ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Κατάταξη σε Γραφήματα Ιστού - Ranking on Web Graphs

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΝΑ ΣΤΕΦΑΝΗ

Α.Μ. 575

ΕΠΙΒΛΕΠΟΝΤΕΣ ΚΑΘΗΓΗΤΕΣ: Π.ΜΠΟΖΑΝΗΣ,

Δ.ΚΑΤΣΑΡΟΣ

Βόλος, Οκτώβριος 2013

Η σελίδα αυτή είναι σκόπιμα λευκή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Κατάταξη σε Γραφήματα Ιστού - Ranking on Web Graphs

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΝΑ ΣΤΕΦΑΝΗ

ΕΠΙΒΛΕΠΟΝΤΕΣ:

ΠΑΝΑΓΙΩΤΗΣ ΜΠΟΖΑΝΗΣ	ΔΗΜΗΤΡΙΟΣ ΚΑΤΣΑΡΟΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ	ΛΕΚΤΟΡΑΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Εγκρίθηκε από την διμελή εξεταστική επιτροπή την ...……..……..

(Υπογραφή**)**

...……..…..

ΠΑΝΑΓΙΩΤΗΣ ΜΠΟΖΑΝΗΣ

ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

(Υπογραφή)

...…….……..

ΔΗΜΗΤΡΙΟΣ ΚΑΤΣΑΡΟΣ

ΛΕΚΤΟΡΑΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

(Υπογραφή**)**

...…….……..

ΑΝΝΑ ΣΤΕΦΑΝΗ

Διπλωματούχος Μηχανικός Ηλεκτρονικών Υπολογιστών, Τηλεπικοινωνιών και Δικτύων του Τμήματος

Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της Διπλωματικής μου Εργασίας επιθυμώ να ευχαριστήσω από καρδιάς όλους εκείνους οι οποίοι συνέβαλλαν σ' αυτό. Ιδιαίτερα αισθάνομαι την ανάγκη να ευχαριστήσω θερμά τον πρώτο επιβλέποντα καθηγητή μου κ. Παναγιώτη Μποζάνη για την ηθική του στήριξη και τη συστηματική του καθοδήγηση κατά τη διάρκεια της προσπάθειάς μου αλλά και τον δεύτερο επιβλέποντα καθηγητή κ. Δημήτριο Κατσαρό για την εποικοδομητική του κριτική και τις δημιουργικές του παρατηρήσεις αναφορικά με τη βελτίωση της εργασίας μου.

Περίληψη

Κατά καιρούς έχουν υλοποιηθεί πολλοί αλγόριθμοι με βάση το θέμα κατάταξης ιστοσελίδων στις διάφορες μηχανές αναζήτησης. Ευρέως διαδεδομένοι είναι οι παραδοσιακοί αλγόριθμοι HITS [Kleinberg, 1997] , PageRank [Sergey Brin και Larry Page, 1998] και SALSA [Lempel και Moran, 2000] οι οποίοι βασίζονται σε link ανάλυση.

Οι αλγόριθμοι αυτοί παρουσιάζουν ορισμένα μειονεκτήματα γι' αυτό γίνεται προσπάθεια βελτιστοποίησης τους. Έτσι έχουμε παραλλαγές των παραδοσιακών αλγορίθμων ή την δημιουργία καινούριων. Παρουσιάστηκαν οι αλγόριθμοι Block Level PageRank και Block Level HITS. Με τους αλγορίθμους αυτούς μια ιστοσελίδα δεν αντιμετωπίζεται ως ενιαίος κόμβος αλλά χωρίζεται σε πολλά blocks.

Μια άλλη προσέγγιση με σκοπό την βελτιστοποίηση της ποιότητας της αναζήτησης στο διαδίκτυο, είναι η χρήση κοινωνικών σχολίων. Στη χρήση κοινωνικών σχολίων βασίζονται οι αλγόριθμοι Social Similarity Ranking και Social Page Ranking. Επιπρόσθετα στην εργασία αυτή γίνεται αναφορά στους αλγορίθμους ταξινόμησης LiftHits , Adaptive PageRank, NetRank, Laplacian Rank και Supervised Random Walk οι οποίοι αποτελούν ειδικές περιπτώσεις των προβλημάτων Framework. Τέλος γίνεται αναφορά σε πέντε είδη συστημάτων επεξεργασίας γράφων μεγάλης κλίμακας (Pegasus, Hama, Pregel, Trinity, Graphor) καθώς και σύγκριση τους.

Abstract

At times many algorithms have been implemented based on web ranking on various search engines. Widely popular are traditional algorithms HITS [Kleinberg, 1997], PageRank [Sergey Brin and Larry Page, 1998] and SALSA [Lempel and Moran, 2000] which are based on link analysis.

These algorithms have some drawbacks so, we attempt engine optimization. Thus, we have variations of traditional algorithms or creating new ones. Then, Block Level PageRank and Block Level HITS are presented. With these algorithms, a website is not treated as a single node but is divided into several blocks.

Another approach to optimize the quality of web search is the use of Social Annotations. Social Similarity Ranking and Social Page Ranking are based in using Social Annotations. In addition to, this work refers to the classification algorithm LiftHits, Adaptive PageRank, NetRank, Laplacian Rank as well as Supervised Random Walk, which are special cases of problems Framework. Finally, we report on five kinds of Large-Graph Processing Systems (Pegasus, Hama, Pregel, Trinity, Graphor) and we compare them.

Εισαγωγή

Ως γνωστόν το διαδίκτυο αποτελεί έναν θησαυρό πληροφοριών και υπηρεσιών που αφορούν μια ποικιλία θεμάτων και παρέχονται στους χρήστες. Το μέγεθος του ξεπερνά τα δύο δισεκατομμύρια καταχωρημένες σελίδες όπως έχουν δείξει τελευταίες μελέτες, ενώ συνεχίζει και αυξάνεται. Η ποσότητα και η ποικιλομορφία των διαθέσιμων δικτυακών τόπων έχουν ως αποτέλεσμα ο χρήστης να δυσκολεύεται να αποχτήσει την πληροφορία που χρειάζεται. Η ανάγκη για την επεξεργασία του τεράστιου όγκου δεδομένων είχε ως αποτέλεσμα τη δημιουργία μηχανών αναζήτησης. Με τη βοήθεια των μηχανών αναζήτησης ο χρήστης μπορεί να αναζητήσει και να βρίσκει σελίδες που τον ενδιαφέρουν. Σήμερα υπάρχουν στο διαδίκτυο περισσότερες από 200 μηχανές αναζήτησης. Ευρέως διαδεδομένες είναι οι εξής : Google, AltaVista, Lycos, LookSmart, Yahoo!, Msn Search, Ask Jeeves, In.gr. Ο χώρος του διαδικτύου αποτελεί την ιδανική πλατφόρμα για την προώθηση όχι μόνο των ιστοσελίδων, αλλά και των προϊόντων, των υπηρεσιών και των επιχειρήσεων.

Κατά την διαδικασία εύρεσης χρήσιμων πληροφοριών από τον χρήστη σε μια μηχανή αναζήτησης υπάρχουν χιλιάδες ιστοσελίδες που προσπαθούν να βγουν στο προσκήνιο, αλλά πολλές από αυτές δεν τα καταφέρνουν, καθώς η δημιουργία μιας καλαίσθητης και σωστά δομημένης ιστοσελίδας δε σημαίνει αυτόματα και επιτυχία. Η επιτυχία μιας ιστοσελίδας είναι πολύ πιθανό να εξασφαλιστεί μόνο από την κατάταξή της στη λίστα αποτελεσμάτων μιας μηχανής αναζήτησης (SERP – Search Engine Result Page). Γι' αυτό το λόγο η βελτιστοποίηση των ιστοσελίδων για τις μηχανές αναζήτησης αποτελεί μία σύγχρονη ανάγκη. Μέσα από αυτήν την εργασία επιχειρείται μια προσέγγιση στην έννοια του Μάρκετινγκ των μηχανών αναζήτησης ή αλλιώς Search Engine Marketing (SEM), μια μορφή Μάρκετινγκ Διαδικτύου που επιδιώκει να προωθήσει τις ιστοσελίδες αυξάνοντας την προβολή στις σελίδες των Μηχανών Αναζήτησης μέσα από ένα σύνολο εργαλείων που συμβάλλουν στην αύξηση της ποιότητας και του αριθμού των επισκεπτών. Εστιάζοντας σε ένα από τα εργαλεία του SEM και ειδικότερα στο Search Engine Optimization υπάρχουν διαδικασίες που πρέπει να γίνουν στη δομή και το περιεχόμενο μιας ιστοσελίδας (On Page SEO) αλλά και εκτός αυτής (Off Page SEO), ώστε να προετοιμαστεί για την αποδοχή των μηχανών αναζήτησης. Στην παρούσα πτυχιακή εργασία θα

παρουσιαστούν οι δημοφιλέστεροι παραδοσιακοί αλγόριθμοι algorithms HITS, PageRank και SALSA καθώς και βελτιστοποιήσεις τους.

Στη συνέχεια παρουσιάζονται κάποιοι επιπλέον βελτιστοποιημένοι αλγόριθμοι οι Social Similarity Ranking και Social Page Ranking, που βασίζονται στα σχόλια των χρηστών και οι Block Level PageRank και Block Level HITS, οι οποίοι δεν αντιμετωπίζουν την ιστοσελίδα σαν ενιαίο κόμβο αλλά ως μια διαίρεση από πολλά blocks. Οι αλγόριθμοι αυτοί αναλύονται ως προς την πολυπλοκότητά τους, την σύγκλισή τους και τα πλεονεκτήματα και μειονεκτήματά τους. Για κάθε αλγόριθμο υπάρχουν σχετικά παραδείγματα εφαρμογής του πάνω σε τυχαία γραφήματα για καλύτερη κατανόηση. Επιπρόσθετα σκοπός των αλγορίθμων αυτών είναι να προσελκύσουν αρχικά τις μηχανές αναζήτησης και κατ' επέκταση τον επισκέπτη που αναζητά πληροφορίες εκεί. Τέλος, γίνεται αναφορά στα πέντε είδη συστημάτων επεξεργασίας γράφων μεγάλης κλίμακας (Pegasus, Hama, Pregel, Trinity, Graphor) αλλά και στους αλγορίθμους κατάταξης που τα συστήματα αυτά υποστηρίζουν.

ΚΕΦΑΛΑΙΟ 1^{ο}

1.1 Τι είναι γράφημα

Γράφημα είναι οτιδήποτε μπορεί να αναπαρασταθεί με σημεία (κορυφές) και γραμμές (ακμές – κατευθυνόμενες ή μη)μεταξύ των σημείων. Τα γραφήματα διακρίνονται σε δύο ειδών κατευθυνόμενα ή μη. Ένα μη- κατευθυνόμενο γράφημα ή γράφος G είναι ένα διατεταγμένο ζεύγος G=(V,E) , όπου V={ u1, u2,…un} είναι το σύνολο των κορυφών του και E={e1,e2,…em} είναι το σύνολο των ακμών του. Κάθε ακμή είναι ένα διμελές σύνολο κορυφών e={u1,u2}, όχι απαραίτητα διαφορετικών μεταξύ τους. Στα κατευθυνόμενα γραφήματα κάθε ακμή είναι ένα διατεταγμένο ζεύγος κορυφών e=(v1,v2). Τα μεγέθη που χαρακτηρίζουν ένα γράφημα G(V,E) είναι ο αριθμός των κορυφών του, συνήθως συμβολίζεται με n ή|V| και ο αριθμός των ακμών του, συνήθως συμβολίζεται με m ή |E|. Η μη κατευθυνόμενη ακμή e={u1u2} λέμε ότι συνδέει τις κορυφές υ1 και u2 οι οποίες ονομάζονται και άκρα της. Η κατευθυνόμενη ακμή e=(u1,u2) λέμε ότι συνδέει την κορυφή u1 με την u2 . Η u1 ονομάζεται ουρά (ή αρχή) της ακμής e και η u2 ονομάζεται κεφαλή (ή τέλος)της e. Δύο κορυφές που συνδέονται με ακμή ονομάζονται γειτονικές. Μια ακμή που τα δύο άκρα της ταυτίζονται (ή η αρχή της ταυτίζεται με το τέλος της αν είναι κατευθυνόμενη) ονομάζεται βρόγχος (loop). Δύο ακμές με κοινά άκρα (ή κοινή αρχή και τέλος αν είναι κατευθυνόμενες) ονομάζονται παράλληλες. Επίσης ένα γράφημα ονομάζεται απλό όταν δεν έχει παράλληλες ακμές και βρόγχους.

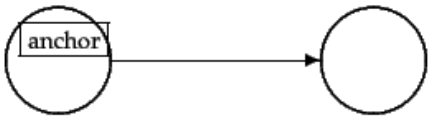
Στην εργασία αυτή γίνεται αναφορά στα ευρέως κλίμακας γραφήματα τα οποία χωρίζονται σε **Social graphs** (κοινωνικά γραφήματα) όπως είναι Messenger, Facebook, Twitter, Entity Cube σε **Endorsement graphs** όπως είναι Web link graph, Paper citation graph, σε **Location graphs** (τοπογραφικά γραφήματα) όπως είναι Map, Power grid, Telephone network και σε **Co-occurrence graphs** όπως είναι Term-document bipartite, Click-through bipartite. Τα γραφήματα αυτά είναι πολύ μεγάλα. Τα πλεονεκτήματα ενός ευρέως κλίμακας γραφήματος είναι ότι είναι πολύ αραιό , πλούσιο σε πληροφορίες όσον αφορά τους κόμβους και τις ακμές καθώς και

πλούσιο σε πληροφορίες. Στη συνέχεια δίνεται η περιγραφή ενός χαρακτηριστικού γραφήματος από κάθε κατηγορία.

1.2 Παραδείγματα Γραφημάτων Ευρέως Κλίμακας

1.2.1 Web link graph

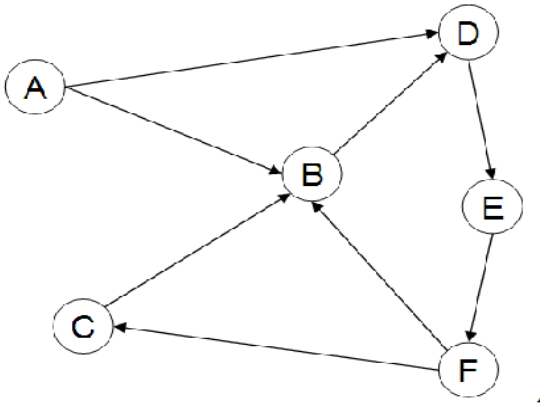
Μια στατική ιστοσελίδα αποτελείται από στατικές σελίδες HTMLοι οποίες μαζί με τις υπερ-συνδέσεις τους συνθέτουν ένα κατευθυνόμενο γράφημα στο οποίο κάθε ιστοσελίδα είναι ένας κόμβος και κάθε υπερσύνδεσμος μια κατευθυνόμενη ακμή. Στο παρακάτω Σχήμα 1 βλέπουμε δύο κόμβους ενός Web link graph. Κάθε κόμβος αντιστοιχεί σε μια ιστοσελίδα με μια υπερσύνδεση από το A στο B. Το σύνολο όλων αυτών των κατευθυνόμενων ακμών και των κόμβων ονομάζεται Web link graph. Συγκεκριμένα ο κόμβος A περιέχει κώδικα (anchor code) και υπάρχει μια υπερσύνδεση μετάβασης από τον κόμβο A στον κόμβο B. Το κατευθυνόμενο αυτό γράφημα δεν είναι άρρηκτα συνδεδεμένο δηλαδή υπάρχουν ζευγάρια σελίδων στα οποία δεν μπορούμε να μεταβούμε από τη μια σελίδα στην άλλη ακολουθώντας τους συνδέσμους (δηλαδή από τον κόμβο B δεν μπορούμε να πάμε στον A). Οι υπερσυνδέσεις σε μια σελίδα ονομάζονται in-links ενώ εκείνες που βρίσκονται εκτός σελίδας out-links . Ο αριθμός των in-links μετά από μελέτες είναι περίπου 8-15. (Christopher D. Manning, 2008 a)



Δύο κόμβοι του γραφήματος web που ενώνονται με ένα σύνδεσμο

Σχήμα1

Ένα άλλο παράδειγμα είναι ο παρακάτω Web link Graph (Σχήμα 2) ο οποίος έχει 6 σελίδες A-F. Η σελίδα B έχει in-links 3 και out-links 1. Το γράφημα αυτό δεν είναι άρρηκτα συνδεδεμένο δηλαδή δεν υπάρχει μονοπάτι από τις σελίδες B, F προς την σελίδα A.



Σχήμα2: Web link Graph

Γενικά ένα Web link graph αποτελείται από ένα ευρετήριο δεκάδων δισεκατομμυρίων κόμβων και πάνω από ένα τρισεκατομμύριο κόμβους οι οποίοι ανακαλύφτηκαν από τις κυριότερες μηχανές αναζήτησης .Κάθε επιμέρους κόμβος στο γράφο αυτό περιέχει πληροφορίες για το μήκος μιας σελίδας και το χρόνο της δημιουργίας της , ενώ κάθε ακμή περιέχει πληροφορίες για τον αριθμό των ακμών και για τις επιμέρους διασυνδέσεις.

1.2.2 Facebook

Το Facebook αποτελεί το πιο χαρακτηριστικό παράδειγμα ενός κοινωνικού γραφήματος. Όπως και όλα τα κοινωνικά δίκτυα έτσι και αυτό δεν είναι ένα ιεραρχημένο δίκτυο αλλά ένα δίκτυο από αλληλεξαρτώμενους, ίσους και ανοιχτούς κόμβους επικοινωνίας . Το Facebook προωθήθηκε τον Φεβρουάριο του 2004 και ιδρύθηκε από τον Mark Zuckerberg καθώς και από τους Eduardo Saverin , Andew McCollum, Dustin Moskovitz και τον Chris Hughes. Μέχρι το Σεπτέμβριο του 2012 το Facebook έχει πάνω από ένα δισεκατομμύριο ενεργά μέλη, τα μισά από τα οποία κάνουν χρήση του Facebook από κινητή συσκευή. Κάθε κόμβος (χρήστης) του κοινωνικού αυτού γραφήματος περιέχει πληροφορίες σχετικά με την ηλικία, τα ενδιαφέροντα και το φύλο των χρηστών ενώ οι ακμές περιέχουν πληροφορίες σχετικά με τη συχνότητα επικοινωνίας ή το χρόνο δημιουργίας του. Οι χρήστες πρέπει να συνδεθούν πριν χρησιμοποιήσουν το site, αφού αρχικά έχουν δημιουργήσει ένα προσωπικό προφίλ. Ο κάθε χρήστης έχει τη δυνατότητα να προσθέτει άλλους

χρήστες ως φίλους με τους οποίους να ανταλλάσει μηνύματα και αυτόματες κοινοποιήσεις όταν ενημερώνει το προφίλ του. Επιπρόσθετα οι χρήστες του Facebook μπορούν να προστεθούν σε ομάδες που τους ενδιαφέρουν. Σε αυτές τις ομάδες δυνατόν να περιλαμβάνονται μαθητές ενός σχολείου, εργαζόμενοι μιας εταιρίας, φοιτητές ενός τμήματος και γενικά άτομα με κοινά χαρακτηριστικά.

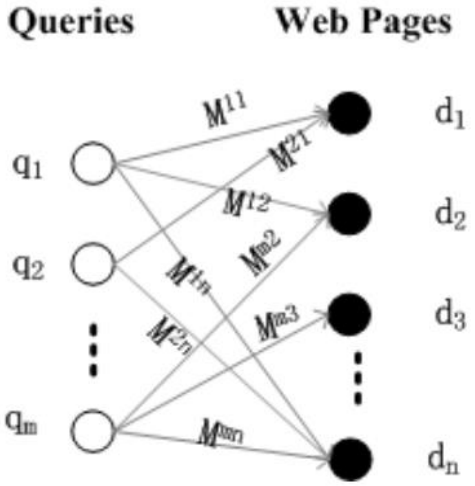
1.2.3 China Telephone Networks

Το τηλεφωνικό δίκτυο της Κίνας ανήκει στην κατηγορία των Τοπογραφικών δικτύων και αποτελείται από 1,1 δισεκατομμύρια κόμβους. Παρέχει εγχώριες και διεθνείς υπηρεσίες οι οποίες είναι όλο και περισσότερο διαθέσιμες για ιδιωτική χρήση. Το δίκτυο αυτό είναι άνισα κατανεμημένο στο εγχώριο σύστημα και εξυπηρετεί αρκετές πόλεις και βιομηχανικά κέντρα. Η Κίνα συνεχίζει την ανάπτυξη των τηλεπικοινωνιακών υποδομών και συνεργάζεται με ξένους παρόχους ώστε να επεκταθεί η εμβέλεια της σε παγκόσμιο επίπεδο. Η Κίνα το Καλοκαίρι του 2008 ξεκίνησε μια μεγάλη αναδιάρθρωση της τηλεπικοινωνιακής βιομηχανίας, με αποτέλεσμα την ενοποίηση των έξι φορέων παροχής υπηρεσιών της σε τρείς. Οι Υπηρεσίες αυτές είναι οι China Telecom, China Mobile and China Unicom οι οποίες παρέχουν σταθερή και κινητή τηλεφωνία. Οι κόμβοι του δικτύου αυτού περιέχουν πληροφορίες για την κατηγορία στην οποία προσφέρουν υπηρεσίες αλλά και για το προφίλ του πελάτη. Οι ακμές περιέχουν πληροφορία για την συχνότητα επικοινωνίας, το εύρος ζώνης και το είδος των κλήσεων.

1.2.4 Click-through Bipartite

Ως Click-through data ορίζεται ένα σύνολο Session, του οποίου συνόλου κάθε μέλος ορίζεται ως το ζεύγος «αίτηση» (query) και «ιστοσελίδα» (web page). Τα Click-through data προέρχονται από μια μεγάλη ποσότητα πληροφορίας που ενδέχεται να εμπεριέχει και άχρηστα δεδομένα όπως είναι εικόνες ή χειρόγραφα. Για την εξαγωγή των Click-through data (τα οποία είναι χρήσιμα για το χρήστη λόγω του συνόλου των πληροφοριών) χρησιμοποιούνται αλγορίθμοι που φιλτράρουν κάθε πληροφορία. Τα Click-through data μπορούν να μοντελοποιηθούν σε ένα κατευθυνόμενο γράφημα με βάρη το οποίο ονομάζεται Click-through Bipartite graph

G=(V,E) .Οι κόμβοι στο V αντιπροσωπεύουν τις ιστοσελίδες και τα αιτήματα. Οι ακμές E αντιπροσωπεύουν τα Click-through που δημιουργούνται από μια αίτηση σε μια ιστοσελίδα και τέλος το σύνολο M που αντιπροσωπεύει το βάρος των ακμών. Το σύνολο V διαιρείται σε δύο υποσύνολα στο Q={q1,q2,..qm} και στο D={d1,d2,…dn} όπου το Q αντιπροσωπεύει τις αιτήσεις και το D τις ιστοσελίδες. Στο γράφημα αυτό αντιστοιχίζονται δύο ή περισσότερες αιτήσεις του χρήστη σε μία ιστοσελίδα. Για παράδειγμα στο Σχήμα 3 οι αιτήσεις q1 και qm αντιστοιχίζονται στην ιστοσελίδα d2. (Wei Wu κ. συν., 2011 a).



Σχήμα3: Click-through Bipartite graph

Τα Click-through data μπορούν να προέλθουν από μια μεγάλη έρευνα και είναι αποτέλεσμα των μηχανών αναζήτησης ιστού. Ο χρήστης αρχικά υποβάλλει μια ερώτηση στη μηχανή αναζήτησης. Από την έρευνα αυτή προκύπτουν οι URL ιστοσελίδες, οι οποίες είναι σχετικές με την ερώτηση .Τις ιστοσελίδες αυτές μπορεί να τις ανοίξει ο χρήστης «κλικάροντας» πάνω αυτές εξού και Click-through data. Αν ένας χρήστης κάνει κλικ σε μια ιστοσελίδα είναι πιθανό αυτή να είναι σχετική ή ως ένα βαθμό σχετική με την αίτηση του χρήστη. Ένα Click-through Bipartite graph περιλαμβάνει αρκετές δισεκατομμύρια αιτήσεις και δεκάδες δισεκατομμύρια διευθύνσεις URL . (Wei Wu κ. συν., 2011 b).

1.3 Ranking on Large-scale Graph (Ταξινόμηση σε μεγάλης κλίμακας γράφημα)

Καθορισμός προβλήματος:

Λαμβάνοντας υπόψη έναν μεγάλης κλίμακας κατευθυνόμενο γράφο πλούσιο σε κόμβους (πληροφορίες) γίνεται προσπάθεια ταξινόμισης των κόμβων του γράφου ως προς την δημοτικότητα ,την σημαντικότητα ή την προτίμηση.

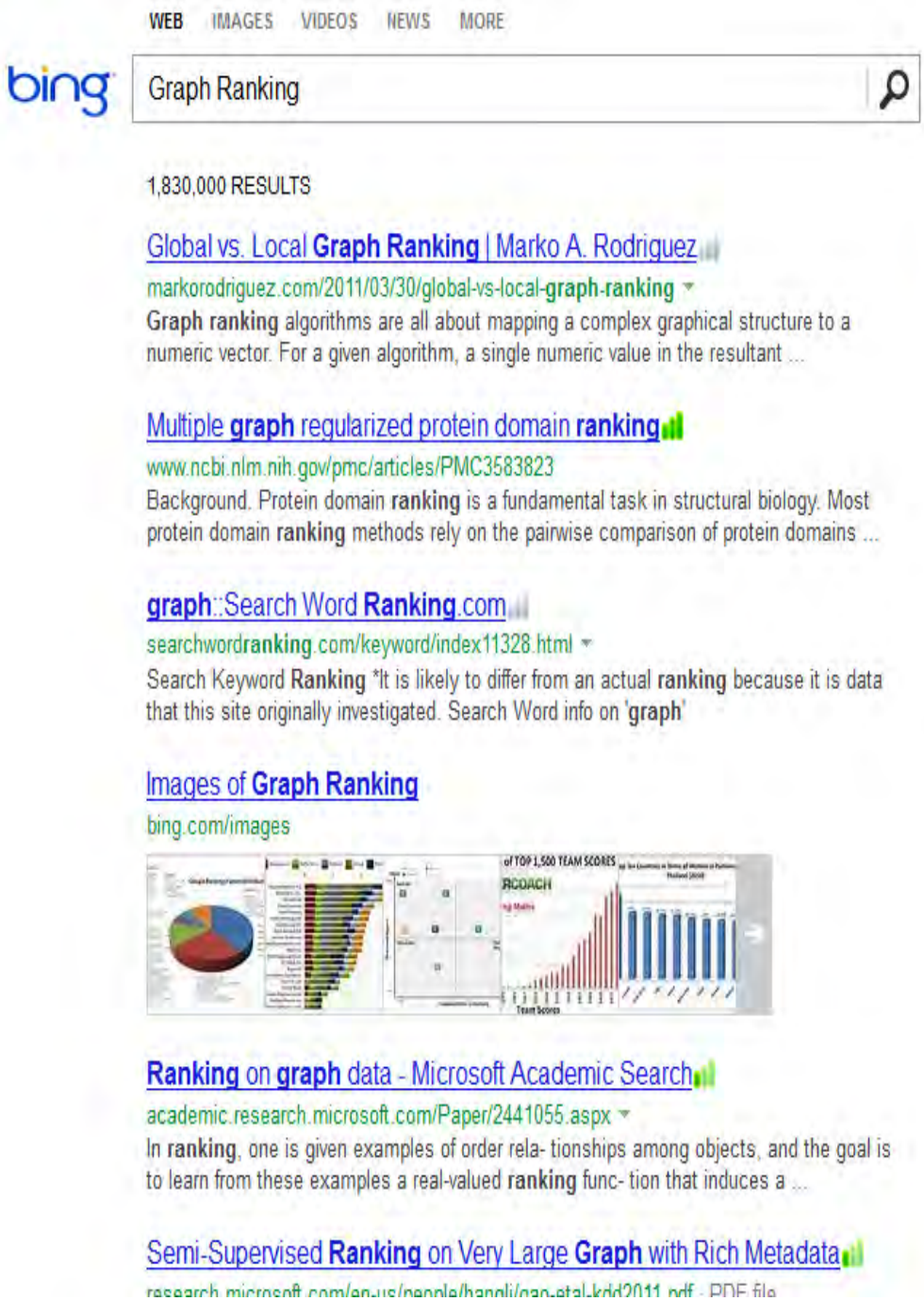
Εφαρμογές:

- Webpage ranking
- Paper ranking
- Entity ranking in social networks

1.3.1 Webpage ranking

Οι παράγοντες που λαμβάνονται υπόψη για το Webpage ranking είναι οι εξής:

- Η ποιότητα της ιστοσελίδας
- Η συχνότητα επίσκεψης από τους χρήστες
- Ο χρόνος παραμονής του χρήστη σε μια ιστοσελίδα
- Η αμοιβαία υποστήριξη μεταξύ των σελίδων (Bin Gao κ.συν. 2011a)

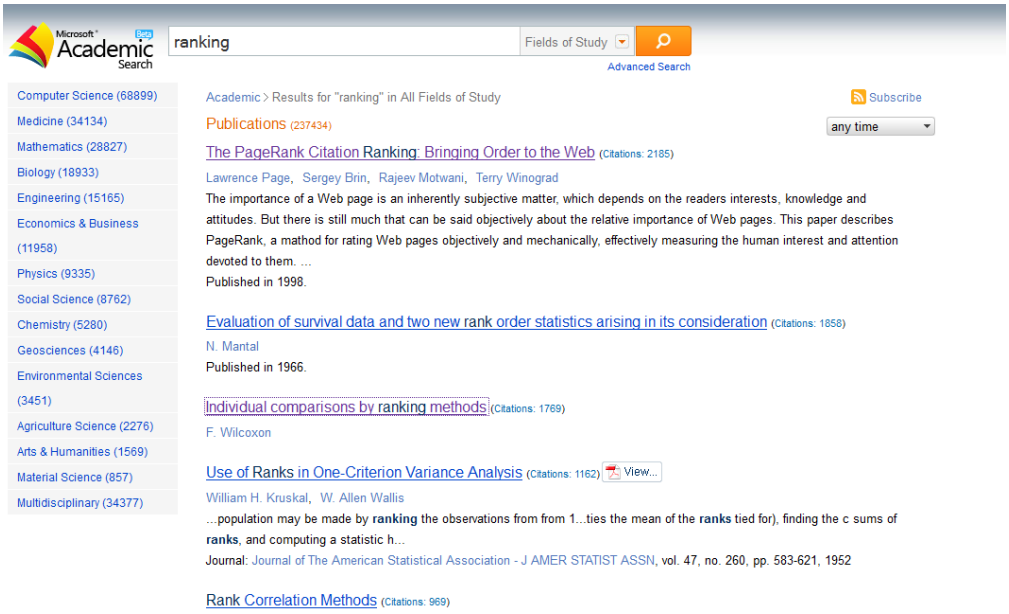

WEB IMAGES VIDEOS NEWS MORE
bing
Graph Ranking
1,830,000 RESULTS
Global vs. Local Graph Ranking | Marko A. Rodriguez
markorodriguez.com/2011/03/30/global-vs-local-graph-ranking
Graph ranking algorithms are all about mapping a complex graphical structure to a numeric vector. For a given algorithm, a single numeric value in the resultant ...
Multiple graph regularized protein domain ranking
www.ncbi.nlm.nih.gov/pmc/articles/PMC3583823
Background. Protein domain ranking is a fundamental task in structural biology. Most protein domain ranking methods rely on the pairwise comparison of protein domains ...
graph::Search Word Ranking.com
searchwordranking.com/keyword/index11328.html
Search Keyword Ranking *It is likely to differ from an actual ranking because it is data that this site originally investigated. Search Word info on 'graph'
Images of Graph Ranking
bing.com/images
Ranking on graph data - Microsoft Academic Search
academic.research.microsoft.com/Paper/2441055.aspx
In ranking, one is given examples of order rela- tionships among objects, and the goal is to learn from these examples a real-valued ranking func- tion that induces a ...
Semi-Supervised Ranking on Very Large Graph with Rich Metadata
RELATED SEARCHES
Tiger Woods Ranking Graph

Σχήμα 4

1.3.2 Paper ranking

Οι παράγοντες που λαμβάνονται υπόψη για το Paper ranking είναι οι εξής:

- Citation
- Συγγραφείς
- Publication venue
- Βραβεία
- Μέρα δημοσίευσης (Bin Gao κ.συν. 2011 b)



Σχήμα 5

1.3.3 Social Entity ranking

Οι παράγοντες που λαμβάνονται υπόψη για το Social Entity ranking είναι οι εξής:

- Δραστηριότητα του λογαριασμού
- Liked or followed από άλλους
- Χρόνος δημιουργίας λογαριασμού
- Friends related (Bin Gao, κ.συν 2011 c)

ΚΕΦΑΛΑΙΟ 2ο

Διάσημοι Παραδοσιακοί Αλγόριθμοι που βασίζονται σε υπερσυνδέσεις

2.1 Αλγόριθμος HITS

Ένας πολύ γνωστός αλγόριθμος, ο HITS (Hypertext Induced Topic Search) ή αλλιώς ο Αλγόριθμος Αναζήτησης Θεματικής ο οποίος προέκυψε από τα Υπερκείμενα παρουσιάστηκε και αναπτύχτηκε από τον Kleinberg το 1998. Ο Kleinberg πρότεινε ότι η σημαντικότητα μιας ιστοσελίδας θα έπρεπε να εξαρτάται από την εκάστοτε επερώτηση του χρήστη πράγμα το οποίο αποτελεί και αδυναμία του αλγορίθμου (query-dependent). Όρισε επίσης ότι για κάθε ιστοσελίδα θα πρέπει να υπάρχουν δύο ξεχωριστές τιμές οι οποίες προσδιορίζουν την αξία της και οι οποίες είναι το **authority ranking ή authority score (αυθεντία) και το hub ranking ή hub score (ομφαλός).** Το **authority ranking** αναφέρεται στους συνδέσμους που εισέρχονται προς την ιστοσελίδα, από άλλες ιστοσελίδες του διαδικτύου ενώ το **hub ranking** αναφέρεται στους συνδέσμους που εξέρχονται από την ιστοσελίδα. Ειδικότερα μια ιστοσελίδα ή ένα έγγραφο χαρακτηρίζεται ως authority ή ως hub όταν έχει πολλούς εισερχόμενους συνδέσμους από άλλες ιστοσελίδες ή πολλούς εξερχόμενους συνδέσμους προς άλλες ιστοσελίδες αντίστοιχα. Διαισθητικά μπορούμε να πούμε ότι οι καλές authorities δείχνονται από καλά hubs και τα καλά hubs δείχνουν σε καλές authorities. (Παπαδόπουλος, 2008a).

