Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Selective Approach To Handling Topic Oriented Tasks On The World Wide Web

Amit C. Awekar
North Carolina State University
Raleigh, NC 27695, USA
Email: acawekar@ncsu.edu

Jaewoo Kang
Korea University
Seoul, Korea
Email: kangj@korea.ac.kr

Abstract— **We address the problem of handling topic oriented tasks on the World Wide Web. Our aim is to find most relevant and important pages for broad-topic queries while searching in a small set of candidate pages. We present a link analysis based algorithm SelHITS which is an improvement over Kleinberg's HITS algorithm. We introduce concept of virtual links to exploit latent information in the hyperlinked environment. Selective expansion of the root set and novel ranking strategy are the distinguishing features of our approach. Selective expansion method avoids topic drift and provides results consistent with only one interpretation of the query. Experimental evaluation and user feedback show that our algorithm indeed distills the most relevant and important pages for broad-topic queries. Trends in user feedback suggests that there exists a uniform notion of quality of search results within users.**

I. INTRODUCTION

Searching information on the World Wide Web (www) is now a common practice. Web surfers fire queries to search engines hoping that they will find what they want. But distributed, heterogeneous and unstructured nature of the www poses many new challenges in searching the right information. Hence traditional information retrieval techniques are not sufficient for dealing with the www [1]. Constantly increasing size, availability and dependency on the www makes the problem even more difficult and important. The www is a networked information space where web pages can be considered as information entities and hyperlinks as association between them [2]. The underlying link structure of the www can be treated as meta data and has tremendous latent information encoded in it. Algorithms and systems that extract and use this latent information have achieved a major success in the past decade [3], [4]. But still we are far from getting completely satisfying answer.

Not all the queries given by users are same. User queries can be classified in two main categories, *specific queries* and *broad-topic queries* [5]. There are very few pages that are relevant for a specific query. Thus specific query has scarcity problem. "Is there any coffee shop on the centennial campus of NCSU?" is an example of a specific query. There will be handful of pages which will help us answer this query. But for broad-topic query we can have thousands or millions of relevant pages. Then the task of the search algorithm is to distill most important pages from this set. Problem with broad-topic query is not scarcity, but abundance. "Information about Data Mining" is an example of broad-topic query. We will restrict our attention to broad-topic queries and solving the abundance problem.

It can be observed that users generally give very short queries to search engines. Query like "Java" can be interpreted as "java programming language" or "java island" or in many other ways. Result set from a search engine might contain a mix of pages relevant to different interpretations of such ambiguous query. At a time user will be interested in single interpretation of the query. It will be very convenient for user if search engine can provide results corresponding to only one interpretation of the query. If user is not satisfied with this result set then search engine should provide another result set corresponding to different interpretation of the query. For example, given query "Java", search engine should provide a set of results relevant to only one interpretation of the query, say "java programming language". If user is not satisfied with this result set then search engine should provide results corresponding to "java island". We will deal with such ambiguous broad-topic queries and aim to provide results corresponding to only one interpretation of the query.

Kleinberg proposed Hyperlink Induced Topic Search (HITS) algorithm to tackle these broad-topic queries [5]. It starts with a focused subgraph of the www for given query, using some existing search system. This initial set of pages is called the root set. Then it adds pages from one link neighborhood of the root set to form base set. Based on page to page connectivity in the base set, iterative, eigenvector based calculation is performed to rank pages in two different categories, namely hubs and authorities. A good authority page should contain authoritative information on the query topic. A good hub should contain links to authoritative pages. There are some weaknesses in this approach which inspire us to propose improvements. HITS considers only page to page connectivity to rank the pages. It ignores other latent information in the hyperlinked environment. Many pages added to the root set during expansion step may not be relevant to the query. Such noise substantially degrades the quality of results. In short our work is an improvement over HITS algorithm based on two features.

- Novel ranking strategy which uses the concept of virtual links
- Selective expansion of the root set

Fig. 1. HITS Algorithm

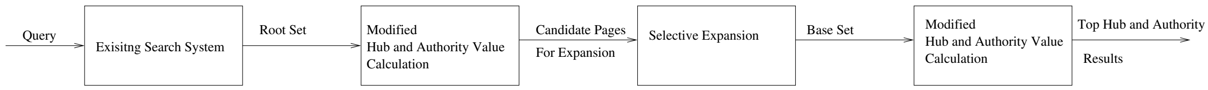

Fig. 2. SelHITS Algorithm

We do not get involved in content analysis of the pages. Rather content analysis based methods suggested by Bharat [6] and Chakrabarti [7] can be used in conjunction with our algorithm.

