

The Compass Filter:

Search Engine Result Personalization using Web Communities

Content Areas: Web personalization, search engines, metrics for personalization effectiveness

Apostolos Kritikopoulos

Dept. of Computer Science
Athens University of
Economics and Business
Patision 76, Athens, T.K.10434,
GREECE
+306977687978
apostolos@kritikopoulos.info

Martha Sideri

Dept. of Computer Science
Athens University of
Economics and Business
Patision 76, Athens, T.K.10434,
GREECE
+302108203149
sideri@aueb.gr

Abstract

It is of interest to *personalize* search engines, so that past interactions of the user with the search engine can be used to improve future search results. We are not aware of many current systematic approaches to this problem. In this paper we propose a simple approach to such personalization based on Web communities [Gibson et al., 1988]. Past 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 preliminary experimental evidence suggesting that our method indeed improves the quality of the returned order. 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 briefly 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 reached formidable levels in terms of the precision and recall of the returned documents, 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*. There are, of course, related domains, such as recommendation systems [Basu et al., 1998; Billsus and Pazzani, 1998; Kumar et al., 1998] and push channel technology [Liao, 2000], in which personalization based on the user’s declared or mined

preferences is the supreme consideration. But we are not aware of current systematic approaches to personalizing search engine results based on data collected from the history of the user’s past interaction with the search engine.

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 [Brin and Page 1998; Ding et al., 2001; Page et al., 1998] and Kleinberg’s hubs and authorities [Kleinberg, 1999] 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 [Achlioptas et al., 2001; Ding et al., 2001; Gibson et al., 1988; Kleinberg, 1999; Kleinberg et al., 1999; Kumar et al., 1999]. One of the most interesting and intriguing observations in this study is the existence of abundant *Web communities* [Gibson et al., 1988, Kumar et al., 1999], that is, small sets of documents that are highly connected, typically as a bipartite clique (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 been emphasized often [Flake et al., 2000; Kumar et al., 2000].

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.

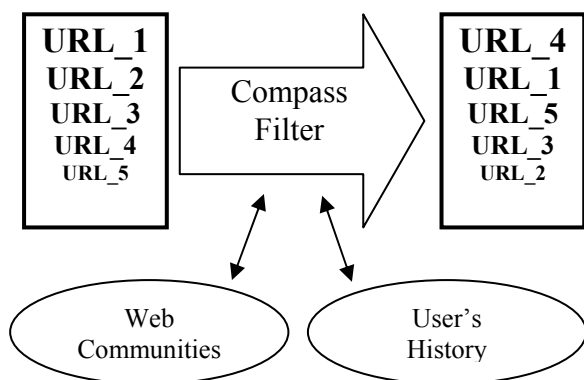


Figure 1. Reordering the result set, using the Compass Filter

The main idea is this: We extract the Web communities (all non-star bipartite graphs in the corpus) and for each community we also determine its neighborhoods (the documents linked to, or from, documents in the community). 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 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. The rating is based on the relevant accuracy that every Web page has in conjunction with the given query:

Table 1. First result set of the example

	URL	MAIN THEME OF THE WEB SITE
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, we know that the user had chosen a Web page (www.hunting_guns.gr/double-barrel/carbines.htm) that belongs at a 2x2 community. We also have found that the fifth page (Table 1) also belongs to the same hyperlinked community (Table 2).

Table 2. 2x2 Community of the example

2x2 COMMUNITY (main theme :HUNTING)	
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

At the end, we re-rank the result set, so that the fifth Web page must ascend the order of the presented Web pages:

Table 3. Final result set of the example

URL	MAIN THEME OF THE WEB SITE
<u>1</u> www.hunter.gr/duck_spots.html	<u>HUNTING</u>
2 www.greek_natural_park.gr/crete/duck.htm	NATURAL PARK
3 www.gastronomy.gr/recipes/duck_potatoes.html	RECIPES
4 www.corfu_island.gr/local_animals/duck.html	ISLAND OF CORFU
5 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* [SpiderWave], 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 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 preliminary 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 unsuccessfully and the one that was processed successfully by the Compass), 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 our 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. For the same reasons, we were not able to use our experiments in order to optimize the parameters of our method.

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

2 Description of the Method

Web communities are complete bipartite graphs of hyperlinks; the surprising prevalence of Web communities is a much-talked about 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. Data mining user access to such small fraction of our corpus would be impossible.

2.1 Step 1 (Preprocessing): Expand the communities

We chose to “expand” the communities in a manner very similar with HITS [Ding et al., 2001; Gibson et al., 1988; Kleinberg, 1999]: we add to the set of the Web pages 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). With this step we expand widely the community S_{CG} , and we signify the community as the sum of the 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.

