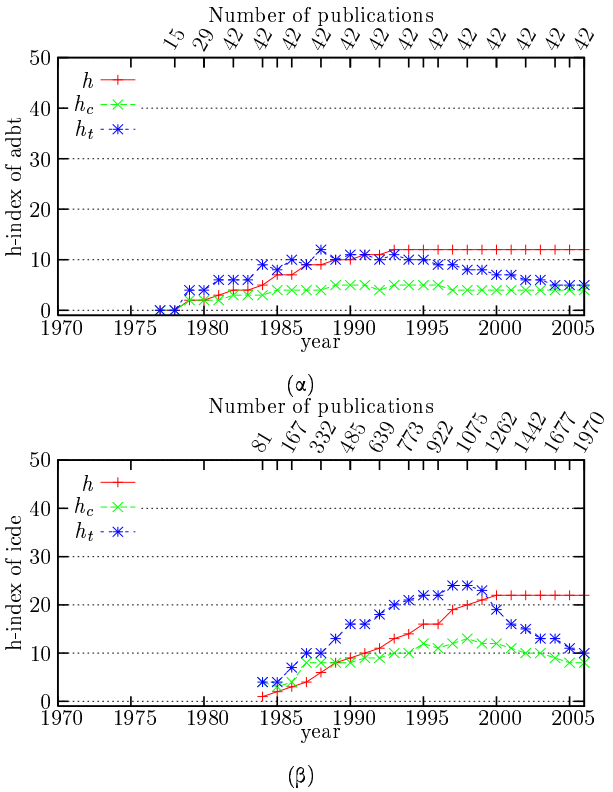
Δεδομένου ότι στα συνέδρια είναι περισσότερο σταθερή η ποιότητα, παρατηρούμε ότι στους Πίνακες 6.8 και 6.9 δεν παρουσιάζονται αλλαγές (ιδίως στις πρώτες θέσεις) σε σύγκριση με τον *h-index*.

Στα Σχήματα 6.3 και 6.4 βλέπουμε την πορεία κάποιων συνεδρίων, με τον



Σχήμα 6.3. *h-index* συνεδρίων της περιοχής των Βάσεων Δεδομένων.

ίδιο τρόπο όπως και στους συγγραφείς. Υπ'όψη ότι, αν και τα σχήματα δείχνουν το *h-index* ανά έτος, αυτό είναι το *h-index* του συνεδρίου που θα υπολογίζαμε αν βρισκόμασταν στο συγκεκριμένο έτος, και δεν διαχωρίζουμε τις διοργανώσεις των συνεδρίων (πχ. VLDB'99, VLDB'98 κ.ο.κ.). Τέλος, υπενθυμίζεται και πάλι ότι τα δεδομένα μας είναι ελλιπή για τα έτη 1999 και μετά. Άρα σε όλα τα



Σχήμα 6.4. *h-index* συνεδρίων της περιοχής των Βάσεων Δεδομένων (συνέχεια).

σχήματα παρουσιάζεται μια σταθεροποίηση του *h-index* και μια πτώση των *trend h-index* και *contemporary h-index*, όπως ήταν αναμενόμενο.

Βλέποντας, λοιπόν στο Σχήμα 6.3(γ) το SIGMOD που κατέχει την πρώτη θέση σύμφωνα με τους πίνακες που παρουσιάσαμε, παρατηρούμε την έντονη ανοδική πορεία, καθώς και το ότι το *trend h-index* παραμένει με διαφορά υψηλότερα από το *h-index* (μέχρι το 1999). Από την άλλη πλευρά, για το PODS στο Σχήμα 6.3(β) βλέπουμε ότι το *trend h-index* αρχίζει να πέφτει από το 1993 και μετά (το οποίο δεν οφείλεται στα ελλιπή δεδομένων) και συνεπώς και η πορεία του *h-index* είναι σταθερά ελαφρώς ανοδική. Το ICDE (δες Σχήμα 6.4(β)), αν και σχετικά νεότερο συνέδριο σε σχέση με τα υπόλοιπα, παρουσιάζει έντονη ανοδική πορεία σύμφωνα με το *trend h-index*. Τέλος, παρουσιάζουμε το ADBT (δες Σχήμα 6.4(α)) για το οποίο δεν υπάρχουν αρκετά δεδομένα. Στον επάνω άξονα $x2$ βλέπουμε ότι το πλήθος των δημοσιεύσεων σταματά να αυξάνεται το 1982, και άρα δεν είναι συγκρίσιμο με τα υπόλοιπα συνέδρια. Για το λόγο αυτό, πλεο-

νεκτεί στην κατάταξη *normalized h-index* του Πίνακα 6.7 έχει πλεονέκτημα και κατατάσσεται πρώτο.

Η άλλη διάσταση της αξιολόγησης των συνεδρίων, είναι αυτή που διατυπώνεται στους Ορισμούς 6.2 και 6.3. Έτσι θα αξιολογήσουμε ξεχωριστά και ανεξάρτητα πχ. το VLDB'95 από το VLDB'94. Προφανώς σε αυτήν την περίπτωση δεν έχει νόημα να προσθέσουμε και δεύτερη χρονική διάσταση (με τους δείκτες *contemporary h-index* και *trend h-index*) εφ'όσον από τη μια το *contemporary h-index* δεν θα έχει καμιά διαφορά - όλες οι εργασίες έχουν δημοσιευθεί την ίδια χρονιά - και από την άλλη δεν μας ενδιαφέρει ένα συνέδριο που διοργανώθηκε το 1980 αν συνεχίζει ακόμη και *φέτος* να δέχεται αναφορές.

Ενδεικτικά, παρουσιάζουμε τους Πίνακες 6.10 και 6.11 οι οποίοι περιέχουν την κατάταξη των συνεδρίων για τα έτη 1995 και 1990 αντίστοιχα. Στο (α) τμήμα των πινάκων αυτών παρουσιάζεται η κατάταξη με βάση το *yearly h-index*

Όνομα	h_{1995}	a	h^n_{1995}	$N_{c,1995}$	$N_{p,1995}$
1.vldb	11	3.57	0.15	432	72
2.sigmod	9	4.62	0.10	375	85
3.icde	6	6.63	0.08	239	68
4.pods	6	4.16	0.20	150	29
5.ssd	5	2.08	0.20	52	24
6.kdd	4	3.81	0.07	61	56
7.cikm	3	6.22	0.05	56	55
8.dood	3	5.88	0.06	53	46
9.icdt	3	3.66	0.08	33	34
10.er	3	3.33	0.06	30	47

(α)

Όνομα	h^n_{1995}	h_{1995}	$N_{p,1995}$
1.ssd	0.20	5	24
2.pods	0.20	6	29
3.cdb	0.2	2	10
4.vldb	0.15	11	72
5.coopis	0.14	3	21
6.artdb	0.11	2	17
7.sdb	0.11	1	9
8.sigmod	0.10	9	85
9.ride	0.10	2	19
10.tdb	0.1	2	20

(β)

Πίνακας 6.10. Κατάταξη των συνεδρίων του 1995.

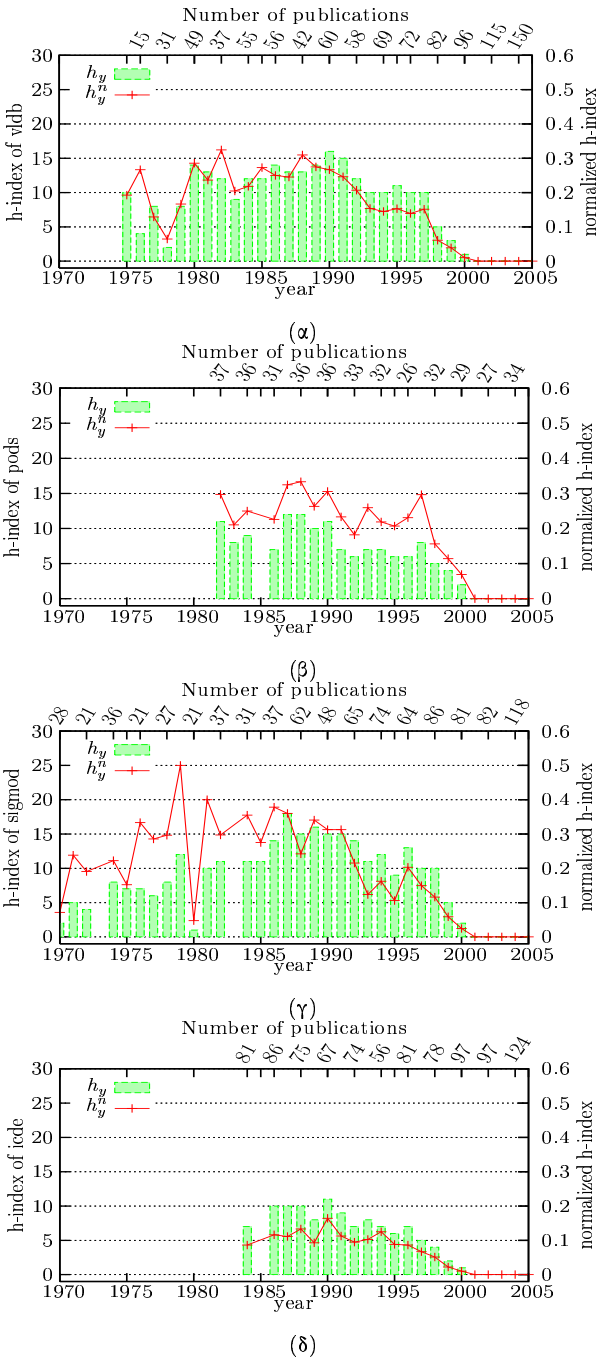
Όνομα	h_{1990}	a	h^n_{1990}	$N_{c,1990}$	$N_{p,1990}$
1.vldb	16	2.57	0.26	659	60
2.sigmod	15	3.44	0.31	776	48
3.icde	11	2.76	0.16	335	67
4.pods	11	2.40	0.30	291	36
5.edbt	7	2.83	0.21	139	32
6.icdt	5	4.32	0.14	108	34
7.dpds	4	3.75	0.22	60	18
8.er	3	4.66	0.08	42	35
9.ds	3	4.11	0.12	37	24
10.ssdbm	3	3	0.16	27	18

(α)

Όνομα	h^n_{1995}	h_{1995}	$N_{p,1995}$
1.sigmod	0.31	15	48
2.pods	0.30	11	36
3.vldb	0.26	16	60
4.dpds	0.22	4	18
5.edbt	0.21	7	32
6.ssdbm	0.16	3	18
7.icde	0.16	11	67
8.icdt	0.14	5	34
9.ds	0.12	3	24
10.ewdw	0.10	3	29

(β)

Πίνακας 6.11. Κατάταξη των συνεδρίων του 1990.



Σχήμα 6.5. *yearly h-index* και *normalized yearly h-index* συνεδρίων ΒΔ.

(y_y). Ο συντελεστής a είναι το δεύτερο κριτήριο ταξινόμησης, και ενδεικτικά συμπεριλαμβάνουμε τις στήλες: h_y^n που είναι ο δείκτης *h-index* διαιρεμένος με το πλήθος των δημοσιεύσεων $N_{p,y}$, και την $N_{c,1995}$ που είναι το πλήθος των αναφορών που έχουν δεχτεί οι δημοσιεύσεις του αντίστοιχου έτους. Στο (β) τμήμα των πινάκων η κατάταξη είναι με βάση το *normalized h-index*. Σημειώνεται ότι αν και φαίνεται από τους πίνακες να έχουμε ισοβαθμίες στο δείκτη h_y^n, τα πραγματικά νούμερα (πχ: $5/24 = 0.20833$, $6/29 = 0.206897$) σχεδόν αποκλείουν αυτό το ενδεχόμενο. Αυτό που παρατηρούμε είναι ότι δεν έχουμε σημαντικές αποκλίσεις στην κατάταξη, ούτε στα δυο ενδεικτικά έτη, αλλά ούτε και με τη μέθοδο *normalized h-index*. Προφανώς όμως, με τη μέθοδο *normalized h-index* τα συνέδρια που κάνουν μικρό πλήθος δημοσιεύσεων έχουν ένα μικρό πλεονέκτημα. Για παράδειγμα, το συνέδριο VLDB που έχει μεγάλο πλήθος δημοσιεύσεων, αν και σχεδόν σταθερά πρώτο με την κατάταξη *yearly h-index*, πολύ δύσκολα θα βρεθεί στην πρώτη θέση με την κατάταξη *normalized yearly h-index*.

Στο Σχήμα 6.5 παρουσιάζουμε τις τιμές των *yearly h-index* (h_y) και *normalized yearly h-index* (h_y^n) για τα συνέδρια VLDB, PODS, SIGMOD και ICDE. Οι τιμές του h_y απεικονίζονται με μπάρες, διότι είναι ανεξάρτητες μεταξύ τους εφ'όσον το καθένα έχει διαφορετικό άνω φράγμα. Το άνω φράγμα προφανώς ορίζεται από το πλήθος των δημοσιεύσεων, το οποίο και απεικονίζεται στον επάνω άξονα $x2$. Από την άλλη, το h_y^n εφ'όσον είναι κανονικοποιημένο, είναι συγκρίσιμη ποσότητα για δυο διαφορετικές διοργανώσεις ενός συνεδρίου, οπότε και απεικονίζεται με την (κόκκινη) καμπύλη και τα σημεία τιμών με το σταυρό. Οι τιμές του δείκτη h_y^n φαίνονται στο δεξιό άξονα $y2$. Θα μπορούσαμε να πούμε ότι η σχέση των δυο αξόνων $y1$ και $y2$ είναι τυχαία ορισμένη (με βάση απλά την εμφάνιση), συνεπώς δεν μπορούμε να συγκρίνουμε τις δυο τιμές μεταξύ τους (h_y^n και h_y). Το μόνο που μπορούμε να ισχυρισθούμε είναι ότι περίπου οι δυο καμπύλες ακολουθούν η μια την άλλη. Αυτό έρχεται σε συμφωνία με τα συμπεράσματα από τους Πίνακες 6.10 και 6.11.

6.6 Πειραματικά Αποτελέσματα για Περιοδικά

Στην περίπτωση των περιοδικών, μπορούμε να χρησιμοποιήσουμε τη βασική μορφή του *h-index*, καθώς και τις παραλλαγές *normalized h-index*, *contemporary h-index* και *trend h-index* που ορίσαμε για τους συγγραφείς καθώς και για τα συνέδρια. Και εδώ, όπως και στην περίπτωση των συνεδρίων, έχει πρωτογενή αξία ο δείκτης *normalized h-index* σε αντίθεση με την περίπτωση των συγγραφέων-ερευνητών.

