Exploiting Popularity to Improve Blog Search

Luiz Guilherme P. Santos
luizgps@dcc.ufmg.br

Marcos André Gonçalves
mgoncalv@dcc.ufmg.br

Alberto H.F. Laender
laender@dcc.ufmg.br

Jussara M. Almeida
jussara@dcc.ufmg.br

Virgílio Almeida
virgilio@dcc.ufmg.br

Departamento de Ciência da Computação, Universidade Federal
de Minas Gerais
Belo Horizonte MG Brasil

ABSTRACT

The blogosphere is a highly dynamic and interconnected subset of the Web that has triggered a lot of interest due to its social and personal nature. In this paper, we present a study of an important social aspect of blogs, namely popularity. This study, based on the most popular blogs from four important blog domains in Brazil, shows that, despite the blogosphere being a social network, popularity has been underexploited by at least the most popular search engines in the context of blog search. In our experiments, queries specifically formulated for retrieving these popular blogs were not capable of ranking them at the top positions (top 100) by the most popular search engines. Besides, their PageRank values are very low. We also provide evidence that explicitly incorporating popularity into the search engine algorithm has the potential to significantly improve the final rankings.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information storage and retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Measurement

Keywords

Blog search, information retrieval

1. INTRODUCTION

The increasing popularity of blogging has created a highly dynamic and interconnected subset of the Web which has become known as the "blogosphere". In fact, the number of blogs has grown exponentially in the last few years [13]. This impressive growth has led to the need to effectively access these blogs, for example, through search engines. Indeed, there are currently a lot of search services offered by many sites on the Web, some of them specialized in blog search (e.g., GoogleBlogSearch and Technorati[1]).

The blogosphere has, by its own nature, a more outgoing content and a very informal language if compared with the open Web in general. Many blogs are created by their authors as a self-expression mechanism. Despite this, many blogs have huge audiences [18], meaning that there is a great deal of interest in them. This also means that there is room for exploiting other services on top of them such as advertisement and recommendation [19].

General Web search engines can obviously be used for finding blogs, especially if one knows exactly the desired blog or its topic. However, a hypothesis investigated in this paper is that specialized blog search engines may potentially better satisfy the needs of blog searches if they provide specific features that consider intrinsic characteristics of the blogosphere.

In fact, a previous analysis of more than 35 million requests made to a large blog service in Brazil concluded that about 46% of the traffic to blogs comes from search engines [6]. In this same study, the authors observed that most of the *popular blogs* are generally easier to be reached from links from other blogs than through search engine results. Although search engines are responsible for most of the traffic into the blogosphere, they were not able to reach the most popular blogs as should be expected. In other words, the intensity of traffic directed to a blog through search engines does not seem to correlate with its real "popularity". As users usually just click on the first results, this might be evidence that search engines are not considering popularity as a major feature in their rankings when blogs are the target. This highlights the need for developing ranking strategies that take into consideration the social attributes of the blogosphere, especially in the context of specialized blog search engines. The integration of social network information with already known search techniques was also suggested in [17] as a means to improve the quality of Web search experience.

To be more precise, popularity is here regarded as an intrinsic relationship between the collective behavior of a given community and an object (e.g., a blog), meaning that a significant portion of that community likes, approves or finds

[1]http://blogsearch.google.com, http://technorati.com

the object suitable in some given context. We assume that a popularity indicator can be associated with this relationship allowing us to quantify the level of popularity of a certain object and to compare multiple objects according to their relative popularity. Examples of such indicators include number of visits, downloads and even socially-oriented aspects such as number of social annotations in user-generated content [4]. For blogs, specifically, other popularity indicators include number of individuals who have subscribed to them, relative click-through ratio [2] and, as considered here, number of times the blog appeared in top lists.

In this paper, we focus on blog search considering the blogosphere as a social network where popularity is an important aspect [1]. We start by analyzing the quality of blog search in actual general Web search engines (restricted to a given blog domain)[2]. We would expect that a successful search in the blogosphere should return not only relevant blogs, but, the most popular ones, as would be expected for any social network. We verify, though, that this is currently not the case. Four important blog domains in Brazil were tracked for some time to extract their most popular blogs. In our experiments, queries specifically formulated by volunteers for retrieving these popular blogs were not capable of ranking them at the top positions (top 100) of popular search engines. Moreover, their PageRank values, as measured by the typical web graph topology, were considered very low.

Additionally, in order to further investigate the potential of exploiting popularity in blog search, we run experiments in which we explicitly incorporate *a popularity factor* into the search engine algorithm. By doing so, we produced rankings that were considered very relevant by volunteers and much better (63% improvement) than the original ones.

