Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 19η

Η μέθοδος BrowseRank

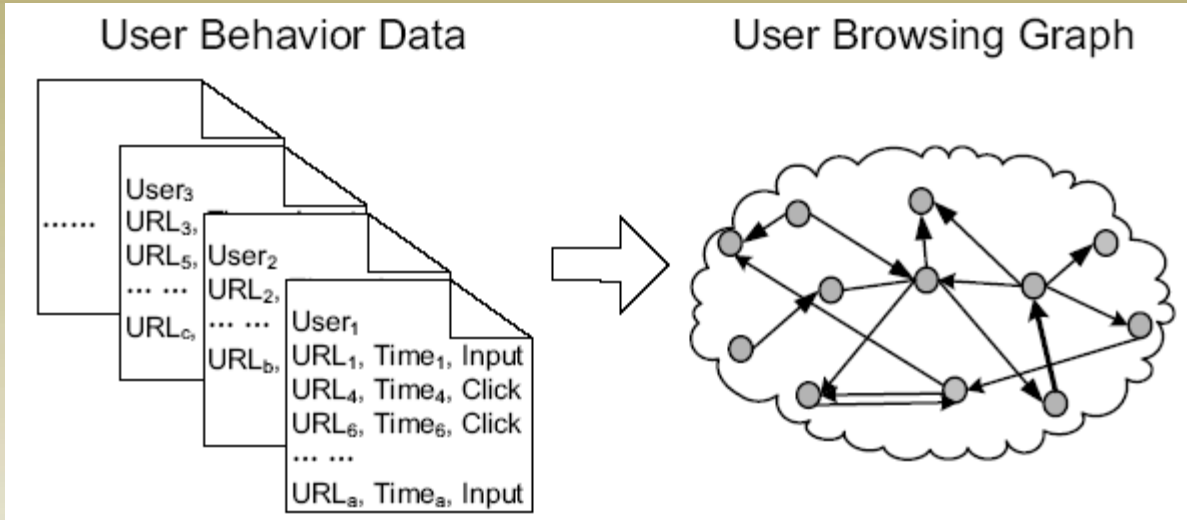
Εισαγωγή

- Η *page importance*, που αναπαριστά την 'αξία' μιας σελίδας του Web, είναι παράγων-κλειδί για την αναζήτηση στο Web, επειδή οι σύγχρονες μηχανές αναζήτησης, ο ερπυσμός (crawling), το indexing, και η διαβάθμιση (ranking) συνήθως καθοδηγούνται από αυτή τη μετρική
- Προς το παρόν, η page importance υπολογίζεται με χρήση του *link graph* του Web και αυτή η διαδικασία λέγεται link analysis
- Παρουσιάσαμε ήδη αλγορίθμους για link analysis: τον *HITS* και *PageRank*

Google PageRank

- Ο PageRank βασίζεται σε μια *discrete-time Markov διαδικασία* πάνω στον Web link graph για να υπολογίσει την page importance, και στην ουσία υλοποιεί έναν τυχαίο περίπατο (random walk) ενός Web surfer πάνω στους υπερσυνδέσμους (hyperlinks) του Web
- Περιορισμοί του PageRank
  - Ο link graph, πάνω στον οποίο βασίζεται ο PageRank, δεν είναι αξιόπιστη πηγή δεδομένων, επειδή τα hyperlinks του Web μπορούν να προστεθούν/διαγραφούν συχνά από τους δημιουργούς περιεχομένου
  - Ο PageRank μοντελοποιεί απλά έναν random walk πάνω στον link graph, αλλά ΔΕΝ λαμβάνει υπόψην του την *διάρκεια του χρόνου που ξοδεύει ο surfer* πάνω στις Web σελίδες κατά την διάρκεια του random walk

User Browsing Graph



- **<u>Μπορούμε να βρούμε μια καλύτερη πηγή δεδομένων αντί του link graph?</u>**
- Χρήση του *user browsing graph*, που προκύπτει από τα user behavior data
- Τα δεδομένα συμπεριφοράς των χρηστών (user behavior data) μπορούν να καταγραφούν από τους browsers και να συλλεγούν από τους web servers