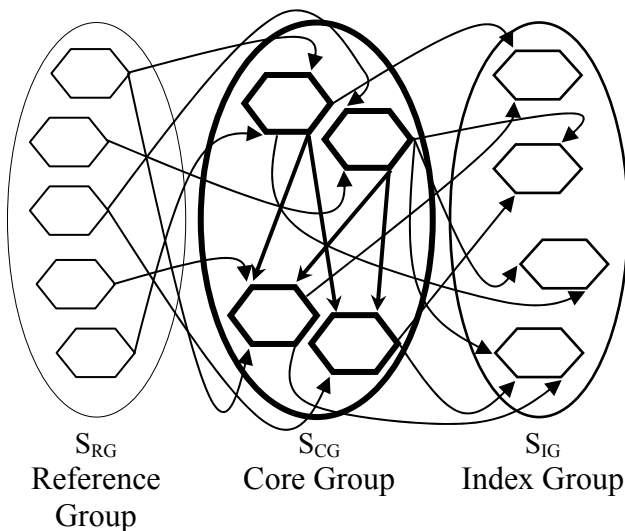


Figure 2. Link graph of the community groups

2.2 Step 2: Calculate the Community Weights of the User

While a user surfs and clicks on query results, we monitor the core, reference, and index groups s/he visits, and we calculate, for each user and 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 decrease 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 numerous. We capture this by the following formula:

$$\text{WEIGHT COMMUNITY} = (\text{Visits } S_{RG} \text{ Community}) + (3 * \text{Visits } S_{CG} \text{ Community})^3 + (2 * \text{Visits } S_{IG} \text{ Community})^2$$

2.3 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.

$$\text{Weight Of Url} = \sum_1^n \text{Weight Of Community It Belongs}$$

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

EXAMPLE: A user 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 are ranked highest by the Compass Filter.

3 Experimental Set-Up and Evaluation Metric

SpiderWave (<http://195.251.252.44>) 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 testbed 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 [ED232 contract no.99, 2001]. 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 the process described in [Kumar et al., 1999] to extract the communities of the Greek Web. The community extraction process traced 1337 communities having in total 11917 Web pages, with dimensions varying from 2x2 to 2x12, and 8x2 to 8x8. Following the first step of the method, we expanded the communities and finally concluded with 1337 expanded communities and 348826 Web pages. Studying these communities gives a very clear idea about the formation of the Greek Web: we found out that many of them represent the sociology of the Greek Web (some main themes of the communities where about 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).

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

$$SI \text{ _ Index} = \frac{1}{n} \sum_{t=1}^n \frac{n-t+1}{d_t * n}$$

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

t is the order that the user selected the URL

d_t is the order that the URL is positioned at the list

The SI score rewards the clicking of high items early on. The inverses of the ranks of the items clicked are weight-averaged, and the weights decrease 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:

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

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

4 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 engine “flipped” a fair coin and decided that the answer will not be filtered. On the contrary, for groups B and C, 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.

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 tells us that the observed difference between the means is significant. We could support a conclusion that the results of group C are substantial better than the results of the other two groups, and that our method considerably improves the quality of the retrieved information.

5 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 preliminary 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 [Kleinberg, 1999] and HITS [Ding et al., 2001; Gibson et al., 1988; Kleinberg, 1999]) 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 hubs and pointed to by the authorities; 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. [1999] 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 behavior. 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

- [Achlioptas et al., 2001] Dimitris Achlioptas, Amos Fiat, Anna Karlin and Frank McSherry. Web Search via Hub Synthesis. Proc. Symp. On Foundations of Computer Science, 2001.
- [Basu et al., 1998] C.Basu, H.Hirsh and W.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.
- [Billsus and Pazzani, 1998] D.Billsus and M.J.Pazzani. Learning Collaborative Information Filters. In Proc. 15th International Conference on Machine Learning.
- [Brin and Page, 1998] Sergey Brin, Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of the WWW7 Conference, 1998.
- [Ding et al., 2001] C.Ding, X.He, P.Husbands, H.Zha, H.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.
- [ED232 contract no.99, 2001] 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.
- [Flake et al., 2000] G.W.Flake, Steve Lawrence, C. Lee Giles, Efficient Identification of Web Communities. In Proc. of the 6th ACM SIGKDD, August 2000, pp.150-160.

[Gibson et al., 1988] David Gibson, Jon Kleinberg, Prabhakar Raghavan. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1988.

[Kleinberg, 1999] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632, 1999

[Kleinberg et al., 1999] J 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

[Kumar et al., 1999] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins. Trawling the web for emerging cyber-communities. WWW8 / Computer Networks, Vol 31, p1481-1493, 1999

[Kumar et al., 1998] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tompkins. Recommender systems: A probabilistic analysis. In Proc. 39th IEEE Symp. Foundations of Computer Science, 1998

[Kumar et al., 2000] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 57-65, 2000.

[Liao, 2000] T. Liao. Global Information Broadcast: An Architecture for Internet Push Channels. IEEE Internet Computing, 4(4):16-25, July/August 2000.

[Page et al., 1998] 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>

[SpiderWave] SpiderWave , <http://195.251.252.44>