The rest of this paper is organized as follows. Section 2 summarizes related work. Section 3 analyzes the quality of blog search in existing search engines, whereas a strategy to incorporate popularity into blog search is presented and evaluated in Section 4. Finally, Section 5 offers conclusions and future work.

2. RELATED WORK

There are a number of studies focused on blogs, several of which aim at improving blog search engines, some exploiting user behavior, others exploiting blog characteristics. Here, we only refer to the most related ones.

In [6], the authors characterize the access patterns in the blogosphere and conclude that the nature of the users' interactions is different from what is observed with traditional Web content. They also conclude that the access pattern to blogs is more dependent on social networks. In our work, we go further by explicitly considering the issue of blog popularity in the quality of the search in the blogosphere.

In [16], the authors present an analysis of a large blog search engine query log, focusing on aspects such as query intent, query topics and user sessions. They conclude that blog search is different in many aspects from Web search, particularly in terms of interest area. Nevertheless, when browsing results, user behavior is similar: users are usually interested only in the first positions of the returned ranked list. Our strategy for blog search explores this principle, by trying to boost the most important (i.e., most popular) blogs to the top of the ranking.

A new search engine that considers the particular characteristics of blogs is proposed in [7]. Different interfaces are proposed, each one with a different focus: topic search, blog author search and reputation search. Nevertheless, none of them exploit popularity into ranking blogs. Intrinsic properties are commonly used in other domains to enhance information retrieval quality like in [20].

More related to our work is BlogRank, a blog ranking method based on applying the traditional PageRank algorithm [5] to an enhanced link graph, extended with links representing author and topic similarity between blogs [11]. Also, according to its patent application, the Google BlogSearch algorithm considers blog popularity, assessed by RSS feed readership, as one possible positive indicator of blog quality [2]. Other positive indicators include click-through ratio and PageRank. Accordingly, a recent study has proposed a new popularity ranking method (BRank) which exploits various social interconnections between bloggers [12].

Mishne [15] exploits several blog properties such as temporal information, level of discussion and level of spam, to improve opinion retrieval in blogs, showing significant gains over state-of-the-art techniques. In [10], the authors propose to explore another blog property, namely credibility, to rank blogs.

Meiss [14] use traffic data to partially validate the PageRank algorithm, but also show differences in navigation patterns not captured by that algorithm. As a result, they suggest that Web traffic data available to an Internet Service Provider (ISP) could be used to induce a ranking measure over all sites to better reflect their relative importance according to the dynamic behavior of the user popularity.

3. ANALYSIS OF BLOG POPULARITY

This section describes two experiments we conducted with the goal of providing evidence that general and blog domain search engines are currently not adequately exploiting popularity to rank blogs.

3.1 Data Collection

We collected, during thirty days, the most popular blogs from four of the most well known blog domains in Brazil: UOL, Blogger, BlogLog and Terra[3]. UOL and Terra are two leading Brazilian Web portals where users can freely create their blogs. Blogger is a blog domain owned by a large Brazilian communication company, which requires (paid) user subscription for blog creation. BlogLog is a restricted blog domain where only invited artists can maintain a blog.

[2]We here focus on blogs, instead of posts within a blog, as it is easier to obtain popularity information at that granularity. Nevertheless, we acknowledge that, for some information needs, a finer granularity may be more useful; we leave this investigation for future work.

[3]blog.uol.com.br, blogger.com.br, bloglog.globo.com, blog.terra.com.br

Each blog domain applies a somewhat distinct strategy to determine its most popular blogs, but all of them consider the role of the users. UOL uses a voting system in which the users give points to the blogs, in a scale from zero to ten, based on their opinions. In Blogger and Terra, popular blogs are the ones with the largest numbers of hits and best recommendations from users. BlogLog uses the number of accesses. In all of them, a list of the most popular blogs is made available daily on the main page of the domain.

During the 30-day period, we gathered, daily, the ten most popular blogs from each domain, thus creating a collection of 30 top-10 lists for each domain. We first ranked the collected blogs from each domain by the number of days they appeared in its top-10 lists. We then selected the ten most highly ranked blogs as the most popular blogs from each domain, thus, ending up with forty blogs for analysis. Most of the selected blogs are personal diaries (around 62,5%). This seems to indicate a social tendency of people enjoying following personal experiences of others, mainly influential ones such as artists, a phenomenon also observed in other blog domains such as the "micro" blog Twitter[4].