Continuous-time Markov chain

- **Τι είδους αλγορίθμους πρέπει να χρησιμοποιήσουμε για να αξιοποιήσουμε την νέα πηγή δεδομένων;**
- Η χρήση μιας discrete-time Markov process δεν είναι πλάον επαρκής
- Ορίζουμε μια *continuous-time Markov process* ως το μοντέλο για τον user browsing graph
- Υποθέτουμε ότι η διαδικασία είναι *time-homogenous*
- Η stationary probability distribution της διαδικασίας μπορεί να χρησιμοποιηθεί για να ορίσουμε την importance των Web pages
- Εφαρμόζουμε τον αλγόριθμο *BrowseRank*, για να υπολογίσουμε αποδοτικά την stationary probability distribution της continuous-time Markov process
- Κάνουμε χρήση ενός μοντέλου προσθετικού θορύβου (additive noise) για να αναπαραστήσουμε τις παρατηρήσεις σε σχέση με την Markov process και για να εκτιμήσουμε τις παραμέτρους τις διαδικασίας
- Υιοθετούμε μια *embedded Markov chain* για να επιταχύνουμε τον υπολογισμό της stationary distribution

User Behavior Data

URL	TIME	TYPE
http://aaa.bbb.com/	2007-04-12, 21:33:05	INPUT
http://aaa.bbb.com/1.htm	2007-04-12, 21:34:11	CLICK
http://ccc.ddd.org/index.htm	2007-04-12, 21:34:52	CLICK
http://eee.fff.edu/	2007-04-12, 21:39:03	INPUT
...	...	...

- Τα user behavior data μπορούν να καταγραφούν και να αναπαρασταθούν με τριάδες της μορφής <URL, TIME, TYPE>
- Από τα δεδομένα, εξάγουμε μεταβάσεις των χρηστών από σελίδα σε σελίδα καθώς και τον χρόνο που ξοδεύουν οι χρήστες στις σελίδες ως ακολούθως:
  - Κατακερματισμός των sessions (διάσπαση με: time rule & type rule)
  - Κατασκευή των URL pair
  - Εκτίμηση της reset probability
  - Εξαγωγή του staying time

Εξαγωγή του staying time

- Για κάθε ζεύγος URL, χρησιμοποιούμε την διαφορά μεταξύ του χρόνου της δεύτερης σελίδας και αυτού της πρώτης σελίδας, ως εκτίμηση του χρόνου παραμονής στην πρώτη σελίδα
- Για την τελευταία σελίδα του session, χρησιμοποιούμε το ακόλουθο ευρεστικό για να εκτιμήσουμε τον χρόνο παραμονής
  - Εάν το session κατακερματιστεί με τον *time rule*, παίρνουμε ένα ***τυχαίο*** (!?) δείγμα από την κατανομή των χρόνων των παρατηρημένων staying time των σελίδων σε όλες τις εγγραφές
  - Εάν η session κατακερματιστεί με τον *type rule*, χρησιμοποιούμε την διαφορά μεταξύ του χρόνου της τελευταίας σελίδας στο session και του χρόνου της πρώτης σελίδας του επόμενου session (INPUT page)

Χτίσιμο ενός user browsing graph

- Κάθε κόμβος στο γράφημα αναπαριστά ένα URL στα user behavior data, και συσχετίζεται με:
  - reset probability, και
  - staying time

  ως μεταδεδομένα
- Κάθε κατευθυνόμενη ακμή αναπαριστά μια μετάβαση μεταξύ δυο κόμβων, και συσχετίζεται με τον αριθμό των μεταβάσεων που αποτελεί το βάρος της
- Ο user browsing graph είναι ένα γράφημα με βάρη στις ακμές που οι κόμβοι του περιέχουν μεταδεδομένα