Σε κάθε ιστοσελίδα-κόμβο ανατίθεται ένα authority score a και ένα hub score h. Πιο συγκεκριμένα για τον κόμβο vi το authority score ορίζεται ως a(vi) και το hub score h(vi). Κάθε σελίδα i θεωρείται ότι διαθέτει ένα βαθμό αυθεντίας xi και έναν βαθμό ομφαλού yi (Kleinberg, 1997a).

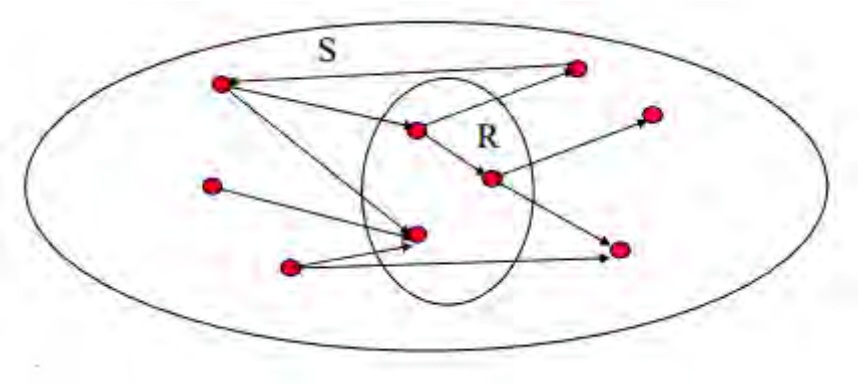
Ο αλγόριθμος HITS υπολογίζει επαναληπτικά τις ποσότητες :

$$x_i^k = \sum_{j:e_{ij}\in E} y_j^{(k-1)} \quad \text{και} \quad y_i^{(k)} = \sum_{j:e_{ij}\in E} x_j^{(k)} \quad \text{για} \quad k = 1, 2, 3, \ldots$$

2.1.1 Εκτέλεση του αλγορίθμου HITS

Σύμφωνα με τον Kleinberg (1997 b) η εκτέλεση του αλγορίθμου HITS περιλαμβάνει τα εξής βήματα. Αρχικά επιλέγεται το σύνολο των σελίδων που περιστρέφονται με τη βοήθεια μιας μηχανής αναζήτησης γύρω από ένα συγκεκριμένο θέμα που προσδιορίζεται από μια επερώτηση Q. Το σύνολο αυτό ονομάζεται R (Root Set). Το Root set βρίσκεται με τη βοήθεια μιας μηχανής αναζήτησης keyword-search. Επιπρόσθετα ως Base Set (S) ορίζεται ο Neighborhood Graph που δημιουργείται σταδιακά.

1) Αρχικοποιούμε S:=R
2) Προσθέτουμε στο S όλες τις σελίδες που δείχνονται από σελίδες του R.
3) Προσθέτουμε στο S όλες τις σελίδες που δείχνουν σε σελίδες του R.



Σχήμα 6: Επέκταση του αρχικού συνόλου σελίδων (Παπαδόπουλος,2008)

Ακόμη και μέσα στο σύνολο S, οι κόμβοι (σελίδες) με μεγάλο βαθμό εισόδου δεν είναι κατ' ανάγκη authorities. Μπορεί απλά να είναι δημοφιλείς σελίδες όπως Amazon ή το yahoo. Για να βρούμε τα πραγματικά authorities θα πρέπει να δούμε από πόσα hubs δείχνεται η κάθε σελίδα. Για να μην μεγαλώσει σε μέγεθος το Base Set που δημιουργείται σταδιακά ορίζεται ένας μέγιστος αριθμός κόμβων (με εισερχόμενους/εξερχόμενους) υπερσυνδέσμους για τον κάθε κόμβο του Root Set τους οποίους ενσωματώνουμε στο Base Set. Από τις σελίδες που ανήκουν στο σύνολο S προκύπτει ένα γράφημα το G[S], στον οποίο κόμβοι είναι οι σελίδες και ακμές είναι οι διασυνδέσεις που συνδέουν τις σελίδες του συνόλου. Στις σελίδες υπάρχουν

διασυνδέσεις οι οποίες δεν μεταφέρουν κάποια σημαντική πληροφορία, αλλά απλά διευκολύνουν την πλοήγηση του χρήστη. Έτσι προτείνεται μία Ευρεστική μέθοδος, η οποία σκοπό έχει να αντισταθμίσει το αποτέλεσμα των διασυνδέσεων αυτών. (Ding κ. συν., 2001a).

Οι ακμές που υπάρχουν στο γράφημα G[S] χωρίζονται σε δύο κατηγορίες. Μία ακμή χαρακτηρίζεται εγκάρσια (transverse) εάν συνδέει δύο σελίδες οι οποίες ανήκουν σε διαφορετικό domain, και φυσική (intrinsic) εάν συνδέει δύο σελίδες που βρίσκονται στο ίδιο domain. Το domain είναι το πρώτο επίπεδο της διεύθυνσης, η οποία σχετίζεται με κάποια σελίδα. Επειδή οι φυσικές ακμές είναι αυτές που διευκολύνουν την πλοήγηση μέσα σε έναν διαδικτυακό κόμβο προκύπτει ότι μεταφέρουν πολύ λιγότερη πληροφορία για τη σπουδαιότητα και την ποιότητα της σελίδας προς την οποία δείχνουν, σε σχέση με τις εγκάρσιες ακμές. Για αυτό το λόγο οι φυσικές ακμές του γραφήματος αφαιρούνται με αποτέλεσμα να μένουν σε αυτόν μόνο οι εγκάρσιες ακμές. Το γράφημα που προκύπτει τελικά είναι το G. Η μέθοδος αυτή της διαγραφής των φυσικών ακμών, είναι μεν ιδιαίτερα απλή, όμως είναι αποτελεσματική. (Kleinberg, 1997 c).

Ο αλγόριθμος αυτός στη συνέχεια υπολογίζει τα βάρη των Hubs και των Authorities. Το γράφημα G που έχει δημιουργηθεί περιέχει πολλές σχετικές με ένα ερώτημα του χρήστη σελίδες και αρκετά σημαντικές. Αυτό που χρειάζεται στη συνέχεια είναι να βρεθούν αυτές οι σημαντικές σελίδες, αναλύοντας τη δομή των ακμών του γραφήματος αυτού. Μία απλή προσέγγιση είναι η ταξινόμηση των σελίδων βάση του in-degree, του αριθμού δηλαδή των ακμών που δείχνουν στη συγκεκριμένη σελίδα. Η ιδέα αυτή είχε απορριφθεί για το σύνολο όλων των σελίδων που περιέχουν το συγκεκριμένο ερώτημα του χρήστη. Σε αυτή τη φάση το γράφημα που έχει δημιουργηθεί είναι χαρακτηριστικά μικρότερο και όμως οι σελίδες που περιέχει είναι σημαντικότερες για το χρήστη. (Ding κ. συν., 2001b).

Παρόλο που αυτή η προσέγγιση δίνει καλύτερα αποτελέσματα για το γράφημα από ότι για το σύνολο όλων των σελίδων εντούτοις αν εφαρμοστεί στο γράφημα μπορεί να δημιουργήσει και σημαντικά προβλήματα. Αυτό συμβαίνει γιατί δεν διαχωρίζει τις σημαντικές σελίδες σε σχέση με το ερώτημα του χρήστη από τις γενικότερα δημοφιλείς σελίδες, καθώς και οι δύο αυτοί τύποι σελίδων έχουν μεγάλο in-degree. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με την παρατήρηση ότι οι authoritative σελίδες που είναι σχετικές με το ερώτημα του χρήστη δεν απαιτείται να

έχουν μόνο μεγάλο in-degree, αλλά και να έχουν αρκετά κοινά χαρακτηριστικά με τα σύνολα των σελίδων που δείχνουν προς αυτές. Επομένως εκτός από τις authoritative σελίδες θα πρέπει να προσδιοριστούν και οι hub σελίδες, οι οποίες έχουν διασυνδέσεις προς τις authoritative σελίδες. Οι σελίδες αυτές συνενώνουν κατά κάποιο τρόπο τις authorities σε ένα κοινό θέμα, αγνοώντας σελίδες που απλά έχουν μεγάλο in-degree. (Παπαδόπουλος, 2008b).

Ο επαναληπτικός αλγόριθμος που θα περιγραφεί στη συνέχεια, και ο οποίος υπολογίζει και ενημερώνει τα βάρη των τιμών hub και authority για κάθε σελίδα, εκμεταλλεύεται αυτή την αμοιβαία σχέση των hubs και authorities σελίδων. (Ding κ. συν., 2001c).

2.1.2 Επαναληπτικός Αλγόριθμος HITS

- Για κάθε σελίδα p που ανήκει στο σύνολο S
  - Authority score: a(p) (πίνακας a)
  - Hub score: h(p) (πίνακας h)
- Αρχικοποίηση a(p)=h(p)=1
- Οπότε τα αρχικά scores θα είναι:

$$\sum_{p \in S} a(p)^2 = 1 \qquad \sum_{p \in S} h(p)^2 = 1$$

Δύο Κανόνες ενημέρωσης σε κάθε επανάληψη:

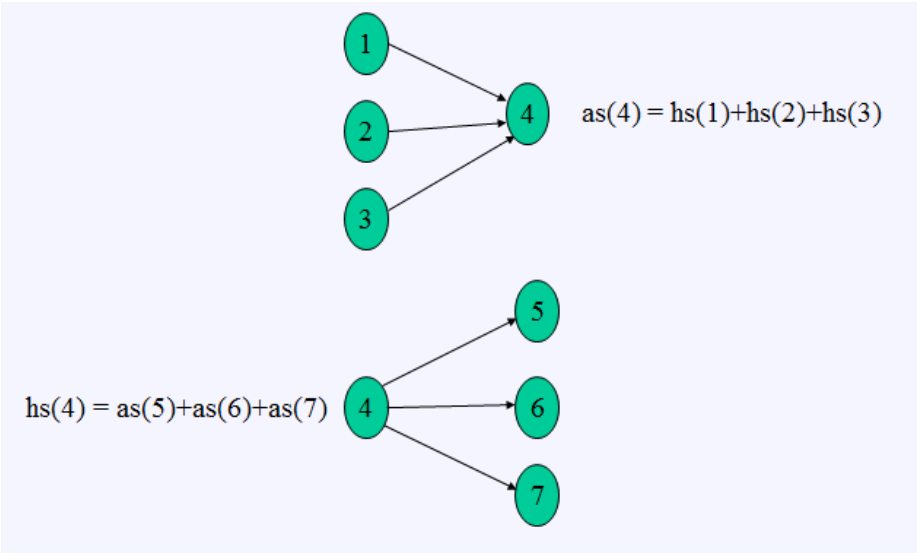
- Αυθεντίες δείχνονται από πολλούς καλούς ομφαλούς

$$a(p) = \sum_{q \in in(p)} h(q)$$

- Ομφαλοί δείχνουν σε πολύ καλές αυθεντίες

$$h(p) = \sum_{q \in out(p)} a(q)$$

Τους δύο αυτούς κανόνες τους βλέπουμε στο Σχήμα 5.



Σχήμα 7

2.1.3 Σε επίπεδο Γραμμικής Άλγεβρας- Βασικός Επαναληπτικός Αλγόριθμος HITS

Θεωρείται ο πίνακας γειτνίασης **Lij** με στοιχεία ίσα με 1, εάν υπάρχει υπερσύνδεσμος από τη σελίδα i στην j και ίσα με 0 στην άλλη περίπτωση. Οπότε οι εξισώσεις του σχήματος 2 μπορούνε να γραφτούνε ως :

$$\mathbf{x}^{(k)} = \mathbf{L}^T \mathbf{y}^{(k-1)}$$
$$\mathbf{y}^{(k)} = \mathbf{L}\mathbf{x}^{(k)},$$

Όπου x και y είναι n* 1 διανύσματα με τους κατά προσέγγιση βαθμούς αυθεντίας και ομφαλού σε κάθε επανάληψη. Αυτό μας οδηγεί στον Βασικό Επαναληπτικό Αλγόριθμο HITS

1) Αρχικοποίηση: $y^{(0)}$= e Όπου e είναι ένα μοναδιαίο διάνυσμα στήλης.

2) Μέχρι να επιτευχθεί η σύγκλιση επανέλαβε:

$x^{(k)}= L^T y^{(k-1)}$ (1.1)

$y^{(k)}= L\, x^{(k)}$ (1.2)

$k=k + 1$

Κανονικοποιήστε τα $x^{(k)}$ $y^{(k)}$

Οι δύο παραπάνω εξισώσεις (1.1) και (1.2) μπορούν να απλοποιηθούν στις επόμενες:

$x^{(k)}= L^T L\, x^{(k-1)}$
$y^{(k)}= L\ L^T y^{(k)}$

Έτσι ορίζεται η επαναληπτική power method για τον υπολογισμό των κυρίαρχων ιδιοδιανυσμάτων των πινάκων $L^T L$ και $L\ L^T$ η οποία είναι παρόμοια με τον υπολογισμό του PageRank. Εντούτοις χρησιμοποιείται διαφορετικός πίνακας συντελεστών. Ο πρώτος πίνακας λέγεται πίνακας authority αφού καθορίζει τα authority scores και ο δεύτερος λέγεται πίνακας hub αφού καθορίζει τα hub scores. Και οι δύο πίνακες είναι συμμετρικοί, θετικοί και Semidefinite. Δεν χρειάζεται να ηλοποιηθεί το κυρίαρχο ιδιοδυάνυσμα για τον πίνακα $\mathbf{L^T L}$ και για τον $\mathbf{L\ L^T}$.Μόνο για τον ένα πίνακα αφού ισχύει: **y=L x (**Ayman Farahat κ.συν. 2006a & Amy N. Langville and Carl D. Meyer ,2004 a**).**

2.1.4 Σύγκλιση του αλγορίθμου HITS

Ο επαναληπτικός αλγόριθμος που χρησιμοποιείται για τον υπολογισμό των διανυσμάτων authorities και hubs είναι ουσιαστικά η μέθοδος δύναμης εφαρμοσμένη στους $\mathbf{L^T L}$ και $\mathbf{L\ L^T}$. Για κάποιο διαγωνιοποιήσιμο πίνακα B n *n του οποίου οι διακριτές ιδιοτιμές είναι {λ1,λ2,…λk} τέτοιες ώστε |λ1|> |λ2|>= |λ3|… >=|λk|, η μέθοδος δύναμης χρησιμοποιώντας ένα αρχικό διάνυσμα $x^{(0)}$ υπολογίζει

$$\mathbf{x}^{(k)} = \mathbf{B}\mathbf{x}^{k-1}, \quad \mathbf{x} \leftarrow \frac{\mathbf{x}^{(k)}}{m(\mathbf{x}^{(k)})},$$

Όπου $m(x^{(k)})$ είναι σταθερά κανονικοποίησης παραγόμενη από το x. Το $m(x^{(k)})$ χρησιμοποιείται ως η προσημασμένη συνιστώσα μέγιστου μεγέθους οπότε συγκλίνει στην πρωτεύουσα ιδιοτιμή λ1, ενώ το $x^{(k)}$ συγκλίνει σε κάποιο συσχετισμένο κανονικοποιημένο ιδιοδιάνυσμα. Στην περίπτωση που απαιτείται μόνο το πρωτεύον ιδιοδιάνυσμα και όχι η ιδιοτιμή λ1, τότε μπορεί να χρησιμοποιηθεί μια κανονικοποίηση της μορφής $m(x^{(k)})=||x^{(k)}||$. Εάν λ1<0 τότε η $m(x^{(k)})=|| x^{(k)}||$ δεν μπορεί να συγκλίνει στην λ1 αλλά το $x^{(k)}$ εξακολουθεί να συγκλίνει σε ένα κανονικοποιημένο ιδιοδιάνυσμα συσχετισμένο με τη λ1. Ο ασυμπωτικός ρυθμός σύγκλισης της μεθόδου δύναμης είναι ο ρυθμός στον οποίο $(|\lambda 2(B)|/ |\lambda 1(B)|)^{\kappa} \to 0$ (Peseric κ.συν., 2009 a).

Οι διακριτές ιδιοτιμές {λ1,λ2,…λκ} των πινάκων $L^T L$ και $L\ L^T$ είναι αναγκαστικά πραγματικοί μη αρνητικοί αριθμοί με λ1> λ2>….λκ>=0 γιατί οι πίνακες

είναι συμμετρικοί θετικώς ημιορισμένοι και μη-αρνητικοί. Η εξειδίκευση της μεθόδου δύναμης από τον HITS βοηθά στην αποφυγή προβλημάτων σύγκλισης αφού ο HITS με κανονικοποίηση συγκλίνει πάντα στο ρυθμό όπου $(\lambda 2(B)|/|\lambda 1(B)|)^{\kappa}$ -> 0. Δυστυχώς δεν μπορεί να βρεθεί για τον αλγόριθμο HITS μια καλή εκτίμηση για τον ασυμπτωτικό ρυθμό σύγκλισης. Για την επίτευξη σύγκλισης ίσως να απαιτούνται περίπου 15 επαναλήψεις .Όμως ίσως να προκύψουν προβλήματα όσον αφορά τη μοναδικότητα των διανυσμάτων authorities και hubs. Ενώ λ1>λ2 η δομή του L μπορεί να επιτρέπει στην λ1 να είναι επαναλαμβανόμενη ρίζα του χαρακτηριστικού πολυωνύμου οπότε και ο σχετικός ιδιοχώρος είναι πολυδιάστατος. Αυτό σημαίνει πως μπορούν να παραχθούν διαφορετικά οριακά διανύσματα authorities και hubs επιλέγοντας διαφορετικά αρχικά διανύσματα (Peseric κ. συν., 2009 b).
Για παράδειγμα:

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad \text{και} \quad \mathbf{L}^T\mathbf{L} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Και οι δυο πίνακες $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$ είναι όμοιοι και έχουν δύο διακριτές τιμές λ1=2 και λ2=0 οι οποίες επαναλαμβάνονται δύο φορές. Για το αρχικό διάνυσμα $x^{(0)}$= $(1/4\ 1/8\ 1/8\ 1/2)^T$ η μέθοδος δύναμης συγκλίνει στο x^0 =$(1/2\ 1/4\ 1/4\ 0)^T$. Langville & Meyer (2005, σ.6)

Αναγώγιμοι- Μη αναγώγιμοι τετραγωνικοί πίνακες

Ένας τετραγωνικός πίνακας B ονομάζεται αναγώγιμος (reducible) όταν υπάρχει πίνακας μεταθέσεων Q ώστε:

$$\mathbf{Q}^T\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$$

Όπου X,Z επίσης τετραγωνικοί.

Αν δεν ισχύει ο παραπάνω τύπος ο πίνακας B είναι μη-αναγώγιμος (irreducible). Η αναγωγιμότητα ενός πίνακα συνεπάγεται την ύπαρξη ενός συνόλου καταστάσεων στις οποίες είναι πιθανό να μεταβεί αλλά μόλις μεταβεί σε αυτές δεν

υπάρχει η δυνατότητα επιστροφής σε κάποια προηγούμενη. Ένας μη αναγώγιμος πίνακας μπορεί από μια κατάσταση να μεταβεί σε οποιαδήποτε προηγούμενη.(Langville & Meyer ,2005 σελ. 6)

Θεώρημα Perron-Frobenius

Ένας μη-αναγώγιμος, μη αρνητικός πίνακας διαθέτει ένα μοναδικό κανονικοποιημένο θετικό πρωτεύον ιδιοδιάνυσμα το οποίο ονομάζεται Perron. Συνεπώς η σύγκλιση του αλγορίθμου HITS σε μη μοναδικές λύσεις οφείλεται στην αναγωγιμότητα του L^T L. Μπορεί ωστόσο να γίνει μια τροποποίηση στο μοντέλο η οποία επιβάλλει τη μη αναγωγιμότητα και επομένως και τη μοναδικότητα του τελικού διανύσματος. Ο τροποποιημένος HITS λέγεται **Exponential HITS** (AMY. N.Langville &Carl D. Meyer,2010a)

Ο πίνακας authority μπορεί να πάρει τη μορφή:

$$\xi \mathbf{L}^T\mathbf{L} + \frac{(1-\xi)}{n}\mathbf{e}\mathbf{e}^T, \text{ όπου } 0 < \xi < 1,$$

Ο πίνακας Hubs παίρνει αντίστοιχα την μορφή:

$$\xi \mathbf{L}\mathbf{L}^T + \frac{(1-\xi)}{n}\mathbf{e}\mathbf{e}^T$$

2.1.5 Query-Independent HITS (Ανεξάρτητος από τα ερωτήματα)

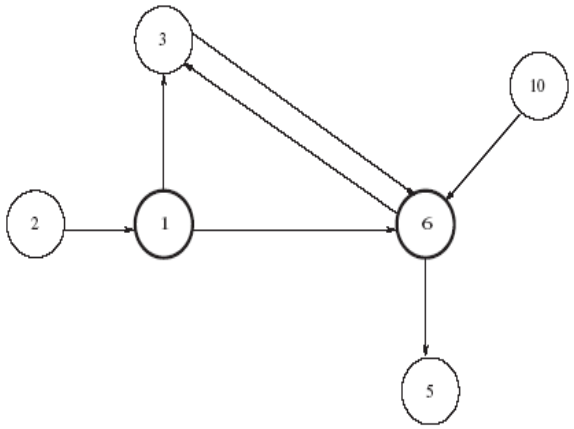
1. Αρχικοποίηση: $y^{(0)}$= e/n όπου e είναι ένα διάνυσμα στήλης με 1.
2. Μέχρι να επιτευχθεί σύγκλιση επανέλαβε.
3. Όρισε ως διάνυσμα authority το $x^{(k)}$ hub το διάνυσμα $y^{(k)}$.

$$\begin{aligned}
\mathbf{x}^{(k)} &= \xi\mathbf{x}^{(k-1)} + (1-\xi)/n\mathbf{e} \\
\mathbf{x}^{(k)} &= \mathbf{x}^{(k)}/\|\mathbf{x}^{(k)}\|_1 \\
\mathbf{y}^{(k)} &= \xi\mathbf{y}^{(k-1)} + (1-\xi)/n\mathbf{e} \\
\mathbf{y}^{(k)} &= \mathbf{y}^{(k)}/\|\mathbf{y}^{(k)}\|_1 \\
k &= k+1
\end{aligned}$$

Συνεπώς ο βασικός αλγόριθμος HITS μπορεί να μετατραπεί ώστε να μην εξαρτάται από τα ερωτήματα υπολογίζοντας ένα καθολικό διάνυσμα authority και hub αντίστοιχα. Έτσι μειώνεται η επιρροή των ανεπιθύμητων συνδέσμων. Στην παραλλαγή αυτή ο L είναι πίνακας γειτνίασης του γραφήματος ολόκληρου του ιστού. Επιπρόσθετα χρησιμοποιούνται οι τροποποιημένοι πίνακες που αναφέρθηκαν πιο πάνω και εγγυώνται μοναδικότητα. (Langville & Meyer,2010b)

2.1.6 Παραδείγματα εφαρμογής HITS

1) Έστω το παρακάτω γράφημα που αποτελείται από 6 κόμβους και το υποσύνολο των κόμβων που περιέχουν τους όρους αναζήτησης είναι το {1,6}.



Σχήμα 8: Γράφημα 6-κόμβων (Langville & Meyer, 2005,σ.7)

Σύμφωνα με τους Langville & Meyer (2005,σ.7) ο πίνακας γειτνίασης L που προκύπτει από το γράφημα του Σχήματος 6 είναι:

$$\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

Οι αντίστοιχοι πίνακες authority και hubs είναι οι εξής:

$$\mathbf{L}^T\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \qquad \mathbf{L}\mathbf{L}^T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 2 & 0 & 1 & 0 & 1 & 1 \end{array} \\ \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Τα κανονικοποιημένα πρωτεύοντα διανύσματα με βαθμούς authority και hub είναι τα εξής:

$\chi^{(T)}$= (0 0 0.3660 0.1340 0.5 0) $y^{(T)}$=(0.3660 0 0.2113 0 0.2113 0.2113)

Σύμφωνα με τα διανύσματα που προκύπτουν, ισοβαθμία μπορεί να δημιουργηθεί στο μηδέν ή σε κάποια θετική τιμή. Οι ισοβαθμίες στο 0 μπορούν να αποφευχθούν με την τροποποίηση θεμελίωσης των πινάκων authorities και hubs. Για μεγαλύτερους πίνακες που μπορεί να προκύψουν η ύπαρξη απόλυτα ίσων τιμών σε κάποιο πρωτεύον ιδιοδιάνυσμα είναι σπάνια. Παρόλα αυτά οποιαδήποτε ισοβαθμία μπορεί να διευθετηθεί μέσω κάποιας πολιτικής άρσης ισοβαθμιών. Χρησιμοποιώντας μια «σειριακή» πολιτική, όπου οι ισοβάθμιοι κόμβοι τοποθετούνται στη σειρά με την οποία προκύπτουν, οι κατατάξεις σελίδων για το παραπάνω γράφημα είναι οι εξής:

Κατάταξη authority: (6 3 5 1 2 10)

Κατάταξη hub: (1 3 6 10 2 5)

Επομένως η σελίδα 6 είναι το καλύτερο authority ενώ η 1 το καλύτερο hub

Εάν υπολογίσουμε και πάλι τα διανύσματα με τους τύπους του τροποποιημένου αλγορίθμου HITS θέτοντας ως **ξ=0.95** έχουμε τα διανύσματα και τις κατατάξεις αντίστοιχα:

$X^{(T)}$=(0.0032 0.0023 0.3634 0.1351 0.4936 0.0023)

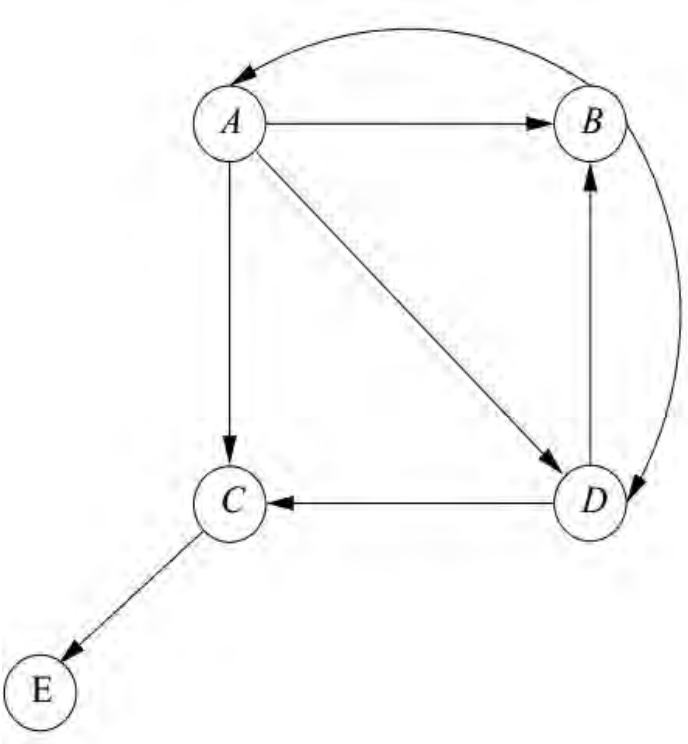
$Y^{(T)}$= (0.3628 0.0032 0.2106 0.0023 0.2106 0.2106)

Κατάταξη authority: (6 3 5 1 2 10)

Κατάταξη hub: (1 3 6 10 2 5)

Όπως παρατηρούμε η τροποποίηση των πινάκων authority και hub δεν αλλάξει τις κατατάξεις, εξαλείφει όμως τις ισοβαθμίες στο 0 και εγγυάται τη σύγκλιση της μεθόδου δύναμης μετά από πεπερασμένο αριθμό επαναλήψεων. (Langville & Meyer, 2005,σ7)

2) Έστω το παρακάτω γράφημα που αποτελείται από 5 κόμβους:



Σχήμα 9: Γράφημα 5-κόμβων (Rajaraman & Ullman,σελ. 192**)**

Ο πίνακας γειτνίασης L και ο ανάστροφός του L^T που προκύπτει από το παραπάνω γράφημα είναι:

$$L = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad L^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Το γεγονός ότι η hubbiness μιας ιστοσελίδας είναι ανάλογη με το άθροισμα των authority των διαδόχων της εκφράζεται με την ισότητα **h= λL a** όπου λ είναι μια άγνωστη σταθερά που αντιπροσωπεύει τον συντελεστή κλίμακας που απαιτείται (scaling factor). Το γεγονός ότι η authority μιας ιστοσελίδας είναι ανάλογη με το άθροισμα των hubbiness των προκατόχων της εκφράζεται με την ισότητα **a= μL^T h** όπου μ είναι ένας άλλος συντελεστής κλίμακας. (Farahat κ συν., 2006 b)

Από τις παραπάνω ισότητες καταλήγουμε στις εξής:

$$\mathbf{h} = \lambda\mu\, \mathbf{L}\mathbf{L}^{\mathbf{T}}\, \mathbf{h}$$

$$\mathbf{a} = \lambda\mu\, \mathbf{L}^{\mathbf{T}}\mathbf{L}\, \mathbf{a}$$

Παρόλα αυτά επειδή οι πίνακες LL^T και L^TL δεν είναι τόσο αραιοί όπως οι L και L^T θα ήταν καλύτερο να υπολογίσουμε το **h** και **a** με αναδρομή. Ξεκινάμε με τον **h,** ένα διάνυσμα με 1 μόνο.