3.2 PageRank Analysis

One of the most effective search engine strategies to rank documents is Google's PageRank [5]. Google's BlogSearch patent application also explicitly mentions PageRank as a possible positive indicator also for blog quality [2]. Thus, our first experiment consists of analyzing the PageRank values of the most popular blogs from the four selected domains. Our goal here is to assess whether there is a correlation between popularity and importance of the blog as measured by PageRank. Despite some aspects of this issue have been discussed in very strict scenarios (Kritikopoulos et al. 2006), we provide clearer evidence of the matter through quantitative measurements specifically for the case of popular blogs, which one might think that could have higher connectivity than non-popular blogs.

The PageRank value was measured for each blog using the Google Toolbar[5] browser plugin, which returns values from zero (least important) to ten (most important). A special value of -1 is used for non existing PageRank values, that is, for pages that are basically invisible to the search engine, according to this criteria. Before showing our results, we should emphasize that the measured PageRank for the main webpage of the UOL blog domain is 6, while the PageRank of the main webpage of the UOL portal, corresponding to the host that contains that blog domain, is 8. Accordingly, the main web pages of the other three blog domains also have PageRank values of 6, whereas their respective hosts have PageRank values of 7 (terra.com.br) and 6 (globo.com). These values imply that the coverage and visibility of the analyzed blog domains by Google are reasonably good. Thus, we should not expect any significant bias in our results due to lack of coverage by Google.

We now turn to the analysis of the measured PageRank values for each individual blog. Figure 1 shows the values for the analyzed blogs, ordered, in the x-axis, by their popularity within the respective domain.

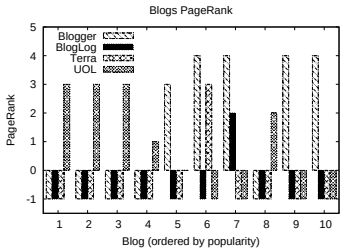


Figure 1: PageRank values for the most popular blogs

We start by noticing that some popular blogs do indeed have PageRank values that are somewhat significant (around 3 and 4) given that these values are close to the ones of their respective blog domains. Moreover, all the blog domains have at least one blog with PageRank value higher than 2. This provides evidence that the search engines have been crawling the blog domains and that, in spite of the different ways of estimating blog popularity, the collected blogs are indeed popular ones within their respective domains.

On the other hand, in a broader perspective, the highest absolute PageRank value was 4, which can be seen as low, given that we are working with the most popular blogs of important domains. Moreover, the vast majority of them (i.e., 27 out of 40) do not have a PageRank value. In fact, the four most popular blogs in Blogger, BlogLog and Terra do not have PageRank values, whereas, for UOL, the four most popular blogs have PageRank values under 3.

In sum, the above results are indications of the low correlation between the importance of the blogs in the Web Graph and their relative popularity. In fact, some of these results are surprising given that Google BlogSearch patent explicitly mentions popularity as a factor that could help blog search [2].

3.3 Analysis of the Ranking

In our second experiment, we recruited five volunteers to analyze twenty blogs randomly chosen from the forty most popular blogs in our collection. Each volunteer was asked to assign six keywords to each analyzed blog. The keywords should well describe the blog content and should be those that they would actually use if they wanted to find that blog by using any existing search engine. Two volunteers analyzed each blog, sorting their selected keywords by their importance. We selected six out of the twelve keywords assigned to each blog, prioritizing keywords assigned by both volunteers and randomly selecting between both of them for the remaining keywords, following the predefined order. We note that, in some cases, the selected keywords were not present in the text of the blog (e.g., "diary", "video", "children") despite accurately describing its content.

We then defined three types of query: (1) queries with the two most important keywords; (2) queries with the three most important keywords; and (3) queries with all six keywords. For the first two types, we made a conservative choice

[4]twitter.com
[5]toolbar.google.com

of discarding keywords that appeared in the URL or in the title of the blog as search engine ranking algorithms use these as strong evidence for retrieval, mainly for queries aimed at finding a specific blog (i.e., navigational queries). In other words, the first two types of query cover scenarios in which the intent is to look for popular blogs about a specific subject, i.e., queries looking for the informational content of the blogs. In contrast, the use of all six keywords, regardless of whether they appear in any part of the blog, covers both, informational and navigational queries.