Όνομα	h	a	$N_{c,tot}$	N_p
1.tods	49	3.88	9329	598
2.tkde	18	4.69	1520	1388
3.is	16	4.71	1208	934
4.sigmod	15	5.07	1142	1349
5.tois	13	4.37	740	378
6.debu	11	7.13	863	877
7.vldb	9	5.03	408	281
8.ipl	8	6.06	388	4939
9.dke	6	8.77	316	773
10.dpd	6	5.25	189	238

Πίνακας 6.12. Κατάταξη περιοδικών με βάση το *h-index*.

Όνομα	h_n	h	a	$N_{c,tot}$	N_p
1.tods	0.08	49	3.88	9329	598
2.tois	0.03	13	4.37	740	378
3.vldb	0.03	9	5.03	408	281
4.dpd	0.02	6	5.25	189	238
5.jiis	0.01	6	4.33	156	318
6.datamine	0.01	3	5.11	46	162
7.is	0.01	16	4.71	1208	934
8.ijcis	0.01	4	3.12	50	255
9.tkde	0.01	18	4.69	1520	1388
10.debu	0.01	11	7.13	863	877

Πίνακας 6.13. Κατάταξη περιοδικών με βάση το *normalized h-index*.

Όνομα	h_c	a_c	h	$N_{c,tot}$	N_p
1.tods	8	6.25	49	9329	598
2.tkde	10	6.40	18	1520	1388
3.sigmod	9	6.17	15	1142	1349
4.debu	6	9.21	11	863	877
5.vldb	6	6.47	9	408	281
6.tois	6	6.09	13	740	378
7.is	5	12.77	16	1208	934
8.dpd	5	4.19	6	189	238
9.jiis	5	3.79	6	156	318
10.dke	4	7.70	6	316	773

Πίνακας 6.14. Κατάταξη περιοδικών με βάση το *contemporary h-index*.

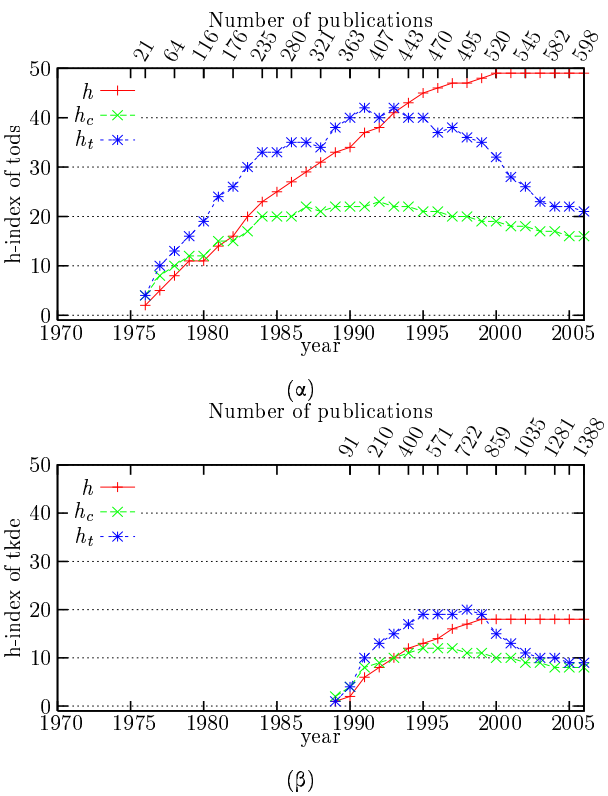
Στις επόμενες σελίδες αυτού του κεφαλαίου, περιλαμβάνονται τα αποτελέσματα της κατάταξης με τους προαναφερθέντες δείκτες για τα Περιοδικά, για τα οποία μπορούμε να πούμε ότι ανήκουν στην περιοχή των Βάσεων Δεδομένων

Όνομα	h_t	a_t	h	$N_{c,tot}$	N_p
1.tods	28	4.93	49	9329	598
2.tkde	13	6.64	18	1520	1388
3.sigmod	12	5.85	15	1142	1349
4.vldb	10	3.75	9	408	281
5.is	9	7.11	16	1208	934
6.debu	9	6.98	11	863	877
7.tois	9	4.83	13	740	378
8.dpd	6	4.88	6	189	238
9.jiis	6	4.75	6	156	318
10.dke	5	8.18	6	316	773

Πίνακας 6.15. Κατάταξη περιοδικών με βάση το *trend h-index*.

σύμφωνα με το [25].

Στους Πίνακες 6.12, 6.13, 6.14 και 6.15 παρουσιάζουμε τα πρώτα 10 περιοδικά στην κατάταξη με βάση τους δείκτες που αναφέραμε. Είναι προφανές ότι τα



Σχήμα 6.6. *h-index* περιοδικών της περιοχής των Βάσεων Δεδομένων.

TODS, TKDE και SIGMOD RECORD είναι σταθερά υψηλά στην κατάταξη με όλους τους δείκτες. Το IS (Information Systems) φαίνεται να χάνει έδαφος από την κορυφή με βάση τον *contemporary h-index* (δες Πίνακα 6.14). Επίσης, παρατηρούμε ότι VLDB JOURNAL και SIGMOD RECORD είναι ανερχόμενα στην κατάταξη *trend h-index* (δες Πίνακα 6.15).

Στα Σχήματα 6.6 και 6.7 παρουσιάζουμε τον υπολογισμό ανά έτος των αντίστοιχων δεικτών. Προφανώς μετά το 2000 τα δεδομένα μας είναι ελλειπή και όλοι οι δείκτες παρουσιάζουν πτώση. Στην περίπτωση όμως του TODS (δες Σχήμα 6.6(α)) και IS (δες Σχήμα 6.7(γ)) είναι εμφανής η πτώση του *trend h-index* μετά το 1993. Επίσης, μπορούμε να παρατηρήσουμε για το SIGMOD (δες Σχήμα 6.7(α)) ότι αν και εκδίδεται από το 1970, μόνο μετά από το 1980 αρχίζουν να αυξάνονται οι δείκτες.

Τέλος, στον Πίνακα 6.16 παρουσιάζουμε την κατάταξη με βάση τους δείκτες *yearly h-index* και *normalized yearly h-index* για το έτος 1995. Τα αποτελέσματα είναι αναμενόμενα και δεν διαφέρουν ουσιαστικά από την συνολική κατάταξη των περιοδικών. Στο Σχήμα 6.8 παρουσιάζουμε τα γραφήματα για τους δείκτες *yearly h-index* και *normalized yearly h-index* των τεσσάρων πρώτων περιοδικών (TODS, TKDE, SIGMOD, VLDB). Αυτό που παρατηρούμε είναι ότι υπάρχει μια συνέχεια και συνέπεια στην πορεία των περιοδικών. Ενδιαφέρον όμως παρουσιάζει η εξ' αρχής υψηλές τιμές του VLDB (δες Σχήμα 6.8(δ)).

Όνομα	h_{1995}	a	h^n_{1995}	$N_{c,1995}$	$N_{p,1995}$
1.tkde	11	4.28	0.28	519	38
2.tods	8	3.14	0.4	201	20
3.sigmod	7	2.61	0.15	128	46
4.tois	5	2.52	0.33	63	15
5.is	4	3.18	0.07	51	53
6.dke	3	4.22	0.15	38	20
7.debu	3	3.66	0.07	33	41
8.ipl	3	3.22	0.01	29	193

(α)

Όνομα	h^n_{1995}	h_{1995}	$N_{p,1995}$
1.tods	0.4	8	20
2.tois	0.33	5	15
3.tkde	0.28	11	38
4.sigmod	0.15	7	46
5.dke	0.15	3	20
6.is	0.07	4	53
7.debu	0.07	3	41
8.ipl	0.01	3	193

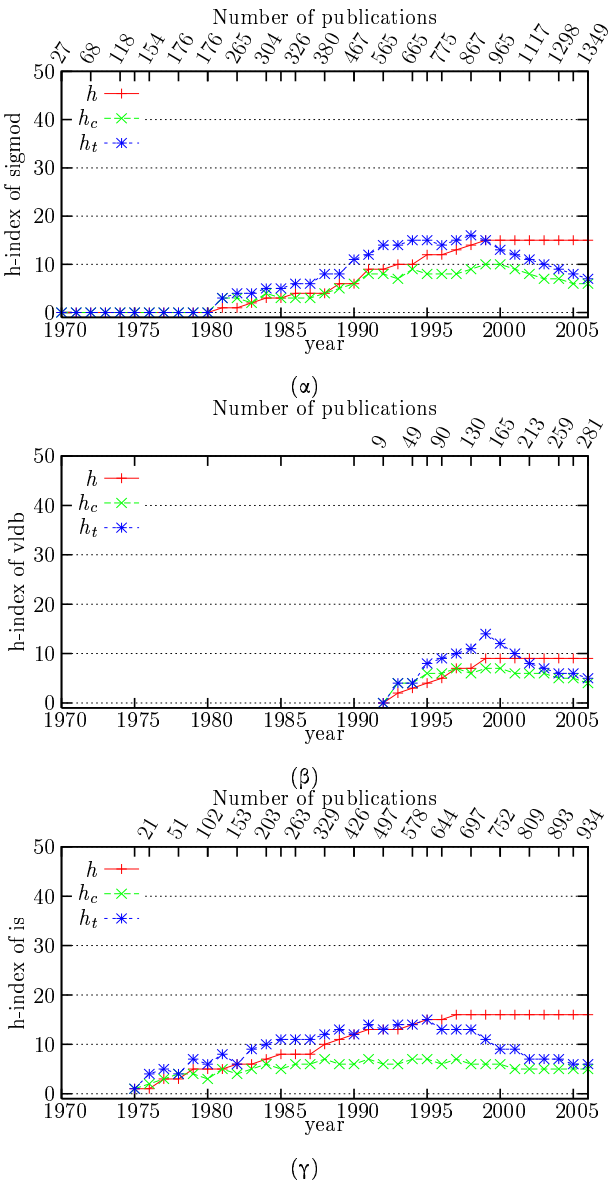
(β)

Πίνακας 6.16. Κατάταξη των εκδόσεων περιοδικών του 1995.

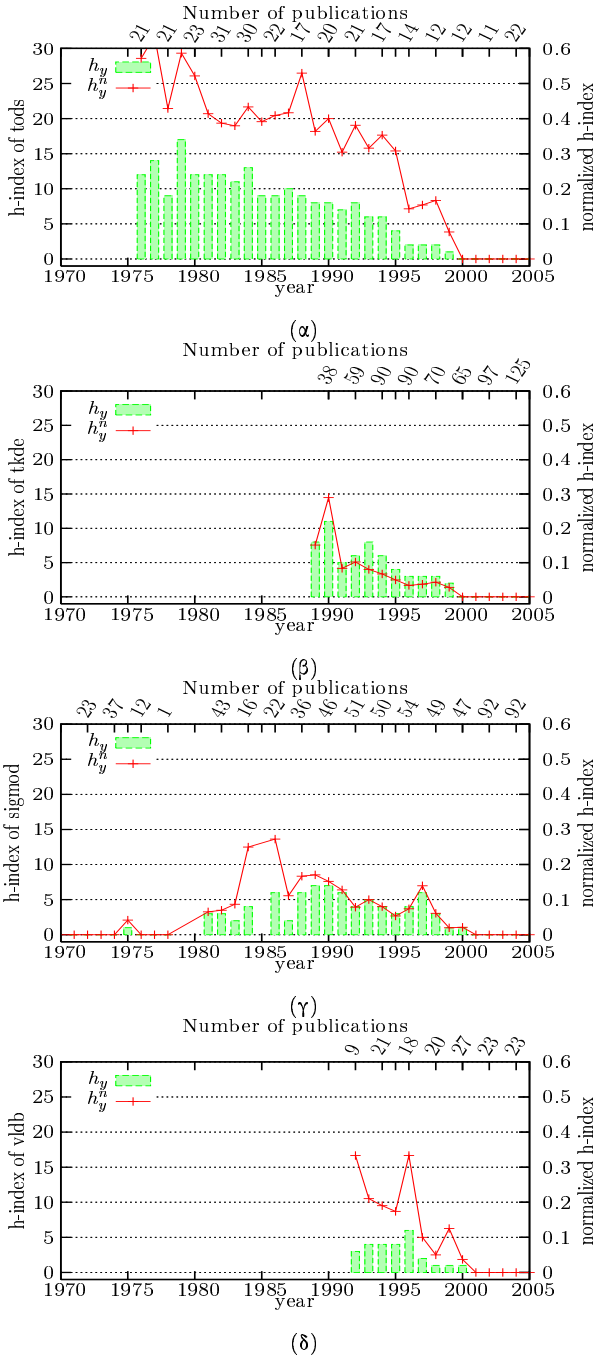
6.7 Συμπεράσματα και Μελλοντική Εργασία

Ο ορισμός του δείκτη *h-index* έχει δώσει νέες προοπτικές στον χώρο της Βιβλιομετρίας. Η έννοια του *h-index* είναι απλή αλλά πάρα πολύ ανθεκτική σε πρόσκαιρες επιτυχίες ερευνητών. Πιστεύουμε ότι έχει πυροδοτήσει νέους δρόμους

για περαιτέρω έρευνα. Η συνεισφορά μας σε αυτό το κεφάλαιο είναι η προσθήκη της χρονικής παραμέτρου στο βασικό δείκτη με τις παραλλαγές *contemporary h-index* και *trend h-index*. Επίσης, κάνουμε εφαρμογή των παραπάνω όχι μόνο για την αξιολόγηση ερευνητών αλλά και συνεδρίων και περιοδικών. Επιπλέον, ειδικά



Σχήμα 6.7. *h-index* περιοδικών της περιοχής των Βάσεων Δεδομένων (συνέχεια).



Σχήμα 6.8. *yearly h-index* και *normalized yearly h-index* Περιοδικών ΒΔ.

για τις περιπτώσεις των περιοδικών και συνεδρίων ορίσαμε τις παραλλαγές *yearly h-index* και *normalized yearly h-index*.

Ενδιαφέρουσα επέκταση της παρούσας έρευνας θα ήταν η απόπειρα μεταφοράς του σκεπτικού του *h-index* από το χώρο της Βιβλιομετρίας στο χώρο της Ιστομετρίας. Με αυτόν τον τρόπο θα μπορούσαμε να κάνουμε αξιολόγηση ιστοχώρους, ανάλογα με το πόσες ιστοσελίδες τους δέχονται περισσότερους από h υπερσυνδέσμους από ιστοσελίδες άλλων ιστοχώρων (κατά αντιστοιχία ιστοσελίδα - δημοσίευση και ιστοχώρος - περιοδικό). Δεδομένης όμως της ιδιαίτερα δυναμικής φύσης του Παγκόσμιου Ιστού δεν είναι δυνατή μια απ'ευθείας αντιστοιχία. Απαιτείται περαιτέρω μελέτη και έρευνα.

