

The Compass Filter: Search Engine Result Personalization Using Web Communities

Apostolos Kritikopoulos and Martha Sideri

Dept. of Computer Science,
Athens University of Economics and Business,
Patission 76, Athens, T.K.10434, Greece
apostolos@kritikopoulos.info
sideri@aueb.gr

Abstract. We propose a simple approach to search engine personalization based on Web communities [14]. User information –in particular, the Web communities whose neighborhoods the user has selected in the past– is used to change the order of the returned search results. We present experimental evidence suggesting that our method indeed improves the quality of the ranking. Our experiments were carried out on a search engine created by our research group and focusing on the Greek fragment of the worldwide Web (1.33 million documents); we also discuss the issue of scaling.

1 Introduction

The worldwide Web has unprecedented size and diversity –both in terms of the *documents* it contains, and in terms of the *users* who access it and depend on it. While search engine technology has advanced tremendously, the criteria used in evaluating the relevance of a document to a particular query do not typically take into account the *user who asked this query* (his/her degree of sophistication, interests and preferences, as evidenced, for example, by the order in which s/he selected the preferred URLs, the groups of URLs that s/he has visited in the past, etc). There are, of course, related domains, such as recommendation systems [4,5,21] and push channel technology [23], in which personalization based on the user’s declared or mined preferences is the supreme consideration. See also the next section for three recent approaches to personalization [2,15,17] based on PageRank [6,9,16,25].

Some of the most successful and elegant approaches to Web information retrieval are based on the realization of the importance of the link structure of the Web. In fact, two of the best known and most successful approaches to www information retrieval, Google’s page rank [6,9,25] and Kleinberg’s hubs and authorities [18] are in principle based exclusively of link structure.

The link structure of the Web has been, of course, the object of extensive study over the past five years [1,9,14,18,19,20]. One of the most interesting and intriguing observations in this study is the existence of abundant *Web communities* [14,20], that

is, small sets of documents that are highly connected, (in other words, small-scale, consensual hubs and authorities in a very specialized subject). The importance of the Web communities to the structure and nature of the worldwide Web has often been emphasized [3,13,22].

Web communities are dense directed bipartite subgraphs of the web graph. A bipartite graph is a graph whose node set can be partitioned into sets, F and C. Every edge in the graph is directed from a node u in F to a node v in C. A bipartite graph is *dense* if many of the $|F| \cdot |C|$ possible edges between F and C are present; it is *complete* (or a *bipartite clique*) if all such edges are present. Without mathematically pinning down density, we proceed with the following hypothesis, proposed in [14]: the dense bipartite graphs that are signatures of web communities contain at least one core, where a core is a complete bipartite subgraph with i nodes from F and j nodes from C. Thus, the community core is an $i \times j$ complete bipartite subgraph of the community, for some small integers i and j greater than one.

In this paper we present a novel approach to search engine result personalization based on Web communities. Our method (Figure 1) filters the results of the search engine to a query, based on its analysis of the frequency with which the user asking the query has in the past visited or selected the (neighborhoods of the) various Web communities in the corpus.

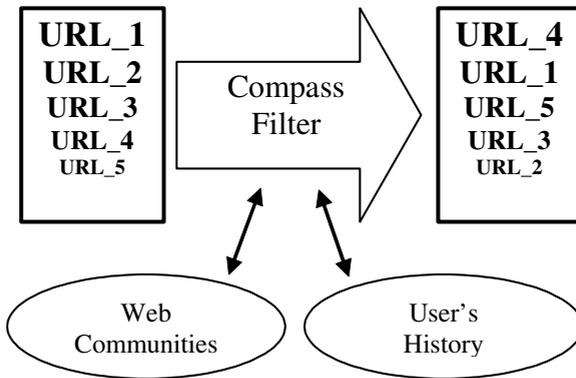


Fig. 1. Reordering the result set, using the Compass Filter

The main idea is as follows: We extract the Web communities (all $i \times j$ complete bipartite graphs in the corpus for all $i, j \geq 2$) and for each such “community core” we also determine its neighborhoods (the documents linked to, or from, documents in the community core; this is a little more general than the original proposal in [14]). When the search engine returns a set of documents in response to a query by the user, we re-evaluate these documents by taking into account the community neighborhoods in which they are involved (and exactly which part of the neighborhood they are involved), and the number of times these community neighborhoods have been visited or selected by the same user in the past. The results are then ordered in decreasing

values of this personalized measure of relevance (the original order used to break ties), and presented to the user.