1. Υπολογίζουμε $\mathbf{a} = \mathbf{L}^{\mathbf{T}}\mathbf{h}$ και μετά κλιμακώνουμε ώστε το μεγαλύτερο στοιχείο του διανύσματος να είναι 1
2. Στη συνέχεια υπολογίζουμε **h=La** και κλιμακώνουμε ξανά

Επαναλαμβάνουμε τα βήματα 1 και 2 μέχρι οι αλλαγές στα δύο διανύσματα να είναι πολύ μικρές- μηδαμινές και δεχόμαστε τις τρέχουσες τιμές ως όριο.
Οι πρώτες δύο επαναλήψεις για τον αλγόριθμο HITS:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1/2 \\ 1 \\ 1 \\ 1 \\ 1/2 \end{bmatrix} \quad \begin{bmatrix} 3 \\ 3/2 \\ 1/2 \\ 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1/2 \\ 1/6 \\ 2/3 \\ 0 \end{bmatrix}$$

$$\mathbf{h} \qquad L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

$$\begin{bmatrix} 1/2 \\ 5/3 \\ 5/3 \\ 3/2 \\ 1/6 \end{bmatrix} \quad \begin{bmatrix} 3/10 \\ 1 \\ 1 \\ 9/10 \\ 1/10 \end{bmatrix} \quad \begin{bmatrix} 29/10 \\ 6/5 \\ 1/10 \\ 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 12/29 \\ 1/29 \\ 20/29 \\ 0 \end{bmatrix}$$

$$L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

Στην πρώτη στήλη γίνεται η αρχικοποίηση του διανύσματος h με 1. Στην δεύτερη υπολογίζουμε την σχετική authority των σελίδων μέσω της σχέσης $\mathbf{a} = \mathbf{L}^{\mathbf{T}}\mathbf{h}$ δίνοντας έτσι σε κάθε σελίδα το άθροισμα των hubbiness των προκατόχων της. Η

τρίτη στήλη μας δίνει τον πρώτο υπολογισμό του a κλιμακούμενη ταυτόχρονα. Στην περίπτωση αυτή διαιρούμε το κάθε στοιχείο της δεύτερης στήλης με το μεγαλύτερο όλων που είναι το 2 ώστε το μεγαλύτερο στοιχείο της τρίτης στήλης να είναι το 1. Στην τέταρτη στήλη παρουσιάζεται το διάνυσμα La δηλαδή υπολογίζουμε το hubbiness κάθε σελίδας αθροίζοντας τα authority καθενός από τους διαδόχους της. Στην περίπτωση αυτή αν διαιρέσουμε κάθε στοιχείο του διανύσματος με το 3 που είναι το μεγαλύτερο στοιχείο της στήλης 4 προκύπτει το h που είναι και η πέμπτη στήλη. Από την στήλη έξι και μετά επαναλαμβάνεται η διαδικασία με πιο ακριβή υπολογισμό των hubbiness και authority. Το όριο αυτής της διαδικασίας ίσως να μην είναι εμφανές αλλά μπορεί να υπολογιστεί με ένα από πρόγραμμα (Rajaraman & Ullman,2011 a).

Τα όρια είναι τελικά:

$$\mathbf{h} = \begin{bmatrix} 1 \\ 0.3583 \\ 0 \\ 0.7165 \\ 0 \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} 0.2087 \\ 1 \\ 1 \\ 0.7913 \\ 0 \end{bmatrix}$$

Το αποτέλεσμα που προκύπτει βγάζει νόημα. Αρχικά παρατηρούμε ότι το hubbiness της Ε είναι σίγουρα 0 από τη στιγμή που δεν οδηγεί πουθενά δηλαδή δεν υπάρχει εξερχόμενη ακμή από τον κόμβο Ε προς κάποιον άλλο κόμβο του γραφήματος. Το hubbiness του C εξαρτάται μόνο από το authority του Ε και αντιστρόφως και γι' αυτό δεν θα πρέπει να μας εκπλήσσει το γεγονός ότι είναι και τα δύο 0. Επιπρόσθετα ο κόμβος Α είναι το καλύτερο hub (κομβικό σημείο) επειδή ενώνει τα τρία μεγαλύτερα authorities B, C, D. Ενώ οι κόμβοι Β και C είναι οι καλύτεροι authorities λόγω του γεγονότος ότι συνδέονται με τα δύο μεγαλύτερα hubs Α και D. Γενικά για Web-sized γραφήματα ο μόνος τρόπος υπολογισμού των hubs και authorities είναι ο επαναληπτικός. Παρόλα αυτά επειδή το παράδειγμά μας είναι μικρό μπορούμε να υπολογίσουμε τη λύση του από την επίλυση εξισώσεων. Μπορούμε να χρησιμοποιήσουμε την ισότητα $\mathbf{h= \lambda\mu\, LL^T\, h}$

Το LL^T είναι:

$$LL^{\mathrm{T}} = \begin{bmatrix} 3 & 1 & 0 & 2 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Θέτουμε v= 1/(λμ) οπότε η παραπάνω σχέση μετατρέπεται σε $\mathbf{vh = LL^T h}$ όπου $\mathbf{h = [a, b, c, d, e]^T}$ άγνωστο διάνυσμα με στοιχεία τους κόμβους του γραφήματος από Α μέχρι Ε αντίστοιχα.
Οι εξισώσεις που προκύπτουν είναι οι εξής:

Va=3a + b+ 2d

Vb=a+2b ⇔b= a/(v-2)

Vc = c

Vd=2a+2d ⇔d=2a/(v-2)

Ve=0 ⇔e=0

Αν αντικαταστήσουμε τα b και d στην πρώτη εξίσωση θα έχουμε va=a(3+5/(v-2)) κάνοντας τις πράξεις καταλήγουμε στη μορφή v^2 -5v +1=0 ⇔ η θετική ρίζα του v είναι v=4.791. Από το αποτέλεσμα ξέρουμε ότι το v δεν είναι ούτε 0 ούτε 1 ενώ c=e=0. Τέλος αν γνωρίζουμε ότι το a είναι η μεγαλύτερη συνιστώσα του h και θέσουμε a=1 παίρνουμε b=0.3583 και d=0.7165 που μαζί με τις τιμές c=e=0 μας δίνουν την οριακή τιμή του h. Η τιμή του a υπολογίζεται από το h πολλαπλασιάζοντας με L^T και κλιμακώνοντας. (Rajaraman & Ullman, 2011b)

2.2 PageRank

2.2.1 Τι είναι PageRank

Ο αλγόριθμος PageRank ενώ εκμεταλλεύεται τον γράφο διασυνδέσεων των ιστοσελίδων διαφέρει από τη προσέγγιση HITS. Πρόκειται για μια καθολική μέθοδο και πιο συγκεκριμένα η βασική ιδέα είναι να αξιολογεί εκ των προτέρων τη σημασία κάθε σελίδας ξεχωριστά. Σε αντίθεση η αξιολόγηση στο HITS εξαρτάται και εφαρμόζεται κάθε φορά που ο χρήστης θέτει ένα συγκεκριμένο ερώτημα για αναζήτηση.

Ο αλγόριθμος αυτός αναπτύχθηκε στο Πανεπιστήμιο του Stanford από τους Sergey Brin και Larry Page ιδρυτές της σημερινής Google. Η ανάπτυξη έγινε στα πλαίσια μιας ερευνητικής προσπάθειας για τη δημιουργία ενός νέου είδους μηχανής αναζήτησης. Το έργο ξεκίνησε το 1995 και το πρώτο λειτουργικό πρωτότυπο ονομάστηκε Google. Λίγο αργότερα, οι Brin και Page ίδρυσαν την εταιρία Google Inc. Αν και ο αλγόριθμος PageRank αποτελεί μόνο έναν από το σύνολο των παραγόντων που προσδιορίζουν την κατάταξη των αποτελεσμάτων στο Google, παραμένει ίσως το πιο σημαντικό και η βάση για τις περισσότερες λειτουργίες της μηχανής αναζήτησης. Όταν πρωτοδημιουργήθηκαν οι μηχανές αναζήτησης, βασιζόταν σε συγκεκριμένους παράγοντες όπως είναι η επανάληψη των λέξεων κλειδιών και τα tags μιας σελίδας. Με τον καιρό παρόλα αυτά, το σύστημα βαθμολογίας της Google ήταν αυτό που έφερε την επανάσταση στον χώρο των μηχανών αναζήτησης και αυτό έδειχνε το πόσο «αξία» είχε μία ιστοσελίδα.

Για να καθορίσει λοιπόν πόσο σημαντική ήταν μία ιστοσελίδα, η Google επέλεξε διάφορες μεγάλες ιστοσελίδες όπως το cnn.com, το dmoz.org κλπ. Αυτές οι ιστοσελίδες ήταν ξεκάθαρο κορυφαίες σελίδες με ποιοτικό περιεχόμενο. Αν λοιπόν μία από αυτές τις κορυφαίες ιστοσελίδες αποφάσιζε να τοποθετήσει μία υπερσύνδεση (link) προς μία άλλη ιστοσελίδα, τότε ένα μέρος της αξίας αυτής της ιστοσελίδας θα μεταφερόταν προς την άλλη. Η ιστοσελίδα αυτή με τη σειρά της αν έβαζε μία υπερσύνδεση (link) προς μία άλλη σελίδα, τότε και αυτή η ιστοσελίδα θα έπαιρνε ένα μέρος της αξίας της προηγούμενης σελίδας. Δημιουργώντας λοιπόν αυτό το σύστημα αξιολόγησης, η Google μετράει πόσο «κύρος» έχει μία ιστοσελίδα. Η αξιολόγηση

έγινε με μία βαθμολογία από 0 - 10 . Μέσω διάφορων εργαλείων οι webmasters μπορούσαν να δουν ποιο είναι το PageRank της ιστοσελίδας τους. Ουσιαστικά ένας αριθμός μεταξύ αυτού του διαστήματος από 0-10 εκφράζει την πιθανότητα που έχει κάποιος χρήστης να επιλέξει στην τύχη συνδέσμους στον Παγκόσμιο Ιστό και να φτάσει στην ιστοσελίδα που επιθυμεί. (Page κ. συν. ,1998a)

2.2.2 Ορισμός του PageRank

Έστω u μία ιστοσελίδα, Lu το σύνολο των σελίδων στις οποίες δείχνει η u και Bu το σύνολο των σελίδων που δείχνουν προς τη σελίδα u (backlinking). Ακόμη Nv είναι ο αριθμός των εξερχόμενων συνδέσμων από την ιστοσελίδα v .Επίσης, έστω Nu= |Lu| ο αριθμός των συνδέσμων από την u και c ένας παράγοντας κανονικοποίησης (έτσι ώστε ο συνολικός βαθμός όλων των ιστοσελίδων να είναι σταθερός) (Andersson & Ekstrom,2004a).

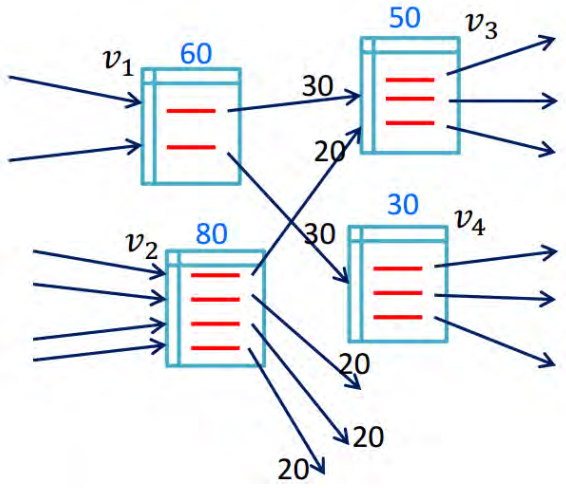
Μια απλουστευμένη μορφή του τύπου του PageRank είναι:

$$Rank(u) = c\sum_{v\in B_u}\frac{Rank(v)}{N_v}$$

Σχήμα 10: Απλουστευμένη Εξίσωση PageRank

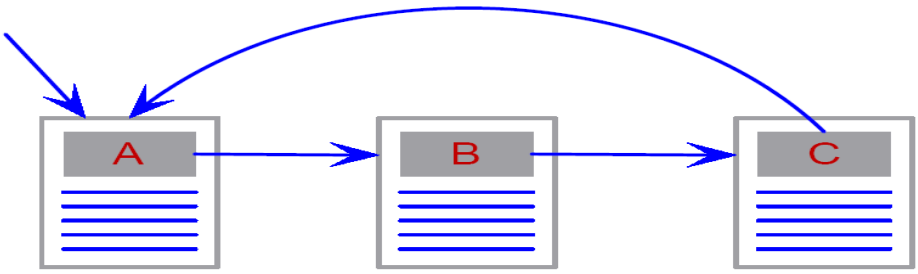
Με τον τρόπο αυτό, το Rank μιας οποιασδήποτε σελίδας διαιρείται εξίσου στους εξερχόμενους συνδέσμους της (εσωτερικούς και εξωτερικούς, δηλαδή όλους τους συνδέσμους που εξέρχονται από τον κόμβο, είτε αυτοί δείχνουν σε εξωτερικές σελίδες είτε δείχνουν σε σελίδες που βρίσκονταν στον ίδιο ιστότοπο) για να συνεισφέρουν στο συνολικό Rank των σελίδων στις οποίες δείχνουν. Ο παραπάνω τύπος περιλαμβάνει το c που ορίζεται ως παράγοντας κανονικοποίησης, ο οποίος βοηθάει στο να μην δημιουργηθεί πρόβλημα σε περίπτωση που έχουμε μια σελίδα που δεν έχει εξερχόμενες ακμές δηλαδή links προς άλλες σελίδες. Ο παράγοντας κανονικοποίησης είναι μικρότερος της μονάδας (**c < 1) .** Το Rank(v) είναι άγνωστο αρχικά. Επιπρόσθετα ο παραπάνω τύπος είναι αναδρομικός, ξεκινάμε από μια τιμή αρχική και επαναλαμβάνουμε μέχρι να επιτευχθεί σύγκλιση. Corso κ.συν 2005a).

Έτσι ξεκινώντας από μια αρχική κατάσταση των σελίδων u1 και u2 με Rank(u1)=60, Rank(u2)=80 μπορούμε να υπολογίσουμε, το Rank των σελίδων u3 και u4. Οπότε Rank(u3)=30+20=50 και Rank(u4)=30.



Σχήμα 11

Ο αλγόριθμος PageRank τερματίζεται όταν επιτευχθεί σύγκλιση δηλαδή όταν όλα τα links (σύνδεσμοι) έχουν την ίδια αξία μετά από έναν αριθμό επαναλήψεων. Οι ερευνητές όμως παρατήρησαν ότι ο παραπάνω τύπος ήταν προβληματικός σε ορισμένες περιπτώσεις. Ειδικότερα όταν δύο ή περισσότερες ιστοσελίδες αντάλλαζαν αποκλειστικά links (συνδέσμους) μεταξύ τους ενώ μία άλλη σελίδα συνδέονταν με μία εκ των δύο αυτών ιστοσελίδων. Με το πέρας των επαναλήψεων οι σελίδες αυτές συγκέντρωναν το μεγαλύτερο PageRank χωρίς να πραγματοποιείται κατανομή του PageRank σε όλες τις ιστοσελίδες. (Page κ.συν.,1998b) Το φαινόμενο αυτό ονομάζεται **«νεροχύτης βαθμού» (rank sink),** όπως φαίνεται παρακάτω:



Σχήμα12: Απλουστευμένο παράδειγμα ενός Rank Sink

Για την αντιμετώπιση του rank sink ορίστηκε μια πηγή βαθμού, ένα είδος αντίμετρου- πηγή βαθμολογίας (rank source) E(u), που συμπεριλήφτηκε στον τρόπο υπολογισμού του PageRank. Εάν E(u) είναι κάποιο διάνυσμα ιστοσελίδων που αντιστοιχεί σε μια πηγή βαθμού (rank source) τότε ο βαθμός PageRank του συνόλου ιστοσελίδων είναι η τιμή που ανατίθεται στις ιστοσελίδες αυτές η οποία ικανοποιεί την αναδρομική σχέση:

$$PageRank(u) = c\sum_{v\in B_u}\frac{PageRank(v)}{N_v} + cE(u)$$

Ο παράγοντας κανονικοποίησης c με αυτόν τον τρόπο μεγιστοποιείται και η L1 νόρμα του PageRank ισούται με τη μονάδα δηλαδή: || PageRank'||1=1. Αν χρησιμοποιήσουμε τον συμβολισμό πινάκων ξαναγράφουμε την παραπάνω σχέση ως εξής: PageRank' = c(A PageRank' + E). Αφού || PageRank '||1= 1 η σχέση ξαναγράφεται ως PageRank' = c(A + E1) PageRank' όπου 1 διάνυσμα αποτελούμενο από μονάδες. Έτσι το PageRank είναι ένα ιδιοδιάνυσμα του (A + E1) και το c είναι η κυρίαρχη ιδιοτιμή. (Andersson & Ekstrom,2004b)

2.2.3 Υπολογισμός του PageRank

Σύμφωνα με τους Brin και Page (1998), στη δημοσίευσή τους «The Anatomy of a Large-Scale Hypertextual Web Search Engine» που πραγματοποιήθηκε στα πλαίσια της προσπάθειάς τους να αναπτύξουν μία νέα βελτιωμένη μηχανή αναζήτησης που αργότερα έγινε γνωστή ως μηχανή αναζήτησης Google, ο βαθμός PageRank δεν περιορίζεται στον υπολογισμό του αριθμού των εισερχόμενων συνδέσμων μίας σελίδας με σκοπό την εκτίμηση της δημοτικότητας της ιστοσελίδας αυτής, αλλά επεκτείνει την ιδέα της οργάνωσης του Διαδικτύου αποδίδοντας διαφορετική βαρύτητα (ποιότητα) σε κάθε σύνδεσμο που δείχνει στη σελίδα και κανονικοποιεί την αξία αυτής με κριτήριο τον αριθμό των συνδέσμων που ξεκινούν από μία ιστοσελίδα. (Page κ.συν.,1998 c).

Ο βαθμός PageRank μιας ιστοσελίδας A, PR(A), υπολογίζεται ως εξής:

$$PR(A) = (1-d) + d[PR(T_1)/C(T_1) + ... + PR(T_v)/C(T_v)]$$

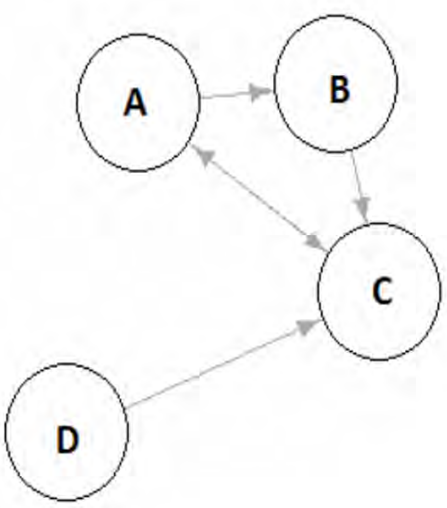
ή αλλιώς:

$$PageRank(A) = (1-d) + d\sum_{i=1}^{n} \frac{PageRank(T_i)}{C(T_i)}$$

Όπου :

- T1, T2, …, Τν είναι οι σελίδες που συνδέουν στη σελίδα Α.
- Η παράμετρος d είναι ένας συντελεστής απόσβεσης που μπορεί να πάρει τιμές στο διάστημα [0,1] συνήθως λαμβάνει την τιμή 0,85
- C(A) ορίζεται ως ο αριθμός των εξερχόμενων συνδέσμων της σελίδας Α.

2.2.4 Παράδειγμα υπολογισμού του PageRank



Σχήμα 13: Γράφημα

Στο Σχήμα 13 βλέπουμε ένα δίκτυο με τις ιστοσελίδες A, B, C και D των οποίων τα εξερχόμενα και τα εισερχόμενα links παρουσιάζονται στο Σχήμα 14.

Σελίδες	Εξερχόμενα Links	Εισερχόμενα links
Page A	Page B, Page C	Page C
Page B	Page C	Page A
Page C	Page A	Page A, Page B , Page D
Page D	Page C	

Σχήμα 14

Το αρχικό PageRank για κάθε ιστοσελίδα είναι 1-d= 0,15 (όπου d=0,85) οπότε το PageRank κάθε ιστοσελίδας για την 1η επανάληψη υπολογίζεται ως εξής:

PR(Page A)= (1-d)+ d * (PR(Page C)/C(Page C))=(1-0.85) + 0.85 *(0.15/1)=0.27750

PR(Page B)= (1-d)+ d * (PR(Page A)/C(Page A))=(1-0.85) + 0.85 *(0.15/2)=0.21375

PR(Page C)= (1-d)+ d * (PR(Page A)/C(Page A)+ PR(Page B)/C(Page B) + PR (Page D)/C(PageD))= (1-0.85) + 0.85 *(0.15/2 +0.15/1 +0.15/1)= 0.46875

PR(Page D)= (1-d)+ d *(0)=(1-0.85) + 0.85 * (0)=0.15

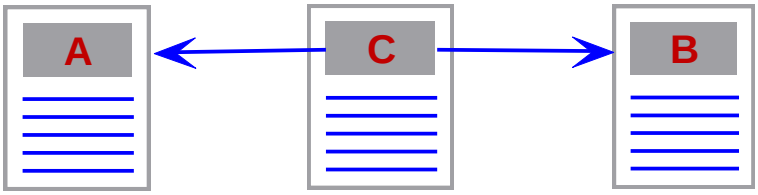
Σελίδες	Την 1η επανάληψη	Μετά από N=100 επαναλήψεις
Page A	0.27750	1.484690
Page B	0.21375	0.780587
Page C	0.46875	1.571180
Page D	0.15000	0.150000

Σχήμα 15

Παρόλα αυτά με βάση τον τελευταίο τρόπο υπολογισμού του PageRank δημιουργείται ένα επιπλέον πρόβλημα . Το πρόβλημα αυτό είναι οι λεγόμενοι **«αδιέξοδοι σύνδεσμοι» (dangling links).** Πρόκειται για συνδέσμους που δείχνουν σε μία σελίδα, η οποία δεν περιέχει εξερχόμενους συνδέσμους. Σε αυτή την κατηγορία περιλαμβάνονται και οι σύνδεσμοι που δείχνουν σε σελίδες που δεν έχουν μεταφορτωθεί και αναλυθεί ακόμη για τον προσδιορισμό των εξερχόμενων συνδέσμων τους. Η αφαίρεση των συνδέσμων αυτών από το γράφημα αποτελεί επίλυση του προβλήματος αυτού. Και ο λόγος είναι ότι με αυτόν τον τρόπο δεν επηρεάζεται ο βαθμός των υπόλοιπων σελίδων.

Στη συνέχεια υπολογίζοντας το PageRank των σελίδων, προσθέτονται ξανά οι σύνδεσμοι στον γράφημα δημιουργώντας μικρές ασήμαντες αλλαγές στην παράμετρο κανονικοποίησης των άλλων links (συνδέσμων) που βρίσκονται στις ίδιες σελίδες με αυτούς. Οι αλλαγές αυτές θεωρούνται αμελητέες. Στο παρακάτω σχήμα απεικονίζεται ένας γράφος στον οποίο οι σελίδες A και B έχουν εισερχόμενα αλλά όχι εξερχόμενα links (συνδέσμους). Ένας επιπλέον τρόπος επίλυσης του προβλήματος αυτού είναι οι σελίδες A και B να έχουν links (συνδέσμους) προς όλες

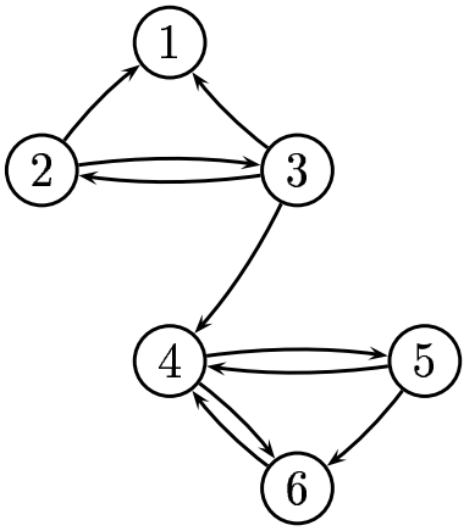
τις σελίδες του γραφήματος με την ίδια πιθανότητα μετάβασης (Langville & Meyer,2010c).



Σχήμα16: Dangling Pages (αιωρούμενες σελίδες)

2.2.5 Matrix model

Ένας άλλος απλούστερος τρόπος υπολογισμού του PageRank είναι με τη χρήση πινάκων. Η διαδικασία αθροίσματος για τον υπολογισμό του βαθμού της κάθε σελίδας ξεχωριστά με τη χρήση πινάκων απλουστεύεται. Έστω το παρακάτω γράφημα (Andersson&Ekstrom,2004, σελ. 3).



Σχήμα 17: Μικρό γράφημα που αποτελείται από 6 ιστοσελίδες P1-P6 (Andersson & Ekstrom,2004)

Σύμφωνα με τους Andersson & Ekstrom (2004 c σελ 3) ορίζεται ο πίνακας Q του οποίου τα στοιχεία υπολογίζονται με βάση τον τύπο:

$$Q_{ij} := \begin{cases} 1/N_i & \text{if } P_i \text{ links to } P_j \\ 0 & \text{otherwise} \end{cases}$$

Οπότε ο πίνακας Q για το παραπάνω γράφημα θα είναι ο παρακάτω:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Ο πίνακας Qij περιγράφει ότι υπάρχει σύνδεσμος από την ιστοσελίδα Pi στην Pj και αυτές διαιρούνται από το Ni (που είναι ο αριθμός των εξερχόμενων συνδέσμων της ιστοσελίδας Pi).

Ο υπολογισμός του **PageRank r** σύμφωνα με την **μέθοδο δύναμης (Power Method)**:

$$\mathbf{r}_{(k+1)}^T = \mathbf{r}_{(k)}^T \mathbf{Q} \qquad , k = 0, 1, \ldots$$

Ο πίνακας Q είναι ένας αραιός (sparse) πίνακας δηλαδή ένα μεγάλο μέρος των στοιχείων του είναι μηδενικά. Οι αραιοί πίνακες απαιτούν μικρό χώρο αποθήκευσης αφού υπάρχουν συστήματα που αποθηκεύουν μόνο τα μη μηδενικά τους στοιχεία. Τα μη μηδενικά στοιχεία της γραμμής i αντιστοιχούν στις σελίδες στις οποίες δείχνει η σελίδα i ενώ τα μη μηδενικά στοιχεία της στήλης i αντιστοιχούν στις σελίδες που δείχνουν στη σελίδα i. Αξίζει να προσθέσουμε ότι ο πίνακας Q μοιάζει με έναν στοχαστικό πίνακα κάποιας αλυσίδας Markov. Οι κόμβοι του γραφήματος που δεν έχουν εξερχόμενους συνδέσμους, δημιουργούν γραμμές γεμάτες με μηδενικά στοιχεία. Οι υπόλοιπες γραμμές που αντιστοιχούν σε κόμβους με εξερχόμενους συνδέσμους, αποτελούν τις στοχαστικές γραμμές του πίνακα Q. Έτσι ο πίνακας Q χαρακτηρίζεται ως υποστοχαστικός (substochastic). Συγκεκριμένα από το γράφημα παρατηρούμε ότι η ιστοσελίδα P1 είναι ένας αιωρούμενος (dangling) κόμβος αφού έχει μόνο εισερχόμενους συνδέσμους. Αυτό επιβεβαιώνεται και από τον πίνακα όπου

η $1^{η}$ γραμμή έχει μόνο μηδενικά στοιχεία. Για την επίλυση του προβλήματος αυτού χρησιμοποιείται η έννοια του μοντέλου του τυχαίου χρήστη (random surfer) όπως ορίστηκε από τους ιδρυτές της σημερινής Google (Corso, κ.συν 2005b).

2.2.6 Το μοντέλο του τυχαίου χρήστη (random surfer)

Οι Brin και Page χρησιμοποίησαν την έννοια **του τυχαίου διαδικτυακού χρήστη** (random surfer) για να περιγράψουν τις αλλαγές τους στο αρχικό μοντέλο και να το βελτιώσουν. Η έννοια αυτή βασίζεται στην υπόθεση ότι υπάρχει ένας τυχαίος χρήστης ο οποίος βρίσκεται σε μια τυχαία ιστοσελίδα και επιλέγει έναν οποιονδήποτε από τους εξερχόμενους συνδέσμους της ιστοσελίδας αυτής με ίση πιθανότητα. Ο χρήστης αυτός επιπλέον συνεχίζει να κάνει κλικ σε links (συνδέσμους) που συναντάει χωρίς να γυρίζει πίσω αλλά τελικά «βαριέται» και ξεκινάει από μια καινούρια διαφορετική τυχαία σελίδα. Μακροπρόθεσμα, ο χρόνος που ξοδεύει ο χρήστης σε μια ιστοσελίδα είναι ένα μέτρο για το πόσο σπουδαίο είναι η ιστοσελίδα αυτή. Ιστοσελίδες όπου ο χρήστης κάνει «κλικ» συχνά είναι σημαντικές γιατί δείχνουν σε άλλες σημαντικές σελίδες (Agirre, 2009a).

Για την επίλυση του προβλήματος της ύπαρξης αιωρούμενων κόμβων και συνάμα της ύπαρξης ολόκληρων γραμμών με μηδενικά στοιχεία, αντικαθίστανται όλα τα μηδενικά στοιχεία με 1/n σε όλες τις μηδενικές γραμμές. Το n είναι η διάσταση του πίνακα στην περίπτωση μας n=6. Έτσι δημιουργείται ένας νέος πίνακας Q' ο οποίος είναι στοχαστικός. Χρησιμοποιώντας την έννοια του τυχαίου χρήστη τώρα πια ο διαδικτυακός χρήστης εάν μεταβεί σε κάποιον αιωρούμενο κόμβο μπορεί να μεταβεί τυχαία σε οποιαδήποτε σελίδα.