We discuss HITS algorithm and related work in Section 2. Our approach is explained in Section 3. Experiments and user feedback is discussed in Section 4. We conclude and discuss future work in Section 5.

II. HITS Algorithm

HITS algorithm [5] is briefly outlined in Figure 1. It associates two scores with each page. *Authority* which is a measure of authoritative information contained in the page and *Hub* which is a measure of links to good authorities. Using some existing search system, HITS algorithm gets a set of relevant pages for user query. This is referred to as the root set. Then all the pages from one link neighborhood of the root set are added to the root set. This is referred to as the base set. Consider a page P_i. Let Par_i be the set of pages which have hyperlink to P_i and are present in the base set. Let Chi_i be the set of pages that P_i has hyperlink to and are present in the base set. Let hub value for P_i be denoted as H_i and authority value as A_i. Then $A_i = \sum_{l \in Par_i} H_l$ and $H_i = \sum_{l \in Chi_i} A_l$. If E is the adjacency matrix for the base set, then authority vector V_a is calculated as $V_a = E^T E V_a$. The vector V_a converges to the principal eigenvector of $E^T E$. The matrix $E^T E$ is real, symmetric and non-negative. Hence its principal eigenvector is real and non-negative. Hub vector V_h is calculated as $V_h = E V_a$.

HITS is further augmented by Bharat [6] and Chakrabarti [7] by combining link analysis with content analysis of pages. Stochastic Algorithm for Link analysis (SALSA) [8] combines HITS with Page Rank [9]. Ng [10] combined random surfer model with HITS to achieve better stability. A comprehensive survey of various modifications proposed for HITS and other eigenvector based algorithms for web information retrieval can be found in [11].

III. SelHITS Algorithm

Selective Hypertext Induced Topic Search (SelHITS) is briefly outlined in Figure 2. It is motivated by following observations.

- If multiple pages from the same host are present in the root set then it indicates that there exists a community of pages on that host which is relevant to the query topic. This context of location of pages is ignored by HITS. Ranking strategy should exploit such latent information in the hyperlinked environment.
- All the pages in the one link neighborhood might not be relevant to the query topic. Such noise will substantially degrade the quality of results. While adding pages from one link neighborhood we should have a selection procedure to avoid irrelevant pages.

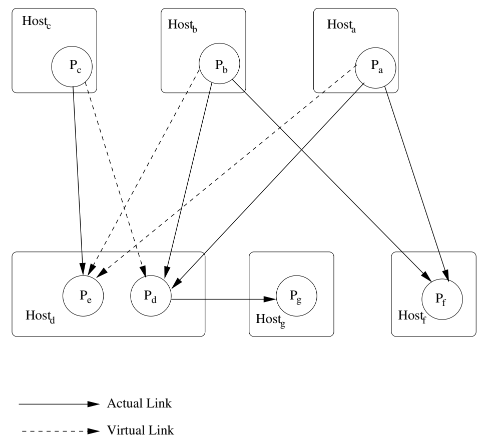

Fig. 3. Example Root Set

Now we will define few terms along with an example (Refer Figure 3). Our root set contains 7 pages P_a, P_b, P_c, P_d, P_e, P_f and P_g. Let P_a and P_b have hyperlinks to both P_d and P_f. Also, P_c has hyperlink to P_e whereas P_d has hyperlink to P_g. Pages P_d and P_e reside on same host. Rest all pages reside on different hosts. Let $Host_p$ indicate the host on which the page p resides. Thus we have $Host_d$ = $Host_e$.

Actual Links are those hyperlinks which are actually present between different pages. Thus link $P_a \rightarrow P_d$ is an actual link.

Virtual Links are the pseudo links that we hypothetically insert. If a page P_i has actual link to page P_j in the root set then we insert virtual links from P_i to all other pages in the root set which reside on the same host as that of P_j. Thus $P_a \rightarrow P_e$ is a virtual link.