For a hypothetical illustrative example, consider the query “duck”. The engine returns the following ranked Web pages, where the ranking depends only on the query terms and the corpus:

Table 1. First result set of the example

	<i>URL</i>	<i>MAIN THEME OF THE WEB SITE</i>
1	www.greek_natural_park.gr/crete/duck.htm	NATURAL PARK
2	www.gastronomy.gr/recipes/duck_potatoes.html	RECIPES
3	www.corfu_island.gr/local_animals/duck.html	ISLAND OF CORFU
4	www.ornithologic_home/duck_in_danger.htm	ECOLOGIC ORGANIZATION
5	www.hunter.gr/duck_spots.html	HUNTING
6	www.greek_encyclopedia/birds/duck.html	ENCYCLOPEDIA

From the history of the user, however, we know that the user had in the past clicked on a Web page (www.hunting_guns.gr/double-barrel/carbines.htm) that belongs to a 2x2 community (where the set F is are the url’s on the first column, and the set C on the second, and in the corpus there are links from both entries of the first column to both entries of the second, see Table 2). We also have found that the fifth page (highlighted in Table 1) belongs to the same hyperlinked community (Table 2).

We next re-rank the result set, adding appropriate weights to the various web pages in a manner explained in Section 3, so that the Web pages appearing in communities, such as the fifth Web page in this example, ends up higher in the order of the presented Web pages:

Table 2. 2x2 Community of the example

<i>2x2 COMMUNITY (main theme :HUNTING)</i>	
www.hunting_guns.gr/double-barrel/carbines.htm	www.bird_chasing.gr/venues.html
www.hunter.gr/duck_spots.html	www.hunting_laws.gr/guns/limitations.html

Table 3. Final result set of the example

	<i>URL</i>	<i>MAIN THEME OF THE WEB SITE</i>
5	www.hunter.gr/duck_spots.html	HUNTING
1	www.greek_natural_park.gr/crete/duck.htm	NATURAL PARK
2	www.gastronomy.gr/recipes/duck_potatoes.html	RECIPES
3	www.corfu_island.gr/local_animals/duck.html	ISLAND OF CORFU
4	www.ornithologic_home/duck_in_danger.htm	ECOLOGIC ORGANIZATION
6	www.greek_encyclopedia/birds/duck.html	ENCYCLOPEDIA

We implemented our method on top of *SpiderWave* [27], a research search engine for the Greek fragment of the Web (about 1.33 million documents, basically the .gr domain) designed by our research group, which can be clicked from the Web site of our University (www.aueb.gr) as an alternative search engine.

SpiderWave totally resides on the server-side, and it was extended to include the capability of tracking the individual user profile (search and navigation history). We call this implementation of our idea “The Compass Filter” (for community-pass). In this paper we present some experimental results to evaluate our method.

Whenever a query is asked, our experiment engine flips a fair coin to decide whether the answer will be filtered through Compass or not. In either case we monitor the user’s response (the results clicked, the order in which they were clicked, and the timing of the clicks—even though we do not use the latter data in our evaluation). We evaluate the user’s response by a formula that rewards early clicking on high-ranking results, and penalizes extra clicks. Comparison between the three suites (the one without the Compass Filter, the one that was processed successfully by Compass and the one that was processed unsuccessfully due to the fact that the user had not visited any relevant communities in the past), followed by a statistical test, suggests that, our method significantly improves the quality of the returned results.

The main limitation of our experiments has been the difficulty to have our system used by enough users long and intensively enough so that the Compass Filter can intervene meaningfully (we believe that these are problems small academic research groups are bound to face, and they do not limit by themselves the applicability of our method). From over 450 users in the period April 2002 to February 2003, only 18 interacted long enough with the system so our method made a difference in the ranking of the results, and they asked a total of 44 queries. Still, a statistical test (see Section 5) indicates that the Compass Filter improves the quality of the user experience in a statistically significant way.

In the next section we describe recent approaches to personalization, in Section 3 we describe our method, in Sections 4 and 5 the experiments and the results, and in Section 6 the research directions suggested by this work.

2 Recent Approaches to Personalization

Recently, several methods for the personalization of web search engines have been proposed; we briefly review these advances in this section.