Υποθέσεις

- Ανεξαρτησία χρηστών και sessions
  - Οι διαδικασίες browsing διαφορετικών χρηστών σε διαφορετικές sessions είναι ανεξάρτητες. Με άλλα λόγια, θεωρούμε το web browsing ως μια στοχαστική διαδικασία, με τα παρατηρούμενα δεδομένα σε κάθε session του κάθε χρήστη να είναι ένα I.I.D. δείγμα από αυτήν την διαδικασία
- Ιδιότητα του Markov
  - Η επόμενη σελίδα που επιλέγει να επισκεφτεί κάποιος χρήστης εξαρτάται μόνο από την τρέχουσα σελίδα, και είναι ανεξάρτητη από τις σελίδες που επισκέφτηκε προηγουμένως
  - Αυτή η υπόθεση είναι επίσης βασική στον PageRank
- Time-homogeneity
  - Οι συμπεριφορές browsing των χρηστών (π.χ., μεταβάσεις και staying time) δεν εξαρτώνται από τον χρόνο. Παρόλο που αυτή η υπόθεση δεν είναι απαραιτήτως αληθής στην πράξη, την υιοθετούμε για τεχνικούς λόγους
  - Αυτή η υπόθεση είναι επίσης βασική στον PageRank

Το continuous-time Markov μοντέλο

- Έστω ένας Web surfer που περιηγείται σε όλες τις Webpages
- Έστω ότι X_s είναι η σελίδα την οποία επισκέπτεται ο surfer την χρονική στιγμή s, s>0
- Τότε, με τις τρεις υποθέσεις, η διαδικασία $X = \{X_s, s \geq 0\}$ σχηματίζει μια continuous-time time-homogenous Markov process
- Έστω ότι $p_{ij}(t)$ είναι η transition probability από την σελίδα i στην j για το χρονικό διάστημα t σε αυτήν την διαδικασία
- Μπορεί ν' αποδειχτεί ότι υπάρχει μια stationary probability distribution **π**, η οποία είναι μοναδική και ανεξάρτητη του t, και συσχετίζεται με την $P(t) = [p_{ij}(t)]_{NXN}$, τέτοια ώστε για οποιονδήποτε $t > 0$

$$\pi = \pi P(t)$$

- Το i^{th} κελί της κατανομής **π** είναι το κλάσμα του χρόνου που ο surfer περνά στην i^{th} σελίδα προς τον χρόνο που περνά σε όλες τις σελίδες όταν το χρονικό διάστημα t τείνει στο άπειρο
- Με αυτήν την λογική, η κατναομή **π** μπορεί ν' αποτελέσει μια μετρική της page importance

Μηχανισμός

- Για να υπολογίσουμε αυτήν την stationary probability distribution, χρειάζεται να εκτιμήσουμε την πιθανότητα κάθε κελιού του matrix P(t)
- Στην πράξη, είναι δύσκολο να έχουμε αυτόν τον πίνακα, επειδή είναι δύσκολο να πάρουμε την πληροφορίας για όλα τα πιθανά χρονικά διαστήματα
- Για να επιλύσουμε αυτό το πρόβλημα, προτείνεται ένας νέος λαόγριθμος που βασίζεται στον *transition rate matrix*
- Ο transition rate matrix ορίζεται ως η παράγωγος της P(t) όταν t τείνει στο 0, εάν υπάρχει

$$\mathbf{Q = P'(0)}$$

- Αποκαλούμε τον πίνακα $Q = (q_{ij})_{NXN}$ ως ο *Q-matrix*

Ο Q-πίνακας

- Όταν ο χώρος καταστάσεων είναι πεπερασμένος, υπάρχει μια ένα-προς-ένα αντιστοιχία μεταξύ του Q-πίνακα και του P(t), και ισχύει $-INF < q_{ii} < +INF$ και $SUM_j\ q_{ij} = 0$
- Εξαιτίας αυτής της αντιστοιχίας, μπορούμε να χρησιμοποιήσουμε την Q-Process για να αναπαραστήσουμε την αρχική continuous-time Markov process, δηλαδή, η browsing process $X = \{X_s, s \geq 0\}$ που ορίστηκε προηγουμένως είναι μια Q-Process εξαιτίας του πεπερασμένου χώρου καταστάσεων
- Τα πλεονεκτήματα της χρήσης του Q-πίνακα
  - Οι παράμετροι του Q-matrix μπορούν να εκτιμηθούν από τα δεδομένα
  - Βασιζόμενοι στον Q-matrix, υπάρχει ένα αποδοτικός τρόπος για να υπολογίσουμε την stationary probability distribution του P(t)
- Η αποκαλούμενη EMC είναι μια discrete-time Markov process που έχει πίνακα πιθανοτήτων μεταβάσεων με μηδενικά σε όλες τις θέσεις της διαγωνίου, και $-q_{ij}/q_{ii}$ στις θέσεις εκτός της διαγωνίου