Adjacency Matrix (E) represents connectivity between pages. If P_i has hyperlink to P_j then we will have $E[i,j] = 1$, otherwise $E[i,j]$ = 0. While building the adjacency matrix, we

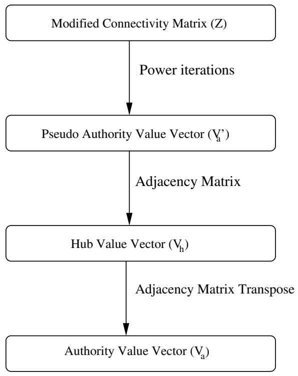

Fig. 4. SelHITS Ranking Strategy

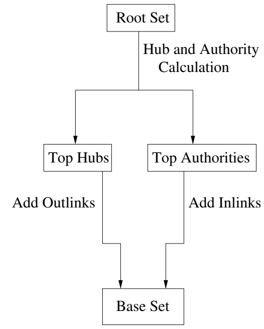

Fig. 5. Selective Expansion Procedure

consider only actual links. E for our example is

	P_a	P_b	P_c	P_d	P_e	P_f	P_g
P_a	0	0	0	1	0	1	0
P_b	0	0	0	1	0	1	0
P_c	0	0	0	0	1	0	0
P_d	0	0	0	0	0	0	1
P_e	0	0	0	0	0	0	0
P_f	0	0	0	0	0	0	0
P_g	0	0	0	0	0	0	0

Modified Adjacency Matrix (Z) is built considering actual links as well as virtual links. $Z[i,\ j]$= 1 if either one of the following holds

- P_i has actual link to P_j
- $\exists q\ for\ which\ (P_i \rightarrow P_q) \wedge (Host_j = Host_q)$

Otherwise $Z[i, j] = 0$. For our example Z is

	P_a	P_b	P_c	P_d	P_e	P_f	P_g
P_a	0	0	0	1	1	1	0
P_b	0	0	0	1	1	1	0
P_c	0	0	0	1	1	0	0
P_d	0	0	0	0	0	0	1
P_e	0	0	0	0	0	0	0
P_f	0	0	0	0	0	0	0
P_g	0	0	0	0	0	0	0

Note that while constructing matrices E and Z, hyperlinks between pages on same host are ignored as these links are considered to be nepotistic [5].

A. Novel Ranking Strategy

If we apply HITS algorithm for our example root set then both P_d and P_f will be rated equally as authority. But multiple pages from the $Host_d$ are present in the root set. It indicates that there exists a community of pages on $Host_d$ about the query topic. We should boost authority scores for the pages in this community and hub scores for the pages having hyperlinks to this community. P_d and P_c will be rated equally as hubs by HITS algorithm. But P_c has hyperlink to page on $Host_d$ which has multiple pages related to query. Considering this context of location, ranking strategy should rate P_d as a better authority than P_f, but it should rate P_c as better hub than P_d. Using virtual links while ranking helps us achieve this goal.

Our ranking strategy is depicted in Figure 4. First we calculate pseudo authority vector ($V_a^{'}$) using modified adjacency matrix Z. It can be represented as

$$V_a^{'} = Z^T Z V_a^{'} \quad (1)$$

$V_a^{'}$ will converge to principal eigenvector of $Z^T Z$ and can be calculated using power iteration method. Note that, with modified adjacency matrix all the pages on same host have same set of inlinks to them. So all pages on same host have same authority value. That is why these are pseudo authority values. In our example, P_d and P_e are on the same host. So they will have same virtual authority value. But now P_d is rated better authority than P_f. Then hub vector is calculated as

$$V_h = E V_a^{'} \quad (2)$$

We consider only actual links while calculating hub score. Now P_c will be rated as better hub than P_d. But still we need to differentiate between authority values of pages on same host. We calculate actual authority values as

$$V_a = E^T V_h \quad (3)$$

This step will differentiate between authority scores of P_d and P_e. We run this ranking strategy twice, first on the root set and then on the base set as illustrated in Figure 2.

B. Selective Expansion

Selective expansion procedure is depicted in Figure 5. All the pages in one link neighborhood of the root set might not be relevant to the query. We have to expand root set selectively so that base set has smaller size and contains only relevant pages. First we calculate hub and authority values for pages in the root set using ranking strategy described in Section III-A. Then we select top few hubs and authorities as candidates for expansion. We consider following factors while adding pages to the root set

- As per definition, hubs should point to good authorities [5]. So target pages of hyperlinks from top hubs in the root set can be good authorities. So we add outlinks from top hubs to the root set.

- As per definition, authorities are pointed by good hubs [5]. So pages having hyperlinks to top authorities in the root set can be good hubs. So we add inlinks to top authorities to the root set.