The method of Topic-Sensitive PageRank [15] proposes to compute a set of PageRank vectors, in principle one per user, each biased by a set of representative topics (which are taken from Open Directory [24]). By using these precomputed vectors, Topic-Sensitive PageRank generates query-specific importance scores for pages at query time. This technique is modified in [17] so that it scales well with the corpus size and the number of users, and can thus be feasibly implemented. “Partial vectors” are shared across multiple personalized views, and their computation and storage costs scale well with the number of views; on the other hand, incremental computation allows the calculation of personalized rankings at query time.

The ranking proposed in [2] also derives from PageRank; the difference is that it takes into consideration user preferences based on URL features such as Internet domains. For instance, a user might favour pages from a specific geographic region, or pages with topical features also captured in Internet domains, or documents from domains such as academic institutions in which pages are more likely to be monitored by experts for accuracy and quality. Users specify interest profiles as binary feature vectors where a feature corresponds to a DNS tree node or node set, and the method pre-computes PageRank scores for each profile vector by assigning a weight to each URL based on the match between the URL and the profile features. A weighted PageRank vector is then computed based on URL weights, and used at query time to rank results.

Paper [3] presents and evaluates a novel ranking technique that combines the content of the objects being retrieved and the interest-based community of the user issuing the search. The theory of Bayesian belief networks is used as the unifying framework of the approach. Interest-based communities are groupings of users that share common interests. They are created using clickthrough data recorded in the form of user surfing behaviour. The method infers communities even for sources that do not explicitly show relationships between the pieces of information provided. The communities that are recognized are not necessarily based on the link information of the Web. Query contextualization is achieved by the juxtaposition of the current user interaction with a set of previous user interactions of all users in a way similar to collaborative filtering.

Other proposals for web search personalization in the recent literature include methods based on syntactic clustering of the Web [7], and on the recording of user preferences for meta-search engine personalization [8]. For two reviews on personalization see [10] and [11].

3 Description of the Method

Web communities are complete bipartite graphs of hyperlinks; the surprising prevalence of Web communities is an important and rather surprising property of the worldwide Web. For example, we shall see in the next section that in our crawl of .gr with 1,329,260 documents we found 1337 communities with a total 11,917 documents –roughly .9% of the crawl.

3.1 Step 1 (Preprocessing): Expand the Communities

We chose to “expand” the communities in a manner very similar with HITS [9,14,18]: we add to the set of the Web pages on either side of the original core community (S_{CG} – Core Group of community, see Figure 2), the group of pages that point to the core (S_{RG} – Reference Group of community), and the group of pages that are pointed to by any page in the core (S_{IG} – Index Group of community). From now on, we understand as “community” the union of the pages in S_{CG} , S_{RG} and S_{IG} . This way, the 11,917 Web pages of the core communities were expanded 30fold to 348,826, almost 32% of the corpus.

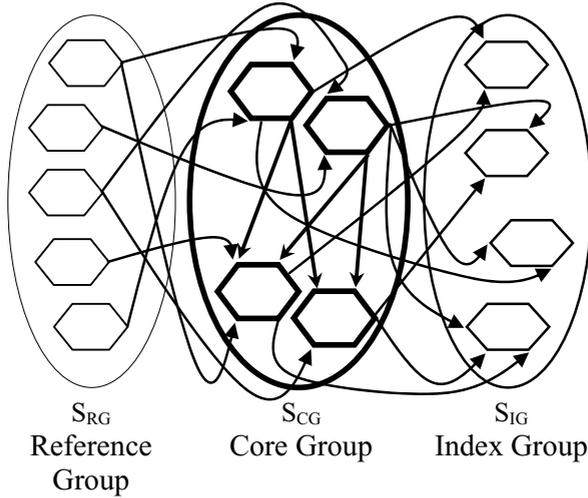


Fig. 2. Link graph of the community groups

3.2 Step 2: Calculate the Community Weights of the User

While a user clicks on query results, we monitor the core, reference, and index groups s/he visits, and we calculate, for each user and each community, the community weight of the user.