Το βασικό θεώρημα

THEOREM 1. *Suppose X is a Q-process, and Y is the Embedded Markov Chain derived from its Q-matrix. Let* $\pi = (\pi_1, \ldots, \pi_N)$ *and* $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_N)$ *denote the stationary probability distributions of the process X and Y, then we have*

$$\pi_i = \frac{\frac{\tilde{\pi}_i}{q_{ii}}}{\sum_{j=1}^{N} \frac{\tilde{\pi}_j}{q_{jj}}} \qquad (2)$$

- Η διαδικασία Y είναι μια discrete-time Markov chain, και έτσι η stationary probability distribution $\tilde{\boldsymbol{\pi}}$ μπορεί να υπολογιστεί από απλές μεθόδους, π.χ., την power method
- Κατόπιν, θα εξηγήσουμε πώς να εκτιμήσουμε τις παραμέτρους στον Q-πίνακα, ή ισοδύναμα, τις παραμέτρους q_{ii} και τις transition probabilities $-q_{ij}/q_{ii}$ ($-q_{ij}/q_{ii} > 0$, αφού $q_{ii}<0$)

Εκτίμηση των q_{ii}

- Για μια Q-Process, ο staying time T_i πάνω στον i^{th} κόμβο καθορίζεται από μια exponential distribution με παραμέτρους q_{ii}:

$$P(Ti > t) = \exp(q_{ii}\ t)$$

- Αυτό υπονοεί ότι μπορούμε να εκτιμήσουμε τα q_{ii} από μεγάλους αριθμούς παρατηρήσεων του staying time στα the user behavior data
- Αυτή η εργασία δεν είναι απλή, επειδή οι παρατηρήσεις των user behavior data συνήθως περιέχουν θόρυβο εξαιτίας της ταχύτητας σύνδεσης του Internet, μέγεθος page, δομή page, και άλλων παραμέτρων, δηλαδή, οι παρατηρούμενες τιμές δεν ικανοποιούν την exponential distribution
- Υποθέτουμε ότι η Z είναι συνδυασμός του πραγματικού staying time T_i και του θορύβου U, δηλαδή: $Z = U + T_i$

Εκτίμηση της Transition Probability στην EMC

- Οι πιθανότητες μετάβασης στην EMC περιγράφουν τις ‘καθαρές’ μεταβάσεις του surfer πάνω στον user browsing graph
- Η εκτίμηση αυτών μπορεί να βασιστεί στις παρατηρημένες μεταβάσεις μεταξύ σελίδων στα user behavior data
- Χρησιμοποιούμε την ακόλουθη μέθοδο για την εκτίμηση

Εκτίμηση της Transition Probability στην EMC

We start with the user browsing graph $G =< V, W, T, \sigma >$. We then add a pseudo-vertex (the $(N+1)^{th}$ vertex) to G, and add two types of edges: the edges from the last page in each session to the pseudo-vertex, associated with the click number of the last page as its weight; and the edges from the pseudo-vertex to the first page in each session, associated with the reset probability. We denote the new graph as $\tilde{G} =< \tilde{V}, \tilde{W}, T, \tilde{\sigma} >$, where $|\tilde{V}| = N + 1$, $\tilde{\sigma} =< \tilde{\sigma}_1, \ldots, \tilde{\sigma}_N, 0 >$. Then we explain the EMC model as the random walk on this new graph $\tilde{G}$. Based on *the law of large number* [19], the transition probabilities in the EMC are estimated as below,

$$-\frac{q_{ij}}{q_{ii}} = \begin{cases} \alpha \frac{\tilde{w}_{ij}}{\sum_{k=1}^{N+1} \tilde{w}_{ik}} + (1-\alpha)\sigma_j, & i \in V, j \in \tilde{V} \\ \sigma_j, & i = N+1, j \in V \end{cases} \quad (8)$$