We started our experiment by trying to run the queries with the specialized blog search engine offered by one of the blog domains, namely UOL[6]. Results for the three types of query are shown in Figure 2. The graph shows the ranking positions of the selected popular blogs, in the results produced for each query type. If the blogs appeared after the 100th position, they were assigned to this position so that we could keep the scale of the graph. Blogs are ordered by their popularity in the x-axis of the graph. In only eight out of the thirty results the blogs were returned in the top 100 list. Moreover, this happened for only three of the 10 most popular blogs of the domain, and in one case (with two keywords) the ranking of the blog was higher than 50. These results are summarized in Table 1 (column 2), which shows the percentage of popular blogs that appeared in the first result page (i.e., top-10 results), and, thus, that would possibly be noted by users. Most of the popular blogs (at least 70% of them) were not returned in the first page.

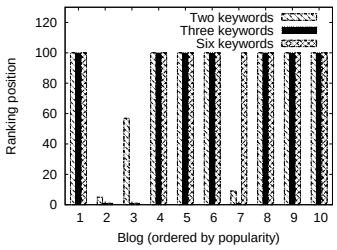


Figure 2: Ranking produced by the UOL search engine

Given our current focus on search within a blog domain, and the fact that specialized blog search engines, such as Google-BlogSearch and Technorati, do not allow us to restrict the search to a specific blog domain, in our next set of experiments we use two of the largest general Web search engines that, in theory, also index a large portion of the blogosphere: Google and Yahoo!. These are usually the entry points of the Web for non-specialized users. Moreover, as mentioned before, we restricted the search performed in each experiment to the specific domain from which the blog was collected (BlogLog were searched only within the BlogLog domain, for example). This set of experiments also allowed us to compare the results across the four blog domains within a common framework.

Similarly to Figure 2, Figures 3 shows the ranking position of each popular blog as they appeared in the Google and Yahoo! rankings. Note that there are several cases (Figure 3(b), 3(c), 3(d), 3(g) and 3(h)) in which the most popular blog appears only on (or after) the 100th position of the ranking. This is true for all three types of query. In fact, considering the four blog domains and the two search engines, the most popular blog was returned in the first page only in three cases (3(a), 3(e) and 3(f)), and even so for only one type of query. Overall, a significant fraction of the most popular blogs appears in very low positions in the ranking.

[6]Terra also offered a search service, but it uses the general Google search engine. This scenario is captured in the last set of experiments in this section.

	UOL	Google	Yahoo!
2 Keywords	10%	52%	42%
3 Keywords	30%	42%	37%
6 Keywords	20%	62%	52%

Table 1: Percentage of blogs in the first result page

Table 1 also summarizes the percentage of popular blogs that appeared in the first result page (i.e., top-10 results) in the Google and Yahoo! rankings. The fraction of popular blogs that do not appear in the first result page of both search engines is quite significant (over 52%). In fact, more than 57% of the blogs do not appear in the first page returned by Yahoo! for two and three keyword queries. Even when we used all six keywords, which should be the easiest situation, since these keywords could appear in the URL or in the title of the blog, we were not able to retrieve approximately one third of the popular blogs in the first page of the Google results. For Yahoo! results, the portion is even lower: almost half of the blogs are not in the first page.

Table 2 presents three examples of blogs to which volunteers assigned good keywords (i.e., we manually verified that the keywords accurately capture the blog subject) but whose ranking positions are low. The third and fourth column shows the ranking position of each blog in the Google and Yahoo! results respectively. Some of them are very low, too far down to catch the user's attention.