Τα στοιχεία του πίνακα Q' προκύπτουν ως εξής:

$$\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{n}\,\mathbf{a}\,\mathbf{e}^{\mathsf{T}}$$

Όπου **e** είναι μία στήλη-πίνακας που τα στοιχεία της είναι μονάδες και **a** είναι μια στήλη- πίνακας που περιγράφει σε ποια γραμμή ο πίνακας Q έχει μηδενικά στοιχεία. Επί της ουσίας ο **a** είναι ένα διάνυσμα αιωρούμενων κόμβων.

$$\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{n}\mathbf{a}\mathbf{e}^{\mathsf{T}} = \mathbf{Q} + \frac{1}{6}\begin{pmatrix}1\\0\\0\\0\\0\\0\end{pmatrix}\begin{pmatrix}1 & 1 & 1 & 1 & 1 & 1\end{pmatrix} =$$

$$= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0\\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 & 0\\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2}\\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0\\ 0 & 0 & 0 & 1 & 0 & 0\end{pmatrix} + \begin{pmatrix}\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix} = \begin{pmatrix}\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 & 0\\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2}\\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0\\ 0 & 0 & 0 & 1 & 0 & 0\end{pmatrix}$$

Συνεπώς η 1η γραμμή του πίνακα Q αντικαθίσταται με το μη μηδενικό στοιχείο 1/6. Ακόμη η ομάδα των κόμβων 4, 5, 6 αποτελεί μια καταβόθρα διαβάθμισης (rank sink) οι ιστοσελίδες αυτές αποκτούν μεγάλη διαβάθμιση με τις επαναλήψεις. Εάν ο τυχαίος χρήστης κατέβει χαμηλά στο γράφημα και τελειώσει με το τμήμα των κόμβων 4, 5 και 6 , δεν υπάρχουν πιθανότητες γι' αυτόν να επιστρέψει στο πάνω κομμάτι του γραφήματος. Με άλλα λόγια δεν υπάρχει εγγύηση για τα επιθυμητά αποτελέσματα σύγκλισης. Η μέθοδος που χρησιμοποιήθηκε στον υπολογισμό του PageRank για την εγγύηση σωστής σύγκλισης και της ελάχιστης πιθανότητας ο δικτυακός χρήστης να «πηδήξει» από μια σελίδα του υπογραφήματος που συμπεριφέρεται ως καταβόθρα προς οποιαδήποτε άλλη σελίδα του συνολικού γραφήματος (έξω από το υπογράφημα αυτό) καλείται **"teleportation" (** Rossi, κ.συν. 2012).

Μαθηματικά η έννοια αυτή περιγράφεται ως εξής:

$$\hat{\hat{\mathbf{Q}}} = \mathrm{d}\hat{\mathbf{Q}} + (1-\mathrm{d})\frac{1}{n}\mathbf{e}\mathbf{e}^{T}$$

Όπου d είναι **damping factor** (συντελεστής απόσβεσης) και κυμαίνεται μεταξύ 0 και 1. Στο μοντέλο αυτό το d είναι μια παράμετρος που ρυθμίζει την αναλογία του χρόνου στο οποίο ο τυχαίος χρήστης ακολουθεί υπερσυνδέσμους σε σχέση με το χρόνο που μεταφέρεται απευθείας σε τυχαίες σελίδες εκτός της δομής. Για παράδειγμα αν d=0.6 σημαίνει ότι στο 60% του χρόνου ο τυχαίος χρήστης ακολουθεί τη δομή των υπερσυνδέσμων του διαδικτύου και στο υπόλοιπο 40% του χρόνου μεταφέρεται απευθείας σε κάποια τυχαία σελίδα. Η απευθείας μεταφορά είναι

τυχαία καθώς ο πίνακας E= (1/n) e e^T είναι ομοιόμορφος, που σημαίνει ότι ο τυχαίος χρήστης μπορεί να μεταβεί σε οποιαδήποτε σελίδα με σχεδόν την ίδια πιθανότητα (Manning, 2008 c).

Ο πίνακας Q'' που προκύπτει είναι στοχαστικός αφού είναι ο κυρτός συνδυασμός των δύο επιμέρους στοχαστικών πινάκων G' και E. Είναι τεχνικός αφού ο G τροποποιήθηκε δύο φορές για να πετύχουμε σύγκλιση. Είναι αμείωτος αφού κάθε μεμονωμένη σελίδα συνδέεται απευθείας με καθεμία από όλες τις υπόλοιπες. Θεωρείται ότι είναι απεριοδικός και θεμελιώδης . Ο G'' είναι θεμελιώδης καθώς G''^{k} >0 για κάποιο k. Επιπρόσθετα ο Q'' είναι απόλυτα πυκνός κάτι που από θέμα μνήμης είναι αρνητικό .Ο G'' μπορεί να γραφτεί ως μια ανανέωση γραμμοβαθμού1 του αραιού πίνακα υπερσυνδέσμων Q. (Andersson, κ .συν.2004d,. σελ 5).

$$G''=d\,G' + (1/n)(1-d)\,e\,e^T$$

$$=d\,(G+ 1/n\,a\,e^T) + (1/n)(1-d)ee^T$$

$$=d\,G + (1/n)\,(da + (1-d)e)e^T$$

Εάν εφαρμόσουμε όλα τα παραπάνω θα έχουμε :

$$\hat{\hat{\mathbf{Q}}} = \mathrm{d}\,\hat{\mathbf{Q}} + (1-\mathrm{d}\,)\frac{1}{n}\mathbf{e}\mathbf{e}^T =$$

$$= 0.85 \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} + (1-0.85)\frac{1}{6} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} =$$

$$= \frac{17}{20} \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} + \frac{1}{40} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 11/12 & 1/60 & 11/12 & 1/60 & 1/60 & 1/60 \\ 19/60 & 19/60 & 1/60 & 19/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 7/15 & 1/60 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

2.2.7 Εξατομίκευση του PageRank(Personalization)

Όπως θα μπορούσε εύκολα να συμπεράνει κανείς σημαντικό ρόλο στην εξίσωση του PageRank παίζει η πηγή βαθμολογίας (rank source) E όπως αναφέρθηκε και παραπάνω. Η πηγή βαθμολογίας E χρησιμοποιήθηκε για να λυθεί το πρόβλημα του rank sinks. Το γεγονός ότι ένας χρήστης «βαριέται» περιοδικά και σταματάει να κάνει διαδοχικά κλικ σε συνδέσμους και επιλέγει μια καινούρια σελίδα να ξεκινήσει την αναζήτησή του αυτό εκφράζεται μέσω του E. Η καινούρια αυτή σελίδα επιλέγεται με βάση την κατανομή του E. Επιλέγεται το διάνυσμα E να έχει ομοιόμορφη κατανομή με σταθερή τιμή 0,15. (|E|=0.15) Η επιλογή αυτής της τιμής θεωρείται πολύ δημοκρατική αφού οι διάφορες ιστοσελίδες αξιολογούνται μόνο και μόνο γιατί υπάρχουν. Τέτοιες ιστοσελίδες θα μπορούσε να ήταν τα αρχεία λιστών ηλεκτρονικού ταχυδρομείου ή οι προειδοποιήσεις περί πνευματικής ιδιοκτησίας. Η τιμή αυτή του E δημιουργεί όμως το πρόβλημα ότι ορισμένες ιστοσελίδες είναι υπερεκτιμημένες. (Agirre, 2009b).

Ακόμη εκτός από την ανάθεση του E σε μία συγκεκριμένη τιμή ενδέχεται και η αντιστοίχηση του E σε μια συγκεκριμένη ιστοσελίδα. Με αυτόν τον τρόπο πετυχαίνεται εξατομίκευση. Οι ιστοσελίδες αυτές μπορεί να είναι σημαντικές για το χρήστη και έτσι είναι πολύ πιθανόν να δοθεί σ' αυτές υψηλότερο PageRank. Όμως αυτό είναι πολύπλοκο για να εφαρμοστεί αφού ο αλγόριθμος που υπολογίζει το PageRank είναι αρκετά χρονοβόρος.

2.2.8 Παράγοντας d

Η σταθερά d χρησιμοποιείται για να υπολογιστεί ο πίνακας $Q'' = d\,Q' + (1-d)E$ όπου $E=1/n\ ee^T$ ο πίνακας μεταφοράς (teleportation matrix) του μοντέλου. Συνεπώς αντίθετα με τον πίνακα E η σταθερά d ελέγχει την προτεραιότητα που δίνεται στην πραγματική δομή του δικτύου. Αρχικά οι Brin και Page προτείνουν τη χρήση d=0.85. Όπως όμως έχει αναφερθεί η παράμετρος d ελέγχει τον ασυμπτωτικό ρυθμό σύγκλισης της δυναμικής μεθόδου. Όπως φαίνεται στον παρακάτω πίνακα καθώς το d τείνει στο 1 ο αναμενόμενος αριθμός επαναλήψεων που απαιτούνται από τη μέθοδο δύναμης για σύγκλιση αυξάνεται δραματικά. Για d=0.5 απαιτούνται μόλις 34 επαναλήψεις προκειμένου η δυναμομέθοδος να συγκλίνει με περιθώριο σφάλματος

10^{-10} . Καθώς d →1, το πλήθος των επαναλήψεων γίνεται απαγορευτικό. Λόγω του μεγέθους των πινάκων και των διανυσμάτων που περιλαμβάνονται στον υπολογισμό. Ακόμα και με την επιλογή d=0,85 χρειάζονται πολλές μέρες υπολογισμού για να επιτευχθεί ικανοποιητική σύγκλιση. Καθώς d→1, η επίδραση του τεχνητού πίνακα τηλεμεταφοράς $E=(1/n)ee^T$ ελαττώνεται, αλλά ο χρόνος υπολογισμού αυξάνεται.

Ενδιαφέρον παρουσιάζει το γεγονός πως η παράμετρος d εκτός από τη σύγκλιση της μεθόδου επηρεάζει την ευαισθησία του παραγόμενου διανύσματος διαβάθμισης σελίδων. Πιο συγκεκριμένα καθώς d-> 1, οι διαβαθμίσεις παρουσιάζουν σημαντικές διακυμάνσεις ακόμη και για μικρές αλλαγές της δομής του διαδικτύου. Η δυναμική φύση του διαδικτύου καθιστά την ευαισθησία του μοντέλου σημαντικό παράγοντα. Στην ιδανική περίπτωση δύναται η παραγωγή μιας διαβάθμισης που θα είναι σταθερή για τέτοιες μικρές αλλαγές .

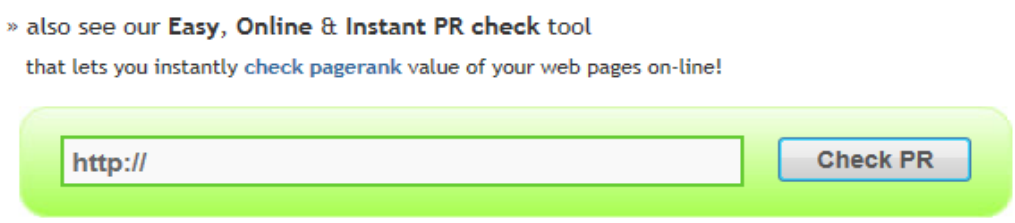
d	Αριθμός επαναλήψεων
0.5	34
0.75	81
0.8	104
0.85	142
0.9	219
0.95	449
0.99	2292
0.999	23015

Σχήμα 18 : Επίδραση της παραμέτρου d στον αναμενόμενο αριθμό επαναλήψεων της δυναμικής μεθόδου. (Power Method) (Langville & Meyer,2010)

2.2.9 Πως βλέπουμε το PageRank μιας ιστοσελίδας

Εάν ένας χρήστης θέλει να δει το PR μιας σελίδας μπορεί να χρησιμοποιήσει πολλούς τρόπους. Ο πιο απλός και συνηθισμένος τρόπος είναι η εγκατάσταση του Google Toolbar στον υπολογιστή. Είναι διαθέσιμο για IE και Firefox. Με την εγκατάσταση του εμφανίζεται μια μπάρα εργαλείων στον Browsertου υπολογιστή. Κάθε φορά που ο χρήστης επισκέπτεται μια σελίδα μόλις ολοκληρωθεί η φόρτωσή της, η μπάρα αυτή γεμίζει με πράσινο χρώμα δείχνοντας το PageRank της συγκεκριμένης σελίδας. Επίσης υπάρχουν και κάποιες ιστοσέλιδες μέσα από τις οποίες μπορούμε να δούμε το Page Rank, για παράδειγμα http://www.prchecker.info/ και http://www.checkpagerank.net. Σε αυτές τις ιστοσελίδες το μόνο που χρειάζεται είναι η εισαγωγή του link της σελίδας της οποίας υπολογίζεται το Page Rank. Τέλος το εργαλείο Page Rank Checker διατίθεται δωρεάν και τοποθετώντας το ο χρήστης στον ιστοχώρο του μπορεί άμεσα να ελέγχει την βαθμολογία όλων των ιστοσελίδων του. Για την εγκατάσταση του Page Rank Checker, ο χρήστης χρειάζεται να προσθέσει ένα κομμάτι HTML κώδικα στις ιστοσελίδες όταν θέλει να ελέγξει την βαθμολογία που έχουν. Το εργαλείο Page Rank Checker θα εμφανίσει στην οθόνη ένα μικρό εικονίδιο με τη βαθμολογία της σελίδας, σύμφωνα με τον αλγόριθμο Page Rank (Paper,2008). Ο κώδικας που θα πρέπει να προστεθεί στο HTML έγγραφο είναι ο εξής:

```
<table align="center" cellspacing="1" cellpadding="5" border="1">
<tr><td><b>Έλεγξε την βαθμολογία του αλγορίθμου Page Rank άμεσα
</b></td></tr>
<form action= http://www.prchecker.info/check_page_rank.php
method="post">
<tr><td><input type=hidden name="action" value="docheck">
<input type="text" value="http://" name="urlo" size="30"
maxlength="300">
<input type="submit" name="do_it_now" value="Check PR">
</td></tr></form>
```



Σχήμα 19

2.2.10 Πόσο Σημαντικό είναι το PageRank Σήμερα - 10 κορυφαίοι παράγοντες κατάταξης της Google

Όταν η Google πρωτοεμφάνισε το PageRank, αυτός ήταν ο κύριος παράγοντας με τον οποίο μια ιστοσελίδα μπορούσε να εμφανίζεται στις πρώτες θέσεις των οργανικών αποτελεσμάτων. Παρόλα αυτά, μόλις η κοινότητα των seoists άρχισε να αναλύει το PageRank, άρχισε να κατανοεί πως λειτουργεί και άρα πώς να φτιάχνονται ιστοσελίδες, οι οποίες θα επηρεάζουν τα οργανικά αποτελέσματα. Αυτό οδήγησε σταδιακά στην ελαχιστοποίηση της σημαντικότητας του παράγοντα PageRank, για αυτό και σήμερα ελάχιστα μετράει στην κατάταξη των ιστοσελίδων. Κάνοντας μια αναζήτηση για έναν όρο, μπορούμε να βρούμε ιστοσελίδες που δεν έχουν καθόλου ή ελάχιστο PageRank και παρόλα αυτά να βρίσκονται σε υψηλότερες θέσεις από άλλες ιστοσελίδες με υψηλό PageRank. Αυτό που έχει επίσης σημασία να τονιστεί, είναι πως η κάθε σελίδα της ιστοσελίδας μας έχει διαφορετικό PageRank και αυτό γιατί η κάθε σελίδα είναι μοναδική. (Paper,2008).

Όπως αναφέρει η Google υπάρχουν πάνω από 200 παράγοντες που χρησιμοποιεί η ίδια για να ταξινομήσει τις ιστοσελίδες στα αποτελέσματα αναζήτησης. Επειδή όμως οι παράγοντες αυτοί και το βάρος σημαντικότητας που έχει το καθένα αποτελούν εφτασφράγιστο μυστικό αυτής της μεγαλύτερης μηχανής αναζήτησης θα ήταν χρήσιμο για αυτούς που τους ενδιαφέρει το search engine optimization (SEO Βελτιστοποίηση Μηχανών Αναζήτησης) και γενικότερα το search engine marketing να γνωρίζουν τους 10 κορυφαίους παράγοντες κατάταξης. Για να παρουσιάσουμε μία τέτοια λίστα θα ήταν καλύτερα να έχουμε την άποψη των ειδικών. Η SEOmoz.com πραγματοποίησε μια εξαιρετική έρευνα που συγκεντρώνει την εμπειρία 37 κορυφαίων στο κόσμο που ασχολούνται με search engine optimization (SEO). Από κοινού αυτοί ψήφισαν διάφορους παράγοντες που εκτιμάται ότι περιέχει ο αλγόριθμος της Google για τις κατατάξεις των αποτελεσμάτων. (Wikipedia) Παρακάτω παρουσιάζουμε τους 10 κορυφαίους παράγοντες όπως τους ψήφισαν οι ειδικοί:

- Λέξεις κλειδιά που χρησιμοποιούνται στο τίτλο της σελίδας
- Το anchor text των εισερχόμενων συνδέσμων

- Το γενικό link popularity ενός ιστοχώρου (PageRank score)
- Η ηλικία του ιστοχώρου
- Το link popularity στην εσωτερική δομή του ιστοχώρου
- Πόσο σχετικοί είναι οι εισερχόμενοι σύνδεσμοι σε μία ιστοσελίδα
- Το link popularity μίας ιστοσελίδας στην τοπική κοινότητα

2.2.11 Ομοιότητες - Διαφορές PageRank και HITS (query-dependent)

Ο Αλγόριθμος HITS όμοια με τον αλγόριθμο PageRank εκμεταλλεύεται τη δομή υπερσυνδέσμων του διαδικτύου για να εξάγει βαθμούς δημοτικότητας ιστοσελίδων. Διαφέρει από τον PageRank σε δύο βασικά σημεία. Ενώ ο PageRank παράγει ένα βαθμό δημοτικότητας για μια σελίδα ο HITS παράγει δύο βαθμούς αυτοί είναι οι **authority score ai** και **hub score hi**. Επιπλέον ο HITS όπως έχει αναφερθεί εξαρτάται από το αντικείμενο αναζήτησης (query-dependent) ενώ ο PageRank είναι ανεξάρτητος από αυτό (query-independent). Εξετάζοντας πιο προσεχτικά τους δύο αυτούς αλγορίθμους παρατηρείται ότι ο PageRank υπερτερεί του HITS γιατί το Query-time κοστίζει λιγότερο (ανεξάρτητος) και είναι λιγότερο επιρρεπής σε τοπικές συνδέσεις (τα λεγόμενα spam). Από την άλλη πλευρά ο HITS υπερτερεί του PageRank αφού είναι ευαίσθητος ως προς το θέμα της αναζήτησης.

Όσον αφορά την σταθερότητα των δύο αυτών αλγορίθμων. Ο αλγόριθμος HITS είναι πολύ ευαίσθητος ακόμα και σε μια μικρή αλλαγή στο κλάσμα των ακμών / κόμβων και στη δομή του συνδέσμου ενώ ο PageRank είναι πιο σταθερός εξαιτίας της χρήσης του «τυχαίου άλματος». Χρησιμοποιώντας μια παραλλαγή του αλγορίθμου HITS σαν **Bidirectional Random Walk** παρουσιάζεται πιο σταθερός σε επιμέρους αλλαγές.

Ειδικότερα :

- Με πιθανότητα d: Τυχαία πιθανότητα p=1/n για να πηδήσει σε έναν κόμβο.

- Με πιθανότητα d-1:
  - odd timestep: Επιλέξτε τυχαία εξωτερικό σύνδεσμο (outlink) από τον τρέχον κόμβο
  - even timestep: Πηγαίνετε προς τα πίσω σε τυχαίους εσωτερικούς συνδέσμους (inlinks) του τρέχοντος κόμβου.

Αυτή η παραλλαγή του HITS φαίνεται πιο σταθερή όταν το d αυξάνεται. (Nidhi Grover ,2012a)

Σχήμα 20

Στην πραγματικότητα τα αποτελέσματα των δύο παραπάνω αλγορίθμων κατάταξης είναι εντελώς διαφορετικά και αυτό οφείλεται στο ότι υπάρχουν και επιπλέον παράγοντες που επηρεάζουν την κατάταξη των αποτελεσμάτων. Αξίζει να σημειωθεί ότι στο διαδίκτυο υπάρχει μια ιστοσελίδα με url http://matalon.org/search-algorithms/ που συγκρίνει τους αλγορίθμους αυτούς προσομοιώνοντας τους. Αρχικά γίνεται επιλογή μιας οποιαδήποτε ιστοσελίδας που επιθυμεί ο χρήστης και με το πρόγραμμα που είναι διαθέσιμο ανιχνεύεται ένα μεγάλο μέρος των διευθύνσεων που συνδέονται άμεσα ή έμμεσα με την αρχική ιστοσελίδα. Τα βήματα που χρησιμοποιεί το πρόγραμμα αυτό είναι τα εξής: Ανίχνευση στο Διαδίκτυο, Εντοπισμός των διευθύνσεων που είναι συνδεδεμένων με την αρχική σελίδα που επιθυμεί ο χρήστης, χτίσιμο και παραγωγή των αντίστοιχων πινάκων που περιέχουν 1 και 0, εμφάνιση των αποτελεσμάτων και στο τέλος σύγκριση των αλγορίθμων. Επιλέγοντας ως

αρχική σελίδα το google και εκτελώντας όλα τα βήματα διαπιστώνεται ότι τα αποτελέσματα των αλγορίθμων είναι εντελώς διαφορετικά (Grover ,2012b).

2.2.12 Ομοιότητες- Διαφορές PageRank και HITS (query-independent)

Ο πίνακας Q του PageRank αναλύεται σε δυο άλλους πίνακες τον D και L.

$$\mathbf{Q= D^{-1}\ L}$$

Όπου **Q** είναι ένας διαγώνιος πίνακας που κρατάει τους βαθμούς των εξερχόμενων σελίδων και **L** ο πίνακας γειτνίασης όλου του γραφήματος με στοιχεία άσσους και μηδενικά. Σε κάθε επανάληψη έχουμε τους πολλαπλασιασμούς πίνακα - διάνυσμα.

Στην περίπτωση του HITS: $\mathbf{L^T L\ x^{(k-1)}}$

Στην περίπτωση του PageRank: $\mathbf{L^T D^{-1}\ x^{(k-1)}}$

Μέθοδος	Πολλαπλασιασμοί	Προσθέσεις
HITS	0	$2nnz(\mathbf{L})$
Τροποποιημένος HITS	0	$4nnz(\mathbf{L}) + 2n$
Τυχαίος περιηγητής PageRank	n	$nnz(\mathbf{L}) + n$
Ευφυής περιηγητής PageRank	$nnz(\mathbf{q})$	$nnz(\mathbf{q}) + n$

Σχήμα 21: Αναφέρονται οι υπολογισμοί που απαιτούνται σε κάθε επανάληψη για τις παραπάνω τέσσερις μεθόδους. Όπου n το μέγεθος ενός πίνακα και nnz τα μη μηδενικά στοιχεία του πίνακα αυτού.

Για τον τροποποιημένο HITS που είναι ανεξάρτητος του ερωτήματος ισχύει **nnz(G)=nnz(L)** ενώ για τον αρχικό HITS ισχύει **nnz(L)<< nnz(G)** όπου G πίνακας υπερσυνδέσμων στο PageRank. Ο τροποποιημένος HITS απαιτεί από δύο έως τέσσερις φορές περισσότερο έργο απ' ότι ο PageRank. (Ding, 2002).

ΘΕΩΡΗΜΑ

Έστω ο $M=\xi L^T L + (1-\xi)/ n \ e \ e^T$ όπου $0<\xi<1$ και M τροποποιημένος πίνακας αυθεντίας.

Έστω $\lambda 1>=\lambda 2>\ldots>=\lambda n$ οι ιδιοτιμές του $L^T L$ και $\gamma 1>=\gamma 2>=\ldots\gamma n$ οι ιδιοτιμές του M.

Τότε, ισχύει:

$$\gamma 1>=\alpha\lambda 1>=\gamma 2>=\alpha\lambda 2>=\ldots..>=\gamma n>=\alpha\lambda n.$$

Επιπλέον, υπάρχουν βαθμωτά μεγέθη $\beta i >=0$, $\Sigma \beta i=1$ τέτοια ώστε $\gamma i=\xi\lambda i + (1-\xi)\beta i$ (Chris H.Q. Ding (2001-2003)).

Μέθοδος	Γενική περίπτωση
HITS	$\frac{\lambda_2}{\lambda_1}$
Τροποποιημένος HITS	$\frac{\xi\lambda_2}{\xi\lambda_1+1-\xi} \leq \frac{\gamma_2}{\gamma_1} \leq \frac{\lambda_2}{\lambda_1} + \frac{(1-\xi)}{\xi\lambda_1}$
Τυχαίος περιηγητής PageRank	d
Ευφυής περιηγητής PageRank	d

Σχήμα 22: Σύγκριση των ασυμπτωτικών ρυθμών σύγκλισης των τεσσάρων μεθόδων

Τα όρια για τον λόγο $\gamma 2/\gamma 1$ εξάγονται από την εξέταση των ακραίων καταστάσεων. Στην καλύτερη περίπτωση η μετατροπή του $L^T L$ αυξάνει μόνο τη $\lambda 2$ κατά το μέγιστο, σε $\lambda 2+ 1-\xi$. Στη χειρότερη περίπτωση μόνο η $\lambda 1$ αυξάνεται σε $\lambda 1+1-\xi$. Στην πράξη πολλά βi αλλάζουν ταυτόχρονα ωστόσο το άθροισμα όλων των β για κάθε ι παραμένει ίσο με τη μονάδα κάνοντας έτσι τις επιπτώσεις λιγότερο εμφανείς από ότι είναι στις ακραίες περιπτώσεις. Ανεξάρτητα από τις ακραίες τιμές των βi για τον τροποποιημένο αλγόριθμο συνήθως επιλέγεται $\xi=1$ οπότε και $\gamma 2/\gamma 1$ περίπου ίσο με $\lambda 2/\lambda 1$. Έτσι ο ασυμπτωτικός ρυθμός σύγκλισης του τροποποιημένου HITS είναι σχεδόν ίσος με αυτόν του βασικού HITS. Διάφορα πειράματα έχουν δείξει πως $\lambda 2/\lambda 1< 0.5$ για τον HITS κάτι το οποίο είναι πολύ μικρότερο από το d=0.85 που

χρησιμοποιείται τυπικά σαν παράγοντας του PageRank. Έτσι ο HITS και ο τροποποιημένος HITS απαιτούν πολύ λιγότερες επαναλήψεις από τον PageRank. Με περίπου διπλάσιο κόστος ανά επανάληψη ο τροποποιημένος HITS απαιτεί λιγότερο από ¼ των επαναλήψεων του PageRank και παράγει 2 διανύσματα στο χρήστη. (Rajaraman, 2001d).

2.3 Stochastic Approach for Link Structure Analysis (SALSA)

Ο αλγόριθμος SALSA δημιουργήθηκε από τον Lempel και Moran το 2000 και αποτελεί συνδυασμό των παραδοσιακών αλγορίθμων HITS και PageRank. Η ομοιότητά του με τον HITS είναι ότι χρησιμοποιεί authority και hub score και δημιουργεί ένα γειτονικό γράφο χρησιμοποιώντας authority και hub σελίδες και links. Η διαφοροποίηση του αλγορίθμου SALSA από το HITS εντοπίζεται στο ότι καταφέρνει να αναγνωρίσει και να ανιχνεύσει περισσότερες σελίδες ως authorities, σε θεματικές ομάδες εγγράφων όπου ο HITS αδυνατεί. Επιπρόσθετα ο αλγόριθμος SALSA θεωρεί λιγότερο στενή τη σχέση ανάμεσα στις authority και hub σελίδες (Moran, 2010 a).

Ο αλγόριθμος SALSA ξεκινάει όπως ο HITS κατασκευάζοντας ένα αρχικό, βασικό σύνολο ιστοσελίδων. Στη συνέχεια πραγματοποιεί έναν τυχαίο περίπατο επιλέγοντας εναλλακτικά είτε α) την μετακίνηση ομοιόμορφα σε μια από τις ιστοσελίδες που δείχνουν στην ιστοσελίδα στην οποία βρίσκεται εκείνη τη στιγμή είτε β) την μετακίνηση ομοιόμορφα σε μια από τις ιστοσελίδες που δείχνονται από την ιστοσελίδα στην οποία βρίσκεται εκείνη τη στιγμή. Τα authority scores προσδιορίζονται ως η stationary distribution της αλυσίδας δύο βημάτων όπου πρώτα εκτελείται το βήμα α) και μετά το β) . Αντίστοιχα τα hub scores προσδιορίζονται ως η stationary distribution της αλυσίδας δύο βημάτων όπου πρώτα εκτελείται το βήμα β) και μετά το α). (Signorini,2005).

Επισήμως η Markov Chain για τις authorities έχει τις ακόλουθες πιθανότητες μετάβασης:

$$P_a(i,j) = \sum_{k:k\in B(i)\cap B(j)} \frac{1}{|B(i)|}\frac{1}{|F(k)|}$$

Θεωρείται ότι Gα =(A, Eα) είναι ένας authority γράφος στον οποίο υπάρχει μία ακμή ανάμεσα σε δύο authorities εάν μοιράζονται ένα hub. Αυτή η Markov αλυσίδα ανταποκρίνεται σε έναν τυχαίο περίπατο για τον authority γράφο Gα, στον οποίο μετακινούμαστε από την authority i στην authority j με πιθανότητα Pα (i , j) . Οι πίνακες H και A για τα authorities και τα hubs αντίστοιχα προέρχονται από τον **adjacency πίνακα L** χρησιμοποιώντας τις μεθόδους του PageRank και του HITS. Ο HITS χρησιμοποιεί μη-σταθμισμένο πίνακα L ενώ ο PageRank χρησιμοποιεί ένα σταθμισμένο ως προς τις γραμμές πίνακα L. Σε αντίθεση με τους προηγούμενους ο SALSA χρησιμοποιεί έναν πίνακα L σταθμισμένο και ως προς γραμμές και ως προς στήλες. Ως Lr ορίζεται ένας πίνακας που αποτελείται από τα στοιχεία του L τα οποία είναι διαιρεμένα με το άθροισμα των γραμμών του. Αντίστοιχα Lc είναι ένας πίνακας που αποτελείται από τα στοιχεία του L διαιρεμένα με το άθροισμα των στηλών του

Συνεπώς ο πίνακας H αποτελείται από τις μη-μηδενικές γραμμές και στήλες του πίνακα $L_r L_c^T$ ενώ ο A αποτελείται από τις μη-μηδενικές γραμμές και στήλες του πίνακα $L_c^T L_r$.Οπότε η stationary distribution είναι το αριστερό ιδιοδυάνυσμα του πίνακα $L_c^T L_r$. Εάν ο γράφος Gα αποτελείται από πολλές συνιστώσες τότε ο αλγόριθμος SALSA επιλέγει έναν κόμβο (ως σημείο εκκίνησης) ομοιόμορφα στην τύχη και εκτελεί μια τυχαία διαδρομή εντός της συνεκτικής συνιστώσας που περιέχει τον συγκεκριμένο κόμβο (Moran,2010 b).