Εκτίμηση της Transition Probability στην EMC

- Η διαισθητική ερμηνεία της μετάβασης έχει ως εξής:
  - Όταν ο surfer περιηγείται πάνω στον user browsing graph, μπορεί να ακολουθήσει έναν σύνδεσμο με ποιθανότητα α, ή να επιλέξει να ξεκινήσει από μια νέα σελίδα με πιθανότητα (1-α)
  - Η επιλογή της νέας σελίδας καθορίζεται από την reset probability
  - Πλεονεκτήματα της χρήσης της Εξίσωσης (8) για την εκτίμηση
    - αυτή η εκτίμηση δεν θα είναι πολωμένη λόγω του περιορισμένου αριθμού των παρατηρημένων μεταβάσεων
    - η αντίστοιχη EMC είναι πρωτογενής, και συνεπώς έχει μια μοναδική stationary distribution
  - Επομένως, μπορούμε να χρησιμοποιήσουμε την power method για να υπολογίσουμε την stationary distribution με αποδοτικό τρόπο

Ο αλγόριθμος BrowseRank

Input: the user behavior data.
Output: the page importance score π
Algorithm:

1. Construct the user browsing graph (see Section 3.1).
2. Estimate q_{ii} for all pages(see Section 3.3.2).
3. Estimate the transition probability matrix of the EMC and then get its stationary probability distribution by means of power method (see Section 3.3.3).
4. Compute the stationary probability distribution of the Q-process by using of equation (2).

Top-20 Websites από τους 3 αλγορίθμους

No.	PageRank	TrustRank	BrowseRank
1	adobe.com	adobe.com	***myspace.com***
2	passport.com	yahoo.com	msn.com
3	msn.com	google.com	yahoo.com
4	microsoft.com	msn.com	***youtube.com***
5	yahoo.com	microsoft.com	live.com
6	google.com	passport.net	***facebook.com***
7	mapquest.com	ufindus.com	google.com
8	miibeian.gov.cn	***sourceforge.net***	ebay.com
9	w3.org	***myspace.com***	***hi5.com***
10	godaddy.com	***wikipedia.org***	***bebo.com***
11	statcounter.com	phpbb.com	***orkut.com***
12	apple.com	yahoo.co.jp	aol.com
13	live.com	ebay.com	***friendster.com***
14	xbox.com	nifty.com	***craigslist.org***
15	passport.com	mapquest.com	google.co.th
16	***sourceforge.net***	cafepress.com	microsoft.com
17	amazon.com	apple.com	***comcast.net***
18	paypal.com	infoseek.co.jp	***wikipedia.org***
19	aol.com	miibeian.gov.cn	***pogo.com***
20	***blogger.com***	***youtube.com***	***photobucket.com***