ΚΕΦΑΛΑΙΟ 7

Επίλογος

Περιεχόμενα

7.1 Συμπεράσματα

Στην σύγχρονη κοινωνία της πληροφορίας υπάρχει η ανάγκη για κατάταξη-αξιολόγηση σε βιβλιομετρικά δεδομένα καθώς και σε δεδομένα του Παγκόσμιου Ιστού. Συγχρόνως, μια σωστή κατάταξη βοηθά τους χρήστες του Παγκόσμιου Ιστού ή ψηφιακών βιβλιοθηκών να βρουν τις σημαντικότερες από τις πληροφορίες που αναζητούν. Από την άλλη, η κατάταξη-αξιολόγηση είναι σημαντική και για τα ίδια τα "αντικείμενα" που αξιολογούνται: μια καλή αξιολόγηση ενός ερευνητή μπορεί να αποφέρει στις ερευνητικές εργασίες του επιπλέον εισοδήματα, όπως και στην περίπτωση ενός ιστοχώρου. Στην πιο απλή περίπτωση, μια καλή αξιολόγηση ενός έργου δίνει ηθική ικανοποίηση στο δημιουργό και έναυσμα για περαιτέρω συνέχιση και επέκταση της εργασίας του. Το έργο αυτό μπορεί να είναι ένας ιστοχώρος, μια ερευνητική δημοσίεση, ένα νέο συνέδριο ή περιοδικό. Τέλος, στον ερευνητικό χώρο, η αξιολόγηση-κατάταξη ενός ερευνητή μπορεί να βοηθήσει (θετικά ή αρνητικά) στην κρίση για την εξέλιξή του ή στην αξιολόγηση του ερευνητικού του έργου.

Επίσης, η ομαδοποίηση δεδομένων, είτε βιβλιογραφικών είτε δεδομένων του Παγκόσμιου Ιστού έχει διάφορες εφαρμογές. Κάποιες από αυτές είναι η ομα-

δοποίηση δεδομένων με σκοπό την έξυπνη τοποθέτησή τους σε ένα κατανεμημένο περιβάλλον διακομιστών, είτε αυτά είναι δεδομένα του Παγκόσμιου Ιστού είτε είναι βιβλιογραφικά δεδομένα. Επίσης, η ομαδοποίση δεδομένων μπορεί να βελτιώσει τη διαδικασία αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό καθώς και στις ψηφιακές βιβλιοθήκες βιβλιογραφικού περιεχομένου, προσθέτοντας και τη διάσταση της "εξόρυξης γνώσης" στην αναζήτηση.

Στο Κεφάλαιο 2 εξετάσθηκε το ζήτημα της αξιολόγησης-κατάταξης συνεδρίων και γενικότερα συλλογών δημοσιεύσεων. Δόθηκε μια καινούργια διάσταση στον Παράγοντα Αντικτύπου ISI - ο Αντεστραμμένος Παράγοντας Αντικτύπου. Επίσης ορίστηκε μια νέα μετρική κατάταξης-αξιολόγησης συλλογών (συνεδρίων ή περιοδικών) στην οποία μπορεί να προστεθεί και η διάσταση του Αντεστραμμένου Παράγοντα Αντικτύπου.

Στα Κεφάλαια 3 και 4 μελετήσαμε το πρόβλημα της κατάταξης-αξιολόγησης στο βασικό επίπεδο γράφου αναφορών ή ιστογράφου. Κάναμε μια επισκόπηση της βιβλιογραφίας και εντοπίσαμε τα αδύνατα σημεία των αλγορίθμων που χρησιμοποιούνται αυτή τη στιγμή για την κατάταξη βιβλιογραφικών δεδομένων. Οι περισσότεροι από αυτούς τους αλγορίθμους είναι εννοιολογικά σχεδιασμένοι για τον Παγκόσμιο Ιστό και, αν και υπάρχει ομοιότητα, όταν εφαρμόζονται σε γράφους αναφορών βιβλιογραφικών δεδομένων παρουσιάζουν προβληματικές συμπεριφορές. Ορίσαμε λοιπόν μια νέα οικογένεια αλγορίθμων για κατάταξη βιβλιογραφικών δεδομένων (δημοσιεύσεων), που την ονομάσαμε SCEAS. Οι μέθοδοί μας παρουσιάζουν καλύτερη συμπεριφορά από τις υπάρχουσες, κατά την εφαρμογή τους σε βιβλιογραφικά δεδομένα. Επίσης, στο Κεφάλαιο 4 εφαρμόσαμε τις ίδιες μεθόδους και σε δεδομένα του Παγκόσμιου Ιστού. Διαπιστώσαμε ότι η μέθοδός μας έχει εξίσου καλά αποτελέσματα με τις υπάρχουσες μεθόδους, υπερτερεί όμως κατά πολύ στην ταχύτητα υπολογισμού.

Στο Κεφάλαιο 5 ορίσαμε μια νέα μέθοδο για ομαδοποίηση δεδομένων που μοντελοποιούνται σε γράφους αναφορών χρησιμοποιώντας την έννοια της ενδιάμεσης κεντρικότητας. Η ενδιάμεση κεντρικότητα μας βοηθά να εντοπίσουμε κόμβους-πυρήνες ομάδων και στη συνέχεια να "χτίσουμε" τις ομάδες γύρω από αυτούς. Χρησιμοποιώντας αυτήν την ιδιότητα των γράφων επιταχύνεται σημαντικά η διαδικασία της ομαδοποίησης.

Τέλος, στο Κεφάλαιο 6 μελετήσαμε το νεο-εμφανιζόμενο δείκτη αξιολόγησης συγγραφέων-ερευνητών με την ονομασία *h-index* ή Hircsh-Index (από το όνομα του εμπνευστή του). Προσθέσαμε σε αυτό το δείκτη τη διάσταση του χρόνου και παρουσιάσαμε νέες παραλλαγές, τις οποίες τις εφαρμόσαμε για να αξιολογήσουμε συγγραφείς, συνέδρια και περιοδικά (με βάση τα δεδομένα που περιέχονται στη

ψηφιακή βιβλιοθήκη DBLP).

7.2 Δρόμοι Μελλοντικής Έρευνας

Στην παράγραφο αυτή παρουσιάζουμε σύντομα μερικές περιοχές και ιδέες για μελλοντική έρευνα, οι οποίες βασίζονται στην έρευνα που έγινε κατά την εκπόνηση της παρούσας διατριβής.

Στο Κεφάλαιο 3 παρουσιάσαμε τους αλγορίθμους για κατάταξη δεδομένων που μοντελοποιούνται σε κατευθυνόμενους γράφους χωρίς βάρη στις ακμές. Μπορούμε να διευρύνουμε τους αλγορίθμους, ώστε να εφαρμόζονται και σε κατευθυνόμενους γράφους με βάρη στις ακμές. Για παράγειγμα, ο αλγόριθμος PageRank για γράφους με βάρη θα μπορούσε να ορισθεί ως:

$$PR_x \;=\; (1-d) \;+\; d\,\frac{\sum\limits_{\forall y \in I_x} \frac{PR_y * w_{y\to x}}{|O_y|}}{\sum\limits_{\forall y \in I_x} w_{y\to x}}$$

Αντίστοιχα ο SCEASRank ορίζεται ως:

$$S_x \;=\; (1-d) + d * \frac{\sum_{\forall y \in I_x} \frac{w_{y\to x} * (S_y + b)}{|O_y|} * a^{-1}}{\sum\limits_{\forall y \in I_x} w_{y\to x}}$$

Κάτι τέτοιο θα μπορούσε να έχει εφαρμογή σε γράφους αναφορών σε επίπεδο συλλογών (πχ. γράφοι αναφορών συνεδρίων). Αναφερόμαστε δηλαδή σε μια εναλακτική μέθοδο αυτών που παρουσιάσθηκαν στο Κεφάλαιο 2. Αν και οι αλγόριθμοι PageRank και HITS υπάρχουν σχεδόν μια δεκαετία, δεν έχουμε διαπιστώσει ακόμη στην βιβλιογραφία εφαρμογή τους σε γράφους με βάρη.

Από την άλλη, στην περίπτωση του Παγκόσμιου Ιστού διαβλέπουμε ότι στο μέλλον η έρευνα θα κινηθεί προς τη χρήση κατευθυνόμενων γράφων με βάρη στις ακμές. Τα βάρη των ακμών (που αντιστοιχούν στους υπερσυνδέσμους) μπορούν να προκύπτουν από τα χαρακτηριστικά του υπερκειμένου που περιέχονται:

1. ΄Ενας υπερσύνδεσμος έχει τόσο μεγαλύτερο βάρος, όσο περισσότερο εμφανής είναι στη σελίδα ανάλογα με τη θέση του. ΄Ετσι, αν ο υπερσύνδεσμος βρίσκεται στο μέσο της σελίδας και προς την κορυφή της, τότε πρέπει να έχει μεγαλύτερο βάρος από έναν άλλον που βρίσκεται στο τέλος της σελίδας - και άρα για να τον δεί ένας χρήστης θα πρέπει να κυλίσει (scroll) τη σελίδα αρκετά προς τα κάτω.

2. Θα μπορούσε επίσης να λαμβάνεται υπ' όψη το μέγεθος των χαρακτήρων του κειμένου που περιέχεται στον υπερσύνδεσμο. Ένας υπερσύνδεσμος που εμφανίζεται με μεγάλα γράμματα "μάλλον" είναι περισσότερο σημαντικός από κάποιον άλλον που εμφανίζεται με μικρά. Ομοίως, αν ο υπερσύνδεσμος αντιστοιχεί σε "κλικ" σε μια εικόνα είναι λογικό το βάρος του υπερσυνδέσμου αυτού να εξαρτάται από το μέγεθος της εικόνας.

3. Τέλος, ίσως είναι σημαντικό και το χρώμα με το οποίο εμφανίζεται ένας υπερσύνδεσμος. Αν διαφέρει αρκετά χρωματικά από τους υπόλοιπους, ίσως αυτό σημαίνει ότι πρέπει να τονισθεί και άρα είναι σημαντικός.

Θα μπορούσαμε να σκεφτούμε και άλλους τρόπους με τους οποίους να υπολογίζεται το βάρος μιας υπερσύνδεσης. Ίσως μάλιστα κάποιοι από τους παραπάνω να έχουν υλοποιηθεί ήδη από μηχανές αναζήτης στον Παγκόσμιο Ιστό - αλλά για λόγους βιομηχανικού ανταγωνισμού να μην έχουν ανακοινωθεί στο κοινό και στην ερευνητική κοινότητα.

Σε σχέση με το *h-index* που μελετήσαμε στο Κεφάλαιο 6, μελλοντική εργασία θα μπορούσε να είναι η εφαρμογή του στον χώρο του Παγκόσμιου Ιστού για αξιολόγηση-κατάταξη ιστοχώρων ή κάποιας άλλης εννοιολογικής ομαδοποίησης. Μια απλή μορφή του *h-index* ενός ιστοχώρου είναι το πλήθος των σελίδων h που περιέχονται σε αυτόν, στις οποίες δείχνουν περισσότεροι από h υπερσύνδεσμοι που προέρχονται από άλλους ιστοχώρους. Κάτι τέτοιο θα αποτελούσε μια αρκετά σταθερή μετρική για την αξιολόγηση ιστοχώρων. Από την άλλη όμως, η περίπτωση του Παγκόσμιου Ιστού διαφέρει από αυτή των βιβλιογραφικών δεδομένων. Οι ιστοχώροι διαφέρουν μεταξύ τους στο μέγεθος, καθώς επίσης πλέον η πλειοψηφία αυτών περιέχει πολλές δυναμικές σελίδες[1]. Έτσι, μια δυναμική σελίδα μπορεί να "μεταλάσσεται" συχνά ή να έχει πάρα πολλές μορφές. Πρέπει να βρεθεί μέθοδος, ώστε να αναγνωρίζονται αυτές οι σελίδες και να γίνεται μια "δίκαιη" καταμέτρησή τους. Πρακτικά, κάτι τέτοιο είναι δύσκολο να υλοποιηθεί σε βραχυπρόθεσμο χρονικό ορίζοντα.

Μια άλλη ιδέα, ώστε να παρακάμψουμε το προηγούμενο τεχνικό πρόβλημα, είναι να οριστεί ο *h-index* ενός ιστοχώρου ως το πλήθος των σελίδων h που περιέχονται σε αυτόν, στις οποίες δείχνουν περισσότεροι από h ιστοχώροι (και όχι ιστοσελίδες). Αυτό τεχνικά είναι εφικτό να υπολογιστεί με τα σημερινά δεδομένα.

[1] Σελίδες που δημιουργούνται κατά την πρόσβαση του χρήστη αντλώντας δεδομένα από κάποιο σύστημα βάσης δεδομένων.

Τέλος, στο Κεφάλαιο 5 ορίσαμε μια νέα μέθοδο για ομαδοποίηση δεδομένων που μοντελοποιούνται σε γράφους μη κατευθυνόμενους χωρίς βάρη. Το επόμενο βήμα σε αυτή την κατευθυνση είναι να γίνει διεύρυνση της μεθόδου ώστε να είναι δυνατή η εφαρμογή της σε γράφους με βάρη στις ακμές. Έτσι θα μπορούσαμε να κάνουμε καλύτερη ομαδοποίση σε περισσότερες περιπτώσεις, πχ. ομαδοποίηση ερευνητών με βάση το γράφο συν-συγγραφέων (co-authorship graph). Σε έναν τέτοιο γράφο το βάρος μιας ακμής από ένα συγγραφέα προς έναν άλλον πρέπει να εμπεριέχει τον βαθμό/ποσοστό συνεργασίας μεταξύ τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Lada A. Adamic and Bernardo A. Huberman. Power-law Distributions of the World Wide Web. *Science*, 287:2115a, 2000.

[2] Réka Albert and Albert-László Barabási. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[3] Y. An, J. Janssen, and E. Milios. Characterizing and Mining the Citation Graph of the Computer Science Literature. *Knowledge and Information Systems*, 6(6):664–678, 2004.

[4] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[5] H. Baumgartner and R. Pieters. The Influence of Marketing Journals: a Citation Analysis of the Discipline and its Sub-areas. Technical report, Tilburg University, 2000. `http://greywww.kub.nl:2080/greyfiles/center/2000/doc/123.pdf`.

[6] P. Berkin. A Survey of Pagerank Computing. *Internet Mathematics*, 2(1):73–120, 2005.

[7] T. Berners-Lee, R. Caililiau, and A. Luotonen. The World-wide Web. *Communications of the ACM*, 37(8):76–82, 1994.

[8] M. Bianchini, M. Gori, and F. Scarselli. Pagerank and Web Communities. In *Proceedings IEEE International Conference on Web Intelligence (WI'2003)*, pages 365–371, 2003.

[9] M. Bianchini, M. Gori, and F. Scarselli. Inside Pagerank. *ACM Transactions on Internet Technology*, 5(1):92–128, 2005.

[10] Ginestra Bianconi and Albert-László Barabási. Bose-einstein Condensation in Complex Networks. *Physical Review Letters*, 86(24):5632–5635, June 2001.