We noticed empirically that the influence on relevance of visits by the user to the core, index, and reference group of the same community decreases rapidly in this order. That is, if the user has visited the core group, everything else can be ignored, and if not the core but the index group, then visits to the reference group is not very significant unless they are extremely numerous. We capture this by the following formula, which appears (and is) rather arbitrary, but whose main point is that visits to S_{CG} are rewarded much more than those to S_{IG} , and those to S_{IG} much more than those to S_{RG} :

$$\begin{aligned}
 \text{COMMUNITY WEIGHT} = & (\text{Visits to the SRG Community}) + \\
 & (2 * \text{Visits SIG Community})^2 + \\
 & (3 * \text{visits to the SCG Community})^3
 \end{aligned}
 \tag{1}$$

3.3 Step 3: Reorder the Result Set

Given that the search engine has returned a ranked result set, we apply the outcome of the previous step, and we identify the URLs that belong to any of the expanded communities. The final weight of every URL is the sum of the weights (for the user) of each expanded community it belongs.

$$\begin{aligned}
 \text{Weight Of Url} = & \text{Sum of Weights of the Communities} \\
 & \text{to which it Belongs .}
 \end{aligned}
 \tag{2}$$

Finally, we use this to reorder the result set in decreasing weight, using the original order to break ties.

EXAMPLE: A user has searched for the word “Asimov.” In the original result set that the search engine produced, all documents from the Web site “www.altfactor.gr” (a leading Greek science fiction site) were ranked very low (31st place and below). Since www.altfactor.gr is part of a science fiction-based Web community, and the user has visited several sites that are referred to by sites of that community (even though s/he had not visited altfactor.gr itself), all pages from altfactor.gr were ranked highest by the Compass Filter.

4 Experimental Set-Up and Evaluation Metric

SpiderWave (<http://spiderwave.aueb.gr>) is a search engine research project whose aim is to determine the structure of the Greek Web (the .gr domain), and to use it as a test-bed for developing new ideas and methods of searching the Web. The crawl of the .gr domain was made with crawler software developed by a sister research group at the University of Patras [12]. The search engine is based on the ORACLE Intermedia Text processor (we also have implementation of HITS but we did not use it for this experiment). The result to every user’s query is a ranked group of Web pages.

We used a process similar to that described in [20] to extract the communities of the Greek Web. This process starts by extracting all $i \times j$ communities (of i fans and j centers) in which the fans have outdegree exactly j , and the centers have indegree exactly i . A fan of degree j (pointing to j centers) is part of an $i \times j$ community if we can find $i - 1$ other fans pointing to the same centers. For small values of i, j we can check this condition easily. After this first step, we enumerate all remaining communities iteratively as follows: for fixed j we find all vertices of outdegree at least j (all $1 \times j$ communities), then we find all $2 \times j$ communities by checking every fan which also cites any center in a $1 \times j$, then we compute all $3 \times j$ communities by checking every fan which cites any center in $2 \times j$, and so on.

The community extraction process traced 1337 communities having in total 11917 Web pages, with dimensions varying from 2×2 to 2×12 , and 8×2 to 8×8 . Following the first step of the method, we expanded the communities and finally concluded with 1337 expanded communities containing a total of 348826 Web pages. Independently of their use in personalization, these communities seemed to us quite informative: by studying them we discovered that they summarize the “sociology” of the Greek Web, focusing on such diverse topics as Stock Market, Greek music, University issues, Linux, automobiles, literature and movies.

For the experiment we set up an extra interface to our search engine. We asked users to use a login name, which is used to trace each user’s selection history. We explained that by doing so they participate in a search engine research project that will log their preferences, and will use them only for the purpose of improving their own search results. Anecdotal evidence tells us that the vast majority of users turned back at this point and selected the plain version. The history of each logged-in user (the weight of the user viz. all expanded communities) was updated with every selection

of a document (it follows from the numbers above that roughly one in three clicks resulted in an update). In our early implementation we did the expansion of the communities on-demand, but we now have a full list of the expanded communities for our crawl, and we update it periodically.

Whenever the user asked a query, with probability 50% (the user was unaware of the results of this flip, or even that a flip was taking place), the results were filtered through Compass. The returned results, an ordered set of documents, reordered by Compass or not, were presented to the user, who proceeded to click some of them. We recorded the documents clicked on, and the order in which they were clicked (as well as the timing of each click, even though we did not use it in our evaluation formula).