Αποτελέσματα-1

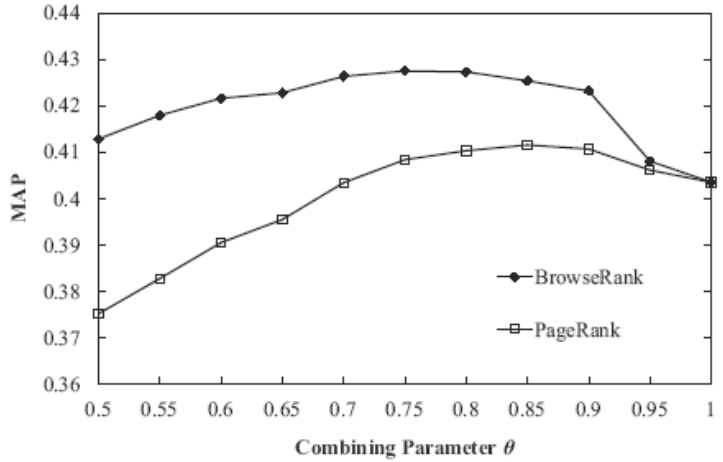


Figure 3: Search performance in terms of MAP for BrowseRank and PageRank

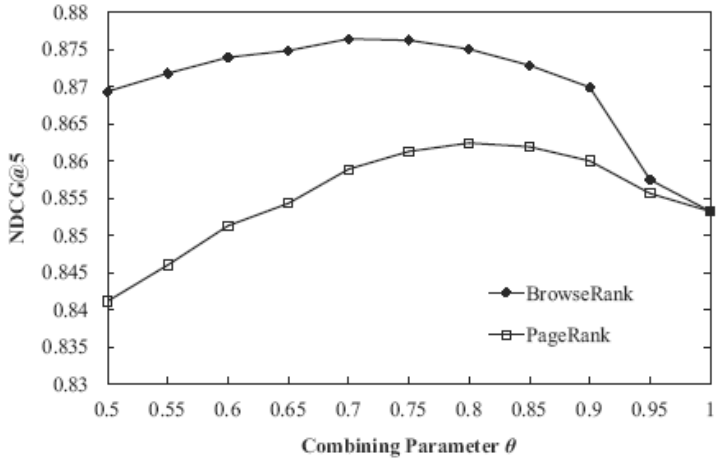


Figure 7: Search performance in terms of NDCG@5 for BrowseRank and PageRank

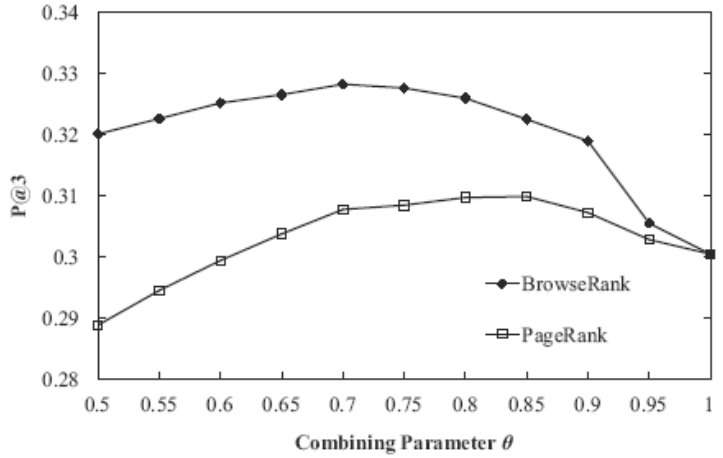


Figure 4: Search performance in terms of P@3 for BrowseRank and PageRank

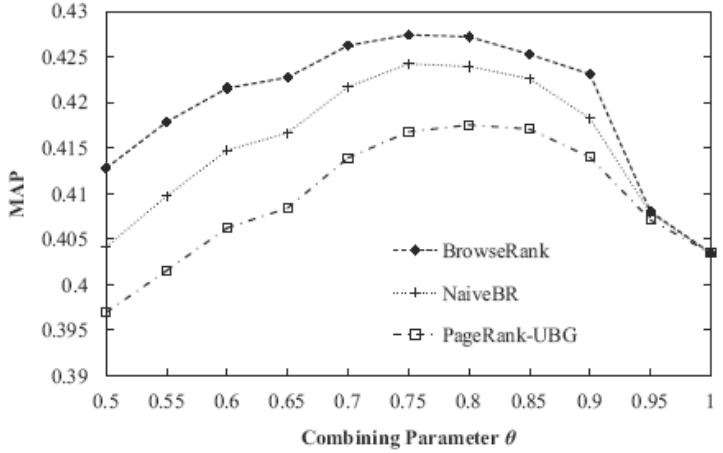


Figure 8: Search performance in terms of MAP for BrowseRank and two simple algorithms