[11] L. Blackert and S. Siegel. Ist in der Wissenschaftlich-technischen Information Platz fuer die Informetrie? *Wissenschaftliche Zeitschrift der Technischen Hochschule Ilmenau*, 25(6):187–199, 1979.

[12] A. Borodin, J. S. Rosenthal, G. O. Roberts, and P. Tsaparas. Link Analysis Ranking: Algorithms, Theory and Experiments. *ACM Transactions on Internet Technologies*, 5(1):231–297, 2005.

[13] S.C. Bradford. Sources on Iinformation on Specific Subjects. *Engineering*, 137:85–86, 1934. Reprinted in: *Collection Management*, 1, 95-103, 1976-77. Also reprinted in: *Journal of Information Science*, 10, 148, 176-180, 1985.

[14] S. Bradshaw and K. Hammond. Using Citations to Facilitate Precise Indexing and Automatic Index Creation in Collections of Research Papers. *Knowledge Based Systems*, 14(1/2):29–35, 2001.

[15] Ulrik Brandes. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[16] Sergey Brin and Lawrence Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings 7th International World Wide Web Conference (WWW'98)*, pages 107–117, 1998.

[17] Soumen Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*, chapter 7.1, pages 205–206. Morgan Kaufmann Publishers, 2003.

[18] D. Dhyani, S. Bhowmick, and W.K. Ng. Deriving and Verifying Statistical Distribution of a Hyperlink-based Web Page Quality Metric. In *Proceedings 13th International Conference on Database and Expert System Applications (DEXA'2002)*, pages 19–28, 2002.

[19] P. Diaconis and R.L. Graham. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Society of Statistics, Series B*, 39:262–268, 1977.

[20] Y. Ding, G. Chowdhury, and S. Foo. Bibliometric Cartography of Information Retrieval Research by using Co-word Analysis. *Information Processing and Management*, 37(6):817–842, 2001.

[21] Pavel Dmitriev, Carl Lagoze, and Boris Suchkov. As We may Perceive: Inferring Logical Documents from Hypertext. In *Proceedings ACM International Conference on Hypertext and Hypermedia (HT'2005)*, pages 66–74, 2005.

[22] L. Egghe. Expansion of the Field of Informetrics: Origins and Consequences (editorial). *Information Processing and Management*, 41:1311–1316, 2005.

[23] L. Egghe and R. Rousseau, editors. *Proceedings 1st International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval (Informetrics'87/88)*, Amsterdam, The Netherlands, 1988.

[24] Nadav Eiron and Kevin S. McCurley. Untangling Compound Documents on the Web. In *Proceedings ACM International Conference on Hypertext and Hypermedia (HT'2003)*, pages 85–94, 2003.

[25] Ergin Elmacioglu and Dongwon Lee. On Six Degrees of Separation in DBLP-DB and More. *ACM SIGMOD Record*, 34(2):33–40, Jun 2005.

[26] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans Coetzee. Self-organization of the Wweb and Identification of Communities. *IEEE Computer*, 35(3):66–71, 2002.

[27] Gary William Flake, Steve Lawrence, and Clyde Lee Giles. Efficient Identification of Web Communities. In *Proceedings 8th ACM International Conference on Knowledge Discovery in Data (KDD'2002)*, pages 150–160, 2002.

[28] Gary William Flake, Robert E. Tarjan, and Kostas Tsioutsiouklis. Graph Clustering and Minimum Cut Trees. *Internet Mathematics*, 1(4):385–408, 2004.

[29] Gary William Flake, Kostas Tsioutsiouliklis, and Leonid Zhukov. Methods for Mining Web Communities: Bibliometric, Spectral, and Flow. In *Web Dynamics*, pages 45–68. Springer, 2004.

[30] E. Garfield. Science Citation Index. `http://www.isinet.com/isi/`.

[31] E. Garfield. Citation Analysis as a Tool in Journal Evaluation. *Essays of an Information Scientist*, 1:527–544, 1972.

[32] E. Garfield. The Impact Factor, 1994. `http://www.isinet.com/isi/hot/essays/JOURNALcitationreports/7.html`.

[33] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. In *Proceedings ACM International Conference on Hypertext and Hypermedia (HT'98)*, pages 225–234, 1998.

[34] David Gibson, Ravi Kumar, and Andrew Tomkins. Discovering Large Dense Subgraphs in Massive Graphs. In *Proceedings 31st International Conference on Very Large Data Bases (VLDB'2005)*, pages 721–732, 2005.

[35] Michelle Girvan and Mark E. J. Newman. Community Structure in Social and Biological Networks. *Proceedings National Academy of Sciences*, 99:7821, 2002.

[36] Andrew V. Goldberg and Robert E. Tarjan. A New Approach to the Maximum Flow Problem. *Journal of the ACM*, 35:921–940, 1988.

[37] R. E. Gomory and T. C. Hu. Multi-terminal Network Flows. *SIAM Journal of Applied Mathematics*, 9(4), Dec 1961.

[38] A. Goodrum, K. McCain, S. Lawrence, and C.L. Giles. Scolarly Publication in the Interent Age: a Citation Analysis of Computer Science Literature. *Information Processing and Management*, 37(2):661–675, 2001.

[39] G. Greco, S. Greco, and E. Zumpano. Web Communities: Models and Algorithms. *World Wide Web Journal*, 7(1):58–82, 2004.

[40] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering Based on Association Rule Hypergraphs. In *Proceedings 2nd Data Mining and Knowledge Discovery Workshop (DMKD'97)*, 1997.

[41] Taher H. Haveliwala. Topic-sensitive Pagerank: a Context-sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, Jul/Aug 2003.

[42] Y. He and S.C. Hui. Mining a Web Citation Database for Author Cocitation Analysis. *Information Processing and Management*, 38(4):491–508, 2002.

[43] S.D. Herring. Use of Electronic Resources in Scholarly Electronic JOURNALs: a Citation Analysis. *College & Research Libraries*, 63(4):334–340, 2002.

[44] J.E. Hirsch. An Index to Quantify an Individual's Scientific Research Output. *Proceedings National Academy of Sciences*, 102(46):16569–16572, Nov 2005.

[45] T. Hult, W. Neese, and E. Bashaw. Faculty Perception of Marketing Journals. *Journal of Marketing Education*, 19(1):37–52, 1997.

[46] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: a Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[47] Sepandar D. Kamvar and Taher H. Haveliwala. The Condition Number of the PageRank Problem. Technical report, Stanford University, 2003.

[48] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel Hypergraph Partitioning: Application in VLSI Domain. In *Proceedings 34th Design Automation Conference (DAC'97)*, pages 526–529, 1997.

[49] Dimitrios Katsaros. FWgen: Generating Web-graphs with Communities. Technical report, Aristotle University. manuscript under preparation.

[50] M.G. Kendall. *Rank Correlation Methods*. Charles Griffin, London, 1970.

[51] Jack P. C. Kleijnen and Willem J. H. Van Groenendaal. Measuring the Quality of Publications: New Methodology and Case Study. *Information Processing and Management*, 36(4):551–570, 2000.

[52] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[53] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The Web as a Graph: Measurements, Models, and Methods. In *Proceedings 5th International Computer and Combinatorics Conference (COCOON'99)*, pages 1–17, 1999.

[54] R. Korobkin. Ranking Journals: Some Thoughts on Theory and Methodology. Technical report, University of Illinois, 1999. `www.law.fsu.edu/JOURNALs/lawreview/downloads/264/koro.pdf`.

[55] A. Langville and C. Meyer. Deeper Inside Pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[56] S. Lawrence. Free Online Availability Substantially Increases a Paper's Impact. *Nature*, 411:521, 2001.

[57] S. Lawrence, K.D. Bollacker, and C.L. Giles. Indexing and Retrieval of Scientific Literature. In *Proceedings 1999 ACM Conference on Information and Knowledge Management (CIKM'99)*, pages 139–146, 1999.

[58] S. Lawrence, C.L. Giles, and K. Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[59] Ronny Lempel and Shlomo Moran. SALSA: the Stochastic Approach for Link-structure Analysis. *ACM Transactions on Information Systems*, 19(2):131–160, 2001.

[60] M. Ley. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *Proceedings SPIRE Symposium*, pages 1–10, 2002.

[61] Wen-Syan Li, K. Selçuk Candan, Quoc Vu, and Divyakant Agrawal. Query Relaxation by Structure and Semantics for Retrieval of Logical Web Documents. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):768–791, 2002.

[62] X. Lin, H. White, and J. Buzydlowski. Real-time Author Co-citation Mapping for Online Searching. *Information Processing and Management*, 39(5):689–706, 2003.

[63] B.A. Lipetz. Aspects of JASIS AUTHORship through Five Decades. *Journal of the American Society for Information Science*, 50(11):994–1003, 1999.

[64] A.J. Lotka. The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–324, 1926.

[65] G. Meghabghab. Discovering Authorities and Hubs in Different Topological Web Graph Structures. *Information Processing and Management*, 38(1):111–140, 2002.

[66] Filippo Menczer. Evolution of Document Networks. In *Proceedings National Academy of Sciences*, volume 101, pages 5261–5265, Apr 2004.

[67] N. Mylonopoulos and V. Theoharakis. Global Perceptions of IS Journals. *Communications of the ACM*, 44(9):29–33, 2001.

[68] O. Nacke. Informetrie: eine neuer Name fuer eine neue Disziplin. *Nachrichten fuer Dokumentation*, 30(6):219–226, 1979.

[69] V.V. Nalimov and Z.M. Mul'cenko. *Naukometrija*. Nauka, Moskva, USSR, 1969.

[70] Mark E.J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

[71] Mark E.J. Newman and Michelle Girvan. Finding and Evaluating Community Structure in Networks. *Physical Review E*, 69:026113, 2004.

[72] Nielsen/NetRatings. Netview Usage Metrics, 2004. `http://www.netratings.com/news.jsp?section = dat_to`.

[73] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida. Analysis and Improvement of HITS Algorithm for Detecting Web Communities. *Systems and Computers in Japan*, 35(13):32–42, 2004.

[74] Melanie J. Norton. *Introductory Concepts in Information Science*. Information Today, New Jersey, 2001.

[75] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, California, CA, 1999.

[76] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners Don't Take All: Characterizing the Competition for Links nn the Web. *Proceedings National Academy of Sciences*, 99(8):5207–5211, 2002.

[77] L. Pretto. A Theoretical Analysis of Google's PageRank. In *Proceedings SPIRE Symposium*, pages 131–144, 2002.

[78] A. Pritchard. Statistical Bibliography or Bibliometrics? *Journal of Documentation*, 25:348–349, 1969.

[79] R.K. Rainer and M.D. Miller. Examining Differences across Journal Rankings. *Communications of the ACM*, 48(2):91–94, 2005.

[80] Antonis Sidiropoulos. Finding Communities in Site Web-graphs and Citation Graphs. In *Proceedings 10th East-European Conference on Advances in Databases and Information Systems (ADBIS'2006)*, 2006.

[81] Antonis Sidiropoulos and Dimitrios Katsaros. Unfolding the Full Potential of h-index for Bibliographic Ranking. In *Proceedings 5th Hellenic Data Management Symposium (HDMS'2006)*, submitted.

[82] Antonis Sidiropoulos and Yannis Manolopoulos. Generalized Comparison of Graph-based Ranking Algorithms for Publications and Authors. *Journal of Systems and Software*. accepted for publication.

[83] Antonis Sidiropoulos and Yannis Manolopoulos. A Citation-based System to Assist Prize Awarding. *ACM SIGMOD Record*, 34(4):54–60, Dec 200.

[84] Antonis Sidiropoulos and Yannis Manolopoulos. Automatically Ranking Scientific Conferences using Digital Libraries. In *Proceedings 9th Panhellenic Conference on Informatics (PCI'2003)*, pages 446–461, Thessaloniki, 2003.

[85] Antonis Sidiropoulos and Yannis Manolopoulos. Generalized Comparison of Ranking Algorithms for Publications and Authors. Technical report, Aristotle University, 2005.

[86] Antonis Sidiropoulos and Yiannis Manolopoulos. A New Perspective to Automatically Rank Scientific Conferences Using Digital Libraries. *Information Processing and Management*, 41(2):289–312, 2005.

[87] Antonis Sidiropoulos, George A. Pallis, Dimitrios Katsaros, Konstantinos Stamos, Athena Vakali, and Yannis Manolopoulos. Prefetching in Content Distribution Networks via Web Communities Identification and Outsourcing. Technical report, Aristotle University, 2005.

[88] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Web Search Engine Query Log. *ACM SIGIR Forum*, 33(1), 1999. `http://www.acm.org/sigir/forum/F99/Silverstein.pdf`.

[89] S.J. Snyder and A. Peterson. The Referencing of Internet Websites in Medical and Scientific Publications. *Brain and Cognition*, 50:335–337, 2002.

[90] A. Spink, S. Ozmutlu, H.C. Ozmutlu, and B.J. Jansen. US versus European Web Searching Trends, 2002. `http://www.acm.org/sigir/forum/F2002/spink.pdf`.

[91] A. Tahai and J. Rigsby. Information Processing using Citations to Investigate Journal Influence in Accounting. *Information Processing and Management*, 34(2/3):341–359, 2002.

[92] Keishi Tajima, Kenji Hatano, Takeshi Matsukura, Ryouichi Sano, and Katsumi Tanaka. Discovery and Retrieval of Logical Information Units in Web. In *Proceedings Workshop on Organizing Web Space (WOWS'99)*, pages 13–23, Berkeley, CA, August 1999.

[93] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[94] Duncan J. Watts and Steven H. Strogatz. Collective Dynamics of 'Small-world' Networks. *Nature*, 393:440–442, 1998.

[95] S. White and P. Smyth. A Spectral Clustering Approach to Finding Communities in Graphs. In *Proceedings SIAM Data Mining Conference (SDM'2005)*, 2005.

[96] Y. Zhang. Scholarly Use of Internet-based Electronic Resources. *Journal of the American Society for Information Science and Technology*, 52(8):628–650, 2001.

[97] G.K. Zipf. *Human Behavior and the Principle of Least Effort*. Cambridge, MA, Addison-Wesley, 1949. Reprinted: Hafner, New York, NY, 1965.

ΠΑΡΑΡΤΗΜΑ Α

Λίστα ερευνητικών εργασιών

Περιοδικά διεθνή με κριτές

1. Sidiropoulos A., Manolopoulos Y.: "A New Perspective to Automatically Rank Scientific Conferences Using Digital Libraries", **Information Processing and Management**, Vol.41, N.2, pp. 22–29, 2005

2. Sidiropoulos A., Manolopoulos Y.: "A Citation-Based System to Assist Prize Awarding", **ACM SIGMOD Record**, Vol.34, N.4, pp. 54–60, 2005.