Generally root set size is of the order of few hundreds. In our experiments we chose top 20 hubs and authorities as candidates for expansion. We are expanding less than 10% of the pages in the root set.As compared to HITS, this selective expansion procedure drastically reduces the base set size and avoids topic drift as irrelevant pages are not added to the base set.

Algorithm 1 SelHITS

Input: User query (u)

Output: Top p hub pages and q authority pages

Begin

1) Get root set from some existing search system for user query (u).
2) Build adjacency matrix (E) and modified adjacency matrix (Z) for the root set.
3) Calculate hub and authority value vector for all pages in the root set using following equations

$$V_a^{'} = Z^T Z V_a^{'}$$

(using power iterations)

$$V_h = E V_a^{'}$$
$$V_a = E^T V_h$$

4) Select top *n* hubs and top *m* authorities.
5) Build base set by adding pages to the root set using selective expansion.
   - Add outlinks of top *n* hub pages with maximum limit Chi_{max} outlinks per page.
   - Add inlinks of top *m* authority pages with maximum limit Par_{max} inlinks per page.
6) Build adjacency matrix (E) and modified adjacency matrix (Z) for the base set.
7) Calculate hub and authority value vectors for the base set using following equations

$$V_a^{'} = Z^T Z V_a^{'}$$

(using power iterations)

$$V_h = E V_a^{'}$$
$$V_a = E^T V_h$$

8) Report top p hubs and top q authorities.

End

C. Ambiguous Queries

If a query is ambiguous then we have corresponding disjoint communities in the link structure [5]. Top hubs and authorities are from the same community as hub and authority values are mutually reinforcing. Selective expansion procedure adds new pages from single community, which further boosts hub and authority values for that community. Hence we are able to give results consistent with only one interpretation of the query. If user is interested in other interpretation of the query then we can simply remove current community from candidate pages and again run our algorithm.

TABLE I
INTERPRETATION OF AMBIGUOUS QUERIES BY SELHITS

Query	Interpretation
$Windows$	Windows Operating System
$Jaguar$	Jaguar Cars
$Socks$	Socks as a Garment
$Gates$	Bill Gates

D. Algorithm

Refer to SelHITS algorithm (Algorithm 1). The values of parameters n, m, p, q have to be tuned. Parameters p and q can be set as per user preference, while n and m should be set depending on how selective one wants to be. Lesser the value of n and m, more selective we are. We conducted experiments with n, m, p and q, all values set to 20. Chi_{max} and Par_{max} control influence of an individual page. We did not put any restriction on Chi_{max} as generally number of hyperlinks from a page are not many. We set Par_{max} to 100.

SelHITS algorithm differs from HITS in terms of selectively expanding the root set and ranking strategy with virtual links. We are changing the order of adjacency matrix and number of non-zero elements in the adjacency matrix. For both the root set and the base set the matrix $Z^T Z$ is still symmetric, positive semidefinite and non-negative. Hence the same proof of convergence and correctness as that of HITS [5] applies to SelHITS. Brief poster version of SelHITS algorithm appeared in [12].

IV. EXPERIMENTS AND RESULTS

We got the root set using Google API [13]. For getting backlinks we observed that results from Google API were not satisfactory and many backlinks were missing. So we used Yahoo! API [14] to get backlinks.

We tested our algorithm for nine different queries. Queries like "*Search Engine*", "*Gates*", "*Jaguar*" were chosen as Kleinberg has evaluated HITS against these queries [5]. We had some queries like "*Abortion*" and "*Iraq War*" for which there are strongly opposing views in the society. For *Abortion*, SelHITS results were a mix of pro-abortion and anti-abortion pages. For *Iraq War*, SelHITS result were mostly anti-war pages and had many results from news as it is a hot topic. We also had ambiguous queries like "*Windows*", "*Jaguar*", "*Socks*"and "*Gates*". Table I lists ambiguous queries and their interpretation by SelHITS. We were able to give results consistent with only one interpretation of ambiguous queries. There was only one exception. All the results for query "*Socks*" were related to "*Socks as a garment*". But there was one result page in top authority list which was about "*Socks as a protocol*". The reason for this exception is that the page might be a so strong authority in in its community that even after boosting authority values for pages of other community,