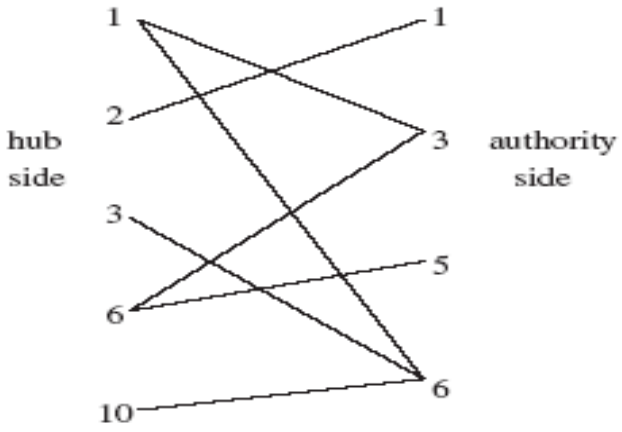
Έστω j μια συνιστώσα που περιέχει τον κόμβο i , τότε Aj ορίζεται ένα σύνολο από authorities με συνιστώσα j και Ej ένα σύνολο από link με συνιστώσα j.
Τότε το βάρος του authority i με συνιστώσα j είναι:

$$a_i = \frac{|A_j|}{|A|} \frac{|B(i)|}{|E_j|}$$

Εάν ο γράφος Gα αποτελείται από μια μόνο συνιστώσα (οι γράφοι αυτοί αναφέρονται ως authority connected graphs) τότε η αλυσίδα Markov είναι irreducible και ο αλγόριθμος μειώνεται σε InDegree αλγόριθμο. Ακόμη και αν ο γράφος Gα δεν είναι συνεκτικός, εάν το σημείο εκκίνησης της τυχαίας διαδρομής επιλέγεται με πιθανότητα αναλογική της δημοτικότητας (in-degree) τότε πάλι ο αλγόριθμος μειώνεται σε InDegree αλγόριθμο (Signorini,2005).

2.3.1 Παράδειγμα εφαρμογής του αλγορίθμου SALSA

Ο Bipartite γράφος του Σχήματος 8 σύμφωνα με τους Langville & C. Meyer (2005 σ.20 & Lempel Moran, 2010 c), παρουσιάζεται στο σχήμα 23 από τον οποίο διεξάγεται ότι το σύνολο των ιστοσελίδων από την hub side είναι το εξής: Vh ={ 1,2,3,6,10} ενώ το σύνολο των ιστοσελίδων από την authority side είναι το εξής: Va={1,3,5,6}.



Σχήμα 23: Bipartite γράφος του Σχήματος 8 (Langville & C. Meyer,2005 σ.20)

Οι πίνακες L, Lr, Lc από τους οποίους προκύπτουν οι $L_r L_c^T$ και $L_c^T L_r$. παρουσιάζονται στα παρακάτω δύο σχήματα:

$$\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \quad \mathbf{L}_r = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array},$$

and

$$\mathbf{L}_c = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \end{array}.$$

Σχήμα 24

$$\mathbf{L}_r\mathbf{L}_c^T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & 0 & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 1 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \end{array}, \quad \mathbf{L}_c^T\mathbf{L}_r = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & \frac{5}{6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Σχήμα 25

Οπότε οι πίνακες H και είναι A οι εξής:

$$\mathbf{H} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 6 & 10 \end{array} \\ \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \end{array}.$$

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 3 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 3 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{pmatrix} \end{array}.$$

Power Method

- $X_{k+1} = AX_k$
- $X_{k+1}^T = X_k^T A$
- Συγκλίνει στο dominant ιδιοδιάνυσμα ($Av = \lambda v$, $\lambda=1$)

Οι πίνακες H και A πρέπει να είναι irreducible για να εφαρμοστεί η power method για να συγκλίνει σε ένα μοναδικό ιδιοδιάνυσμα δίνοντας οποιαδήποτε αρχική τιμή. Εάν ο γειτονικός γράφος Gα είναι συνεκτικός τότε και οι δύο πίνακες H και A είναι irreducible. Στην περίπτωση που δεν είναι irreducible η εφαρμογή της power method δεν θα έχει ως αποτέλεσμα τη σύγκλιση σε ένα dominant ιδιοδιάνυσμα. (Chang, 2011b)

Στο παράδειγμα που εξετάζεται είναι φανερό ότι ο γράφος είναι μη-συνεκτικός καθώς η σελίδα 2 του hub set συνδέεται μόνο με τη σελίδα 1 στο authority set και αντίστροφα. Συνεπώς οι πίνακες H και A είναι reducible και επομένως αποτελούνται από πολλές irreducible συνεκτικές συνιστώσες. Συγκεκριμένα, ο **H** περιέχει τις εξής δύο συνεκτικές συνιστώσες: **C={2} και**

D={1,3,6,10} ενώ ο **A** τις **E={1} και F={3,5,6}** (Langville & C. Meyer,2005σ 21 & Lempel Moran, 2010d)

Εφαρμόζοντας την Power Method για κάθε συνιστώσα του H και A προκύπτει ότι:

$$\pi_h^T(C) = \begin{matrix}2\\(1)\end{matrix}, \quad \pi_h^T(D) = \begin{matrix}1 & 3 & 6 & 10\\(\frac{1}{3} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6})\end{matrix},$$

$$\pi_a^T(E) = \begin{matrix}1\\(1)\end{matrix}, \quad \pi_a^T(F) = \begin{matrix}3 & 5 & 6\\(\frac{1}{3} & \frac{1}{6} & \frac{1}{2})\end{matrix}.$$

Συνενώνοντας τις δύο επιμέρους συνιστώσες σε μια πολλαπλασιάζοντας κάθε στοιχείο με το κατάλληλο βάρος. (Langville & C. Meyer,2005σ 21) Εάν το εφαρμοστεί αυτό για τον H και A αντίστοιχα προκύπτει:

$$\pi_h^T = \begin{matrix}1 & 2 & 3 & 6 & 10\\(\frac{4}{5}\cdot\frac{1}{3} & \frac{1}{5}\cdot 1 & \frac{4}{5}\cdot\frac{1}{6} & \frac{4}{5}\cdot\frac{1}{3} & \frac{4}{5}\cdot\frac{1}{6})\end{matrix}$$
$$= \begin{matrix}1 & 2 & 3 & 6 & 10\\(.2667 & .2 & .1333 & .2667 & .1333)\end{matrix}.$$

$$\pi_\alpha^T = \begin{matrix}1 & 3 & 5 & 6\\(\frac{1}{4}\cdot 1 & \frac{3}{4}\cdot\frac{1}{3} & \frac{3}{4}\cdot\frac{1}{6} & \frac{3}{4}\cdot\frac{1}{2})\end{matrix}$$
$$= \begin{matrix}1 & 3 & 5 & 6\\(.25 & .25 & .125 & .375)\end{matrix}.$$

2.3.2 Σύγκριση του Salsa με τους PageRank HITS

Επειδή ο αλγόριθμος SALSA αναπτύχθηκε ύστερα από τους HITS και PageRank προσπάθησε να συνδυάσει τα καλύτερα χαρακτηριστικά και των δύο. Αντίθετα με τον HITS, δεν είναι τόσο ευαίσθητος στο ζήτημα του topic drift. Επιπλέον ο SALSA αντιμετωπίζει με επιτυχία το πρόβλημα του spamming εξαιτίας

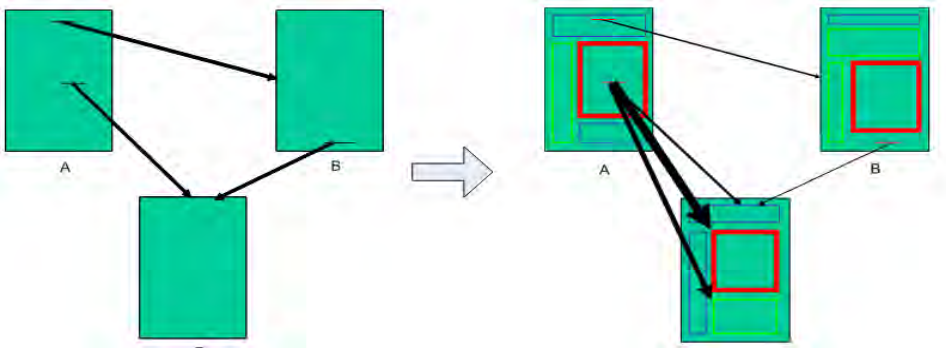
της αλληλεξάρτησης μεταξύ των authority και hub βαρών,(σε σύγκριση με τον HITS) , αφού μειώνεται η αλληλεξάρτησή τους. Όμως ο PageRank είναι ο πιο αποτελεσματικός και από τους δύο στην αντιμετώπιση αυτού του προβλήματος.

Επιπρόσθετα ένα ακόμα πλεονέκτημα του SALSA, όπως και του HITS, είναι ότι δίνει διπλές ταξινομημένες λίστες. Σημαντικό μειονέκτημα του SALSA είναι η εξάρτηση του αποτελέσματος από την επερώτηση. Κατά το χρόνο της διεξαγωγής της επερώτησης θα πρέπει να σχηματιστεί ο γράφος γειτνίασης και να επιλυθούν τα δύο προβλήματα της αλυσίδας Markov. Τέλος μειονεκτεί ως προς τη σύγκλιση, η οποία μοιάζει με αυτήν του HITS. Επειδή κανένας τους δεν εξασφαλίζει την irreducibility του γράφου, τα διανύσματα που προκύπτουν μπορεί να μην είναι μοναδικά. Για την αντιμετώπιση του προβλήματος γίνεται η χρήση μιας μικρής αλλαγής στο γράφημα (Najork, 2007) .

ΚΕΦΑΛΑΙΟ 3ο

Block-level Link Analysis

Οι περισσότεροι από τους υπάρχοντες link analysis αλγορίθμους αντιμετωπίζουν την ιστοσελίδα σαν έναν ενιαίο κόμβο στο γράφο ιστού (web graph). Ωστόσο στις περισσότερες περιπτώσεις μια ιστοσελίδα μπορεί να περιέχει πολλαπλή σημασιολογία και ως εκ τούτου δεν θα μπορούσε να θεωρηθεί ως ενιαίος κόμβος. Θεωρώντας επομένως ότι μια ιστοσελίδα είναι χωρισμένη σε πολλά blocks και χρησιμοποιώντας τον αλγόριθμο vision-based page segmentation με την ανάλυση page-to-block και block-to-page σχέσεις μπορεί να κατασκευαστεί ένας page graph και ένας block graph και γενικότερα ένας σημασιολογικός γράφος πάνω από το διαδίκτυο (www) ώστε κάθε κόμβος να αντιπροσωπεύει ένα συγκεκριμένο σημασιολογικό θέμα. Έχοντας ως βάση την block ανάλυση σε επίπεδο σύνδεσης προτίνοται δύο καινούριοι αλγορίθμοι ο Block Level PageRank και ο Block Level HITS ως διαφορετική προσέγγιση των παραδοσιακών αλγορίθμων PageRank και HITS (Jiang κ.συν, 2004)



Σχήμα 26: Μια ιστοσελίδα μπορεί να χωριστεί σε πολλά blocks

3.1 VISION-BASED PAGE SEGMENTATION

Ο Vision-Based Page Segmentation (VIPS) αλγόριθμος στοχεύει να εξάγει σημασιολογική δομή μιας ιστοσελίδας βασισμένη σε οπτική παρουσίαση. Μια τέτοια σημασιολογική δομή είναι μια δομή δέντρου όπου κάθε κόμβος στο δέντρο αντιστοιχεί σε ένα block. Σε κάθε κόμβο δίνεται μια τιμή (Βαθμός Συνοχής) η οποία δείχνει πόσο συνεκτικό είναι το περιεχόμενο του μπλοκ βάση της οπτικής αντίληψης. Ο αλγόριθμος αυτός χρησιμοποιεί πλήρως τις page-to-block σχέσεις. Αρχικά εξάγει όλα τα κατάλληλα μπλοκ από το html DOM (Document Object Model) δέντρο το οποίο είναι ένας προγραμματισμός API (Application Programming Interface) για έγγραφα. Το dom μοιάζει πολύ με τη δομή των εγγράφων που μοντελοποιεί. Στη συνέχεια ο αλγόριθμος βρίσκει τους διαχωρισμούς μεταξύ αυτών των μπλοκ. Οι διαχωρισμοί αυτοί ορίζονται από οριζόντιες και κάθετες γραμμές που οπτικά διασταυρώνονται σε μια ιστοσελίδα χωρίς κανένα block. Βασιζόμενοι σ' αυτούς τους διαχωρισμούς κατασκευάζεται το σημασιολογικό δέντρο μιας ιστοσελίδας.

Έτσι μια ιστοσελίδα παρουσιάζεται ως ένα σύνολο από μπλοκ (τα φύλλα-κόμβοι του σημασιολογικού δέντρου). «Θορυβώδεις» πληροφορίες όπως είναι η πλοήγηση, οι διαφημίσεις και η διακόσμηση μπορούν πολύ εύκολα να μετακινηθούν επειδή τοποθετούνται συχνά σε συγκεκριμένες θέσεις σε μια ιστοσελίδα. Περιεχόμενα με διαφορετικά θέματα διακρίνονται σαν διαφορετικά μπλοκ. Στην παρακάτω εικόνα παριστάνεται ένα δείγμα από μια ιστοσελίδα (news.yahoo.com). Η σελίδα αυτή αποτελείται από διαφορετικής σημασίας μπλοκ (διαφορετικό χρώμα περιγράμματος). Τα διαφορετικά μπλοκ έχουν διαφορετική σημαντικότητα για τη συγκεκριμένη σελίδα (Cai κ.συν., 2004).

Σχήμα 27 Οι σύνδεσμοι διαφορετικών μπλοκ δείχνουν σε σελίδες με διαφορετικά θέματα (Cai κ.συν, 2004)

Έστω P το σύνολο όλων των ιστοσελίδων P={p1,p2,…pk} όπου k είναι ο αριθμός όλων των ιστοσελίδων και B το σύνολο όλων των blocks , B={ b1, b2,..bn} όπου n ο αριθμός των blocks. Για κάθε block υπάρχει μόνο μια σελίδα που το περιέχει, δηλαδή bi ανήκει pi σημαίνει ότι το block i περιέχεται στην σελίδα j (Cai, κ.συν, 2004)

Block-to-page matrix Z

Ορίζοντας Z έναν Block-to-page matrix με διαστάσεις n*k ως:

$$Z_{ij} = \begin{cases} 1/s_i & \text{if there is a link from block } i \text{ to page } j \\ 0 & \text{otherwise} \end{cases}$$

Όπου si είναι ο αριθμός των σελίδων όπου το block i συνδέεται και Zij μπορεί να θεωρηθεί ως η πιθανότητα άλματος από το block i στην σελίδα j.

Page-to-block matrix X

Είναι γνωστό ότι Page-to-block σχέσεις περιέχονται σε μια ανάλυση σελίδας. Ειδικότερα ορίζοντας έναν πίνακα Page-to-block με διαστάσεις k*n ως:

$$X_{ij} = \begin{cases} 1/s_i & if\ b_j \in p_i \\ 0 & otherwise \end{cases}$$

Όπου si ο αριθμός των blocks που περιέχονται σε μια σελίδα i.

Ο παραπάνω τύπος αναθέτει τιμή ίσιας σημαντικότητας για κάθε block μιας σελίδας. Αυτό είναι απλό αλλά λιγότερο πρακτικό. Μερικά blocks με μεγάλο μέγεθος και κεντρική τοποθέτηση στην σελίδα μπορεί να είναι πιο σημαντικά από κάποια άλλα με μικρό μέγεθος και που είναι στο περιθώριο. Αυτή η παρατήρηση οδηγεί στον παρακάτω τύπο:

$$X_{ij} = \begin{cases} f_{P_i}(b_j) & if\ b_j \in p_i \\ 0 & otherwise \end{cases}$$

Όπου F είναι η συνάρτηση η οποία αναθέτει σε κάθε block b που περιέχεται σε μια σελίδα p, μια τιμή σημαντικότητας. Ειδικότερα η fp(b) είναι τόσο σημαντική όσο είναι το αντίστοιχο block b. Εμπειρικά προκύπτει:

$$f_p(b) = \beta \frac{\text{size of block } b \text{ in page } p}{\text{dist. from the center of } b \text{ to the center of screen}}$$

Όπου β είναι ένας παράγοντας κανονικοποίησης ώστε το άθροισμα των fp(b) για κάθε block που ανήκει στη σελίδα p να είναι 1.

$$\sum_{b \in p} f_p(b) = 1$$

Η fp(b) μπορεί να θεωρηθεί ως η πιθανότητα κατά την οποία ο χρήστης να επικεντρωθεί στο block b όταν «παρατηρεί» τη σελίδα p. Cai , κ.συν.,2004 σελ 2).

3.2 Block Level Page Rank

Ο Block Level Page Rank (BLPR) είναι παρόμοιος με τον αρχικό αλγόριθμο PageRank ως προς το πνεύμα. Η βασική διαφορά μεταξύ τους είναι ότι ο παραδοσιακός αλγόριθμος PageRank βασίζεται σε ανάλυση επιπέδου σελίδας ενώ ο Block Level Page Rank σε ανάλυση επιπέδου block.

Ως A=XZ ορίζεται ένας πίνακας βαρών του γραφήματος G που αναφέρθηκε παραπάνω. Αρχικά κατασκευάζεται ως πίνακας πιθανοτήτων μετάβασης (probability transition) ο πίνακας M ξανα-κανονικοποιώντας (renormalizing) κάθε γραμμή του A ώστε να έχει άθροισμα τη μονάδα. Στη συνέχεια γίνεται αναφορά σε έναν τυχαίο περιηγητή ο οποίος σε κάθε χρονικό βήμα βρίσκεται σε κάποια ιστοσελίδα και αποφασίζει ποια σελίδα να επισκεφτεί στο επόμενο βήμα ως εξής:

- Με πιθανότητα 1-ε επιλέγει τυχαία μία από τις υπερσυνδέσεις της τρέχουσας σελίδας και πραγματοποιεί άλμα στην ιστοσελίδα με την οποία συνδέεται.
- Με πιθανότητα ε ο τυχαίος περιηγητής «resets» με άλμα σε μια ιστοσελίδα που διαλέγει ομοιόμορφα και τυχαία από τη συλλογή. Όπου ε κατάλληλη παράμετρος.

Η διαδικασία αυτή ορίζει Markov Chain πάνω σε ιστοσελίδες με πίνακα μετάβασης εU +(1-ε)M όπου U είναι ένας πίνακας μετάβασης με ομοιόμορφη πιθανότητα μετάβασης (Uij= 1/n για κάθε i, j). Ο πίνακας των PageRank scores ορίζεται από την σταθερή κατανομή της Markov Chain δηλαδή το αριστερό ιδιοδιάνυσμα του πίνακα μετάβασης.

Μαθηματικά BLPR μπορεί να υπολογιστεί ως ακολούθως:

$$(\varepsilon U + (1-\varepsilon)M)^T \mathbf{p} = \mathbf{p}$$

Όπου p είναι ένα διάνυσμα του οποίου το i-οστο στοιχείο είναι το PageRank σε επίπεδο block της i-στης ιστοσελίδας. (Deng Cai κ.συν ,2004σελ 3)

3.3 Block Level HITS

Ο Block Level HITS (BLHITS) διαφέρει από τον Block Level Page Rank (BLPR) στο γεγονός ότι αποδίδει δύο τιμές σε κάθε σελίδα (authority value και hub value) ενώ ο BLPR μόνο μία. Όπως αναλύσαμε προηγουμένως υπάρχουν πολλαπλές σημασιολογικές περιοχές σε μια ιστοσελίδα. Μερικοί υπερσύνδεσμοι όπως banners, navigation Panels και διαφημίσεις σε μια σελίδα δεν μεταφέρουν περιεχόμενο επιθυμητό για το χρήστη είναι άχρηστοι. Έτσι παρουσιάζουμε τον Block Level HITS αλγόριθμο βασιζόμενος στην ίδια ιδέα με τον παραδοσιακό αλγόριθμο HITS με τη διαφορά ότι στον Block Level HITS μια ιστοσελίδα θα έχει μόνο authority score και ένα block μόνο hub score.

Η σχέση μεταξύ authority και hub στον BLHITS περιγράφεται ως εξής:

$$A = Z^T H \qquad H = ZA$$

Όπου A ορίζεται το διάνυσμα που περιέχει τις τιμές authority για τις σελίδες, H ορίζεται ως το διάνυσμα που περιέχει τις τιμές των hub για τα blocks και Z ο block-to-page πίνακας που ορίσαμε πιο πάνω. (Deng Cai κ.συν 2004, σελ3)

3.4 Σύγκριση Block Level HITS (BLHITS) και HITS

- Η ανάλυση του BLHITS είναι σε επίπεδο block και εστιάζει στους υπερσυνδέσμους από block σε pages ενώ ο HITS εστιάζει στους υπερσυνδέσμους από pages σε pages.
- Το root set στον BLHITS αποτελείται από blocks με κορυφαία κατάταξη (rank) σε αντίθεση με το αντίστοιχο του HITS το οποίο αποτελείται από κορυφαίες σε κατάταξη σελίδες. Όταν ένα αίτημα υποβάλλεται στο σύστημα που εξετάζει ένας χρήστης αρχικά γίνεται η ανάκτηση των κορυφαίων σε κατάταξη σελίδων. Στο στάδιο αυτό τα «θορυβώδη» block (όπως διαφημιστικά blocks) εξαιρούνται. Στο σύστημα που εξετάστηκε όλες οι σελίδες είναι ταξινομημένες από πριν σε επίπεδο μπλοκ οπότε γίνεται απευθείας η λήψη της κορυφής των ταξινομημένων μπλοκ χωρίς επιπλέον υπολογισμούς.
- Κατά την επέκταση του root set χρησιμοποιούνται μόνο οι εξερχόμενες συνδέσεις που περιέχονται στα block που βρίσκονται στην κορυφή της

ταξινόμησης. Ο HITS επεκτείνει όλες τις συνδέσεις μέσα στις σελίδες, οι οποίες εισάγουν αναπόφευκτα θορυβώδεις σελίδες στο βασικό σύνολο (base set). Ομοίως προστίθενται μόνο οι ομάδες από block οι οποίες περιέχουν συνδέσμους με τις σελίδες στο root set αντί για ολόκληρο το σύνολο σελίδων.

- Οι Bharat and Henzinger απέδειξαν ότι ο συνδυασμός της σύνδεσης και της ανάλυσης του περιεχομένου (κλάδεμα αυτών των κόμβων που είναι σχετικοί με το διευρυμένο ερώτημα) είναι πολύ αποτελεσματικός στην ενίσχυση της απόδοσης του HITS. Στον BLHITS αλγόριθμο οι κόμβοι είναι block, έτσι το μέτρο σχετικότητας είναι μεταξύ των block και του διευρυμένου ερωτήματος κάτι το οποίο έχει μεγαλύτερο νόημα.
- Οι Bharat and Henzinger αναθέτουν authority weight and hub weight σε κάθε ακμή για να λυθεί το πρόβλημα των αμοιβαίων σχέσεων. Το πρόβλημα αυτό μπορεί να συμβεί μόνο όταν υπάρχουν κ ακμές από έγγραφα ενός πρώτου host σε ένα single έγγραφο ενός δεύτερου host και να υπάρχουν t ακμές από ένα single έγγραφο του πρώτου host σε ένα σύνολο από έγγραφα του δευτέρου host. Παρόλα αυτά επεκτείνεται η ιδέα των authority weight and hub weight μιας ακμής πολλαπλασιάζοντας ένα επιπρόσθετο βάρος για κάθε ακμή. Το επιπρόσθετο αυτό βάρος προέρχεται από το βάρος σημαντικότητας του μπλόκ στην αρχική σελίδα. Αυτό το βάρος υπολογίζεται από την αναλογία ανάμεσα στην αξία σημαντικότητας του block που περιέχεται αυτό το link και τη μέγιστη αξία σημαντικότητας του block στη σελίδα αυτή. Έτσι το βάρος είναι 1 για το σημαντικότερο block στην σελίδα. Με τον τρόπο αυτό οι σύνδεσμοι που δεν μεταφέρουν χρήσιμη πληροφορία για τον χρήστη έχουν μικρή επιρροή στον υπολογισμό. (Cai, κ συν, 2004 σελ 4).

ΚΕΦΑΛΑΙΟ 4ο

Αλγόριθμοι Βελτιστοποίησης του Web Search με τη χρήση Κοινωνικών Σχολίων

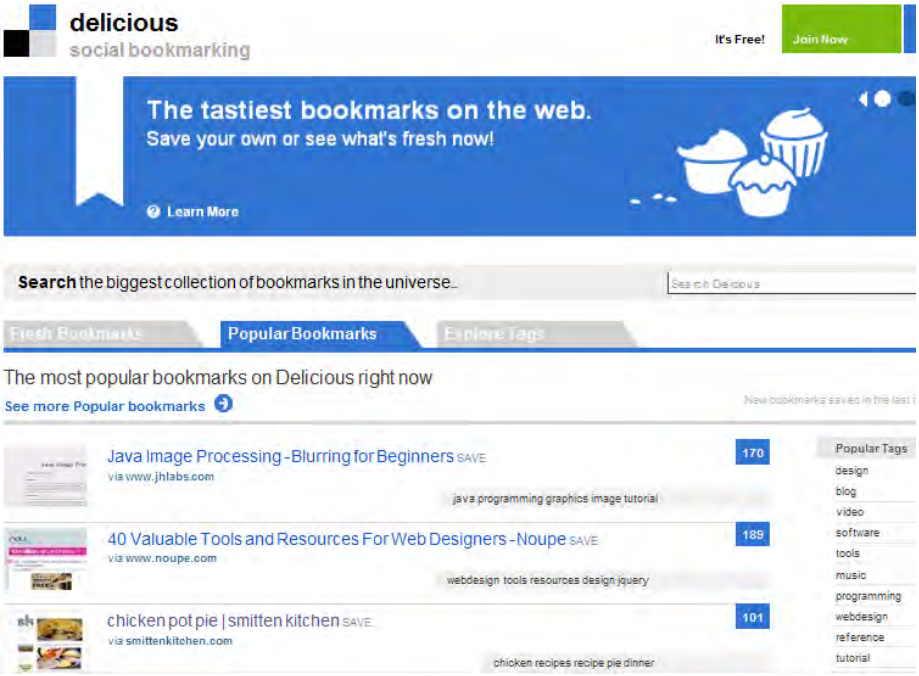
4.1 Εισαγωγή – Ορισμοί

(Optimizing Web Search Using Social Annotations)

Κατά την τελευταία δεκαετία, πολλές μελέτες έχουν γίνει για τη βελτίωση της ποιότητας της αναζήτησης στο διαδίκτυο. Οι περισσότερες από αυτές εστιάζονται σε δύο σημεία:

1) Την κατάταξη των σελίδων βάσει της ομοιότητάς τους με μια ερώτηση (**similarity ranking**). Κατά καιρούς έχουν προταθεί πολλές μέθοδοι για εύρεση πληροφορίας ως πρόσθετα μεταδεδομένα ώστε να ενισχυθεί η απόδοση της κατάταξης με βάση την ομοιότητα όπως document title , anchor text και users' query logs.
2) Την κατάταξη των ιστοσελίδων βάση της ποιότητάς τους. Γνωστή και ως query-independent ranking, ή ως στατική κατάταξη (**static ranking**). Για μεγάλο χρονικό διάστημα, η στατική κατάταξη βασιζόταν στην link analysis χρησιμοποιώντας αλγορίθμους όπως PageRank, HITS.

Πρόσφατα με την ανάπτυξη των τεχνολογιών Web 2.0 οι χρήστες του Διαδικτύου συνηθίζουν να δημιουργούν σχόλια για ιστοσελίδες με απίστευτη ταχύτητα. Για παράδειγμα η περίφημη κοινωνική bookmark ιστοσελίδα, **del.icio.us** έχει πάνω από 1 εκατομμύριο εγγεγραμμένους χρήστες αμέσως μετά την τρίτη του επέτειο και ο αριθμός των χρηστών της αυξήθηκε κατά περισσότερο από 200% στους τελευταίους εννέα μήνες. Τα σχόλια των χρηστών για τη σελίδα του Amazon στο **del.icio.us** είναι αγορά, μουσική, βιβλία και κατάστημα. Αν και τα κοινωνικά σχόλια χρησιμοποιούνται ποικιλοτρόπως για παράδειγμα στην οπτικοποίηση, folksonomy, επιχειρηματική έρευνα. (Muralidharan κ.συν, 2012)



Σχήμα 28: *Η κοινωνική bookmark ιστοσελίδα del.icio.us*

Επίσημα ως κοινωνικό σχόλιο ορίζουμε την τετράδα **{ q,u,c,v}** όπου q είναι το ερώτημα, u το περιεχόμενο, c το κοινωνικό δίκτυο σύνδεσης και v τον βαθμό ενδιαφέροντος της σύνδεσης ως like, dislike ή share. Παρακάτω παρουσιάζεται ένα παράδειγμα κοινωνικού σχολίου όπου q =“maui hotels”, u= είναι σχετικό με την σελίδα Expedia hotel , c=”Tim Harrington”και v=like (Σχήμα 29).

Σχήμα 29: Παράδειγμα κοινωνικού σχολίου (Pantel, κ.συν. 2012, σελ 1)

4.1.1 Query (Ερωτήματα)

- **Query Intent (QA-INT):** Στην κατηγορία αυτή τα ερωτήματα χωρίζονται με βάση το αν είναι πλοήγησης ή μη. Αναμένεται ότι τα κοινωνικά σχόλια θα είναι λιγότερο χρήσιμα εάν ο χρήστης απλά ψάχνει για το Url μιας ιστοσελίδας στην οποία θέλει να φτάσει.
- **Query Class (QA-CLS):** Η χρησιμότητα ενός κοινωνικού σχολίου μπορεί να επηρεαστεί από το είδος-τάξη του θέματος του ερωτήματος. Εστιάζοντας στις ακόλουθες κατηγορίες με βάση το είδος του ερωτήματος καλύπτεται έτσι η πλειοψηφία των κοινωνικών σχολίων που εκμαιεύεται από μια εμπορική μηχανή αναζήτησης.

    - **Commerce (com):** Ερωτήματα σχετικά με τιμές, συγκρίσεις, συναλλαγές και σχόλια.
    - **Health (hea):** Ερωτήματα πάνω σε θέματα που σχετίζονται με την υγεία όπως συμπτώματα, διαδικασίες, θεραπείες.
    - **Movies (mov):** Ερωτήματα σχετικά με τίτλους ταινιών.
    - **Music (mus):** Ερωτήματα για μουσικούς, μπάντες και lyrics από τραγούδια.
    - **Restaurant (res):** Τοπικά ερωτήματα σχετικά με εστιατόρια και καφετέριες.

(Patrick Pantel κ.συν 2012, σελ. 3)

4.1.2 Social Connection (Κοινωνική Δικτύωση):

Η κοινωνική σύνδεση με το διαδίκτυο μέσω ιστοσελίδων κοινωνικής δικτύωσης είναι η βάση των κοινωνικών σχολίων. Έστω το ερώτημα «Korean restaurant» και ένα σύνολο από συνδέσμους-αποτελέσματα για Korean restaurants. Τότε μια σύνδεση κάποιου ο οποίος είναι έμπειρος σε κορεάτικη κουζίνα ίσως αυξήσει την χρησιμότητα του σχολίου καθώς και μια σύνδεση κάποιου που ζει κοντά στην τοπική γειτονιά, από κάποιον που είναι μακριά. Όπως και η γνώμη ενός αγαπητού φίλου που ενδιαφέρεται για ένα εστιατόριο ίσως να έχει μεγαλύτερο βάρος από κάποιον πιο μακρινό συνάδελφο.