3. Sidiropoulos A., Manolopoulos Y.: "Generalized Comparison of Graph-based Ranking Algorithms for Publications and Authors", **Journal for Systems and Software**, accepted for publication.

4. Sidiropoulos A., Pallis G., Katsaros D., Stamos K., Vakali A., Manolopoulos Y.: "Prefetching in Content Distribution Networks via Web Communities Identification and Outsourcing", **World Wide Web Journal**, submitted.

5. Pallis G., Stamos K., Vakali A., Katsaros D., Sidiropoulos A., Manolopoulos Y.: "CDNsim: A Simulation Tool for Content Distribution Networkds", **Software Practice and Experience**, submitted.

Λήμματα σε εγκυκλοπαίδειες με κριτές

1. Sidiropoulos A., Katsaros D., Manolopoulos Y.: "Existence and Discovery of Communities in the World Wide Web", **Encyclopedia of Networked**

and Virtual Organizations, 2006.

2. Sidiropoulos A., Katsaros D., Manolopoulos Y.: "A Survey of Link Analysis Ranking", **Encyclopedia of Networked and Virtual Organizations**, 2006.

Δημοσιεύσεις σε συνέδρια με κριτές

1. Sidiropoulos A., Katsaros D.: "Updating Web Views Distributed over Wide Area Networks", **In Proceedings 1st Balkan Conference on Informatics (BCI'2003)**, pp.267-279, Thessaloniki, 2003.

2. Sidiropoulos A., Manolopoulos Y.: "Automatically Ranking Scientific Conferences using Digital Libraries", **In Proceedings 9th Panhellenic Conference on Informatics (PCI'2003)**, pp.446-461, Thessaloniki, 2003.

3. Pallis G., Vakali A., Stamos K., Sidiropoulos A., Katsaros D., Manolopoulos Y.: "A Latency-based Object Placement Approach in Content Distribution Networks", **In Proceedings 3rd Latin American Web Congress (LA-Web)**, pp. 140-147, Buenos Aires, Argentina, 2005

4. Pallis G., Vakali A., Stamos K., Sidiropoulos A., Katsaros D., Manolopoulos Y.: "Replication based on Objects Load under a Content Distribution Network", **In Proceedings 2nd International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)** (In conjunction with ICDE'06), Atlanta, Georgia, Thessalini, 2006.

5. Sidiropoulos A.: "Finding Communities in Site Web-Graphs and Citation Graphs", **In Proceedings Tenth East-European Conference on Advances in Databases and Information Systems (ADBIS'2006)**, accepted, 2006.

6. Sidiropoulos A., Katsaros D.: "Unfolding the full potential of H-Index for Bibliographic Ranking", **5th Hellenic Data Management Symposium (HDMS 2006)**, submitted.

ΠΑΡΑΡΤΗΜΑ Β

Συμπληρωματικά Πειράματα Κεφαλαίου 3

Β.1 Πλήρη Αποτελέσματα Κατάταξης Συγγραφέων

Εδώ παρουσιάζουμε την πλήρη έκδοση του Πίνακα 3.19 ώστε να απεικονίσουμε τις θέσεις αξιολόγησης μη βραβευμένων συγγραφέων. Είναι προφανές ότι οι συγγραφείς με τη μεγαλύτερη πιθανότητα να βραβευθούν στο μέλλον είναι εκείνοι που αξιολογούνται σχετικά ψηλά (ειδικά με τη μέθοδο SCEAS) και δεν έχουν ήδη βραβευθεί. Αυτό μπορεί να είναι βοηθητικό στην ελάττωση του πλήθους των υποψηφίων για βράβευση στα επόμενα χρόνια (δημιουργία short list). Μία τελευταία αξιόλογη παρατήρηση είναι ότι ο Edgar F. Codd παραμένει στην 1η θέση σε όλες σχεδόν τις μεθόδους αξιολόγησης.

	CC	SCC	PageRank	SALSA_A	SSALSA_A
1	**D.J. DeWitt**	**Jim Gray**	E.F. Codd	**D.J. DeWitt**	E. F. Codd
2	**M. Stonebraker**	E.F. Codd	**Jim Gray**	**M. Stonebraker**	**M. Stonebraker**
3	**Jim Gray**	**M. Stonebraker**	R.A. Lorie	**Jim Gray**	R.A. Lorie
4	R.A. Lorie	**D.J. DeWitt**	**M. Stonebraker**	R.A. Lorie	**Jim Gray**
5	J.D. Ullman	R.A. Lorie	**D.D. Chamberlin**	J.D. Ullman	**P.A. Bernstein**
6	**P.A. Bernstein**	J.D. Ullman	**P.A. Bernstein**	**P.A. Bernstein**	J.D. Ullman
7	E.F. Codd	**P.A. Bernstein**	J.D. Ullman	E.F. Codd	**D.D. Chamberlin**
8	Won Kim	P.P. Chen	**D.J. DeWitt**	Won Kim	**D.J. DeWitt**
9	**D. Maier**	**D.D. Chamberlin**	E. Wong	**D. Maier**	**R. Fagin**
10	Y. Sagiv	**D. Maier**	**R. Fagin**	Y. Sagiv	E. Wong
11	F. Bancilhon	Won Kim	C. Beeri	F. Bancilhon	C. Beeri
12	C. Beeri	**R. Agrawal**	P.P. Chen	C. Beeri	**D. Maier**
13	**S. Abiteboul**	C. Beeri	**D. Maier**	**S. Abiteboul**	Y. Sagiv
14	**D.D. Chamberlin**	F. Bancilhon	Won Kim	**D.D. Chamberlin**	P. P. Chen
15	**R. Agrawal**	U. Dayal	Y. Sagiv	**R. Agrawal**	Won Kim
16	M. J. Carey	**H. Garcia-Molina**	**R. Agrawal**	M.J. Carey	N. Goodman
17	**R. Fagin**	Y. Sagiv	U. Dayal	**R. Fagin**	F. Bancilhon
18	U. Dayal	**R. Fagin**	F. Bancilhon	U. Dayal	U. Dayal
19	R. Ramakrishnan	**S. Abiteboul**	N. Goodman	R. Ramakrishnan	S.B. Yao
20	N. Goodman	M.J. Carey	**H. Garcia-Molina**	N. Goodman	M.J. Carey
21	**H. Garcia-Molina**	E. Wong	M.J. Carey	**H. Garcia-Molina**	**S. Abiteboul**
22	J. Widom	R. Ramakrishnan	B.G. Lindsay	J. Widom	A.V. Aho
23	E. Wong	J. Widom	**S. Abiteboul**	E. Wong	B.G. Lindsay
24	H. Pirahesh	N. Goodman	**R. Bayer**	H. Pirahesh	R. Ramakrishnan
25	P.P. Chen	H. Pirahesh	H. Pirahesh	P.P. Chen	**P.G. Selinger**
26	C. Faloutsos	J.F. Naughton	R. Ramakrishnan	C. Faloutsos	A.O. Mendelzon
27	B.G. Lindsay	B.G. Lindsay	S.B. Yao	B.G. Lindsay	C. Zaniolo
28	R. Hull	T. Imielinski	**P.G. Selinger**	R. Hull	**R. Agrawal**
29	**C. Mohan**	C. Faloutsos	T. Imielinski	**C. Mohan**	H. Pirahesh
30	N. Roussopoulos	R.H. Katz	J. Widom	N. Roussopoulos	**R. Bayer**
31	A.O. Mendelzon	S.B. Navathe	N. Roussopoulos	A.O. Mendelzon	N. Roussopoulos
32	J.F. Naughton	**C. Mohan**	A.V. Aho	J.F. Naughton	R.H. Katz
33	G. Graefe	R. Hull	R.H. Katz	G. Graefe	R. Hull
34	P. Buneman	N. Roussopoulos	J.F. Naughton	P. Buneman	C. Faloutsos
35	R.H. Katz	A. Shoshani	C. Zaniolo	R.H. Katz	G. M. Lohman
36	S.B. Navathe	G. Graefe	D. McLeod	S.B. Navathe	**H. Garcia-Molina**
37	C. Zaniolo	A.O. Mendelzon	C. Faloutsos	C. Zaniolo	D. McLeod
38	**P. G. Selinger**	H. Kriegel	G.M. Lohman	**P. G. Selinger**	P. Larson
39	M.Y. Vardi	**P.G. Selinger**	S.B. Navathe	M.Y. Vardi	S.B. Navathe
40	H. Kriegel	A.P. Sheth	**C. Mohan**	H. Kriegel	**C. Mohan**
missed	73. R. Bayer	51. R. Bayer		74. R. Bayer	

Πίνακας Β.1. Κατάταξη συγγραφέων με τις καλύτερες 25 δημοσιεύσεις (μέρος α).

	PS	BPS	EPS	BEPS	SCEAS
1	E. F. Codd	E. F. Codd	E. F. Codd	**Jim Gray**	**Jim Gray**
2	**M. Stonebraker**	**Jim Gray**	**M. Stonebraker**	E. F. Codd	E. F. Codd
3	**D.D. Chamberlin**	R.A. Lorie	R.A. Lorie	R.A. Lorie	R.A. Lorie
4	R. A. Lorie	**D.D. Chamberlin**	**D.D. Chamberlin**	**M. Stonebraker**	**M. Stonebraker**
5	**P.A. Bernstein**	**M. Stonebraker**	**Jim Gray**	**D.J. DeWitt**	**D.J. DeWitt**
6	**Jim Gray**	**P.A. Bernstein**	**P.A. Bernstein**	J.D. Ullman	J.D. Ullman
7	E. Wong	J.D. Ullman	E. Wong	**D.D. Chamberlin**	**D.D. Chamberlin**
8	J.D. Ullman	**D.J. DeWitt**	**D.J. DeWitt**	**P.A. Bernstein**	**P.A. Bernstein**
9	**R. Fagin**	E. Wong	J.D. Ullman	P.P. Chen	P.P. Chen
10	C. Beeri	**R. Fagin**	C. Beeri	Won Kim	Won Kim
11	**D.J. DeWitt**	C. Beeri	**R. Fagin**	**D. Maier**	**D. Maier**
12	N. Goodman	P.P. Chen	N. Goodman	**R. Agrawal**	**R. Agrawal**
13	Y. Sagiv	Y. Sagiv	**D. Maier**	C. Beeri	C. Beeri
14	**D. Maier**	**D. Maier**	Y. Sagiv	**R. Fagin**	**R. Fagin**
15	S.B. Yao	Won Kim	Won Kim	E. Wong	Y. Sagiv
16	D. McLeod	U. Dayal	S.B. Yao	Y. Sagiv	F. Bancilhon
17	**R. Bayer**	**R. Agrawal**	F. Bancilhon	U. Dayal	E. Wong
18	A. V. Aho	F. Bancilhon	**P.G. Selinger**	F. Bancilhon	U. Dayal
19	**P. G. Selinger**	N. Goodman	U. Dayal	**H. Garcia-Molina**	**H. Garcia-Molina**
20	F. Bancilhon	**R. Bayer**	D. McLeod	M.J. Carey	M.J. Carey
21	Won Kim	**H. Garcia-Molina**	B.G. Lindsay	N. Goodman	**S. Abiteboul**
22	U. Dayal	M.J. Carey	**S. Abiteboul**	**S. Abiteboul**	N. Goodman
23	D. Tsichritzis	B.G. Lindsay	A.V. Aho	R. Ramakrishnan	R. Ramakrishnan
24	P. P. Chen	**S. Abiteboul**	**R. Bayer**	H. Pirahesh	J. Widom
25	H. Schek	S.B. Yao	M.J. Carey	J. Widom	H. Pirahesh
26	**S. Abiteboul**	**P.G. Selinger**	C. Zaniolo	B.G. Lindsay	B.G. Lindsay
27	H.A. Schmid	H. Pirahesh	P.P. Chen	J.F. Naughton	J.F. Naughton
28	B.G. Lindsay	A.V. Aho	H. Schek	T. Imielinski	T. Imielinski
29	C. Zaniolo	N. Roussopoulos	N. Roussopoulos	C. Faloutsos	C. Faloutsos
30	V.Y. Lum	R. Ramakrishnan	G.M. Lohman	R.H. Katz	R.H. Katz
31	M. Schkolnick	T. Imielinski	R.H. Katz	N. Roussopoulos	N. Roussopoulos
32	M.J. Carey	J. Widom	P. Buneman	**P.G. Selinger**	S.B. Navathe
33	N. Roussopoulos	R.H. Katz	R. Hull	S.B. Navathe	**C. Mohan**
34	G.M. Lohman	C. Zaniolo	**C. Mohan**	**C. Mohan**	**P.G. Selinger**
35	P. Buneman	D. McLeod	L.A. Rowe	S.B. Yao	S.B. Yao
36	R. Hull	D. Tsichritzis	M. Schkolnick	**R. Bayer**	A. Shoshani
37	L.A. Rowe	G.M. Lohman	H. Pirahesh	A. Shoshani	G. Graefe
38	R.H. Katz	J.F. Naughton	G. Graefe	G. Graefe	R. Hull
39	A.O. Mendelzon	C. Faloutsos	V.Y. Lum	A.O. Mendelzon	A.O. Mendelzon
40	S. Y.W. Su	S.B. Navathe	G.P. Copeland	R. Hull	**R. Bayer**
missed	41. C. Mohan 69. H. Garcia-Molina 70. R. Agrawal	43. C. Mohan	45. R. Agrawal 50. H. Garcia-Molina		

Πίνακας Β.2. Κατάταξη συγγραφέων με τις καλύτερες 25 δημοσιεύσεις (μέρος β).

Β.2 Κατάταξη Δημοσιεύσεων του SIGMOD'1995 και VLDB'1995

Εδώ παρουσιάζουμε τα αποτελέσματα αξιολόγησης κάποιων μεθόδων για τα πρακτικά του συνεδρίου SIGMOD'1995 και VLDB'1995. Αξίζει να προσεχθεί ότι το σύνολο δεδομένων περιλαμβάνει αναφορές έως το 2000. Συνεπώς, δημοσιεύσεις του 1995 θα μπορούσαν να λάβουν αναφορές κατά τη διάρκεια μόνο 5 χρόνων. Όπως μπορούμε να δούμε από όλες τις μεθόδους (Πίνακες Β.3-Β.13 για το SIGMOD'1995 και Β.14-Β.24 για το VLDB'1995) τρεις δημοσιεύσεις εναλλάσσονται για τη νικήτρια θέση. Αυτό είναι επίσης προφανές από τους Πίνακες Β.25 και Β.25. Οι Πίνακες Β.26 και Β.25 υπολογίζονται με βάση τη μέθοδο σύνοψης που εξηγείται στην Παράγραφο 3.5.1. Σε αυτούς τους πίνακες παρουσιάζουμε τις κορυφαίες 5 δημοσιεύσεις των παραπάνω συνεδρίων για τα έτη 1995-1998 (υποψήφιοι για τα 10-Year Awards της περιόδου 2005-2008).