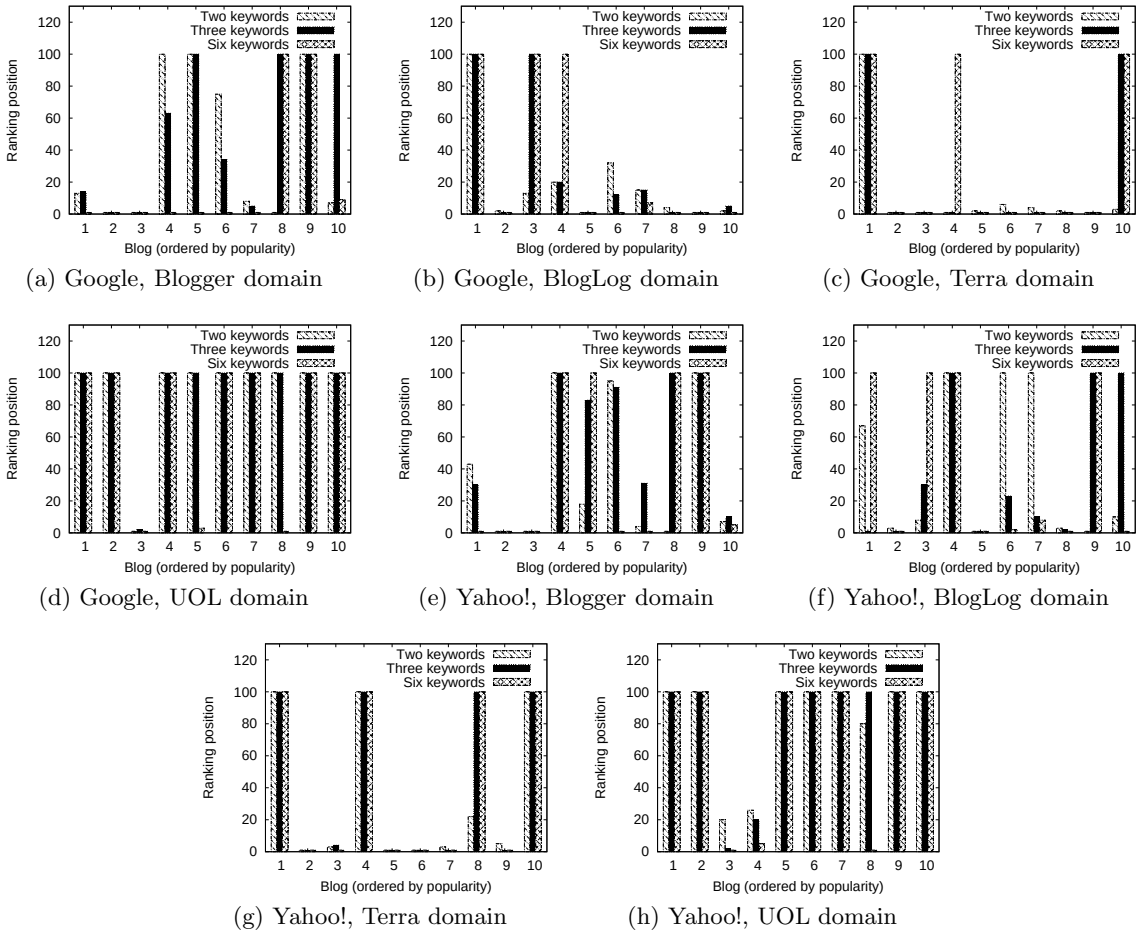
4. USING POPULARITY TO RANK BLOGS

In this section, we propose a new search strategy for blogs based on their popularity. We show how to use this important feature to better rank query results and improve the user experience in blog search. The idea here is to incorporate popularity as a factor in the ranking formula. We then contrast the results of several blog searches with and without the popularity factor by checking the improvement in the overall rankings, based on relevance judgments provided by volunteers. We should stress that our goal here is not to propose the "best possible" ranking strategy that exploits popularity but to provide evidence that popularity can indeed be beneficial in the task of blog search and enhance the user experience as a whole.

4.1 Experimental Setup

Due to the lack of popularity information (as used by us) in standard collections such as the TREC Blog Track [18][7] as well as of information from query logs of real blog search en-

[7]The TREC Blogs06 collection does in fact have a number of top blogs but the popularity information is not explicit[13].



(a) Google, Blogger domain (b) Google, BlogLog domain (c) Google, Terra domain

(d) Google, UOL domain (e) Yahoo!, Blogger domain (f) Yahoo!, BlogLog domain

(g) Yahoo!, Terra domain (h) Yahoo!, UOL domain

Figure 3: Query ranks by Google and Yahoo! for the most popular blogs of each studied domain.

Blog	Keywords	Google Rank	Yahoo! Rank
paidegemeos.zip.net	Diary, TwinChildren	> 100	80
anotacoescinefilo.blogger.com.br	Cinema, Festival	13	43
diariodeleticia.zip.net	Diary, Children	> 100	26

Table 2: Three examples of good keywords with bad ranking positions

gines from where we could extract real queries that retrieved *popular blogs*, a hard task even if we had the logs, we relied again on our set of collected popular blogs and the keywords assigned to them. For these experiments, we also collected a sample of blogs from the UOL. This domain was chosen mainly because its strategy to estimate blog popularity, described in the previous section, takes into consideration the users' opinion over a fine-grained scale (0-10).

We used a crawler to collect 15,000 blogs from the UOL domain. These blogs were indexed using the Lucene API[8]. We incorporated the popularity of the blogs into the index using methods available in Lucene. We chose to crawl and index our own blog collection to facilitate the experimental evaluation, since it is very difficult to conduct this kind of experiment using any commercial search engine.

We define a popularity factor (PF) for each blog of the collection that is proportional to its importance in the domain estimated by the number of days it appeared in the top-10 list during our 30-day collection.