Προκύπτουν οι εξής κατηγορίες:

- **Circle (SA-CIR):** Ο κύκλος μιας σύνδεσης αναφέρεται στη σχέση της σύνδεσης με αυτόν που ψάχνει. Διαισθητικά, το ενδιαφέρον ενός συναδέλφου για ένα άρθρο που σχετίζεται με τον χώρο εργασίας του ίσως έχει μεγαλύτερη αξία από το ενδιαφέρον ενός μέλους της οικογένειας του ή ενός φίλου του. Οπότε θεωρούμε ότι υπάρχουν οι ακόλουθοι κύκλοι: **work colleague** (wkc), **family member** (fam) και **friend**(frn). Κάποιοι άλλοι κύκλοι εξίσου ίδιου ενδιαφέροντος θα μπορούσαν να είναι school friends, college friends, church friends και sport club friends.
- **Affinity(SA-AFF):** Η συγγένεια που υπάρχει μεταξύ του ερευνητή και μας σύνδεσης αναφέρεται ως βαθμός εγγύτητας. Υπάρχουν δύο είδη συγγένειας **close** (cls) και **distant** (dst).
- **Expertise(SA-EXP):** Όταν μια σύνδεση είναι είτε **expert**(exp) είτε **non-expert**(nex) όσον αφορά το θέμα αναζήτησης μπορεί να επηρεάσει την αξία των κοινωνικών σχολίων.
- **Geographical Distance(SA-GEO):** Αναφέρεται στην γεωγραφική απόσταση. Για τοπικά ερωτήματα αναμένουμε ότι η σύνδεση κοντά στην τοποθεσία-στόχο θα προσθέσει περισσότερη αξία από αυτήν που είναι μακριά. Για μη τοπικά ερωτήματα η αξία δεν ορίζεται (not applicable)
- **Interest Valence(SA-INT):** Με αυτόν τον τρόπο εκφράζεται το ενδιαφέρον του χρήστη για την σύνδεση. Ο χρήστης μπορεί να κάνει **like** ή **dislike** υποδηλώνοντας με αυτόν τον τρόπο ότι του αρέσει ή δεν του αρέσει η σελίδα αντίστοιχα καθώς και να μοιραστεί ένα αρχείο με κάποιον επιλέγοντας **share.**

Σχήμα 30: *Social Connection (*Patrick Pantel κ.συν 2012 σελ 4)

4.1.3 Content (περιεχόμενο):

Τέλος τα κοινωνικά σχόλια επηρεάζονται από τη συνάφεια του εγγράφου με το ερώτημα του χρήστη. Χωρίζονται στις εξής κατηγορίες:

- **Perfect:** Η σελίδα είναι κεντρική και καλύπτει στο μέγιστο το ερώτημα του χρήστη
- **Excellent**: Η σελίδα ικανοποιεί κατά πολύ το ερώτημα του χρήστη
- **Good:** Η σελίδα καλύπτει μέτρια το ερώτημα του χρήστη
- **Fair:** Η σελίδα καλύπτει «ασθενικά» το ερώτημα του χρήστη
- **Bad:** Η σελίδα δεν ικανοποιεί το ερώτημα του χρήστη
- **Detrimental:** Η σελίδα περιλαμβάνει περιεχόμενο ακατάλληλο για το γενικό κοινό. (Patrick Pantel κ.συν 2012, σελ 3)

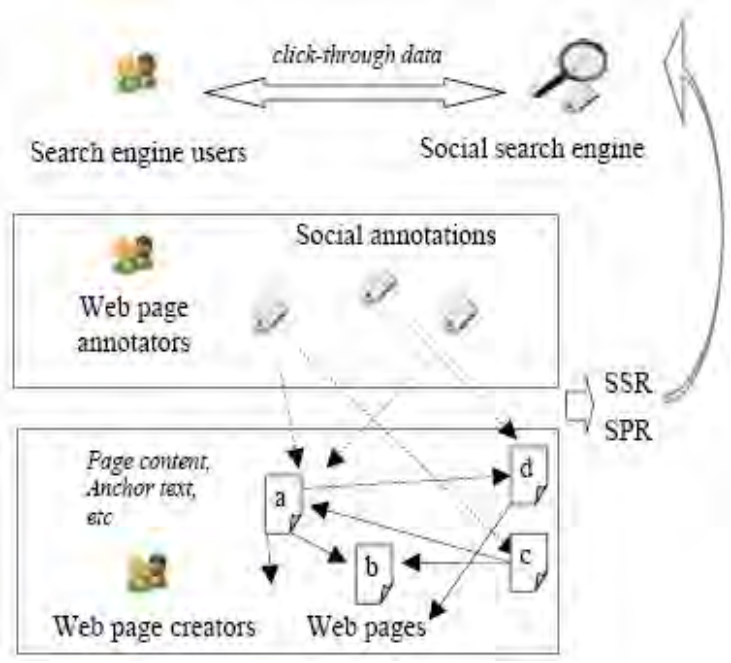
4.2 Similarity ranking και Static ranking

- Όσον αφορά το **similarity ranking**, που ορίζει την εκτιμώμενη ομοιότητα μεταξύ μιας σελίδας και μιας ερώτησης (query-document similarity), τα σχόλια των χρηστών για τη σελίδα προσφέρονται ως επιπλέον μεταδεδομένα γι αυτή την εκτίμηση. Επειδή όμως μια ερώτηση ενδέχεται να περιέχει όρους

που δεν ταιριάζουν πλήρως με κάποια σχόλια (όπως “shop” και “shopping”) χρειάζεται μια επιπλέον βελτίωση του similarity ranking. Έτσι παρουσιάζεται ο αλγόριθμος **SocialSimRank** (SSR) που λύνει το πρόβλημα.

- Όσον αφορά το **static ranking** ο αριθμός των σχολίων που αναφέρονται σε μια ιστοσελίδα υποδηλώνουν τη δημοτικότητα της και ενισχύουν την ποιότητά της μέχρι ως ένα σημείο. Επίσης διαφορετικά σχόλια, πιθανώς να έχουν διαφορετικό βάρος στον καθορισμό της δημοτικότητας. Οι παραδοσιακοί αλγόριθμοι του static ranking όπως PageRank δεν έχουν τρόπο για να μετρήσουν αυτού του είδους την ποιότητα. Έτσι παρουσιάζεται ο αλγόριθμος **SosialPageRank** για να μετρήσει την δημοτικότητα μιας ιστοσελίδας χρησιμοποιώντας κοινωνικά σχόλια**.** (Vu Thanh Nguyen,2009).

Υπάρχουν τρείς ομάδες χρηστών που σχετίζονται με το web search, οι **δημιουργοί των σελίδων** (*web page creators),* **οι σχολιαστές** (*web page annotators*) και **οι χρήστες** (*search engine users*) όπως φαίνεται και στην παρακάτω εικόνα όπου κάθε μία ομάδα παρέχει διαφορετική πληροφορία και παίζει διαφορετικό ρόλο. (Σχήμα 31).



Σχήμα 31: *Απεικόνιση των τριών ομάδων χρηστών (*Vu Thanh Nguyen,2009 σ3*)*

- **<u>Δημιουργοί των σελίδων</u>** : Δημιουργούν σελίδες και συνδέουν τις σελίδες μεταξύ τους για να κάνουν πιο εύκολη την περιήγηση στο διαδίκτυο για τους χρήστες . Παρέχουν τη βάση για την αναζήτηση στο διαδίκτυο.
- **<u>Σχολιαστές:</u>** Χρησιμοποιούν σχόλια για να οργανώσουν, να απομνημονεύσουν και να μοιραστούν τις αγαπημένες τους ιστοσελίδες στο διαδίκτυο.
- **<u>Χρήστες:</u>** Χρησιμοποιούν τις μηχανές αναζήτησης για να πάρουν πληροφορίες από το διαδίκτυο. Μπορούν επίσης να γίνουν και σχολιαστές μιας σελίδας αν τη σχολιάσουν ή αν την σώσουν στα αγαπημένα τους.

4.3. <u>Social Similarity Ranking</u>

Όπως το anchor text συμπληρωματικά συνεισφέρει στην εκτίμηση μιας ερώτησης με μια σελίδα, έτσι και τα σχόλια προσφέρονται ως χρήσιμα μεταδεδομένα. Η ομοιότητα μιας ερώτησης $q=\{q_1,q_2,\ldots,q_n\}$ με ένα σύνολο σχολίων $A(p)=\{\alpha_1,\ldots\alpha_n\}$ ορίζεται ως:

$$sim_{TM}(q,p)=\frac{|q\cap A(p)|}{|A(p)|}$$

Ο αλγόριθμος SSR βασίζεται στην παρατήρηση ότι χρήστες με κοινά ενδιαφέροντα αποδίνουν όμοια σχόλια σε παρόμοιες σελίδες. Αυτό επίσης μεταφράζεται και ως : Η ομοιότητα μεταξύ δύο σχολίων καθορίζεται από τις κοινές σελίδες που περιγράφουν. Έτσι, το πρόβλημα της αγνόησης όρων με παρόμοια σημασία με κάποιους από τους όρους της ερώτησης, επιλύεται με τον υπολογισμό της ομοιότητας μεταξύ των σχολίων (Nguyen,2009).

Ο αλγόριθμος Social Similarity Ranking :

Step 1 ***Init:*** ***Let*** $S_A^0(a_i, a_j) = 1$ for each $a_i = a_j$ otherwise 0

$S_P^0(p_i, p_j) = 1$ for each $p_i = p_j$ otherwise 0

Step 2 ***Do*** {

For each annotation pair (a_i, a_j) ***do***

$$S_A^{k+1}(a_i, a_j) = \frac{C_A}{|P(a_i)||P(a_j)|} \sum_{m=1}^{|P(a_i)|} \sum_{n=1}^{|P(a_j)|} \frac{\min(M_{AP}(a_i, p_m), M_{AP}(a_j, p_n))}{\max(M_{AP}(a_i, p_m), M_{AP}(a_j, p_n))} S_P^k(P_m(a_i), P_n(a_j)) \quad (2)$$

For each page pair (p_i, p_j) ***do***

$$S_P^{k+1}(p_i, p_j) = \frac{C_P}{|A(p_i)||A(p_j)|} \sum_{m=1}^{|A(p_i)|} \sum_{n=1}^{|A(p_j)|} \frac{\min(M_{AP}(a_m, p_i), M_{AP}(a_n, p_j))}{\max(M_{AP}(a_m, p_i), M_{AP}(a_n, p_j))} S_A^{k+1}(A_m(p_i), A_n(p_j)) \quad (3)$$

}***Until*** $S_A(a_i, a_j)$ converges.

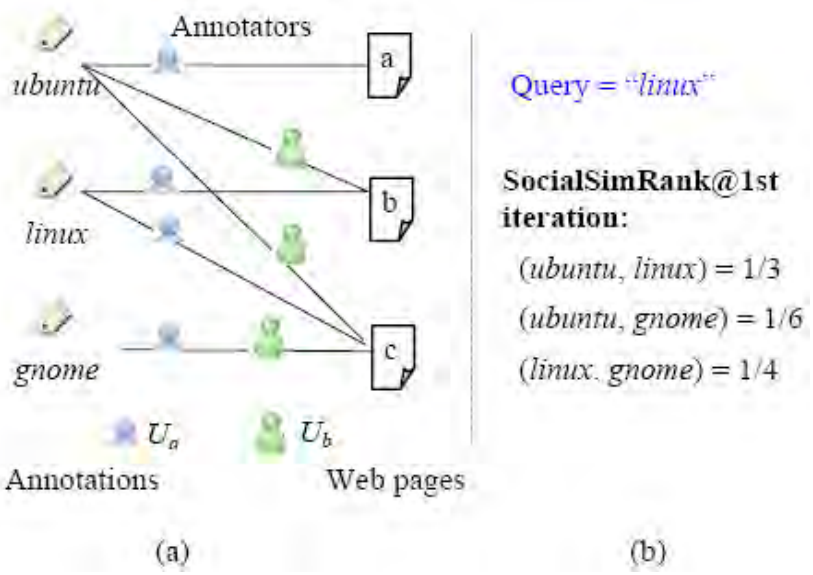
Step 3 ***Output:*** $S_A(a_i, a_j)$

Τα στοιχεία που χρησιμοποιεί είναι τα εξής:

- $\mathbf{N_A}$ = Πλήθος των σχολίων
- **Np**= Πλήθος των ιστοσελίδων
- $\mathbf{N_U}$ =Πλήθος των χρηστών
- $\mathbf{M_{AP}=N_A * N_P}$ Πίνακας με κάθε στοιχείο $M_{AP}(a_x, p_y)$ να δείχνει το πλήθος των χρηστών που έβαλαν το σχόλιο a_x στην σελίδα p_y.
- $\mathbf{S_A=N_A * N_A}$ Πίνακας με κάθε στοιχείο $S_A=(a_i, a_j)$ να δείχνει τον βαθμό ομοιότητας μεταξύ των σχολίων a_i και a_j.
- **Sp=Np *Np** Πίνακας με κάθε στοιχείο να δείχνει την ομοιότητα μεταξύ των δύο ιστοσελίδων
- $\mathbf{C_A, C_P}$ =Damping παράγοντες της διάδοσης της ομοιότητας για τα σχόλια και τις σελίδες αντίστοιχα (C_A, C_P=0.7)

- **$P(a_i)$=** Το σύνολο των ιστοσελίδων που είχαν σχόλιο το ai
- **$A(p_i)$=** Το σύνολο των σχολίων που δίνονται σε μια σελίδα pi
- **$P_m(a_i)$**=m-κοστή σελίδα που σχολιάστηκε από ai
- **$A_m(p_i)$**=m-κοστή annotation assigned to page pi (Vu Thanh Nguyen,2009)

4.3.1 Παράδειγμα εφαρμογής του SSR- Πολυπλοκότητα



Σχήμα 32: Σχέση μεταξύ σχολίων από τους χρήστες Ua και Ub με βάση τις ιστοσελίδες a, b και c (Bao. K. συν.,2007).

Από το Σχήμα 32 παρατηρούμε ότι τη b σελίδα χαρακτηρίζουν τα σχόλια "Ubuntu" και "Linux" από τους Ua και Ub αντίστοιχα. Οπότε υπάρχει ένας βαθμός ομοιότητας μεταξύ τους. Στο (b) οι ομοιότητες προκύπτουν μετά από την $1^η$ επανάληψη του αλγορύθμου θέτοντας $C_A=C_P=1$. Πιο συγκεκριμένα για την ιστοσελίδα b ο χρήστης Ua θα προτιμήσει να χρησιμοποιήσει ως σχόλιο το "Linux" ενώ ο Ub το σχόλιο "Ubuntu". Επομένως τα σχόλια "Linux" και "Ubuntu" ίσως να έχουν κάποια σημασιολογική σχέση. Όσον αφορά τη σελίδα c τα σχόλια "Linux" και "gnome" δίνονται από τον χρήστη Ua οπότε αυτά τα δύο σχόλια ίσως να σχετίζονται μεταξύ τους ως κάποιο βαθμό. Σε κάποιες περιπτώσεις μερικές σελίδες ίσως περιέχουν μόνο το σχόλιο "Ubuntu" όπως αυτό συμβαίνει για την ιστοσελίδα a. Τότε

δίνοντας ως ερώτημα το "Linux" σελίδα που έχει μόνο ως σχόλιο το "Ubuntu" θεωρείται ως ακατάλληλη για τη μέθοδο ταιριάσματος και φιλτράρεται εσφαλμένα. Ακόμη και αν η σελίδα περιέχει και τα δύο σχόλια "Ubuntu" και "Linux" δεν είναι κατάλληλο να υπολογίσουμε την ομοιότητα ανάμεσα στο ερώτημα και στο αρχείο χρησιμοποιώντας μόνο την λέξη "Linux". (Shenghua Bao κ.συν 2007).

Η ομοιότητα μεταξύ **ερώτησης-σελίδας** βασισμένοι στον αλγόριθμο SSR δίνεται από την παρακάτω σχέση όπου A(p)={a1,a2,..am} το σύνολο των σχολίων μιας ιστοσελίδας p και q={q1,q2..qn} ένα ερώτημα με n όρους.

Για κάθε βήμα του αλγορίθμου απαιτείται χρόνος $O(N_A^2 N_P^2)$. Η συνολική πολυπλοκότητα του αλγορίθμου εξαρτάται από τα βήματα που απαιτούνται ώστε ο αλγόριθμος να συγκλίνει.

$$sim_{SSR}(q,p) = \sum_{i=1}^{n}\sum_{j=1}^{m} S_A(q_i, a_j)$$

4.4 Social Page Ranking

Ο αλγόριθμος SPR εκτιμά την ποιότητα μιας σελίδας βασιζόμενος στην παρακάτω ενδιαφέρουσα σχέση μεταξύ των ιστοσελίδων, των χρηστών και των σχολίων: ενημερωμένοι χρήστες προτιμούν δημοφιλείς σελίδες, οι οποίες αποκομίζουν θερμά σχόλια.

Step 1 ***Input:***

Association matrices M_{PU}, M_{AP}, and M_{UA} and the random initial SocialPageRank score P_0

Step 2 ***Do:***

$$U_i = M_{PU}^T \cdot P_i \quad (5.1)$$
$$A_i = M_{UA}^T \cdot U_i \quad (5.2)$$
$$P_i^{'} = M_{AP}^T \cdot A_i \quad (5.3)$$
$$A_i^{'} = M_{AP} \cdot P_i^{'} \quad (5.4) \qquad (5)$$
$$U_i^{'} = M_{UA} \cdot A_i^{'} \quad (5.5)$$
$$P_{i+1} = M_{PU} \cdot U_i^{'} \quad (5.6)$$

Until P_i converges.

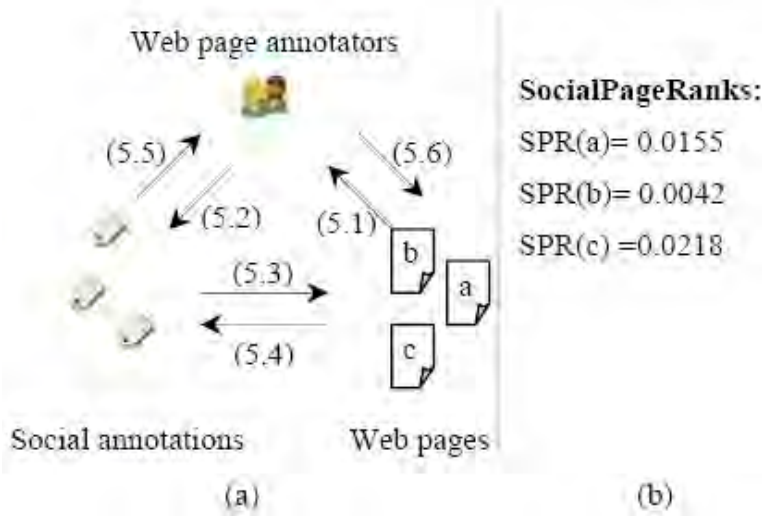
Step 3: ***Output:***

P^*: the converged SocialPageRank score.

Τα Στοιχεία που χρησιμοποιεί είναι τα εξής:

- $\mathbf{N_A}$ = Πλήθος των σχολίων
- **Np**= Πλήθος των ιστοσελίδων
- $\mathbf{N_U}$ =Πλήθος των χρηστών
- $\mathbf{M_{PU}=N_p * N_U}$ Πίνακας συσχετισμών μεταξύ των ιστοσελίδων και των χρηστών με κάθε στοιχείο M_{PU}(pi,uj) να δείχνει το πλήθος των σχολίων που χρησιμοποιούνται από το χρήστη uj για τη σελίδα pi.
- $\mathbf{M_{AP}=N_A * N_p}$ Πίνακας συσχετισμών μεταξύ των σχολίων και των ιστοσελίδων , με κάθε στοιχείο $M_{AP}(a_i, p_j)$ να δείχνει το πλήθος των χρηστών που χρησιμοποιούν το σχόλιο ai για την σελίδα pj
- $\mathbf{M_{UA}=N_U * N_A}$ Πίνακας συσχετισμών μεταξύ των χρηστών και των σχολίων με κάθε στοιχείο M_{UA}(ui,aj) να δείχνει το πλήθος των ιστοσελίδων που έχουν το σχόλιο aj από το χρήστη ui
- **Pi, Ui, Ai** Είναι τα διανύσματα δημοτικότητας των σελίδων, χρηστών και σχολίων αντίστοιχα στη i-επανάληψη (με τόνους αντίστοιχα αποτελούν τις ενδιάμεσες τιμές) (Shenghua Bao κ.συν.2007)

4.4.1 Παράδειγμα εφαρμογής του SPR-Πολυπλοκότητα



Σχήμα 33 (Shenghua Bao κ.συν ,2007)

Στο σχήμα 33 σκιαγραφούνται τα βήματα του αλγορίθμου που αφορούν τους χρήστες, τα σχόλια και τις σελίδες. Συγκεκριμένα:

- (5.1) η δημοτικότητα των χρηστών αντλείται από τις σελίδες που σχολιάζουν.
- (5.2) η δημοτικότητα των σχολίων από τη δημοτικότητα των χρηστών.
- (5.3)των σελίδων από των σχολίων.
- (5.4) των σχολίων από των σελίδων.
- (5.5)των χρηστών από των σχολίων.
- (5.6) ξανά των σελίδων από των χρηστών.

Τελικά η έξοδος του αλγορίθμου μετά τη σύγκλιση είναι το P*.
Σε κάθε επανάληψη του αλγορίθμου η πολυπλοκότητα χρόνου είναι $O(N_U N_P + N_A N_P + N_U N_A)$.

4.5 Αποτελέσματα πειραμάτων-Εκτίμηση αποτελεσματικότητας Αλγορίθμων

4.5.1 Δυναμική Μέθοδος Κατάταξης

Τα χαρακτηριστικά της κοινωνικής έρευνας διακρίνονται σε δυο μεγάλες κατηγορίες query-document features και document static features με υποκατηγορίες που παρουσιάζονται στον παρακάτω πίνακα (Σχήμα 34).

A: query-document features	
DocSimilarity	Similarity between query and page content
TermMatching (TM)	Similarity between query and annotations using the term matching method.
SocialSimRank (SSR)	Similarity between query and annotations based on SocialSimRank.
B: document static features	
GooglePageRank (PR)	The web page's PageRank obtained from the Google's toolbar API.
SocialPageRank (SPR)	The popularity score calculated based on SocialPageRank algorithm.

Σχήμα 34 (Shenghua Bao κ.συν ,2007)

Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν (Σχήμα 33) στην προσπάθεια εκτίμησης της αποτελεσματικότητας των δύο αλγορίθμων. Επίσης γίνεται σύγκριση των αλγορίθμων με άλλους που χρησιμοποιούνται για τον ίδιο σκοπό και τέλος περιγράφεται ο τρόπος με τον οποίο οι αλγόριθμοι SSR και SPR συνεισφέρουν στην ανάκτηση των σχετικών σελίδων από μια ερώτηση.

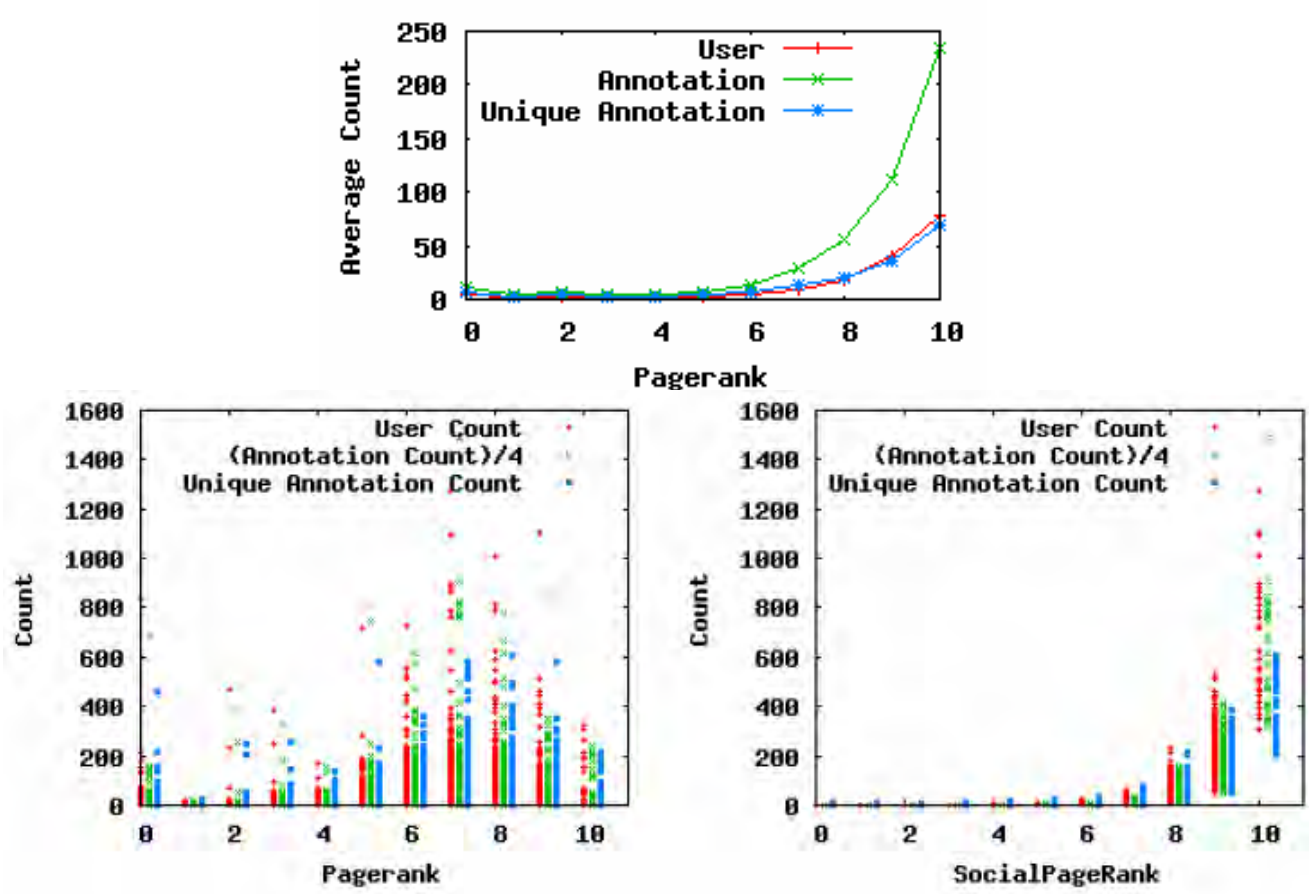
Τα πειράματα που πραγματοποιήθηκαν βασίστηκαν σε ένα σύνολο 1.736.628 σελίδων και 269.566 σχολίων που προέρχονται από το del.icio.us το Μάιο του 2006. Επειδή κάποια σχόλια έχουν μη έγκυρη για τους αλγορίθμους μορφή (π.χ. java.programming ή java/programming), έγινε χωρισμός τους σε συγκεκριμένους όρους με τη βοήθεια του WordNet πριν την εκτέλεση των πειραμάτων. Στα πειράματα για την σύγκλιση του SSR χρειάστηκαν 12 επαναλήψεις . Στον παρακάτω πίνακα παρουσιάζονται σχόλια από τέσσερις κατηγορίες και τα τέσσερα κορυφαία σχόλια που έχουν μεγαλύτερη σημασιολογική ομοιότητα με αυτά.

Technology related:	
dublin	metadata, semantic, standard, owl
debian	distribution, distro, ubuntu, linux
Economy related:	
adsense	sense, advertise, entrepreneur, money
800	number, directory, phone, business
Entertainment related:	
album	gallery, photography, panorama, photo
chat	messenger, jabber, im, macosx
Entity related:	
einstein	science, skeptic, evolution, quantum
christian	devote, faith, religion, god

Σχήμα 35 (Shenghua Bao κ.συν ,2007)

Ο αλγόριθμος SSR για την εύρεση της ποιότητας των σελίδων, (δηλαδή μέχρι την σύγκλισή του) εκτέλεσε 7 επαναλήψεις. Το PageRank κάθε σελίδας υπολογίστηκε επίσης από το API του Google toolbar. Στη συνέχεια έγινε χρήση του PageRank για την περιγραφή του PageRank του Google. Η σύγκρισή του με τον αλγόριθμο SPR θα βοηθήσει στην εκτίμηση της αποδοτικότητας του Social PageRank (Shenghua Bao κ.συν ,2007).

4.5.2 SPR VS PageRank



Σχήμα 36 (Shenghua Bao κ.συν, 2007)

Στο σχήμα 36 παρουσιάζεται η σχέση μεταξύ των σελίδων, των σχολίων και των χρηστών που τα αποδίδουν στις σελίδες, με βάση το σκορ που προκύπτει από τον PageRank. Από τις γραφικές παραστάσεις συμπεραίνεται ότι στις περισσότερες περιπτώσεις, είναι πιθανόν πολλοί χρήστες να αφιερώνουν πολλά σχόλια σε σελίδες με υψηλότερο PageRank. Η unique **annotation** γραμμή αναπαριστά το πλήθος των σχολίων που είναι διαφορετικά με κάθε άλλο. Επιπρόσθετα συμπεραίνεται πως σελίδες με το ίδιο PageRank διαφέρουν πολύ στο πλήθος των σχολίων και των χρηστών που τις χαρακτηρίζουν. Επίσης όσον αφορά σελίδες με μικρότερο PageRank είναι πιθανό να χαρακτηρίζονται από περισσότερα σχόλια και χρήστες από σελίδες με μεγαλύτερο PageRank. Για παράδειγμα, σελίδες με PageRank 0 έχουν περισσότερα σχόλια και χρήστες από σελίδες με PageRank 10. Εφαρμόζοντας τον αλγόριθμο SPR

στα δεδομένα και ύστερα από κανονικοποίηση των σκορ ώστε να έχουν τιμή μεταξύ του 0 και του 10, τα αποτελέσματα που προκύπτουν απεικονίζονται στο Σχήμα 36. Όπως γίνεται εύκολα αντιληπτό, ο αλγόριθμος υπολογίζει με επιτυχία τη δημοτικότητα μιας σελίδας σύμφωνα με το πλήθος των σχολίων και των χρηστών που την προσδιορίζουν (Bao, κ.συν.,2007).

Μια ενδεικτική σύγκριση των αλγορίθμων SPR και PageRank φαίνεται στον παρακάτω πίνακα (Σχήμα 37). Κάποιες σελίδες έχουν μικρό PageRank και μικρό SPR ενώ κάποιες άλλες το αντίθετο. Βέβαια υπάρχουν και σελίδες που έχουν την ίδια τιμή ως αποτέλεσμα. Σαν αποτέλεσμα συμπεραίνεται πως η προτίμηση των δημιουργών των σελίδων διαφέρει από εκείνη των χρηστών (σχολιαστών), η οποία χαρακτηρίζεται πλήρως από τον SPR.

Web Pages	PR	SPR
http://www.sas.calpoly.edu/asc/ssl/procrastination.html	0	10
http://37signals.com/papers/introtopatterns/	0	10
http://www.lcs.mit.edu/	10	0
http://www.macromedia.com/software/coldfusion/	10	0
http://www.w3.org/	10	10
http://www.nytimes.com/	10	10
http://www.cientologia-lisbon.org/	0	0
http://users.tpg.com.au/robotnik/	0	0

Σχήμα 37: Σύγκριση των αλγορίθμων SPR και PageRank για κάθε ιστοσελίδα(Bao, κ.συν, 2007)

4.6 Πλεονεκτήματα-Μειονεκτήματα των Κοινωνικών σχολίων

Τα πλεονεκτήματα της χρήσης κοινωνικών σχολίων, είναι πολλά. Ενδεικτικά αναφέρονται ορισμένα. Το σημαντικότερο πλεονέκτημα είναι η σημαντική βελτίωση των αποτελεσμάτων από την αναζήτηση. Επίσης εξαιτίας των αραιών πινάκων οι δύο αλγόριθμοι συγκλίνουν γρήγορα και η πολυπλοκότητά τους είναι πολύ μικρότερη από την εκτιμώμενη. Τέλος, η προσέγγιση που δόθηκε εκμεταλλεύεται την συνεχή αύξηση των χρηστών-σχολιαστών.