Αποτελέσματα-2

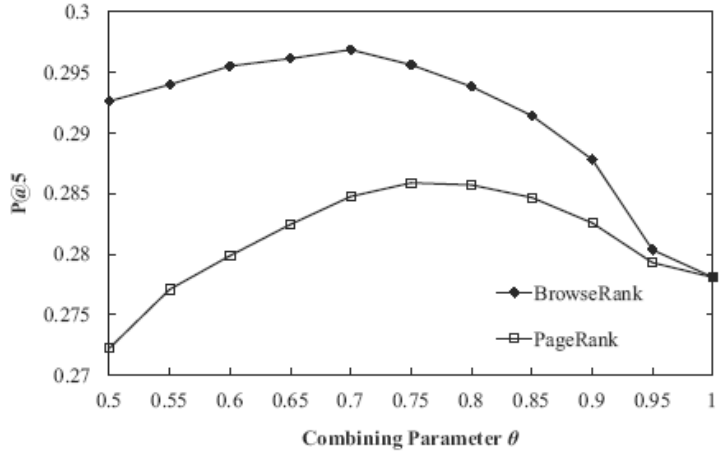


Figure 5: Search performance in terms of P@5 for BrowseRank and PageRank

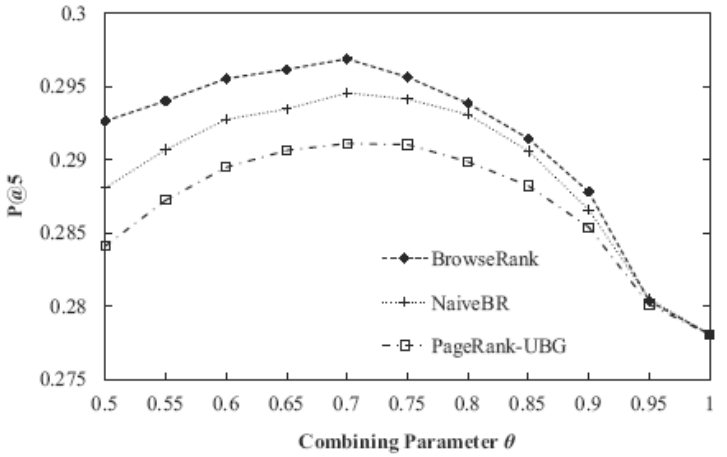


Figure 9: Search performance in terms of P@5 for BrowseRank and two simple algorithms

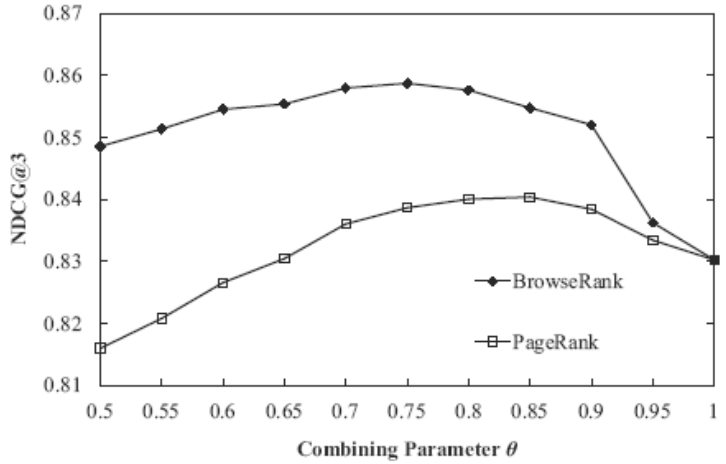


Figure 6: Search performance in terms of NDCG@3 for BrowseRank and PageRank

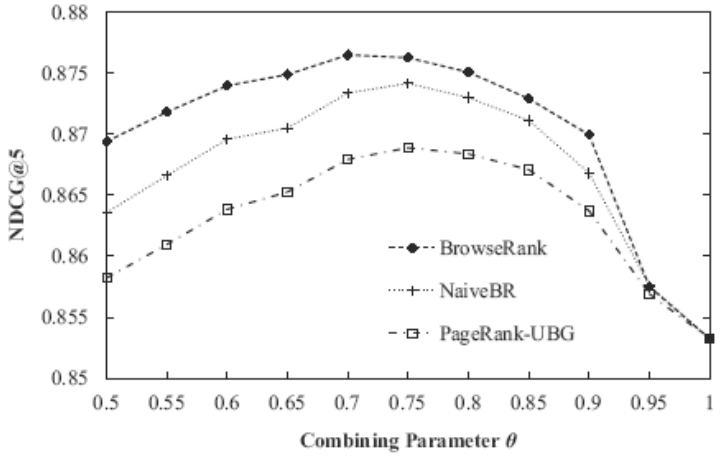


Figure 10: Search performance in terms of NDCG@5 for BrowseRank and two simple algorithms