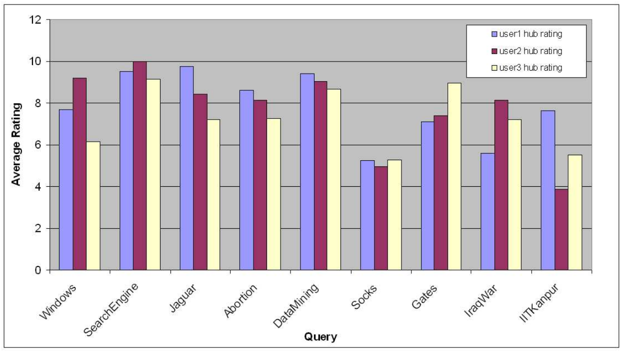

Fig. 6. Average Hub Ratings by Different Users

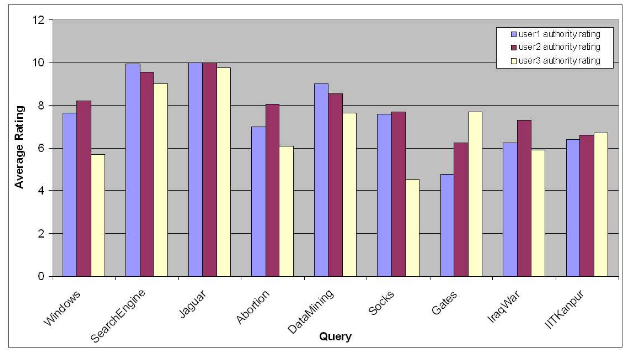

Fig. 7. Average Authority Ratings by Different Users

TABLE II
NUMBER OF HUBS AND AUTHORITIES FOUND THROUGH SELECTIVE EXPANSION IN TOP 20 HUB AND TOP 20 AUTHORITY RESULTS

No.	Query	Hub	Authority
1	*Windows*	19	1
2	*Search Engine*	19	3
3	*Jaguar*	19	4
4	*Abortion*	18	9
5	*Data Mining*	18	7
6	*Socks*	10	14
7	*Gates*	20	7
8	*Iraq War*	20	13
9	*IIT Kanpur*	18	13

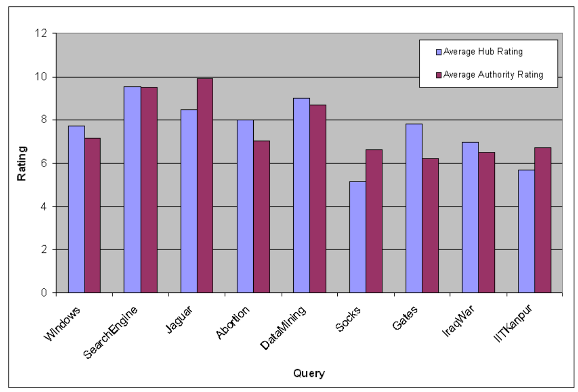

Fig. 8. Overall Average Ratings for Hub and Authorities

it still appears in top authority list. Queries "*Data Mining*" and "*IIT Kanpur*" were chosen because users giving feedback were familiar with these topics. Sample result URLs can be found in Appendix. More details about the experiments are available on author's home page [15].

For each query we generated top 20 hubs and top 20 authorities by SelHITS algorithm. For each query, these 40 URLs were evaluated by three different users on the scale of zero to ten. Zero being worst and ten being best rating. We had different set of users for evaluating each query. First we calculated average of hub ratings and authority ratings per query, per user. Refer Figure 6 and Figure 7. Then we calculated average hub ratings and authority ratings per query, by taking arithmetic mean of average rating by individual users. Refer Figure 8. Overall user feedback was satisfactory. Within a query, average scores for hubs and authority are similar (Figure 8). We also noticed that scores of different users for a query were also similar (Figure 6 and Figure 7). This uniformity in human evaluation indicates that there exists a uniform notion of quality of search results within users.

We evaluated effectiveness of selective expansion. Refer Table II. It shows out of top 20 hubs and top 20 authorities, how many pages are non-root set pages. That is, these pages were not present in the root set and we added them to the base set by selective expansion. Almost all the hubs were found through selective expansion. Also some good authorities were also from non-root set pages. It shows that selective expansion chooses good candidates for expansion.