Βέβαια, απ' την άλλη μεριά υπάρχουν και κάποια μειονεκτήματα. Ένα από αυτά είναι το γεγονός ότι οι χρήστες και τα σχόλια έχουν εκθετικό ρυθμό ανάπτυξης, κάτι που οδηγεί σε πιο αργή σύγκλιση των αλγορίθμων. Δύο ακόμη μειονεκτήματα είναι επίσης η έλλειψη σχολίων για νέες σελίδες και η αδυναμία αποσαφήνισης σχολίων με δισημία ***(***Bao, κ.συν., 2007***)***.

4.7 Επεκτάσεις

Ένα θέμα στο οποίο μπορεί να γίνει έρευνα είναι να επιτευχθεί βελτιστοποίηση του αλγορίθμου SSR για την αντιμετώπιση των εκθετικά αυξανόμενων σχολίων και χρηστών. Ακόμη η χρήση σχολίων ορισμένες φορές είναι κακοπροαίρετη έτσι για να αντιμετωπιστούν σχόλια- spam θα μπορούσε πριν από την εκτέλεση των βασικών αλγορίθμων να γίνεται γλωσσολογική ανάλυση και αγνόησή τους.

ΚΕΦΑΛΑΙΟ 5ο

Semi-Supervised Framework για την κατάταξη γράφου

5.1 Ορισμοί

Παραδοσιακά ένας γράφος ορίζεται ως η τριάδα G'(V,E,W) όπου V είναι ένα σύνολο από κόμβους, Ε σύνολο από ακμές και W πίνακας από βάρη για τις ακμές του γράφου. Σ' αυτόν τον ορισμό μόνο ο σκελετός του γράφου μπορεί να περιγραφεί ενώ οι απαραίτητες πληροφορίες για τους κόμβους και τις ακμές δεν εκφράζονται. Λύνοντας αυτό το πρόβλημα ορίζεται ένας γράφος χρησιμοποιώντας την ακόλουθη τετράδα G (V,E,X,Y). Αυτός είναι ένας γράφος που ακόμα περιέχει ένα σύνολο κόμβων V με n κόμβους και ένα σύνολο ακμών Ε με m ακμές. Επιπρόσθετα ορίζεται ένα σύνολο από χαρακτηριστικά ακμών $X=\{x_{i,j}\}$ και χαρακτηριστικά κόμβων $Y=\{y_i\}$ όπου κωδικοποιούνται πληροφορίες στο γράφο. Πιο συγκεκριμένα για κάθε ακμή από τον κόμβο i στον κόμβο j υπάρχει ένα l-διάστατο χαρακτηριστικό διάνυσμα $x_{ij}=(x_{ij1}, x_{ij2},\ldots x_{ijl})^T$ ενώ για κάθε κόμβο i υπάρχει ένα h-διάστατο χαρακτηριστικό διάνυσμα $y_i=(y_{i1},y_{i2},\ldots y_{ih})^T$. Συνήθως τα l, h είναι μικροί αριθμοί συγκρινόμενοι με το scale του γραφήματος. Έτσι στο Framework ένας γράφος περιλαμβάνει το σύνολο κόμβων V, το σύνολο ακμών Ε ,τα χαρακτηριστικά ακμών και των κόμβων X και Y αντίστοιχα. Η δομή του γράφου ορίζει την global σχέση μεταξύ των κόμβων, τα χαρακτηριστικά ακμών αντιπροσωπεύουν τις τοπικές σχέσεις μεταξύ οποιοδήποτε δύο κόμβων ενώ τέλος τα χαρακτηριστικά κόμβων περιγράφουν τις ιδιότητες κάθε κόμβου ξεχωριστά. (Gao, κ.συν., 2011 σελ2).

Ορίζουμε ένα Unified Framework ως εξής:

$$\min_{\omega\geq 0,\phi\geq 0,\pi\geq 0} R(\pi; f(\omega, X), g(\phi, Y))$$
$$s.t.\ S(\pi; B, \mu) \geq 0.$$

Πρόβλημα 1

Τα σκορ κατάταξης των κόμβων στο γράφο G(V,E,X,Y) αναπαριστώνται ως ένα n-διάστατο διάνυσμα π. Επιπλέον τα ω και φ που περιέχονται στις συναρτήσεις f(ω,X) και g(φ,Y) αποτελούν τα διανύσματα-παράμετροι για τα χαρακτηριστικά των κορυφών και των ακμών αντίστοιχα. Οι περιορισμοί για τις παραμέτρους ω, φ και π είναι ω>=0, φ>=0 και π>=0. Η χρήση του «>=0» σημαίνει ότι όλα τα στοιχεία είναι μη αρνητικά και τουλάχιστον ένα στοιχείο είναι θετικό.

5.1.1 Αντικειμενική Συνάρτηση:

$$R(\pi; f(\omega, X), g(\phi, Y)).$$

Αυτή η αντικειμενική συνάρτηση θεωρείται σαν graph-based smoothing συνάρτηση για σκορ κατάταξης π. Για παράδειγμα μπορούμε χρησιμοποιείται το τυχαίο Markov walk μοντέλο για να χτιστεί μια smoothing συνάρτηση. Η transition πιθανότητα της διαδικασίας Markov ορίζεται ως f(ω,X) το οποίο είναι παραμετρικό μοντέλο βασιζόμενο σε χαρακτηριστικά ακμών ενώ η reset πιθανότητα ορίζεται ως g(φ,Y) το οποίο είναι παραμετρικό μοντέλο βασιζόμενο σε χαρακτηριστικά κορυφών. Έτσι η smoothing συνάρτηση απαιτεί τα σκορ κατάταξης να είναι κοντά σε μια stationary κατανομή της παραμετρικής διαδικασίας Markov. (Gao, κ.συν., 2011 σελ 3).

5.1.2 Περιορισμοί:

Οι περιορισμοί για το framework φαίνονται από τον παρακάτω τύπο:

$$S(\pi; B, \mu) \geq 0,$$

Όπου ο πίνακας B κωδικοποιεί πληροφορίες-εποπτείας (supervision information's) και το μ υποδηλώνει τα βάρη για δείγματα πληροφορία-εποπτείας(supervision information). Η μάθηση με αναφορά στους περιορισμούς είναι ένα εποπτικό θέμα σε αντίθεση με τη μάθηση με αναφορά στην αντικειμενική συνάρτηση που είναι ένα μη-εποπτικό θέμα. Επιπλέον ο πίνακας B μπορεί να αποτελείται από διαφορετικά ήδη πληροφορίας-εποπτείας (supervision information)

όπως είναι: δυαδικές ετικέτες, ζεύγη προτίμησης (pairwise preference), μερική διάταξη ακόμα και συνολική κατάταξη. Για παράδειγμα τα ζεύγη προτίμησης μπορούν να επισημαίνονται από ανθρώπινα σχόλια. Στην περίπτωση αυτή ο πίνακας B είναι r,n με 1, -1, 0 ως στοιχεία όπου το r είναι ο αριθμός των ζευγών προτίμησης. Κάθε γραμμή του B αντιπροσωπεύει ένα ζεύγος προτίμησης u>v σημαίνοντας ότι ο κόμβος u είναι προτιμότερος από τον κόμβο v. Η αντίστοιχη γραμμή του πίνακα B έχει 1 στην u-στήλη και -1 στην v-στήλη και 0 στις άλλες στήλες. Συμπερασματικά οι περιορισμοί μπορούν να καθοριστούν ως εξής όπου e είναι ένα r-διάστατο διάνυσμα με όλα τα στοιχεία ίσα με 1.

$$S(\pi; B, \mu) = -\mu^T (e - B\pi) \geq 0.$$

Για δυαδικές ετικέτες (web search, spam σελίδες και junk σελίδες αντιστοιχίζονται με την ετικέτα 0 ενώ οι καλές σελίδες αντιστοιχίζονται με την ετικέτα 1 μερική κατάταξη ή συνολική κατάταξη δεν είναι δύσκολο να μετατραπούν σε ένα αριθμό σχέσεων ζευγών προτίμησης και με περιορισμούς παρόμοιους με τους παραπάνω (Gao κ.συν., 2011 σελ 3).

5.1.3 Ισοδύναμο πρόβλημα βελτιστοποίησης

Για πιο εύκολους υπολογισμούς οι περιορισμοί μετατρέπονται σε μια συνάρτηση απώλειας (loss function) η οποία προστίθεται στην αντικειμενική συνάρτηση

$$\min_{\omega \geq 0, \phi \geq 0, \pi \geq 0} \alpha R(\pi; f(\omega, X), g(\phi, Y)) - \beta S(\pi; B, \mu),$$

Πρόβλημα 2

Όπου α και β είναι και οι δύο μη αρνητικές συντελεστές.

5.2 Semi-Supervised PageRank (SSP)

Προτείνεται ένας αποτελεσματικός αλγόριθμος που ονομάζεται Semi-Supervised PageRank (SSP) . Αυτός ο αλγόριθμος μπορεί να αποδειχτεί ως μια

βιτρίνα για την επιτυχή αντιμετώπιση της κατάταξης του γραφήματος που αναφέρθηκε προηγουμένως. Στον αλγόριθμο αυτό ορίζεται ως αντικειμενική συνάρτηση χρησιμοποιώντας την ίδια αρχή με τον αλγόριθμο PageRank. Ειδικότερα ένα βήμα στο τυχαίο Markov μοντέλο του PageRank μπορεί να γραφτεί ως εξής:

$$\tilde{\pi} = dP_0^T \pi + (1-d)r_0,$$

Όπου P_0 είναι ένας transition πίνακας, r_0 είναι reset πιθανότητα (πιθανότητα επαναφοράς) και d συντελεστής απόσβεσης. Παρουσιάζοντας τις παραμέτρους και για τον transition πίνακα για τον συντελεστή απόσβεσης και ορίζοντας την απώλεια του $\|\pi' - \pi\|^2$ ως αντικειμενική συνάρτηση προκύπτει η εξής:

$$R(\pi; f(\omega, X), g(\phi, Y)) = \|df(\omega, X)\pi + (1-d)g(\phi, Y) - \pi\|^2$$

Στον παραπάνω τύπο η συνάρτηση f (ω , X) παίζει τον ίδιο ρόλο όπως κάνει ο πίνακας P_0^T στον αλγόριθμο PageRank. Αυτό μπορούμε να ξαναγραφτεί ως $P^T(\omega, X)$ (ή μερικές φορές ως $P^T(\omega)$ ή P^T) για να αποδειχθεί ότι είναι ένας παραμετρικός πίνακας transition πιθανότητας. Κάθε στοιχείο $p_{I,j}(\omega)$ του P(ω) αντιπροσωπεύει την transition πιθανότητα από τον κόμβο I στον κόμβο j, κάτι το οποίο καθορίζεται από το μοντέλο των χαρακτηριστικών ακμής με παράμετρο ω.

Για παράδειγμα χρησιμοποιείται ο παρακάτω γραμμικός συνδυασμός:

$$p_{ij}(\omega) = \begin{cases} \frac{\sum_k \omega_k x_{ijk}}{\sum_j \sum_k \omega_k x_{ijk}}, & \text{if there is an edge from } i \text{ to } j, \\ 0, & \text{otherwise.} \end{cases}$$

Σχήμα 38

Αυτή είναι η transition πιθανότητα για μια υπάρχουσα ακμή στον γράφο που είναι μη-μηδενική και η τιμή προσδιορίζεται από τα χαρακτηριστικά των ακμών. Αλλάζει το βάρος των υπαρχουσών ακμών συμπεριλαμβανομένου την ανάθεση μηδενικού βάρους χωρίς την προσθήκη καινούριων ακμών στον γράφο. Αυτό έχει ως

αποτέλεσμα τη δημιουργία αραιού πίνακα. Εδώ η συνάρτηση g(φ,Y) παίζει τον ίδιο ρόλο όπως το r_0 στον αρχικό PageRank αλγόριθμο. Αυτό μπορεί να ξαναγραφτεί σαν r(φ,Y) (ή μερικές φορές ως r(φ) ή σκέτο r) για να οριστεί ότι αυτό είναι ένα παραμετρικό διάνυσμα επαναφοράς πιθανότητας. Κάθε στοιχείο $r_i(\varphi)$ στο r(φ) αντιπροσωπεύει την πιθανότητα επαναφοράς του I κόμβου η οποία καθορίζεται από το μοντέλο των χαρακτηριστικών κορυφών με παράμετρο φ. Για παράδειγμα χρησιμοποιώντας για μια ακόμα φορά τον γραμμικό συνδυασμό.

$$r_i(\phi) = \phi^T y_i.$$

Υποθέτοντας ότι οι περιορισμοί βασίζονται σε ζεύγη προτιμήσεων προκύπτει το παρακάτω βελτιστοποιημένο πρόβλημα:

$$\min_{\omega \geq 0, \phi \geq 0, \pi \geq 0} \{\alpha \| dP^T(\omega)\pi + (1-d)r(\phi) - \pi\|^2 + \beta \mu^T (e - B\pi)\}.$$

Λύνοντας αυτό το πρόβλημα καθορίζονται τα βέλτιστα ω , φ και π τα οποία χρησιμοποιούνται για την ταξινόμιση των κόμβων του γράφου. (Bin Gao, 2011,σελ3).

5.3 Λύνοντας το Πρόβλημα Βελτιστοποίησης

Σύμφωνα με τους Gao, κ.συν., (2011) για πιο εύκολη υλοποίηση χρησιμοποιείται η G(ω,φ,π) για να υποδηλώθεί η αντικειμενική συνάρτηση όπου G(ω,φ,π) ισούται με:

$$G(\omega, \phi, \pi) = \alpha \| dP^T(\omega)\pi + (1-d)r(\phi) - \pi\|^2 + \beta \mu^T (e - B\pi).$$

Χρησιμοποιείται η μέθοδος κλίσης καθόδου για να την ελαχιστοποίηση της G(ω,φ,π). Οι μερικές παράγωγοι της συνάρτησης G ως προς ω, φ και π υπολογίζονται παρακάτω.

$$\frac{\partial G}{\partial \omega} = 2\alpha d[P^T\pi \otimes \pi - \pi \otimes \pi + (1-d)r \otimes \pi]^T \frac{\partial vec(P)}{\partial \omega^T},$$

$$\frac{\partial G}{\partial \phi} = 2\alpha(1-d)[(1-d)r + dP^T\pi - \pi]\frac{\partial r}{\partial \phi},$$

$$\begin{aligned}\frac{\partial G}{\partial \pi} = 2\alpha[&(dPP^T - dP - dP^T + I)\pi \\ &-(1-d)(I-dP)r] - \beta B^T\mu.\end{aligned}$$

Το σύμβολο $\otimes$ αντιπροσωπεύει το προϊόν Kronecker ενώ το σύμβολο vec(.) αντιπροσωπεύει την επέκταση ενός πίνακα σε ένα διάνυσμα από της στήλες του. Από τις παραπάνω συναρτήσεις προκύπτει:

$$\frac{\partial vec(P)}{\partial \omega^T} = \begin{pmatrix} \frac{\partial p_{11}}{\partial \omega_1} & \cdots & \frac{\partial p_{11}}{\partial \omega_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{n1}}{\partial \omega_1} & \cdots & \frac{\partial p_{n1}}{\partial \omega_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{1n}}{\partial \omega_1} & \cdots & \frac{\partial p_{1n}}{\partial \omega_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{nn}}{\partial \omega_1} & \cdots & \frac{\partial p_{nn}}{\partial \omega_l} \end{pmatrix} \text{ and } \frac{\partial r}{\partial \phi} = \begin{pmatrix} \frac{\partial r}{\partial \phi_1} \\ \vdots \\ \frac{\partial r}{\partial \phi_i} \\ \vdots \\ \frac{\partial r}{\partial \phi_h} \end{pmatrix}$$

Εάν η $p_{i,j}$ είναι γραμμική συνάρτηση των χαρακτηριστικών των ακμών οι μερικές παράγωγοι της $p_{i,j}$ ως προς ω_κ θα είναι:

$$\frac{\partial p_{ij}}{\partial \omega_k} = \frac{x_{ijk}\sum_j\sum_k \omega_k x_{ijk} - (\sum_k \omega_k x_{ijk})(\sum_j x_{ijk})}{(\sum_j\sum_k \omega_k x_{ijk})^2}$$

Με τα παραπάνω παράγωγα ενημερώνονται επαναληπτικά το ω, το φ και το π μέσω της μεθόδου καθόδου κλίσης. Ο αντίστοιχος αλγόριθμος δίνεται παρακάτω όπου ρ είναι το ποσοστό εκμάθησης και ε ελέγχει την συνθήκη τερματισμού.

Input: X, Y, B, μ, l, h, n, ρ, ϵ, α, β.
Output: Node ranking score π^*

Algorithm:

1. Set $s = 0$, initialize $\pi_i^{(0)}$ $(i = 1, \cdots, n)$, $\omega_k^{(0)}$ $(k = 1, \cdots, l)$, and $\phi_t^{(0)}$ $(t = 1, \cdots, h)$.
2. Calculate $P^{(s)} = P(\omega^{(s)})$, $r^{(s)} = r(\phi^{(s)})$, and $G^{(s)} = G(\omega^{(s)}, \phi^{(s)}, \pi^{(s)})$.
3. Update $\pi_i^{(s+1)} = \pi_i^{(s)} + \rho \frac{\partial G^{(s)}}{\partial \pi_i^{(s)}}$, $\omega_k^{(s+1)} = \omega_k^{(s)} + \rho \frac{\partial G^{(s)}}{\partial \omega_k^{(s)}}$, and $\phi_t^{(s+1)} = \phi_t^{(s)} + \rho \frac{\partial G^{(s)}}{\partial \phi_t^{(s)}}$.
4. Normalize $\pi_i^{(s+1)} \leftarrow \frac{\pi_i^{(s+1)}}{\sum_{j=1}^{n} \pi_j^{(s+1)}}$, $\omega_k^{(s+1)} \leftarrow \frac{\omega_k^{(s+1)}}{\sum_{j=1}^{l} \omega_j^{(s+1)}}$, and $\phi_t^{(s+1)} \leftarrow \frac{\phi_t^{(s+1)}}{\sum_{j=1}^{h} \phi_j^{(s+1)}}$.
5. Calculate $G^{(s+1)} = G(\omega^{(s+1)}, \phi^{(s+1)}, \pi^{(s+1)})$, if $G^{(s)} - G^{(s+1)} < \epsilon$, stop and output $\pi^* = \pi^{(s+1)}$; else $s = s + 1$, jump to step 2.

5.4 Αποτελεσματική Εφαρμογή

Στη συνέχεια εφαρμόζεται αποτελεσματικά ο αλγόριθμος χρησιμοποιώντας το γεγονός ότι οι πίνακες είναι αραιοί. Ορίζοντας ως $\pi' = P^T \pi$ και ως $\pi'' = \pi' - \pi$ και με κάποιες μαθηματικές τροποποιήσεις αναπαράγεται η μερική παράγωγος της συνάρτησης G ως προς π.

$$\frac{\partial G}{\partial \pi} = 2\alpha[d(P\pi'' - \pi'') + (1-d)(\pi - r + dPr)] - \beta B^T \mu$$

Για τον υπολογισμό της μερικής παραγώγου μόνο τρείς πολλαπλασιασμοί (πίνακα-διανύσματος) χρειάζονται: Pπ'', Pr, P^Tπ. Παρομοίως οι σχέσεις για τον

υπολογισμό της μερικής παραγώγου της συνάρτησης G ως προς ω και φ με τη βοήθεια το π’ και π’’ μπορούν να απλουστευθούν ως εξής:

$$\frac{\partial G}{\partial \omega} = 2\alpha d\{[\pi'' + (1-d)r] \otimes \pi\}^T \frac{\partial vec(P)}{\partial \omega^T}.$$

$$\frac{\partial G}{\partial \phi} = 2\alpha(1-d)[(1-d)r + d\pi' - \pi]\frac{\partial r}{\partial \phi}.$$

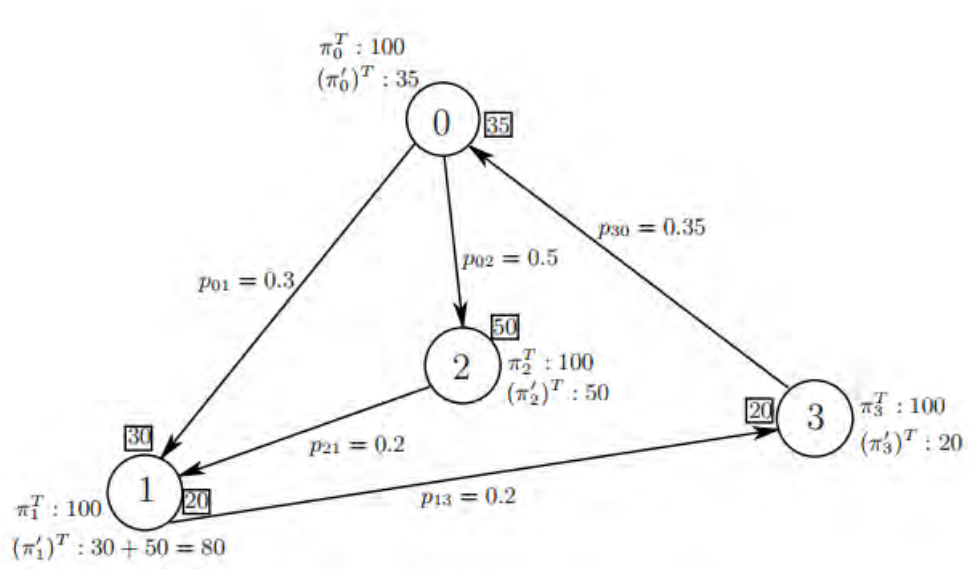
Λόγω του γεγονότος ότι ο πίνακας είναι αραιός, για τον υπολογισμό της μερικής παραγώγου της συνάρτησης G ως προς ω υπολογίζονται τα μη-μηδενικά μπλόκ στο προϊόν Kronecker και η μερική παράγωγος του διανύσματος P ως προς ω^T. Υποθέτοντας ότι υπάρχουν m ακμές στο γράφο τότε το κόστος είναι ανάλογο του m. Γενικότερα η πολυπλοκότητα του αλγορίθμου Semi-Supervised PageRank είναι O(ml +n). (Gao, κ.συν,2011σελ 4).

5.4.1 MapReduce

Είναι ένα μοντέλο προγραμματισμού για παράλληλους υπολογισμούς μεγάλης κλίμακας σε ένα κατανεμημένο σύμπλεγμα υπολογιστών. Η λειτουργία του ταιριάσματος είναι ότι παίρνει ένα ζεύγος «κλειδί, τιμή» και εκπέμπει έναν ή περισσότερους ενδιάμεσα ζεύγη «κλειδί, τιμή». Στη συνέχεια όλες οι τιμές με το ίδιο ενδιάμεσο κλειδί ομαδοποιούνται σε ένα ζεύγος «κλειδί, λίστα τιμών» όπου η λίστα τιμών περιέχει όλες τις τιμές που σχετίζονται με το ίδιο κλειδί. Μια μειωμένη λειτουργία είναι η ανάγνωση ενός ζεύγους «κλειδί, λίστα τιμών» και η εκπομπή ενός η περισσοτέρων ζευγών «κλειδί, τιμή». Υπάρχουν δύο βασικά είδη μεγάλης κλίμακας υπολογισμού στο SSP. Τα είδη αυτά είναι **matrix-vector multiplication** και **Kronecker προϊόν των διανυσμάτων** σε ένα αραιό γράφημα τα οποία αναλύονται παρακάτω. (Kang, κ.συν. ,2011).

5.4.2 Matrix-Vector Multiplication-Παράδειγμα

Αρχικά γίνεται χρήση της ισότητας $\pi' = P^T\pi$, που μπορεί να γραφτεί ισοδύναμα ως $(\pi')^T = \pi^T P$. Ο υπολογισμός αυτός μπορεί να εκπονηθεί από ένα γράφημα βασιζόμενος στη διαδικασία πολλαπλασιασμού. Δηλαδή πρώτα γίνεται η λήψη του π^T σαν μετα-δεδομένα από τον κόμβο i και μετά μεταφέρεται το π^T από τον κόμβο i σε καθένα από τους συνδεδεμένους κόμβους j πολλαπλασιάζοντας με $p_{i,j}$. Αθροίζοντας όλες τις εισερχόμενες τιμές από κάθε κόμβο j προκύπτει το νέο $(\pi')^T$. Η διαδικασία περιγράφεται παρακάτω: (Gao, κ.συν. ,2011 σελ 5)



Σχήμα 39

Ο πίνακας P δημιουργείται από τον παραπάνω γράφο (Σχήμα 39), κάθε στοιχείο $p_{i,j}$ του οποίου ορίζεται ως το κανονικοποιημένο βάρος ακμής από έναν κόμβο I στον κόμβο j. Το προϊόν μπορεί να ερμηνευτεί ως αποτέλεσμα της διαδικασίας πολλαπλασιασμού για τα δεδομένα του γράφου. $\pi^T \Rightarrow (\pi')^T$

Map: Ως είσοδο παίρνει το graph record $<I, \{j, p_{ij}\},\ j \in O_i>$ και το $<I, \pi_\iota>$. Map είσοδος για j, έτσι ώστε οι πλειάδες με το ίδιο j ανακατεύονται στην ίδια μηχανή με μορφή $< j, \{\pi_i p_{ij}\},\ \forall j \in O_i >$

Reduce: Παίρνει το $<j, \{\pi_i p_{i,j}\}, \forall\ I \in I_j>$ και εκπέμπει $< j, \pi'_j>$, όπου

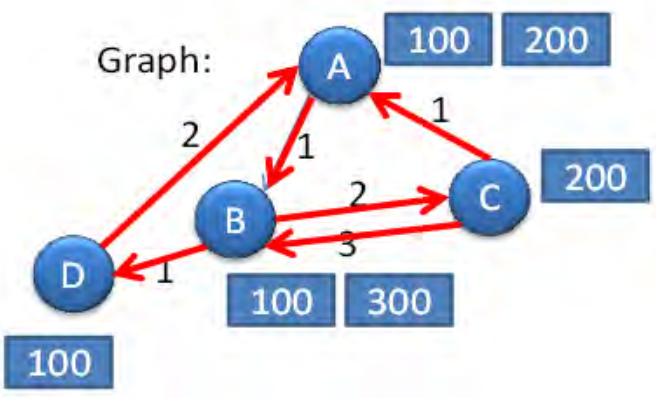
$$\pi'_j = \sum_{\forall i \in I_j} \pi_i p_{ij}$$

Παράδειγμα: (Πολλαπλασιασμού Matrix-Vector)

Δίνεται ο παρακάτω πίνακας P και το διάνυσμα x=[100,100,100,100]^T για τον υπολογισμό του γινομένου πίνακα με διάνυσμα εφαρμόζονται τα εξής:

$$P = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 1 & 0 & 3 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Αρχικά πολλαπλασιάζονται τα στοιχεία του x με τα στοιχεία του γράφου που ορίζεται από τον πίνακα P και στη συνέχεια αθροίζονται οι πολλαπλασιασμένες τιμές για κάθε κόμβο.



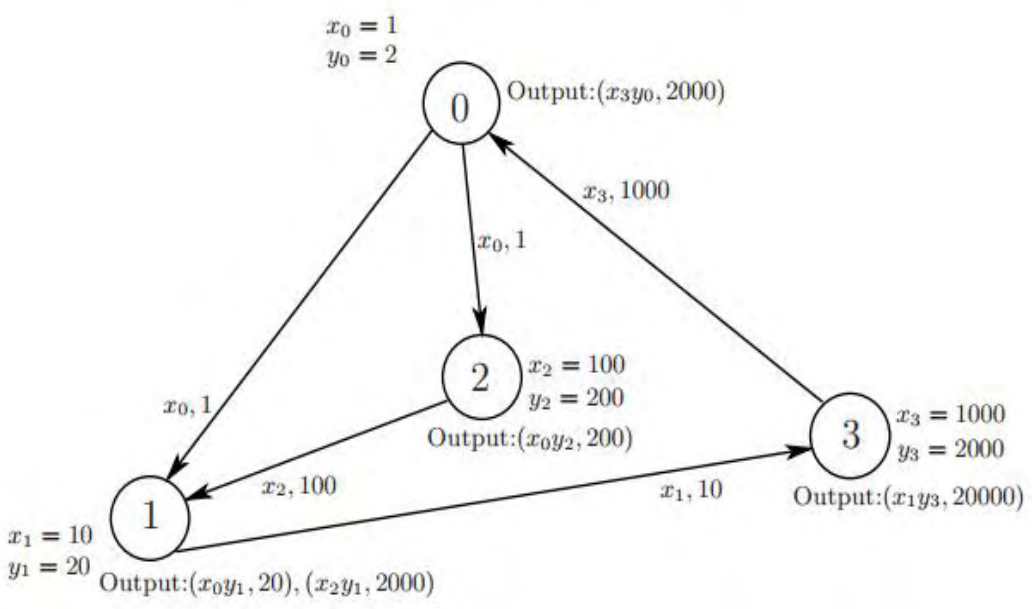
Σχήμα 40: Διαδικασία πολλαπλασιασμού Matrix-Vector (Rajaraman, 2011)

Οπότε έχουμε P x = [300, 400, 200, 100] ^{T}. Οι τιμές 300, 400, 200 και 100 αντιστοιχούν στους κόμβους A, B, C, D αντίστοιχα. (Rajaraman, 2011)

5.4.3 Kronecker προϊόν για διανύσματα σε αραιό γράφο-Παράδειγμα

Έστω ότι X και Y είναι δύο n-διάστατα διανύσματα και στόχος είναι ο υπολογισμός του προϊόντος kronecker z = x ⊗ y (όπου z είναι ένας n^2-διάστατος

πίνακας) από αυτά ,σε ένα αραιό γράφημα. Ειδικότερα γίνεται ο υπολογισμός των x_i y_j μόνο εάν υπάρχει ακμή από τη σελίδα I στη σελίδα j στο γράφημα. Δεν χρειάζεται να γίνεται αναφορά στον αραιό γράφο για να προσδιοριστεί εάν χρειάζεται να υπολογιστεί ή όχι το xi yj. Η λύση είναι η λήψη των xi, yi ως μετα-δεδομένα από τον κόμβο I του γραφήματος και η μετάβαση του xi από τον κόμβο I στους κόμβους με τους οποίους συνδέεται. Πολλαπλασιάζοντας τα yi με όλα τα xi που έχουν ληφθεί από τον κόμβο j, προκύπτουν όλα τα απαραίτητα xiyj. (Gao, κ.συν. , 2011 σελ .5).



Σχήμα 41: Το Kronecker προϊόν των δύο διανυσμάτων προσπαθεί να υπολογίσει το xiyj κάτω από τους περιορισμούς της δομής του γραφήματος. Μπορούμε να διαδώσουμε το xi κατά μήκος των ακμών του γράφου ώστε να φιλτράρουμε το απαραίτητο xi και να το πολλαπλασιάσουμε με το yi.