We then evaluated the user’s response using a metric we call SI (for Success Index), a number between 0 and 1:

$$SI = \frac{1}{n} \sum_{t=1}^n \frac{n - t + 1}{d_t * n} \tag{3}$$

where: **n** is the total number of the URLs selected by the user

d_t is the order in the list of the **t**-th URL selected by the user

The SI score rewards the clicking of high items early on. The reverse ranks of the items clicked are weight-averaged, with weights decreasing linearly from 1 down to 1/n with each click. For example, suppose n = 2 and the documents ranked 2 and 10 were clicked. If 2 is clicked first, then the SI score is bigger (27.5%); if second, smaller (17.5%). More controversially, SI penalizes many clicks; for example, the clicking order 2-1-3 has higher score than 1-2-3-4 (see the table below). Absence of clicks (the empty set) are scored zero –even though there were no such instances. Some examples of **d_t** sequences and their SI scores:

Table 4. Examples of the SI score

<i>Selection Order</i>	1	2	1	3	5	7	10	3	1	2
<i>SI score</i>	100%	42,59%			10,10%			38,88%		

Table 5. Examples of the SI score

<i>Selection Order</i>	1	2	3	4	4	3	2	1	5	8	7	2	1
<i>SI score</i>	40,10%				25%				15,71%				

5 Experimental Results

General

Time period of the Experiment: 23 April 2002 - 21 February 2003

Number of logged-in users: 460

Number of users for which the Compass Filter changed the order in a query: 18

Group A) Queries in the control (no Compass Filter) group

(Note: These queries were randomly selected not to be treated by Compass)

Number of queries:	508
Average SI score:	48.58%
Variance:	13.98%

Group B) Queries in the group processed unsuccessfully, because Compass had no community information

Number of queries:	476
Average SI score:	46.29%
Variance:	13.01%

Group C) Queries in the group processed successfully by Compass Filter

Number of queries:	44
Average SI score:	57.70%
Variance:	9.86%

For group A the coin flip determined that the answer not be filtered. For groups B and C, in contrast, the engine tried to filter the results, but succeeded only for group C. The Compass changed the order of the results only for group C; users that their queries belonged to group A or B, didn't had the chance to see the results ordered by the Compass Filter.

Table 6. t-Tests Results

t-Test Results	
Groups to compare:	P-value
A and B	16.48% (>>5%)
A and C	3.74% (<5%)
B and C	1.35% (<5%)

Submitting these results to the t-Test (one-tailed) statistical analysis method (see Table 6) tells us that the observed difference between the means is significant, supporting the conclusion that the results of group C are substantial better that the results of the other two groups, and that our method appears to significantly improve the quality of the retrieved information.

6 Discussion

We have proposed a method for using communities to personalize and therefore enhance Web information retrieval, and a metric on click sequences for evaluating user satisfaction. Our experimental results are quite encouraging. Much more *experimental evaluation* of our method, as well as *tuning of its parameters* (especially the calculation of weights), is needed. Our SI metric could also use more refinement and justification.

Our way of extending the communities (not unlike that in Kleinberg's algorithm [18] and HITS [9,14]) results in a wealth of documents, but is not the only possibility. For example, a more modest approach would only include the documents pointing to the authorities (centers) and pointed to by the hubs (fans) as in [20]; the quality and relevance of the resulting group may compensate for the loss of volume. This is worth experimenting with.

We developed and tested our method in the context of a very modest fragment of the Web. This scaled-down experimentation and prototyping may be an interesting methodology for quickly testing information retrieval ideas, and for expanding the realm of research groups, especially academic groups lacking strong industrial contacts, that are in a position to conduct search engine research.

But does our method scale to the whole Web? It is based on the fact that Web communities seem to be prevalent in the Greek Web. Ravi Kumar et al. [20] report 191629 communities in a Web with 200,000,000 documents, comprising a total of 3823783 documents belonging to a community, or 1.91% of the whole (compared to our .9%). The degree structure of the Greek Web is not too different from the Web's, and so a 30fold increase by extending the communities is plausible in the Web as well. Hence, the user's clicking history would again present ample community information. The other premise on which the success of our approach depends is that, in the Greek Web, the queries asked by a user are apparently quite often relevant to the communities visited by the same user in the past. How this phenomenon scales is much harder to predict.