The popularity factor is computed using Equation 1 where N represents the number of days the blog appeared in the top-10 list, M is the maximum number of days any single blog made it to the top-10 list, and K is an empirically chosen scaling factor (20 in our experiments).

$$PF = K * \frac{N}{M} + 1 \quad (1)$$

[8]http://lucene.apache.org

Lucene uses the traditional term frequency-inverse document frequency (TF-IDF) weighting scheme and the vector space model [3] to decide the importance of keywords to describe a document. We use *PF* as an additional multiplicative factor in the TF-IDF formula to obtain the final score to rank the blog. It is always worth to stress that we could use other weighting schemes such as BM25 [9] or other query independent features (also called document priors) such as link information. This for sure could provide gains. However, as explicitly stated before, our goal is not to propose the "best" ranking strategy that exploits popularity but to show that there is potential for using this type of evidence and, in a certain extent, *to quantify* this potential. Our contributions rely in part in this explicit quantification. This is even more important given the increasing number of blog search engines that claim to use some type of blog influence into account and the small number of works that indeed try to measure the effects of this influence in blog retrieval.

4.2 Effectiveness of the Popularity Factor

In this section we analyze the effectiveness of the proposed popularity factor by comparing the results obtained when using our popularity factor with the original ranking. The idea is not only to check whether the popular blogs were considered relevant and boosted to the top positions of the rankings but also to assess the overall impact of these modifications in the ranking. In other words, we want to verify whether we are in fact improving the original ranking by boosting the popular blogs (when these have some similarity with the query) without removing other results that may be actually more relevant instead. As some keywords are very general in nature (see Table 2, for example), this is a very possible situation.

We used the same keywords previously defined by the first set of volunteers for the ten most popular blogs from UOL to perform queries in two search engines: one indexed with the popularity factor and the other without it. Like in the previous section, we submitted three types of query to each search engine.

The first ten result blogs of the two rankings, the original one and the one modified by the popularity factor, were put in a joint pool, shuffled, and then presented to a new set of volunteers (different from those who specified the keywords) for evaluation. These volunteers should label each result blog into three categories: very relevant (relevance level = 3), relevant (relevance level = 2), or irrelevant (relevance level = 1) given the specified query and the blog content. Each pair (query, result blog) was evaluated by exactly two different volunteers. To be more precise, two volunteers evaluated the queries and results related to the first five target blogs and two different ones evaluated the queries and results of the other five. Notice that the very broad nature of some of our queries, mainly the queries with two keywords (e.g., "travel diary", "twin parents", "cinema festival" and "writer thoughts"), which reflect general interests and could retrieve a large number of blogs not only the popular ones, may reduce any previously existent bias towards any of the rankings. The level of agreement of the volunteers was around 80%, considering "very relevant" and "relevant" as a unique category; disagreements were handled by averaging the evaluation metrics produced by each individual evaluator's ranking, as we shall see next.

This experiment produced sixty results: 10 target blogs × 3 types of query × 2 evaluations. We evaluated them using the Normalized Discounted Cumulative Gain metric [8], defined in Equation 2 as

$$NDCG = \frac{1}{N_i} \sum_{i=1}^{k} \frac{2^{label(j)} - 1}{\log_2 (1 + i)} \quad (2)$$

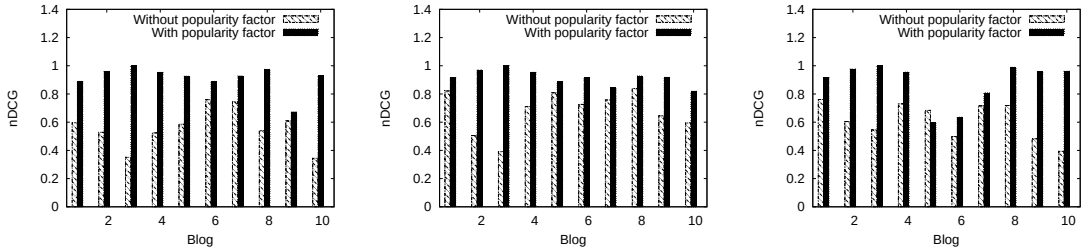
where N_i is a normalization constant calculated based on a perfect ordering of the results for the query q_i and $label(j)$ is the gain value associated with the label of the document at the j^{th} position of the ranked list. For instance, $label(j)$ is equal to 3 if the document is considered very relevant, equal to 2 if considered relevant and equal to 1 if considered irrelevant. In the NDCG formula, the sum computes the cumulative information gain to the user from the already inspected documents and $log_b(1 + i)$ is a discounting function that reduces the document's gain value as its rank increases. The base of the logarithm, b, controls the amount of the reduction. We used $b = 2$ in our experiments which corresponds to a sharper discount.