Score	Title
49	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
46	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
44	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
37	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
28	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
22	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
21	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
14	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
14	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
10	A Critique of ANSI SQL Isolation Levels. H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, P. E. O'Neil

Πίνακας Β.3. Κατάταξη δημοσιεύεσεων του SIGMOD'95 με τη μέθοδο CC.

Score	Title
4.612	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
4.594	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
3.4	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
3.369	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
3.237	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
2.626	A Critique of ANSI SQL Isolation Levels. H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, P. E. O'Neil
2.551	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
2.236	Fault Tolerant Design of Multimedia Servers. S. Berson, L. Golubchik, R. R. Muntz
1.626	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
1.598	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross

Πίνακας B.4. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο BCC.

Score	Title
2.941	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
2.775	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
2.004	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
1.884	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
1.763	A Critique of ANSI SQL Isolation Levels. H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, P. E. O'Neil
1.712	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
1.664	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
1.325	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
1.204	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
1.182	Fault Tolerant Design of Multimedia Servers. S. Berson, L. Golubchik, R. R. Muntz

Πίνακας B.5. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο PR.

Score	Title
1.908	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
1.186	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
1.075	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
0.9303	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
0.6591	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
0.5741	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
0.5189	Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-trees. D. Papadias, Y. Theodoridis, T. K. Sellis, M. J. Egenhofer
0.4233	Towards an Effective Calculus for Object Query Languages. L. Fegaras, D. Maier
0.3906	A General Solution of the n-dimensional B-tree Problem. M. Freeston
0.3874	The LyriC Language: Querying Constraint Objects. A. Brodsky, Y. Kornatzky

Πίνακας B.6. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο HA.

Score	Title
2.887e-103	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
2.26e-103	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
1.615e-103	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
1.422e-103	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
4.387e-104	Applying Update Streams in a Soft Real-Time Database System. B. Adelberg, H. Garcia-Molina, B. Kao
5.339e-105	Parallel Database Systems 101. J. Gray
8.276e-144	Hypergraph Based Reorderings of Outer Join Queries with Complex Predicates. G. Bhargava, P. Goel, B. R. Iyer
7.289e-145	A Database Interface for File Updates. S. Abiteboul, S. Cluet, T. Milo
4.505e-145	Data Extraction and Transformation for the Data Warehouse. C. Squire
4.505e-145	Copy Detection Mechanisms for Digital Documents. S. Brin, J. Davis, H. Garcia-Molina

Πίνακας Β.7. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο P.

Score	Title
8.113	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
7.616	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
7.285	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
6.126	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
4.636	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
3.642	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
3.477	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
2.318	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
2.318	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
2.117	The Lotus Notes Storage System. K. Moore

Πίνακας Β.8. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο SA.

Score	Title
0.2548	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
0.1271	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
0.1218	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
0.08248	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
0.07995	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
0.07879	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
0.07783	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
0.06473	Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-trees. D. Papadias, Y. Theodoridis, T. K. Sellis, M. J. Egenhofer
0.0357	Towards an Effective Calculus for Object Query Languages. L. Fegaras, D. Maier
0.03546	The LyriC Language: Querying Constraint Objects. A. Brodsky, Y. Kornatzky

Πίνακας Β.9. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο BHA.

Score	Title
2.342	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
2.334	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
1.743	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
1.677	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
1.519	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
0.881	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
0.7384	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
0.7267	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
0.723	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
0.491	Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-trees. D. Papadias, Y. Theodoridis, T. K. Sellis, M. J. Egenhofer

Πίνακας B.10. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο BSA.

Score	Title
1.485	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
1.377	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
1.264	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
1.009	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
0.7907	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
0.7243	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
0.5426	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
0.5329	Efficient Maintenance of Materialized Mediated Views. J. J. Lu, G. Moerkotte, J. Schü, V. S. Subrahmanian
0.4068	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
0.2959	Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-trees. D. Papadias, Y. Theodoridis, T. K. Sellis, M. J. Egenhofer

Πίνακας B.11. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο EPS.

Score	Title
4.082	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
3.905	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
2.791	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
2.596	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
2.487	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
2.263	A Critique of ANSI SQL Isolation Levels. H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, P. E. O'Neil
2.167	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
1.571	Fault Tolerant Design of Multimedia Servers. S. Berson, L. Golubchik, R. R. Muntz
1.386	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross
1.38	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin

Πίνακας B.12. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο BEPS.

Score	Title
4.657	An Effective Hash Based Algorithm for Mining Association Rules. J. S. Park, M. Chen, P. S. Yu
4.494	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom
3.349	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent
3.152	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin
3.071	Broadcast Disks: Data Management for Asymmetric Communications Environments. S. Acharya, R. Alonso, M. J. Franklin, S. B. Zdonik
2.781	A Critique of ANSI SQL Isolation Levels. H. Berenson, P. A. Bernstein, J. Gray, J. Melton, E. J. O'Neil, P. E. O'Neil
2.682	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y. E. Ioannidis, V. Poosala
2.107	Fault Tolerant Design of Multimedia Servers. S. Berson, L. Golubchik, R. R. Muntz
1.873	FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. C. Faloutsos, K. Lin
1.859	Adapting Materialized Views after Redefinitions. A. Gupta, I. S. Mumick, K. A. Ross

Πίνακας Β.13. Κατάταξη δημοσιεύσεων του SIGMOD'95 με τη μέθοδο `SCEAS_B1`.

Score	Title
48	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
47	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
39	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
38	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
37	Mining Generalized Association Rules. R. Srikant, R. Agrawal
28	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
24	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
20	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
19	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
16	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos

Πίνακας Β.14. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο `CC`.

Score	Title
5.877	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
5.511	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
5.014	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
3.919	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
3.703	Mining Generalized Association Rules. R. Srikant, R. Agrawal
3.086	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
2.603	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
2.12	Redo Recovery after System Crashes. D. B. Lomet, M. R. Tuttle
2.055	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
1.834	Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. L. Gravano, H. Garcia-Molina

Πίνακας B.15. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο BCC.

Score	Title
3.756	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
3.64	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
3.014	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
2.91	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
2.254	Mining Generalized Association Rules. R. Srikant, R. Agrawal
2.138	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
1.629	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
1.371	A Data Transformation System for Biological Data Sources. P. Buneman, S. B. Davidson, K. Hart, G. C. Overton, L. Wong
1.305	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
1.302	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson

Πίνακας B.16. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο PR.

Score	Title
1.163	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
1.013	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
0.7487	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos
0.7214	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
0.6064	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
0.5647	A Data Transformation System for Biological Data Sources. P. Buneman, S. B. Davidson, K. Hart, G. C. Overton, L. Wong
0.4842	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
0.4497	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
0.4336	The hBP-tree: A Modified hB-tree Supporting Concurrency, Recovery and Node Consolidation. G. Evangelidis, D. B. Lomet, B. Salzberg
0.409	Near Neighbor Search in Large Metric Spaces. S. Brin

Πίνακας B.17. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο HA.

Score	Title
5.575e-104	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
3.817e-144	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
3.022e-157	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
2.885e-157	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
1.411e-158	A Data Transformation System for Biological Data Sources. P. Buneman, S. B. Davidson, K. Hart, G. C. Overton, L. Wong
3.176e-159	Type Classification of Semi-Structured Documents. M. Tresch, N. Palmer, A. Luniewski
7.685e-161	Query Processing in Tertiary Memory Databases. S. Sarawagi
3.113e-161	Benchmarking Spatial Join Operations with Spatial Output. E. G. Hoel, H. Samet
2.286e-161	DB2 Query Parallelism: Staging and Implementation. Y. Wang
5.903e-167	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos

Πίνακας Β.18. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο P.

Score	Title
7.947	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
7.781	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
6.457	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
6.291	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
6.126	Mining Generalized Association Rules. R. Srikant, R. Agrawal
4.636	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
3.974	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
3.311	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
3.146	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
2.649	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos

Πίνακας Β.19. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο SA.

Score	Title
0.1373	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
0.06079	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
0.05258	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos
0.04252	High-Concurrency Locking in R-Trees. M. Kornacker, D. Banks
0.0419	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
0.03097	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
0.02972	A Data Transformation System for Biological Data Sources. P. Buneman, S. B. Davidson, K. Hart, G. C. Overton, L. Wong
0.02787	Type Classification of Semi-Structured Documents. M. Tresch, N. Palmer, A. Luniewski
0.01754	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
0.0174	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim

Πίνακας Β.20. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο BHA.

Score	Title
2.474	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
1.73	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
1.599	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
1.551	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
1.373	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
1.279	Mining Generalized Association Rules. R. Srikant, R. Agrawal
1.232	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
0.9245	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
0.8265	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
0.7978	Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension. A. Belussi, C. Faloutsos

Πίνακας B.21. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο BSA.

Score	Title
1.585	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
1.386	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
1.127	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
1.089	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
0.9529	Mining Generalized Association Rules. R. Srikant, R. Agrawal
0.8986	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
0.6357	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
0.5317	Eager Aggregation and Lazy Aggregation. W. P. Yan, P. Larson
0.5214	A Data Transformation System for Biological Data Sources. P. Buneman, S. B. Davidson, K. Hart, G. C. Overton, L. Wong
0.4555	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim

Πίνακας B.22. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο EPS.

Score	Title
4.83	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
4.747	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
4.451	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
3.464	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
3.131	Mining Generalized Association Rules. R. Srikant, R. Agrawal
3.124	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
2.338	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
1.597	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
1.551	Redo Recovery after System Crashes. D. B. Lomet, M. R. Tuttle
1.466	Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. L. Gravano, H. Garcia-Molina

Πίνακας B.23. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο BEPS.

Score	Title
5.416	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu
5.37	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass
5.1	W3QS: A Query System for the World-Wide Web. D. Konopnicki, O. Shmueli
3.825	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes
3.712	An Efficient Algorithm for Mining Association Rules in Large Databases. A. Savasere, E. Omiecinski, S. B. Navathe
3.692	Mining Generalized Association Rules. R. Srikant, R. Agrawal
2.903	Generalized Search Trees for Database Systems. J. M. Hellerstein, J. F. Naughton, A. Pfeffer
2.108	Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. R. Agrawal, K. Lin, H. S. Sawhney, K. Shim
2.075	Redo Recovery after System Crashes. D. B. Lomet, M. R. Tuttle
1.964	Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. L. Gravano, H. Garcia-Molina

Πίνακας Β.24. Κατάταξη δημοσιεύσεων του VLDB'95 με τη μέθοδο `SCEAS_B1`.

Year	Title	Pos	Points
	View Maintenance in a Warehousing Environment. Y. Zhuge, H. Garcia-Molina, J. Hammer, J. Widom	1	44
2005 (1995)	An Effective Hash Based Algorithm for Mining Association Rules. J.S. Park, M. Chen, P.S. Yu	2	40
	Nearest Neighbor Queries. N. Roussopoulos, S. Kelley, F. Vincent	3	36
	Incremental Maintenance of Views with Duplicates. T. Griffin, L. Libkin	4	19
	Balancing Histogram Optimality and Practicality for Query Result Size Estimation. Y.E. Ioannidis, V. Poosala	5	4
	Implementing Data Cubes Efficiently. V. Harinarayan, A. Rajaraman, J.D. Ullman	1	50
2006 (1996)	A Query Language and Optimization Techniques for Unstructured Data. P. Buneman, S.B. Davidson, G.G. Hillebrand, D. Suciu	2	40
	BIRCH: an Efficient Data Clustering Method for Very Large Databases. T. Zhang, R. Ramakrishnan, M. Livny	3	27
	Improved Histograms for Selectivity Estimation of Range Predicates. V. Poosala, Y.E. Ioannidis, P.J. Haas, E.J. Shekita	4	23
	Data Access for the Masses through OLE DB. J.A. Blakeley	5	5
	Online Aggregation. J.M. Hellerstein, P.J. Haas, H.J. Wang	1	50
	An Array-Based Algorithm for Simultaneous Multidimensional Ag-	2	40
2007 (1997)	gregates. Y. Zhao, P. Deshpande, J.F. Naughton		
	Improved Query Performance with Variant Indexes. P.E. O'Neil, D. Quass	3	22
	Dynamic Itemset Counting and Implication Rules for Market Basket Data. S. Brin, R. Motwani, J.D. Ullman, S. Tsur	4	11
	The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. N. Katayama, S. Satoh	5	10
	Catching the Boat with Strudel: Experiences with a Web-Site Management System. M.F. Fernandez, D. Florescu, J. Kang, A.Y. Levy,	1	45
2008 (1998)	D. Suciu		
	Your Mediators Need Data Conversion! S. Cluet, C. Delobel, J. Siméon, K. Smaga	2	35
	New Sampling-Based Summary Statistics for Improving Approximate Query Answers. P. B. Gibbons, Y. Matias	3	24
	Integrating Mining with Relational Database Systems: Alternatives and Implications. S. Sarawagi, S. Thomas, R. Agrawal	4	23
	Exploratory Mining and Pruning Optimizations of Constrained Association Rules. R.T. Ng, L.V.S. Lakshmanan, J. Han, A. Pang	5	8

Πίνακας B.25. *SIGMOD Test of Time Award*: πρόβλεψη για τα έτη 2005-2008 (1995-1998).

Year	Title	Pos	Points
	Aggregate-Query Processing in Data Warehousing Environments. A. Gupta, V. Harinarayan, D. Quass	1	46
2005 (1995)	Discovery of Multiple-Level Association Rules from Large Databases. J. Han, Y. Fu	2	35
	W3QS: a Query System for the World-Wide Web. D. Konopnicki, O. Shmueli	3	30
	Sampling-Based Estimation of the Number of Distinct Values of an Attribute. P.J. Haas, J.F. Naughton, S. Seshadri, L. Stokes	4	23
	Mining Generalized Association Rules. R. Srikant, R. Agrawal	5	8
	Querying Heterogeneous Information Sources Using Source Descriptions A.Y. Levy, A. Rajaraman, J.J. Ordille	1	47
2006 (1996)	On the Computation of Multidimensional Aggregates. S. Agarwal, R. Agrawal, P. Deshpande, A. Gupta, J.F. Naughton, R. Ramakrishnan, S. Sarawagi	2	39
	The X-tree : an Index Structure for High-Dimensional Data S. Berchtold, D.A. Keim, H. Kriegel	3	28
	Querying Multiple Features of Groups in Relational Databases. D. Chatziantoniou, K.A. Ross	4	16
	Answering Queries with Aggregation Using Views. D. Srivastava, S. Dar, H.V. Jagadish, A.Y. Levy	5	8
	Optimizing Queries across Diverse Data Sources. L.M. Haas, D. Kossmann, E.L. Wimmers, J. Yang	1	47
2007 (1997)	Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources. M.T. Roth, P.M. Schwarz	2	37
	DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. R. Goldman, J. Widom	3	35
	Selectivity Estimation Without the Attribute Value Independence Assumption. V. Poosala, Y.E. Ioannidis	4	20
	To Weave the Web. P. Atzeni, G. Mecca, P. Merialdo	5	4
	Optimal Histograms with Quality Guarantees. H.V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K.C. Sevcik, T. Suel	1	31
2008 (1998)	Hash Joins and Hash Teams in Microsoft SQL Server. G. Graefe, R. Bunker, S. Cooper	2	28
	Clustering Categorical Data: An Approach Based on Dynamical Systems. D. Gibson, J.M. Kleinberg, P. Raghavan	3	20
	A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. R. Weber, H. Schek, S. Blott	4	16
	Using Schema Matching to Simplify Heterogeneous Data Translation. T. Milo, S. Zohar	5	12

Πίνακας Β.26. *VLDB 10 Year Award*: πρόβλεψη για τα έτη 2005-2008 (1995-1998).