Finally, a very challenging question (for this and many other approaches to Web information retrieval) is to develop a realistic mathematical *user model*, predicting on the basis of few parameters the user's needs, expectations and behaviour. Such a model would help evaluate and optimize novel approaches to personalized information retrieval, and suggest more principled metrics for evaluating a search engine's performance.

Acknowledgments

We are grateful to Stelios Psarakis for his help with the statistical tests. We also thank Christos Amanatidis, Serafeim Karapatis, Kiriakos Tassopoulos, Giorgos Papachrysanthou and Christos Papadimitriou for many useful discussions.

References

1. Dimitris Achlioptas, Amos Fiat, Anna Karlin and Frank McSherry: Web Search via Hub Synthesis. Proc. Symp. On Foundations of Computer Science (2001)
2. Mehmet S. Aktas, Mehmet A. Nacar, Filippo Menczer: Personalizing PageRank Based on Domain Profiles. WebKDD 2004
3. Rodrigo B. Almeida, Virgilio A. F. Almeida: A Community-Aware Search Engine. WWW2004

4. Chumki Basu, Haym Hirsh and William Cohen. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, pages 714-720 (1998)
5. Daniel Billsus and Michael J.Pazzani: Learning Collaborative Information Filters. In Proc. 15th International Conference on Machine Learning (1998)
6. Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of the WWW7 Conference (1998)
7. A. Broder, S. Glassman, M. Manasse, and G. Zweig: Syntactic clustering of the web. In Sixth International World Wide Web Conference, pages 391-404, 1997
8. Lin Deng, Xiaoyong Chai, Qingzhao Tan, Wilfred Ng, Dik Lee: Spying Out Real User Preferences for Metasearch Engine Personalization. WebKDD 2004
9. Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha , Horst Simon: PageRank, HITS and a Unified Framework for Link Analysis. Lawrence Berkeley National Lab Tech Report 49371 (www.neresc.gov/~cding.page.ps) (Nov. 2001)
10. S. T. Dumais (1999). Beyond content-based retrieval: Modeling domains, users and interaction. Keynote address at IEEE: Advances in Digital Libraries'99, May 19-21, 1999. Powerpoint slides
11. M.Eirinaki, M.Vazirgiannis, "Web Mining for Web Personalization", ACM Transactions on Internet Technologies (ACM TOIT), Vol.3 Issue 1
12. Final Report of Decision Making in Microeconomics using Data Mining and Optimization Techniques (Project PENED 99, under contract no. 99 ED232): General Secretariat for Research and Technology, Hellenic Ministry of Development, Greece (September 2001)
13. Gary Flake, Steve Lawrence, C. Lee Giles: Efficient Identification of Web Communities. In Proc. of the 6th ACM SIGKDD, pp.150-160 (August 2000)
14. David Gibson, Jon Kleinberg, Prabhakar Raghavan: Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia (1998)
15. Taher Haveliwala: Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference (2002)
16. Taher Haveliwala, Sepandar Kamvar and Glen Jeh: An Analytical Comparison of Approaches to Personalizing PageRank. Stanford University Technical Report (2003)
17. Glen Jeh and Jennifer Widom: Scaling personalized web search. In Proceedings of the Twelfth International World Wide Web Conference (2003)
18. Jon M. Kleinberg.:Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632 (1999)
19. Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew S. Tomkins: The Web as a graph: measurements, models and methods. Proceedings of the 5th International Computing and combinatorics Conference (1999)
20. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins: Trawling the web for emerging cyber-communities. WWW8 / Computer Networks, Vol 31, p1481-1493 (1999)
21. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins: Recommender systems: A probabilistic analysis. In Proc. 39th IEEE Symp. Foundations of Computer Science (1998)
22. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins and Eli Upfal: Stochastic models for the Web graph. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 57-65 (2000)
23. Tie Liao: Global Information Broadcast: An Architecture for Internet Push Channels. IEEE Internet Computing, Volume 4, Issue 4:16-25, (July/August 2000)
24. The Open Directory Project: Web directory. <http://www.dmoz.org/>

25. Larry Page, Sergey Brin, R. Motwani, T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford (Santa Barbara, CA 93106, January 1998). <http://www.db.stanford.edu/~backrub/pageranksub.ps>
26. Skiena, S: "Coloring Bipartite Graphs." §5.5.2 in *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Reading, MA: Addison-Wesley, p. 213, 1990
27. SpiderWave , <http://spiderwave.aueb.gr>