In the context of our study, a higher value of NDCG for the version with the popularity factor, for instance, means that we are substituting less relevant blogs in the first positions of the ranking by more relevant ones, allowing to evaluate the impact of the popularity factor in the ranking. Notice that NDCG is normalized by the best possible ranked list that can be obtained, represented by the normalization factor N_i. In our case, this rank is calculated based on the relevance judgments obtained for both types of query, with and without the popularity factor. The same normalization factor is used for the calculation of both NDCG values.

Figure 4 shows the average of NDCG values of the two volunteers for queries with two, three and six keywords, considering the top 10 results for each type of query. We can see that for all cases but one (query for blog 5 with six keywords), there were improvements when we used the popularity factor. In fact, in several cases the NDCG values of the version without popularity were very low (under 0.6) when compared with the ideal rank, showing the difficulty of performing blog search with traditional strategies. Ignoring the experiment with blog 5 with 6 keywords, the improvements varied from 9.65% up to 184.91%. The average NDCG results, when we consider all blogs and the different types of query, are shown in Table 3. The overall gains of the strategy that considers popularity are up to 63% for two-keywords queries, 34% for three-keywords queries and 43% for six-keywords queries. All results were found to be statistically significant with 99.9% confidence (t-test).

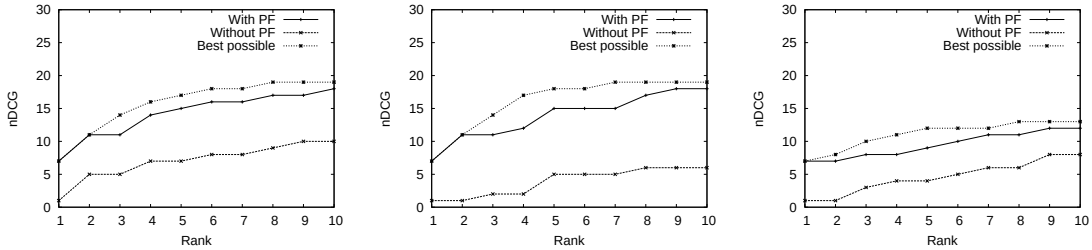
	2 keywords	3 keywords	6 keywords
With PF	0.912	0.915	0.879
Without PF	0.558	0.679	0.613

Table 3: Overall results for NDCG

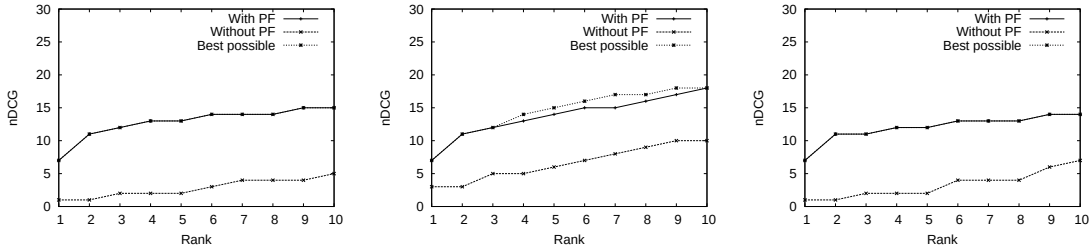


(a) (b) (c)

Figure 4: NDCG with and without popularity factor for two (a), three (b) and six (c) keywords



(a) deborahbllom.zip.net, 2 keywords (b) espallhamerda.zip.net, 2 keywords (c) jessy.valim.zip.net, 3 keywords



(d) diariodeleticia.zip.net, 3 keywords (e) deborahbllom.zip.net, 6 keywords (f) diariodeleticia.zip.net, 6 keywords

Figure 5: Cumulative NDCG for queries with the largest gains

Figure 5 shows the cumulative NDCG in each position of the rank for the six experiments on which the improvements from using the popularity factor are the largest (see labels of Figure 5 to check which ones). As can be noticed, in these cases the cumulative gain is quite higher when compared to the case without popularity, being in one of the cases equal to the best possible NDCG for that query. We should stress that improvements in NDCG could only be obtained if we are in fact substituting less relevant blogs by more relevant ones in the top positions of the rankings. Thus, these results seem to suggest that, if there is *some* textual similarity between a query and a popular blog, in many cases, at least the ones we studied here, it is worth to give some additional boosting for the popular ones. However, the exact tradeoff between the similarity level and the strength of the popularity boosting for specific collections is something that should be better studied in future work. As said before, here we are only concerned with providing evidence of the potential of using the popularity in blog search.