ΠΑΡΑΡΤΗΜΑ Γ

Συμπληρωματικά Πειράματα Κεφαλαίου 4

Γ.1 Αποτελέσματα Κατάταξης στον Παγκόσμιο Ιστό

Ενδεικτικά, σε αυτό το Παράρτημα παρουσιάζουμε τα αποτελέσματα κατάταξης για κάποια από τα ερωτήματα (queries) που ορίσθηκαν στο Κεφάλαιο 4. Πιο συγκεκριμένα, παρουσιάζουμε αναλυτικές λίστες με τις πρώτες 20 ιστοσελίδες ανά ερώτημα και ανά αλγόριθμο κατάταξης. Για λόγους οικονομίας χώρου, περιλλαμβάνουμε μόνο τους πίνακες που αντιστοιχούν στα ερωτήματα:

- "antonis sidiropoulos": Πίνακες Γ.1 - Γ.8
- "movies": Πίνακες Γ.9 - Γ.16
- "tsunami indian ocean": Πίνακες Γ.17 - Γ.24

Οι Πίνακες Γ.1, Γ.9 και Γ.17 περιέχουν την σειρά κατάταξης που δόθηκε από την Google για τα αντίστοιχα ερωτήματα.

Pos	URL
1	http://skyblue.csd.auth.gr/members/asidirop.html
2	http://delab.csd.auth.gr/~dimitris/dkatsaro.htm
3	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
4	http://aetos.it.teithe.gr/~asidirop/
5	http://users.auth.gr/~asidirop/
6	http://aetos.it.teithe.gr/~asidirop/submission/
7	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
8	http://www.robotstxt.org/wc/active/html/dienstspider.html
9	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
10	http://delab.csd.auth.gr/~asidirop/
11	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html
12	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
13	http://www.hostsun.com/gr/bots_index3.php
14	http://www.cs.kuleuven.ac.be/~dirk/ada-belgium/events/03/030612-wdas.html
15	http://www.iei.pi.cnr.it/DELOS/TOM/list.html
16	http://www.bitmechanic.com/mail-archives/dbi-users/Jan1999/author.html
17	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
18	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
19	http://www.faqchest.com/prgm/dbi-l/dbi-99/dbi-9901/dbi-990117/dbi99012818_28999.html
20	http://www.geocrawler.com/mail/msg.php3?msg_id=1070176

Πίνακας Γ.1. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον GOOGLE.

Score	URL
1.41e-08	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
6.49e-09	http://www.robotstxt.org/wc/active/html/dienstspider.html
6.49e-09	http://skyblue.csd.auth.gr/members/asidirop.html
6.49e-09	http://aetos.it.teithe.gr/~asidirop/
6.49e-09	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
6.49e-09	http://portal.acm.org/citation.cfm?id=1055766.1055774
6.49e-09	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
6.49e-09	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100
6.49e-09	http://groups.yahoo.com/group/gimpwin-users/message/215
6.49e-09	http://dke.cti.gr/SSTD03/officers.htm
6.49e-09	http://www.robotstxt.org/wc/active/db/dienstspider.txt
2.14e-10	http://www.hostsun.com/gr/bots_index3.php
1.32e-10	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
3.2e-11	http://www.spiders.pl/builder.php?which_page=baza_browse_details&cat=baza&robot_name=DienstSpider
2.5e-11	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
1.11e-11	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
1.55e-12	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html
1.11e-12	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
8.36e-13	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
6.44e-13	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos

Πίνακας Γ.2. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον HITS_A.

Score	URL
1.23	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.629	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.554	http://users.auth.gr/~asidirop/
0.476	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
0.423	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
0.292	http://www.csd.uch.gr/~sotirop/homepages.html
0.214	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
0.194	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
0.177	http://delab.csd.auth.gr/~dimitris/dkatsaro.htm
0.174	http://www.robotstxt.org/wc/active/html/dienstspider.html
0.169	http://skyblue.csd.auth.gr/members/asidirop.html
0.168	http://aetos.it.teithe.gr/~asidirop/submission/
0.166	http://www.iei.pi.cnr.it/DELOS/TOM/list.html
0.161	http://www.hostsun.com/gr/bots_index3.php
0.159	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
0.155	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html
0.152	http://aetos.it.teithe.gr/~asidirop/
0.152	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
0.152	http://portal.acm.org/citation.cfm?id=1055766.1055774
0.152	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf

Πίνακας Γ.3. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον PageRank.

Score	URL
1.11e-32	http://www.robotstxt.org/wc/active/html/dienstspider.html
1.36e-36	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
5.65e-39	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html
1.22e-39	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
6.34e-41	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
2.29e-41	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
9.3e-44	http://skyblue.csd.auth.gr/members/asidirop.html
9.3e-44	http://aetos.it.teithe.gr/~asidirop/
9.3e-44	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
9.3e-44	http://portal.acm.org/citation.cfm?id=1055766.1055774
9.3e-44	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
9.3e-44	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100
9.3e-44	http://groups.yahoo.com/group/gimpwin-users/message/215
9.3e-44	http://dke.cti.gr/SSTD03/officers.htm
9.3e-44	http://www.robotstxt.org/wc/active/db/dienstspider.txt
3.77e-46	http://users.auth.gr/~asidirop/
3.77e-46	http://aetos.it.teithe.gr/~asidirop/submission/
1.39e-299	http://www.hostsun.com/gr/bots_index3.php
0	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
0	http://www.spiders.pl/builder.php?which_page=baza_browse_details&cat=baza&robot_name=DienstSpider

Πίνακας Γ.4. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον Prestige.

Score	URL
0.00035	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.000287	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
9.03e-05	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
4.13e-05	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
1.7e-05	http://www.csd.uch.gr/~sotirop/homepages.html
1.53e-05	http://users.auth.gr/~asidirop/
8.57e-06	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
8.28e-06	http://www.robotstxt.org/wc/active/html/dienstspider.html
6.43e-06	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
5.09e-06	http://aetos.it.teithe.gr/~asidirop/submission/
3.9e-06	http://skyblue.csd.auth.gr/members/asidirop.html
2.96e-06	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Sidiropoulos:Antonis.html
2.17e-06	http://www.hostsun.com/gr/bots_index3.php
2.16e-06	http://www.spiders.pl/builder.php?which_page=baza_browse_details&cat=baza&robot_name=DienstSpider
2.02e-06	http://aetos.it.teithe.gr/~asidirop/
2.02e-06	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
2.02e-06	http://portal.acm.org/citation.cfm?id=1055766.1055774
2.02e-06	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
2.02e-06	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100
2.02e-06	http://groups.yahoo.com/group/gimpwin-users/message/215

Πίνακας Γ.5. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον SALSA_A.

Score	URL
0.343	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.307	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.245	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
0.243	http://users.auth.gr/~asidirop/
0.213	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
0.198	http://www.csd.uch.gr/~sotirop/homepages.html
0.171	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
0.164	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
0.16	http://delab.csd.auth.gr/~dimitris/dkatsaro.htm
0.157	http://aetos.it.teithe.gr/~asidirop/submission/
0.156	http://www.iei.pi.cnr.it/DELOS/TOM/list.html
0.155	http://skyblue.csd.auth.gr/members/asidirop.html
0.153	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
0.152	http://www.robotstxt.org/wc/active/html/dienstspider.html
0.152	http://www.hostsun.com/gr/bots_index3.php
0.151	http://aetos.it.teithe.gr/~asidirop/
0.151	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
0.151	http://portal.acm.org/citation.cfm?id=1055766.1055774
0.151	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
0.151	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100

Πίνακας Γ.6. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον SCEAS_D0.85B0.

Score	URL
1.59	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
1.26	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.763	http://users.auth.gr/~asidirop/
0.757	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
0.528	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
0.39	http://www.csd.uch.gr/~sotirop/homepages.html
0.173	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
0.123	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
0.086	http://delab.csd.auth.gr/~dimitris/dkatsaro.htm
0.0628	http://aetos.it.teithe.gr/~asidirop/submission/
0.056	http://www.iei.pi.cnr.it/DELOS/TOM/list.html
0.0513	http://skyblue.csd.auth.gr/members/asidirop.html
0.0346	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
0.0285	http://www.robotstxt.org/wc/active/html/dienstspider.html
0.0262	http://www.hostsun.com/gr/bots_index3.php
0.0155	http://aetos.it.teithe.gr/~asidirop/
0.0155	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
0.0155	http://portal.acm.org/citation.cfm?id=1055766.1055774
0.0155	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
0.0155	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100

Πίνακας Γ.7. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον SCEAS_D0.99.

Score	URL
1.59	http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Manolopoulos:Yannis.html
1.25	http://www.sigmod.org/dblp/db/indices/a-tree/m/Manolopoulos:Yannis.html
0.755	http://users.auth.gr/~asidirop/
0.748	http://www.sigmod.org/dblp/db/journals/ipm/ipm41.html
0.52	http://deslab.mit.edu/DesignLab/new_deslab/new-person.html
0.38	http://www.csd.uch.gr/~sotirop/homepages.html
0.164	http://www.mail-archive.com/wget@sunsite.dk/msg03302.html
0.113	http://www.mail-archive.com/wget@sunsite.dk/msg03396.html
0.076	http://delab.csd.auth.gr/~dimitris/dkatsaro.htm
0.0528	http://aetos.it.teithe.gr/~asidirop/submission/
0.046	http://www.iei.pi.cnr.it/DELOS/TOM/list.html
0.0414	http://skyblue.csd.auth.gr/members/asidirop.html
0.0246	http://citeseer.ist.psu.edu/cis?q=Antonis+Sidiropoulos
0.0187	http://www.robotstxt.org/wc/active/html/dienstspider.html
0.0163	http://www.hostsun.com/gr/bots_index3.php
0.00546	http://aetos.it.teithe.gr/~asidirop/
0.00546	http://ii.pmf.ukim.edu.mk/boi2000/teams.html
0.00546	http://portal.acm.org/citation.cfm?id=1055766.1055774
0.00546	http://www.ing.unlpam.edu.ar/laweb05/Full_Short_Accepted.pdf
0.00546	http://groups.yahoo.com/group/gimpwin-users/messages/201?viscount=100

Πίνακας Γ.8. Αναλυτικά Αποτελέσματα Κατάταξης του "antonis sidiropoulos" με βάση τον SCEASRank.

Pos	URL
1	http://movies.yahoo.com/
2	http://www.reel.com/
3	http://www.imdb.com/
4	http://www.lordoftherings.net/
5	http://www.brainpop.com/
6	http://www.hollywood.com/
7	http://www.romanm.ch/
8	http://www.mgm.com/
9	http://www.rottentomatoes.com/
10	http://dmoz.org/Arts/Movies/
11	http://movies.aol.com/
12	http://www.onwisconsin.com/movies/
13	http://www.apple.com/trailers/
14	http://www.hd.net/
15	http://www.foxmovies.com/index1.html
16	http://www.archive.org/details/movies
17	http://www.wmm.com/
18	http://filmforce.ign.com/
19	http://www.nytimes.com/ref/movies/1000best.html
20	http://www.killermovies.com/

Πίνακας Γ.9. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον GOOGLE.

Score	URL
0.00397	http://en.wikipedia.org/wiki/Film
0.000108	http://www.imdb.com/
5.26e-05	http://www.apple.com/trailers/
5.03e-05	http://www.rottentomatoes.com/
2.05e-05	http://www.ifilm.com/
2e-05	http://movies.yahoo.com/
1.38e-05	http://www.lordoftherings.net/
1.02e-05	http://dmoz.org/Arts/Movies/
6.59e-06	http://www.ucmp.berkeley.edu/geology/tectonics.html
4.09e-06	http://hitchhikers.movies.go.com/
3.87e-06	http://www.foxmovies.com/
3.78e-06	http://www.afi.com/tvevents/100years/movies.aspx
3.77e-06	http://www.mgm.com/
2.83e-06	http://www.allmovie.com/
2.28e-06	http://movies.go.com/
2.26e-06	http://www.metroactive.com/
2.23e-06	http://mirrors.creativecommons.org/
2.06e-06	http://movies.aurum3.com/
1.93e-06	http://www.hd.net/
1.83e-06	http://www.tnt.tv/

Πίνακας Γ.10. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον HITS_A.

Score	URL
89.9	http://www.imdb.com/
30.1	http://www.rottentomatoes.com/
17	http://www.apple.com/trailers/
11.3	http://www.rottentomatoes.com/movies/
9.46	http://filmforce.ign.com/
5.36	http://en.wikipedia.org/wiki/Film
5.16	http://www.allmovie.com/
3.77	http://dmoz.org/Arts/Movies/
3.34	http://movies.go.com/
2.72	http://www.chicagoreader.com/movies/
2.72	http://movies.aol.com/
2.41	http://www.triangle.com/movies/
2.38	http://www.fandango.com/
2.23	http://www.foxmovies.com/
2.2	http://movies.yahoo.com/
1.87	http://www.mgm.com/
1.83	http://www.startribune.com/movies/
1.82	http://movies.yahoo.com/mv/upcoming/
1.74	http://www.lordoftherings.net/
1.46	http://www.ifilm.com/

Πίνακας Γ.11. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον PageRank.

Score	URL
7.19e-10	http://www.imdb.com/
6.78e-10	http://www.rottentomatoes.com/
2.45e-12	http://www.hd.net/
1.83e-12	http://www.apple.com/trailers/
1.82e-12	http://www.ifilm.com/
1.22e-12	http://www.angryalien.com/
7.92e-13	http://www.rottentomatoes.com/movies/
6.5e-13	http://mirrors.creativecommons.org/
6.34e-13	http://www.metroactive.com/
6.2e-13	http://en.wikipedia.org/wiki/Film
6.12e-13	http://movies.yahoo.com/
6.04e-13	http://www.time.com/time/2005/100movies/the_complete_list.html
6.03e-13	http://www.time.com/time/2005/100movies/
6.01e-13	http://www.afi.com/tvevents/100years/movies.aspx
2e-14	http://www.nostalgiacentral.com/features/20moviethings.htm
8.86e-15	http://filmforce.ign.com/
6.81e-15	http://www.lordoftherings.net/
6.29e-15	http://movies.aol.com/
6.21e-15	http://rogerebert.suntimes.com/
5.56e-15	http://www.hollywood.com/

Πίνακας Γ.12. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον Prestige.