V. CONCLUSION AND FUTURE WORK

We discussed SelHITS algorithm which is an improvement over HITS algorithm. Main conclusions of our work are as follows

- HITS algorithm does not use context of location of the page in hyperlinked environment. To overcome this problem we proposed novel ranking strategy using concept of virtual links that exploits the latent information in the hyperlinked environment.
- Simple one link neighborhood expansion adds many irrelevant pages to the base set and degrades the quality of results. To avoid this problem we proposed selective expansion method that drastically reduces the size of base set and improves the quality of results.
- For ambiguous queries we were able to give results consistent with only one interpretation of the query. It is also possible to give results consistent with other interpretation of the query.
- Feedback from real users shows that our algorithm indeed distills most important and relevant pages for broad-topic queries.
- Within users there exists a uniform notion of quality of search results.

Currently we are working on applying approach of SelHITS for clustering hypertext documents. We believe that SelHITS can also be useful to other topic oriented tasks like focused crawling. We can improve quality of results by blending SelHITS with content analysis based methods. Other possible direction is to experiment with different criteria for selective expansion. We can also explore different notions of virtual links. Currently we are using criteria of same hostname for inserting virtual links. But one can define virtual links based on other notion of being close like distance in the web graph.

ACKNOWLEDGMENT

We would like to thank Rupesh Mehta from Yahoo!, India and Dr. Pabitra Mitra from IIT, Kharagpur for useful discussion and comments.

REFERENCES

[1] M. R. Henzinger, "Web information retrieval - an algorithmic perspective," in *European Symposium on Algorithms*, 2000, pp. 1–8.
[2] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal, "The web as a graph," in *PODS '00: Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM Press, 2000, pp. 1–10.
[3] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.
[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.
[5] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5–7, pp. 604–632, 1999.
[6] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 1998, pp. 104–111.
[7] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 65–74, 1998.
[8] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," *Computer Networks (Amsterdam, Netherlands: 1999)*, vol. 33, no. 1–6, pp. 387–401, 2000.
[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Tech. Rep., 1998.
[10] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Stable algorithms for link analysis," in *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2001, pp. 258–266.
[11] A. Langville and C. Meyer, "A survey of eigenvector methods of web information retrieval," *SIAM Rev.*, vol. 47, no. 1, pp. 135–161, 2005.
[12] A. C. Awekar, P. Mitra, and J. Kang, "Selective hypertext induced topic search," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2006, pp. 1023–1024.
[13] Google code home page. [Online]. Available: http://code.google.com/
[14] Yahoo! developer network home page. [Online]. Available: http://developer.yahoo.com/
[15] "More details about experiments on SelHITS algorithm. `htttp://www4.ncsu.edu/~acawekar/`SelHITSResults/."

APPENDIX

Here we give top five authority and top five hub results for four queries. We cannot list all the result URLs here for lack of space. All result URLs and details of user feedback are available on author's home page [15].

Query "Search Engine"

Top Authorities

1. http://www.google.com/
2. http://searchenginewatch.com/
3. http://www.yahoo.com/
4. http://www.altavista.com/
5. http://www.alltheweb.com/

Top Hubs

1. http://www.psyche.com/psyche/links/reference.html
2. http://www.seobook.com/archives/2004_02.shtml
3. http://www.geocities.com/altmartinuk/search.html
4. http://www.landenart.com/links2.htm
5. http://www.andilinks.com/sea.htm

Query "Jaguar"

Top Authorities

1. http://www.jaguar.com/
2. http://www.jag-lovers.org/
3. http://www.jaguar.com/uk/
4. http://www.jagweb.com/
5. http://www.jaguar.ca/

Top Hubs

1. http://www.europeanautomotive.com/links/
2. http://directory.google.com/Top/Recreation/Autos/Makes_and_Models/Jaguar/
3. http://helpforcars.net/Makes/Jaguar/Main.htm
4. http://www.nationsonline.org/oneworld/united_kingdom.htm
5. http://wiccac.org/webscat_tot.html

Query Data Mining

Top Authorities

1. http://www.kdnuggets.com/
2. http://www.the-data-mine.com/
3. http://www.acm.org/sigkdd/
4. http://www.spss.com/
5. http://www.ics.uci.edu/ mlearn/MLRepository.html

Top Hubs

1. http://aima.cs.berkeley.edu/ai.html
2. http://www.andilinks.com/esw3.htm
3. http://www.dicarlolaw.com/MyFavoriteBookmarks.htm
4. http://www.medinformatics.org/mdi207/bookmarks.shtml
5. http://www.kimsoft.com/kimsoft.htm