Η μέθοδος μπορεί να υλοποιηθεί με δύο MapReduce διαδικασίες:

Map-I: Ως είσοδο παίρνει το graph record <I, {j, p_{ij}}, ∀ j ∈ O_i> και το <I, $x_ι$ >. Map είσοδος για j, έτσι ώστε οι πλειάδες με το ίδιο j ανακατεύονται στην ίδια μηχανή με μορφή < j, (i, x_i) >

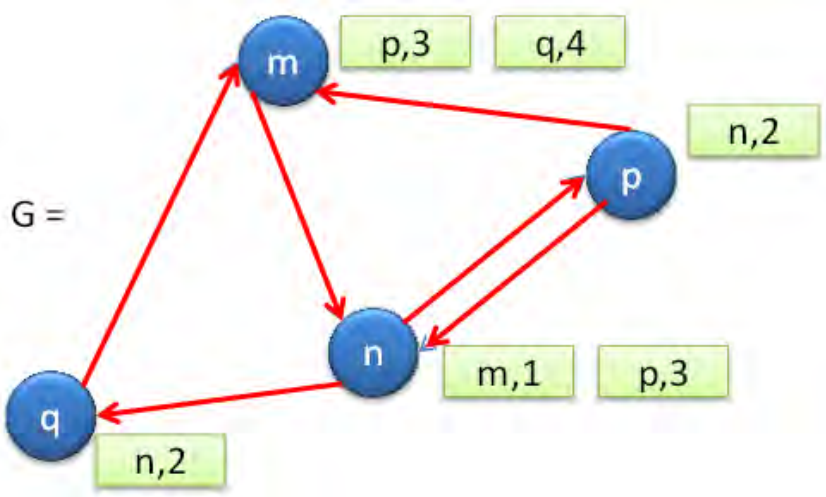
Reduce-I: Παίρνει το <j, (I,xi) , ∀ I ∈ I_j > και εκπέμπει < j, {(I,xi), I ∈ I_j>

Map-II: Γίνεται map το <j, {(i , xi), i ∈ Ij } > και το <j, yj > στο j, αυτές οι πλειάδες με το ίδιο j ανακατεύονται στο ίδιο μηχάνημα με τη μορφή <j, {yj, (i,xi) i∈ Ij } >

Reduce-II: Παίρνουμε <j , {yj, (I,xi), I ∈ Ij} και εκπέμπουμε <i, j , xiyj >

Παράδειγμα: (Kronecker Product)

Σύμφωνα με τον Anand Rajaraman, (2011) για τον υπολογισμό του Kronecker Product $x \otimes_G y$ αρχικά διαδίδεται το διάνυσμα y μέσω του γραφήματος G. Στη συνέχεια για κάθε κόμβο πολλαπλασιάζεται το χ μαζί με τις λαμβανόμενες y τιμές. Όπου $x= [5,6,7,8]^T$ και $y=[1,2,3,4]^T$



Σχήμα 42: Διαδικασία υπολογισμού του Kronecker Product *(* Rajaraman, 2011)

Από τον παραπάνω γράφο (Σχήμα 42) εξάγεται ο παρακάτω πίνακας από τον οποίο προκύπτει το τελικό αποτέλεσμα. (Σχήμα 43)

M	**N**	**P**	**Q**
[p,3]	[m,1]	[n,2]	[n,2]
[q,4]	[p,3]		

Σχήμα 43

Result=

[mp,15] [mq,20]	[nm,6] [np,18]	[pn,14]	[qn,16]

5.5 Αλγόριθμοι ταξινόμησης γράφου που αποτελούν ειδικές περιπτώσεις των προβλημάτων Framework

5.5.1 LiftHITS

Έστω ότι ένας πίνακας M καθορίζει τη σύνδεση μεταξύ των κόμβων. Για παράδειγμα $M_{i,j}=0$ εάν δεν υπάρχει σύνδεσμος από τον κόμβο i στον κόμβο j αλλιώς είναι διαφορετικό του μηδενός. Τότε τα authority scores π μπορούν να υπολογιστούν επαναληπτικά από τη σχέση $\pi'=M^TM\ \pi$ στο HITS. Εάν χρησιμοποιηθεί η ισότητα $S(\pi; B,\mu)= B(\pi'-\pi)$ και ο τύπος (d) της αντικειμενικής συνάρτησης (Σχήμα 44) $R(\pi)= ||\pi'-M^T\ M\ \pi||$, τότε το Πρόβλημα 1 γίνεται:

$$\begin{array}{ll} \min_{\pi\geq 0} & \|\pi' - M^T M\pi\| \\ s.t. & B(\pi' - \pi) \geq 0, \end{array}$$

Όπου B πίνακας που καθορίζει τους κόμβους που θα πρέπει να "lifted" στην επόμενη επανάληψη. Ο αλγόριθμος HITS είναι μια ειδική περίπτωση του γενικού framework στον οποίο χρησιμοποιούνται ο τύπος (d) της αντικειμενικής συνάρτησης και περιορισμοί δυαδικών ετικετών.

5.5.2 Adaptive PageRank

Χρησιμοποιώντας τη σχέση $S(\pi;B,\mu)=B\pi-\xi$, όπου $\xi>0$ είναι ένα slack διάνυσμα και υιοθετώντας τον τύπο (d) της αντικειμενικής συνάρτησης $R\equiv R(\pi(\eta))= ||\pi(\eta)- \pi'||$ όπου π και π' υπολογίζονται από τη δομή του γράφου, τότε το Πρόβλημα 1 γίνεται:

$$\begin{array}{ll} \min_{\pi\geq 0} & \|\pi(\eta) - \bar{\pi}\| \\ s.t. & B\pi \geq \xi,\ \xi \geq 0 \\ & \bar{\pi} = (1-d)(I-dP)^{-1}e \\ & \pi(\eta) = (1-d)(I-dP)^{-1}\eta \\ & \eta \geq 0. \end{array}$$

Όπου P είναι ένας transition πίνακας που προέρχεται από το γράφο , d είναι ο damping factor του PageRank και η ένα n-διάστατο διάνυσμα για τους κόμβους του

γράφου. Ο αλγόριθμος Adaptive PageRank είναι μια ειδική περίπτωση του γενικού framework στον οποίο χρησιμοποιούνται ο τύπος (d) της αντικειμενικής συνάρτησης και περιορισμοί των ζευγών προτίμησης .

5.5.3 NetRank

Υποθέτοντας ότι ο Transition πίνακας μπορεί να παραμετρικοποιηθεί με βάση τους διαφορετικούς τύπους των ακμών τότε η ακμή (i , j) ανήκει στον τύπο t(i, j) . Εάν ο τύπος t σχετίζεται με το βάρος w(t), τότε το βάρος της ακμής θα είναι w(t(I,j)). Επομένως μια πιθανή διαμόρφωση του πίνακα είναι:

$$P_{ij}(t) = \frac{w(t(i,j))}{\sum_j w(t(i,j))}.$$

Θέτοντας α=0, β=1 και $S(\pi;B,\mu)= -e^T(e-B\pi)$ τότε το Πρόβλημα 1 γίνεται:

$$\begin{array}{ll} \min_{\pi \geq 0} & \sum_{i \prec j} (1 + \pi_i - \pi_j) \\ s.t. & \pi = (P^T(t))^H \pi^{(0)}. \end{array}$$

Όπου $\pi^{(0)}$ είναι ένα διάνυσμα με στοιχεία ίσα με 1/n και H είναι ο αριθμός επανάληψης του πολλαπλασιασμού πινάκων. Ο αλγόριθμος NetRank είναι μια ειδική περίπτωση του γενικού framework στον οποίο χρησιμοποιούνται ο τύπος (d) της αντικειμενικής συνάρτησης και περιορισμοί των ζευγών προτίμησης(Gao,κ.συν., 2011 σελ. 6).

5.5.4 Laplacian Rank

Ορίζοντας τον τύπο (d)της αντικειμενικής συνάρτησης και χρησιμοποιώντας Laplacian στο γράφο $R(\pi)=\pi^T L\pi$ προκύπτει ο Laplacian (L) για ένα κατευθυνόμενο γράφο

$$L = I - \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2}$$

Όπου Π ένας διαγώνιος πίνακας με $\Pi_{ii} = \pi'_{i..}$ Χρησιμοποιώντας τους περιορισμούς των ζευγών προτίμησης και θέτοντας α=1/2 το Πρόβλημα 2 γίνεται:

$$\begin{array}{ll} \min_{\pi \geq 0} & \frac{1}{2}\pi^T L\pi + \beta \sum_{(i,j)} \mu_{ij}\xi_{ij} \\ s.t. & \pi_i - \pi_j \geq 1 - \xi_{ij}, \xi_{ij} \geq 0. \end{array}$$

Ο αλγόριθμος Laplacian Rank είναι μια ειδική περίπτωση του γενικού framework στον οποίο χρησιμοποιούνται ο τύπος (d) της αντικειμενικής συνάρτησης και περιορισμοί των ζευγών προτίμησης.

5.5.5 Supervised Random Walk

Υποθέτοντας ότι κάθε ακμή από έναν κόμβο i στον κόμβο j έχει l- διάστατο διάνυσμα $x_{i,j}=(x_{ij1}, x_{ij2,}\ldots\ldots x_{ijl})^T$ το οποίο περιγράφει τα γνωρίσματα κόμβων στο i , τα γνωρίσματα κόμβων στο j και τα γνωρίσματα αλληλεπίδρασης όσο αφορα τις ακμές από τον κόμβο i στον κόμβο j. Χρησιμοποιώντας τον Σχήμα 38 και υιοθετώντας τον τύπο (b) της αντικειμενικής συνάρτησης $R(\pi;f(\omega,X)) \equiv R(\omega)=\|\omega\|^2$ καθώς και τους περιορισμούς των ζευγών προτίμησης $S(\pi;B,\mu)= -\mu B^T \pi$ μπορούμε να πάρουμε ένα καινούριο πρόβλημα από το Πρόβλημα 2.

$$\min_{\omega \geq 0, \pi \geq 0} \alpha\|\omega\|^2 + \beta\mu^T B\pi,$$

Ο αλγόριθμος Supervised Random Walk είναι μια ειδική περίπτωση του γενικού framework στον οποίο χρησιμοποιούνται ο τύπος (b) της αντικειμενικής συνάρτησης και περιορισμοί των ζευγών προτίμησης(Gao, κ.συν .,2011 σελ .7)

TYPE	EDGE	NODE	FORM	EXAMPLE METHODS
(a)	√	√	$R(\pi; f(\omega, X), g(\phi, Y))$	Semi-Supervised PageRank
(b)	√		$R(\pi; f(\omega, X))$	Supervised Random Walks
(c)		√	$R(\pi; g(\phi, Y))$	
(d)			$R(\pi)$	LiftHITS, Adaptive PageRank, NetRank, Laplacian Rank

Σχήμα 44 (Gao, κ.συν., 20

ΚΕΦΑΛΑΙΟ 6ο

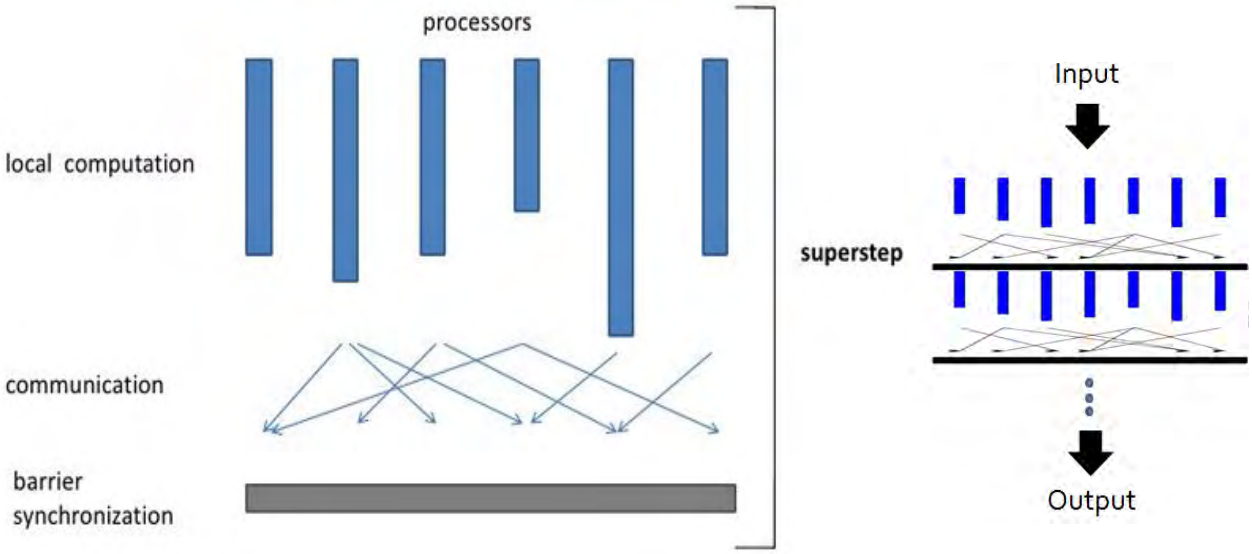
Graph Processing Systems

Semi-Supervised Framework για την κατάταξη γράφου

6.1 Ορισμοί των πέντε ειδών συστήματος επεξεργασίας γράφων

- ✓ **Pegasus:** Είναι ένα σύστημα εξόρυξης δεδομένων που επεξεργάζεται γραφήματα μεγάλης κλίμακας γραμμένο πλήρως σε Java. Το σύστημα αυτό τρέχει με παράλληλα κατανεμημένο τρόπο βασιζόμενο στο Hadoop. Το Hadoop είναι μια πλατφόρμα για cloud computing όπως επίσης και ανοιχτός κώδικας εκτέλεσης του MapReduce πλαισίου(Kang, κ.συν., 2011). Επιπρόσθετα το Pegasus υποστηρίζει τους εξής αλγορίθμους: *PageRank, Random walk with restart, Graph diameter computing, Graph components mining.* Το σημαντικότερο πλεονέκτημά του είναι το γεγονός ότι χάρη στο Pegasus αναλύονται ένα από τα μεγαλύτερα διαθέσιμα στο κοινό γραφήματα που προσφέρει η Yahoo, το οποίο αποτελείται από 6.7 δισεκατομμύρια ακμές (Kang, κ.συν. ,2010).

- ✓ **Hama:** Είναι βιβλιοθήκη επεξεργασίας γραφήματος σε Hadoop. Ως μοντέλο υπολογισμού χρησιμοποιεί MapReduce για να χειριστεί υπολογισμούς μεταξύ πινάκων αλλά και BSP (Bulk Synchronous Parallel) για να χειριστεί την επεξεργασία άλλων γράφων. Υποστηρίζει τους εξής αλγορίθμους: Πολλαπλασιασμό μεγάλης κλίμακας πινάκων, Εύρεση μικρότερου μονοπατιού και PageRank (wikipidia*).*



Σχήμα 45: *Bulk Synchronous Parallel (*Rob H.Bisseling, 2004*)*

- ✓ **Pregel:** Ένα σύστημα για επεξεργασία γράφων μεγάλης κλίμακας, της Google. Το μοντέλο υπολογισμού βασίζεται στον Bulk Synchronous Parallel και στο Ram-Based System. Υποστηρίζει τους αλγορίθμους *PageRank, Μοναδική πηγή κοντινότερου μονοπατιού και για την εύρεση στοιχείων γραφήματος.*

Το σύστημα Pregel σύμφωνα με τον Rob H.Bisseling,(2004a) κάνει χρήση της έννοιας superstep το οποίο είναι χαρακτηριστικό του BSP αλγορίθμου που χρησιμοποιείται ως μοντέλο υπολογισμού για το συγκεκριμένο σύστημα.
Ένα superstep διακρίνεται σε δύο είδη: computation superstep και communication superstep

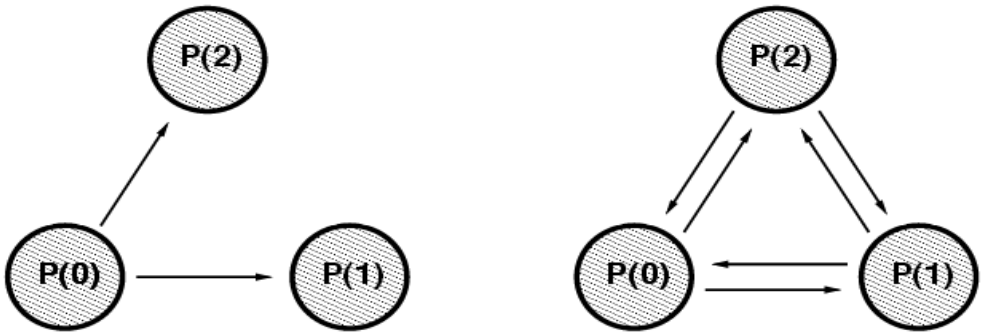
Ένας **υπολογισμός superstep** (computation superstep) αποτελείται από μικρά steps τα οποία μπορεί να είναι floating-point πράξεις (flops) όπως: πολλαπλασιασμός, πρόσθεση, αφαίρεση και διαίρεση. Στην επιστήμη των υπολογιστών ως flop ορίζουμε την κοινή μονάδα για έκφραση κόστους υπολογισμού. Επιπρόσθετα μια **επικοινωνία superstep** (communication superstep) περιλαμβάνει βασικές λειτουργίες όπως είναι η μεταφορά λέξεων-δεδομένων ως ακεραίους από τον ένα επεξεργαστή

στον άλλον. Σε θεωρητικό επίπεδο γίνεται διάκριση ανάμεσα στα δύο αυτά διαφορετικά ήδη superstep και αυτό βοηθάει στο σχεδιασμό και στην ανάλυση παράλληλων αλγορίθμων. Αντίθετα όσον αφορά το πρακτικό επίπεδο η διάκριση μεταξύ των δύο ειδών superstep μπορεί να παραληφθεί.

Αξίζει να αναφερθεί ότι ως **h-relation** είναι μια επικοινωνία superstep στην οποία κάθε επεξεργαστής στέλνει και λαμβάνει το πολύ h λέξεις- δεδομένα.

$$\mathbf{h = max \{ hs ,hr\}}$$

όπου hs είναι ο μέγιστος αριθμός από δεδομένα λέξεις που στέλνονται από έναν επεξεργαστή και hr είναι ο μέγιστος αριθμός δεδομένων λέξεων που λαμβάνονται από έναν επεξεργαστή. Στο παρακάτω γράφημα απεικονίζεται μια 2-relation επικοινωνία superstep (Bisseling,2004b).



Σχήμα 46 (Bisseling, 2004 σ5)

Έτσι για κάθε κορυφή γίνονται οι παρακάτω ενέργειες:

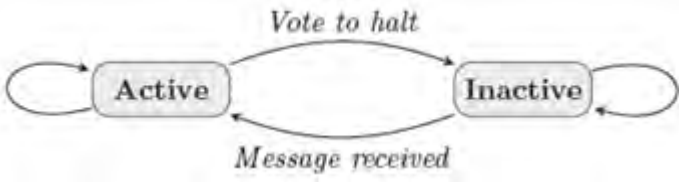
- Λαμβάνονται τα μηνύματα που στέλνονται στο προηγούμενο superstep.
- Σε κάθε κορυφή εκτελείται η ίδια συνάρτηση.
- Κάθε κορυφή τροποποιεί την δικιά της αξία ή την αξία των εξερχομένων ακμών του.
- Κάθε κορυφή στέλνει μηνύματα προς τις άλλες κορυφές τα οποία λαμβάνονται στο επόμενο superstep .
- Κάθε κορυφή επηρεάζει στην αλλαγή της τοπολογίας του γράφου.

Η διαδικασία τερματίζεται όταν όλες οι κορυφές είναι ταυτόχρονα μη-ενεργές και δεν υπάρχουν μηνύματα να μεταδοθούν (Malewicz, κ.συν. , 2010a).

6.1.1 Σύγκριση Pregel με MapReduce

Το σύστημα Pregel διατηρεί τις ακμές και τις κορυφές στη μηχανή που εκτελεί τους υπολογισμούς και χρησιμοποιεί το διαδίκτυο μόνο για τη μεταφορά των μηνυμάτων. Σε αντίθεση η MapReduce περνάει όλη την κατάσταση του συστήματος από το ένα στάδιο στο επόμενο ενώ ταυτόχρονα χρειάζεται να συντονίσει τα βήματα μιας «αλυσοδεμένης» MapReduce (Malewicz, κ.συν. , 2010b).

- ✓ **Trinity:** Είναι βάση δεδομένων ενός γράφου και πλατφόρμα υπολογισμών από τη Microsoft Research. Το μοντέλο υπολογισμού βασίζεται στον Bulk Synchronous Parallel με ασύγχρονο τρόπο όσον αφορά την ανταλλαγή μηνυμάτων και στο Ram-based system. Υποστηρίζει τους αλγορίθμους *PageRank και Breadth first search*. (wikipidia)
- ✓ **Graphor:** Είναι μια μηχανή υπολογισμού γράφου. Το μοντέλο υπολογισμού βασίζεται στον αλγόριθμο MapReduce και στη πρόσθετη λογική για να διατηρήσει την τοπικότητα του γράφου εμπνευσμένο από τον BSP αλγόριθμο. Υποστηρίζει τους αλγορίθμους PageRank, Matrix-vector multiplication και Graph- based Kronecker product των κορυφών. (wikipidia).



Σχήμα 48 (Rajaraman, κ.συν.,2001)

6.2 Σύγκριση των πέντε συστημάτων επεξεργασίας γράφου

System	Model	Fault tolerance	Supported algorithms
Pegasus	MapReduce	MapReduce	PageRank, graph components finding
Hama	MapReduce/BSP	MapReduce	PageRank, matrix vector multiplication
Pregel	BSP	Self designed	PageRank, shortest path, graph components finding
Trinity	BSP +	None	PageRank, breadth first search on graph
Graphor	MapReduce+BSP	MapReduce	PageRank, multiplication of matrix and vector, graph based vector kronecker product

Σχήμα 47

Συμπεράσματα - Επίλογος

Συνοψίζοντας συμπεραίνονται τα εξής: Αρχικά ότι η link ανάλυση είναι μια κλασική μέθοδος κατάταξης (ranking) σε ένα γράφημα και ότι οι κόμβοι και οι ακμές που είναι πλούσιοι σε πληροφορίες μπορούν να βοηθήσουν στην κατάταξη (ranking). Για τη διευκόλυνση της εύρεσης των χρήσιμων πληροφοριών από τον χρήστη είναι επιτακτική ανάγκη η δημιουργία βελτιωμένων αλγορίθμων κατάταξης στις μηχανές αναζήτησης. Οι αλγόριθμοι αυτοί μπορεί να είναι βελτιωμένη έκδοση των παραδοσιακών αλγορίθμων ή η υλοποίηση καινούριων. Επιπρόσθετα πολλά συστήματα έχουν αναπτυχθεί για τα μεγάλα σε κλίμακα κατάταξη γραφήματα και τα συστήματα που έχουμε αναφέρει στο τελευταίο κεφάλαιο εφαρμόζονται ευρέως.

Μελλοντικοί στόχοι:

Όσον αφορά την μελλοντική μελέτη πάνω στους αλγορίθμους δύο είναι τα βασικά σημεία που θα πρέπει να ληφθούν υπόψη. Τα σημεία αυτά είναι οι διάφορες διεργασίες Markov και η θεωρία μάθησης πάνω στην διαδικασία κατάταξης (ranking). Επίσης θα πρέπει να γίνει προσπάθεια υλοποίησης αλγορίθμων οι οποίοι να εφαρμόζουν κατάταξη (ranking) σε μια χρονική σειρά γραφημάτων ή και σε ετερογενή γραφήματα. Όσον αφορά την εκτέλεση των αλγορίθμων στόχος μας θα πρέπει να είναι η αποτελεσματικότητα η ευελιξία και η αξιοπιστία.

Βιβλιογραφία

- Aditi Muralidharan , Zoltan Gyongyi and Ed H. Chi "Social Annotations in Web Search" 2012
- Agarwal et al. Learning to rank networked entities. KDD, 2006.
- Agarwal Ranking on graph data. ICML, 2006.
- Alessio Signorini "A Survey of Ranking Algorithms" Department of Computer Science University of Iowa (September 11, 2005)
- AMY. N.Langville-Carl D. Meyer "Η Μέθοδος PageRank της Google και άλλα συστήματα κατάταξης" (Google's PageRank and Beyond: The Science of Search Engine Rankings) 2010
- Amy N. Langville and Carl D. Meyer "The Use of the Linear Algebra by Web Search Engines", 2004
- A. Langville and C. Meyer, "A Survey of Eigenvector Methods for Web Information Retrieval,"SIAM Review, vol. 47, no. 1,pp. 135–161, 2005
- Anand Rajaraman , Jeffrey David Ullman "Mining of Massive Datasets" (December 2011) The Stanford University InfoLab,
- Ayman Farahat , Thomas Lofaro, Joel C. Miller, Gregory Rae, Lesley A. Ward "Authority Rankings From Hits, PageRank and Salsa: Existence, Uniqueness and effect of initialization Vol. 27, No. 4, pp. 1181–1201, 2006
- Brin. S, & Paget (1998) "The Anatomy of Large-Scale Hypertextual Web Search Engine"
- Bin Gao, Tie –Yan Liu ,Weiwei, Taifeng Wang, Hang Li "Semi-Supervised Ranking on very Large Graph with Rich Metadata" Microsoft Research Technical Report, MSR TR 2011-36, 2011.
- Cai et al. Block level link analysis. SIGIR, 2004.
- Chakrabarti et al. Learning parameters in entity relationship graphs from ranking preference. PKDD, 2006.
- Chang et al. Learning to create customized authority lists. ICML, 2000.
- Christopher D. Manning, Prabhakar Raghakar and Hinrich Schutze *Introduction to Information Retrieval*, Cambridge University Press. 2008.
- Chris Ding, Hongyuan Zha, Xiaofeng He, Parry Husbands and Horst D.Simon "Link Analysis: Hubs and authorities on the world wide web" , 2001ανανέωση 2003

- Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha and Horst Simo "PageRank, HITS and a Unified Framework for Link Analysis", 2001 ανανέωση 2002
- Daxin Jiang, Jian Pei ,Hang Li (30 Απριλίου 2010) "Search and Browse Log Mining for Web Information Retrieval: Challenges, Methods, and Applications"
- Deng Cai , Xiaofei He,Ji-Rong Wen, Wei-Ying Ma (2004) "Block-level Link Analysis"
- D.Gibson, J.Kleinberg and P.Raghavan "Inferring Web Communities from Link Topology" Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998
- Eneko Agirre Personalized PageRank over WordNet for Similarity and Word Sense Disambiguation ,Google 2009
- Enoch Peserico, Luca Pretto - D.E.I., Univ. Padova, Italy "HITS can converge slowly, but not too slowly, in score and rank" 2009
- Erik Andersson, Per-Anders Ekstrom "Investigating Google's PageRank Algorithm" Report in Scientific Computing, Spring, 2004 Cuppsalo University
- Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski Google, Inc. "Pregel: A System for Large-Scale Graph Processing" (6 Ιουν 2010)
- G. Del Corso, A. Gull´ı, and F. Romani, "Fast PageRank Computation via a Sparse Linear System,"Internet Mathematics, 2005
- Gao et al. A general Markov framework for page importance computation. CIKM, 2009.
- Haveliwala et al. An analytical comparison of approaches to personalizing PageRank. Stanford University Technical Report, 2003
- Kang et al. PEAGSUS: a peta scale graph mining system implementation and observations. ICDM, 2009
- Kang et al. PEAGSUS: mining peta scale graphs, knowledge and information systems. DOI: 10.1007/s10115-010-0305-0, 2010.
- Kang, Duen Horng Chau, and Christos Faloutsos Mining Large Graphs: Algorithms, Inference, and Discoveries IEEE International Conference on Data Engineering (ICDE) 2011, Hannover, German

- U Kang, Spiros Papadimitriou, Jimeng Sun, and Hanghang Tong «Centralities in Large Networks: Algorithms and Observations» SIAM International Conference on Data Mining (SDM) 2011, Mesa, Arizona, USA
- Kleinberg. Authoritative sources in a hyperlinked environment. IBM Research Report RJ 10076, 1997.
- Liu et al. BrowseRank: Letting Web users vote for page importance. SIGIR, 2008.
- Created by Lempel Moran in 2000 and represented by A.Simkins «Stochastic Approach for Link Structure Analysis»
- Malewiczet al. Pregel: a system for large-scale graph processing. PODC 2009.
- Malewicz et al.Pregel: a System for large-scale graph processing. SIGMOD, 2010.
- Marc Najork «Comparing the Effectiveness of HITS and SALSA» Microsoft Research, 2007
- Nidhi Grover "Comparative Analysis Of Pagerank And HITS Algorithms" International Journal of Engineering Research & Technology (IJERT) ,Vol. 1 Issue 8, October – 2012
- Patrick Pantel, Michael Gamon Microsoft Research , Omar Alonso, Kevin Haas Microsoft Corp Paper, (2012) "Social Annotations: Utility and Prediction Modeling"
- Page et al. The PageRank citation ranking: bringing order to the Web. Stanford Digital Library Technologies Project, 1998 .
- Rob H.Bisseling "Parallel Scientific Computation: A Structured Approach using BSP and MPI " Oxford (May 6, 2004)
- Ryan A. Rossi and David F. Gleich "Dynamic PageRank using Evolving Teleportation",2012
- Seo et al. HAMA: an efficient matrix computation with the MapReduce framework. IEEE CloudCom Workshop, 2010.
- Shenghua Bao , Xiaoyuan Wu, Ben Fei, Guirong Xue, Zhong Su, Yong Yu. "Optimizing Web Search Using Social Annotations" (www, 2007)
- Tien Hao Chang (Darby Chang) Numerical Analysis 2011
- Tsoi et al. Adaptive ranking of Web pages. WWW, 2003.

- Vu Thanh Nguyen “Using social annotation and web log to enhance search engine” International Journal of Computer Science Issues, Vol. 6, No. 2, 2009
- Wei Wu,Hang Li and Jun Xu “Learning Query and Document Similarities from Click-through Bipartite Graph with Metadata” ,2011
- Xue et al. Exploiting the Hierarchical Structure for Link Analysis. SIGIR, 2004
- X. Wu, L. Zhang, and Y. Yu. Exploring Social Annotations for the Semantic Web. In: *Proc. of WWW 2006*, pp. 417-426, May 23.26, 2006
- Xue-Mei Jiang, Gui-Rong Xue, Wen-Guan Song,, Hua-Jun Zeng, Zheng Chen Wei-Ying Ma «Exploiting PageRank at Different Block Level», 2004
- Απόστολος Ν. Παπαδόπουλος, Ανάκτηση στον παγκόσμιο ιστό (Πανεπιστήμιο Πατρών) 2008
- Wikipidia: http://wiki.apache.org/hama/
- Wikipidia: http://en.wikipedia.org/wiki/Trinity
- Wikipidia:http://el.wikipedia.org/wiki/%CE%92%CE%B5%CE%BB%CF%84%CE%B9%CF%83%CF%84%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7_%CE%B3%CE%B9%CE%B1_%CF%84%CE%B9%CF%82_%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%AD%CF%82_%CE%B1%CE%BD%CE%B1%CE%B6%CE%AE%CF%84%CE%B7%CF%83%CE%B7%CF%82
- Paper: Οδηγός έναρξης για τη βελτιστοποίηση στη μηχανή αναζήτησης του Google Έκδοση 1.1 Δημοσιεύτηκε στις 13 Νοεμβρίου 2008