Score	URL
0.0414	http://www.imdb.com/
0.0162	http://www.rottentomatoes.com/
0.00701	http://www.rottentomatoes.com/movies/
0.0054	http://filmforce.ign.com/
0.00297	http://www.apple.com/trailers/
0.0025	http://en.wikipedia.org/wiki/Film
0.000861	http://www.ifilm.com/
0.000805	http://movies.yahoo.com/
0.000766	http://www.lordoftherings.net/
0.00074	http://www.allmovie.com/
0.000718	http://www.killermovies.com/
0.000697	http://movies.go.com/
0.00054	http://www.hollywood.com/
0.000521	http://www.badmovies.org/
0.000504	http://www.angryalien.com/
0.000391	http://www.fandango.com/
0.000368	http://www.pocketmovies.net/
0.000366	http://www.baltimoresun.com/entertainment/movies/
0.000365	http://www.movietickets.com/
0.000343	http://www.christianitytoday.com/movies/

Πίνακας Γ.13. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον SALSA_A.

Score	URL
21.6	http://www.imdb.com/
7.65	http://www.rottentomatoes.com/
3.37	http://www.rottentomatoes.com/movies/
3.29	http://www.apple.com/trailers/
2.48	http://filmforce.ign.com/
1.89	http://www.allmovie.com/
1.37	http://en.wikipedia.org/wiki/Film
0.962	http://movies.go.com/
0.75	http://www.fandango.com/
0.645	http://movies.yahoo.com/
0.593	http://movies.aol.com/
0.572	http://www.lordoftherings.net/
0.554	http://www.foxmovies.com/
0.51	http://movies.yahoo.com/mv/upcoming/
0.5	http://www.ifilm.com/
0.473	http://www.mgm.com/
0.455	http://www.killermovies.com/
0.412	http://www.hollywood.com/
0.408	http://dmoz.org/Arts/Movies/
0.387	http://www.angryalien.com/

Πίνακας Γ.14. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον SCEAS_D0.85B0.

Score	URL
174	http://www.imdb.com/
60.6	http://www.rottentomatoes.com/
26.1	http://www.apple.com/trailers/
25.7	http://www.rottentomatoes.com/movies/
18.8	http://filmforce.ign.com/
13.7	http://www.allmovie.com/
9.84	http://en.wikipedia.org/wiki/Film
6.55	http://movies.go.com/
4.84	http://www.fandango.com/
4	http://movies.yahoo.com/
3.71	http://movies.aol.com/
3.4	http://www.lordoftherings.net/
3.34	http://www.foxmovies.com/
2.95	http://movies.yahoo.com/mv/upcoming/
2.82	http://www.ifilm.com/
2.71	http://www.mgm.com/
2.47	http://www.killermovies.com/
2.33	http://dmoz.org/Arts/Movies/
2.12	http://www.hollywood.com/
1.93	http://www.angryalien.com/

Πίνακας Γ.15. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον SCEAS_D0.99.

Score	URL
174	http://www.imdb.com/
60.8	http://www.rottentomatoes.com/
26.2	http://www.apple.com/trailers/
25.8	http://www.rottentomatoes.com/movies/
18.9	http://filmforce.ign.com/
13.7	http://www.allmovie.com/
9.85	http://en.wikipedia.org/wiki/Film
6.55	http://movies.go.com/
4.84	http://www.fandango.com/
4	http://movies.yahoo.com/
3.72	http://movies.aol.com/
3.4	http://www.lordoftherings.net/
3.35	http://www.foxmovies.com/
2.95	http://movies.yahoo.com/mv/upcoming/
2.81	http://www.ifilm.com/
2.71	http://www.mgm.com/
2.46	http://www.killermovies.com/
2.34	http://dmoz.org/Arts/Movies/
2.12	http://www.hollywood.com/
1.93	http://www.angryalien.com/

Πίνακας Γ.16. Αναλυτικά Αποτελέσματα Κατάταξης του "movies" με βάση τον SCEAS-Rank.

Pos	URL
1	http://europa.eu.int/comm/press_room/presspacks/tsunami_asia/index_en.htm
2	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
3	http://www.noaa.gov/tsunamis.html
4	http://www.ngdc.noaa.gov/spotlight/tsunami/tsunami.html
5	http://www.pmel.noaa.gov/tsunami/sumatra20041226.html
6	http://www.usaid.gov/locations/asia_near_east/tsunami/
7	http://walrus.wr.usgs.gov/tsunami/srilanka05/
8	http://serc.carleton.edu/NAGTWorkshops/visualization/collections/tsunami.html
9	http://www.ngdc.noaa.gov/seg/hazard/tsuintro.shtml
10	http://wcatwc.arh.noaa.gov/IndianOSite/IndianO12-26-04.htm
11	http://www.prh.noaa.gov/ptwc/bulletins.htm
12	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
13	http://www.alertnet.org/thenews/emergency/SA_TID.htm
14	http://www.waveofdestruction.org/
15	http://www.firstgov.gov/Citizen/Topics/Asia_Tsunamis.shtml
16	http://www.asil.org/insights/2005/01/insight050118.htm
17	http://www.waxy.org/archive/2004/12/28/amateur_.shtml
18	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
19	http://www.lib.utexas.edu/maps/tsunami_2004.html
20	http://iri.columbia.edu/~lareef/tsunami/

Πίνακας Γ.17. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον GOOGLE.

Score	URL
0.000727	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
8.76e-05	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
7.22e-05	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
7.05e-05	http://www.waveofdestruction.org/
6.57e-05	http://www.guardian.co.uk/tsunami/
6.57e-05	http://news.bbc.co.uk/1/hi/in_depth/world/2004/asia_quake_disaster/default.stm
6.56e-05	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
6.53e-05	http://www.pmel.noaa.gov/tsunami/sumatra20041226.html
6.33e-05	http://wcatwc.arh.noaa.gov/IndianOSite/IndianO12-26-04.htm
6.25e-05	http://www.noaanews.noaa.gov/stories2004/s2358.htm
6.17e-05	http://www.ukabc.org/tsunamis.htm
1.77e-05	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
6.2e-06	http://en.wikinews.org/wiki/Tsunami_Help
5.44e-06	http://www.yusufislam.org.uk/
4.93e-06	http://www.ausaid.gov.au/hottopics/topic.cfm?Id=9562_2054_7529_7688_4864
4.84e-06	http://www.tsunamiassist.gov.au/
4.19e-06	http://dmoz.org/Science/Earth_Sciences/Geophysics/Earthquakes/Past_Earthquakes/Indian_Ocean_2004/Disaster_Relief_and_Recovery/
3.94e-06	http://dir.yahoo.com/Science/Earth_Sciences/Geology_and_Geophysics/Seismology/Historic_Earthquakes/Indian_Ocean__December_26__2004/Relief_Efforts/
2.46e-06	http://iri.columbia.edu/~lareef/tsunami/
2.28e-06	http://dmoz.org/Science/Earth_Sciences/Geophysics/Earthquakes/Past_Earthquakes/Indian_Ocean_2004/

Πίνακας Γ.18. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον HITS_A.

Score	URL
3.47	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
2.42	http://www.usaid.gov/locations/asia_near_east/tsunami/
1.97	http://ioc3.unesco.org/itic/
1.78	http://www.yusufislam.org.uk/
1.37	http://usinfo.state.gov/gi/global_issues/recovery.html
1.2	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
1.08	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
1.05	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
1.04	http://www.waveofdestruction.org/
0.807	http://www1.kaiho.mlit.go.jp/sumatra/index_e.html
0.766	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
0.703	http://www.prh.noaa.gov/ptwc/bulletins.htm
0.698	http://en.wikinews.org/wiki/Tsunami_Help
0.674	http://www.pbs.org/wgbh/nova/tsunami/
0.666	http://www.tsunamiassist.gov.au/
0.659	http://www.noaa.gov/tsunamis.html
0.644	http://www.international.ucla.edu/tsunami/
0.641	http://www.newscientist.com/channel/earth/tsunami
0.618	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
0.589	http://www.pata.org/patasite/index.php?id=1112

Πίνακας Γ.19. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον PageRank.

Score	URL
1.7e-09	http://www.usaid.gov/locations/asia_near_east/tsunami/
7.92e-12	http://www.firstgov.gov/Citizen/Topics/Asia_Tsunamis.shtml
1.75e-12	http://www.usaid.gov/our_work/humanitarian_assistance/disaster_assistance/countries/indian_ocean/et_index.html
7.97e-15	http://www.nature.com/news/specials/tsunami/
6.42e-15	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
3.5e-17	http://www.nature.com/news/2005/051121/full/051121-3.html
1.54e-17	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
9.22e-18	http://www.newscientist.com/channel/earth/tsunami
8.55e-18	http://www.penmachine.com/techie/learn_about_tsunamis_2005-01.html
8.48e-18	http://wcatwc.arh.noaa.gov/IndianOSite/IndianO12-26-04.htm
7.73e-18	http://www.pmel.noaa.gov/tsunami/sumatra20041226.html
6.86e-18	http://www.waveofdestruction.org/
6.82e-18	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
6.81e-18	http://www.guardian.co.uk/tsunami/
6.8e-18	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
6.74e-18	http://news.bbc.co.uk/1/hi/in_depth/world/2004/asia_quake_disaster/default.stm
6.71e-18	http://www.noaanews.noaa.gov/stories2004/s2358.htm
6.7e-18	http://www.ukabc.org/tsunamis.htm
5.56e-18	http://www.noaa.gov/tsunamis.html
3.53e-18	http://www.prh.noaa.gov/ptwc/bulletins.htm

Πίνακας Γ.20. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον Prestige.

Score	URL
0.00133	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
0.000603	http://www.waveofdestruction.org/
0.000577	http://www.yusufislam.org.uk/
0.000494	http://www.usaid.gov/locations/asia_near_east/tsunami/
0.000409	http://www.newscientist.com/channel/earth/tsunami
0.000366	http://usinfo.state.gov/gi/global_issues/recovery.html
0.000271	http://www.prh.noaa.gov/ptwc/bulletins.htm
0.000263	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
0.000243	http://en.wikinews.org/wiki/Tsunami_Help
0.000232	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
0.000183	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
0.000176	http://www.waxy.org/archive/2004/12/28/amateur_.shtml
0.000172	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
0.000163	http://blogs.sun.com/roller/page/MortazaviBlog?entry=numerical_simulation_of_indian_ocean
0.000152	http://iri.columbia.edu/~lareef/tsunami/
0.000137	http://www.livescience.com/forcesofnature/tsunami_special_report.html
0.000133	http://ioc3.unesco.org/itic/
0.000132	http://news.bbc.co.uk/1/hi/in_depth/world/2004/asia_quake_disaster/default.stm
0.000131	http://www.alertnet.org/thenews/emergency/SA_TID.htm
0.000124	http://www.noaanews.noaa.gov/stories2004/s2358.htm

Πίνακας Γ.21. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον SALSA_A.

Score	URL
0.968	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
0.664	http://www.usaid.gov/locations/asia_near_east/tsunami/
0.606	http://www.yusufislam.org.uk/
0.49	http://ioc3.unesco.org/itic/
0.479	http://usinfo.state.gov/gi/global_issues/recovery.html
0.404	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
0.399	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
0.352	http://www.waveofdestruction.org/
0.324	http://www.tsunamiassist.gov.au/
0.31	http://www.newscientist.com/channel/earth/tsunami
0.309	http://www.prh.noaa.gov/ptwc/bulletins.htm
0.307	http://en.wikinews.org/wiki/Tsunami_Help
0.304	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
0.289	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
0.286	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
0.284	http://www.pbs.org/wgbh/nova/tsunami/
0.281	http://www.noaa.gov/tsunamis.html
0.27	http://www.pata.org/patasite/index.php?id=1112
0.26	http://www.international.ucla.edu/tsunami/
0.258	http://www.noonsite.com/Members/doina/R2005-01-05-1

Πίνακας Γ.22. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον SCEAS_D0.85B0.

Score	URL
6.6	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
4.16	http://www.usaid.gov/locations/asia_near_east/tsunami/
3.66	http://www.yusufislam.org.uk/
2.84	http://ioc3.unesco.org/itic/
2.64	http://usinfo.state.gov/gi/global_issues/recovery.html
2.04	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
2	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
1.64	http://www.waveofdestruction.org/
1.39	http://www.tsunamiassist.gov.au/
1.28	http://www.newscientist.com/channel/earth/tsunami
1.28	http://www.prh.noaa.gov/ptwc/bulletins.htm
1.26	http://en.wikinews.org/wiki/Tsunami_Help
1.25	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
1.18	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
1.1	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
1.09	http://www.pbs.org/wgbh/nova/tsunami/
1.07	http://www.noaa.gov/tsunamis.html
0.973	http://www.pata.org/patasite/index.php?id=1112
0.917	http://www.international.ucla.edu/tsunami/
0.873	http://www.noonsite.com/Members/doina/R2005-01-05-1

Πίνακας Γ.23. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον SCEAS_D0.99.

Score	URL
6.61	http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
4.16	http://www.usaid.gov/locations/asia_near_east/tsunami/
3.66	http://www.yusufislam.org.uk/
2.84	http://ioc3.unesco.org/itic/
2.64	http://usinfo.state.gov/gi/global_issues/recovery.html
2.03	http://www.cnn.com/SPECIALS/2004/tsunami.disaster/
2	http://en.wikipedia.org/wiki/Donations_for_victims_of_the_2004_Indian_Ocean_earthquake
1.63	http://www.waveofdestruction.org/
1.38	http://www.tsunamiassist.gov.au/
1.27	http://www.prh.noaa.gov/ptwc/bulletins.htm
1.27	http://www.newscientist.com/channel/earth/tsunami
1.25	http://en.wikinews.org/wiki/Tsunami_Help
1.24	http://en.wikinews.org/wiki/2004_Indian_Ocean_Tsunami
1.17	http://staff.aist.go.jp/kenji.satake/Sumatra-E.html
1.09	http://www.crisp.nus.edu.sg/tsunami/tsunami.html
1.08	http://www.pbs.org/wgbh/nova/tsunami/
1.06	http://www.noaa.gov/tsunamis.html
0.965	http://www.pata.org/patasite/index.php?id=1112
0.911	http://www.international.ucla.edu/tsunami/
0.865	http://www.noonsite.com/Members/doina/R2005-01-05-1

Πίνακας Γ.24. Αναλυτικά Αποτελέσματα Κατάταξης του "tsunami indian ocean" με βάση τον SCEASRank.