In order to further investigate these issues, we manually checked the rankings and the respective relevance judgments and verified that the popularity factor was able to bring the specific blogs we were looking for to the top positions of the rankings, that these were considered very relevant by the volunteers and that in general they substituted or removed irrelevant blogs from the original rankings. Perhaps even more interesting, we verified that several other popular blogs that had textual similarity with the query were also boosted by the popularity factor and these were also considered very relevant, even for queries not specifically designed to retrieve them. This happened most probably due to the broad nature of the specified queries. The only case in which there were losses (query for blog 5 with six keywords) is exactly a case when too many popular blogs were boosted to the top of the rank (keywords were *"Peace Love Magic Images Religion Jesus"*). Despite only one in thirty possible cases, this indicates that we should further investigate the situations in which we may not want to use the popularity boosting factor, i.e., simply using a high scaling factor in the popularity factor formula for all cases indiscriminately would not work, as many irrelevant and popular blogs would always be in the top positions, independently of the query. We leave for future work the study of all these tradeoffs.

5. CONCLUSIONS AND FUTURE WORK

In this paper we focused on exploiting the potential of social network features in blog search, more specifically popularity. Our study revealed some interesting findings, which includes the fact that, in the context of blog search, widely used search engines do not retrieve the most popular blogs of a particular domain in the first positions of the ranking. Besides, these blogs usually present very low PageRank values. Considering that the blogosphere is a social network, popularity should be considered as an evidence to rank according to user queries. We constructed a search engine that uses the popularity factor to improve the ranking of the blogs. Our experiments, with volunteers, show that this strategy has the potential to improve the quality of the blog search process and the satisfaction of the users.

Our work also raises interesting issues: since some of the popularity statistics within a blog domain would not be directly acessible by some search engines, their use would require some collaboration from the blog domains in order to obtain this information. Moreover, some popularity measures should be standardized so that blogs coming from different domains could be compared. This issues could be solved by establishing metadata standards and harvesting protocols similar to the ones of the OAI that would allow the blog domains to periodically export usage and popularity statistics so that search engines can incorporate this information and use it to improve their blog rankings.

As future work, we would like to run additional experiments, with samples of top blogs from other "regions of the blogosphere" (e.g., from English speaking countries) to check whether our observations would still hold. Other experiments could also help to better understand when the popularity boosting is more beneficial and when not to use it.

6. ACKNOWLEDGMENTS

This work is partially supported by the projects INCT-Web (MCT/CNPq grant 57.3871/2008-6) and by the authors' individual grants and scholarships from CNPq, FAPEMIG, and CAPES.

7. REFERENCES

[1] N. Ali-Hasan and L. A. Adamic. Expressing social relationships on the blog through links and comments. In *Proc. Int'l Conference on Weblogs and Social Media*, 2007.

[2] S. Baehni, R. Guerraoui, B. Koldehofe, and M. Monod. Towards fair event dissemination. In *Proc. Int'l Conference on Distributed Computing Systems Workshops*, page 63, 2007.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. Int'l World Wide Web Conference*, pages 501–510, 2007.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *Proc. Int'l Conference on Weblogs and Social Media*, 2007.

[7] K. Fujimura, H. Toda, T. Inoue, N. Hiroshima, R. Kataoka, and M. Sugizaki. Blogranger-a multi-faceted blog search engine. *Institute of Electronics, Information and Communication Engineers technical report*, 105(650):19–24, 2006.

[8] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proc. Special Interest Group on Information Retrieval*, pages 41–48, 2000.

[9] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *IP&M*, 36(6):779–808, 2000.

[10] A. Juffinger, M. Granitzer, and E. Lex. Blog credibility ranking by exploiting verified content. In *Workshop on Information Credibility on the Web*, pages 51–58, 2009.

[11] A. Kritikopoulos, M. Sideri, and I. Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *Proc. Third Int'l Workshop on Advanced Architectures and Algorithms for Internet DElivery and Applications*, page 8, 2006.

[12] C.-L. Lin, H.-L. Tang, and H.-Y. Kao. Utilizing social relationships for blog popularity mining. In *Proc. Asia Information Retrieval Symposium*, pages 409–419, 2009.

[13] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. *DCS Technical Report Series*, 2006.

[14] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *Proc. Int'l Conference on Web Search and Data Mining*, pages 65–76, 2008.

[15] G. Mishne. Using blog properties to improve retrieval. In *Proc. Int'l Conference on Weblogs and Social Media*, 2007.

[16] G. Mishne and M. de Rijke. A study of blog search. In *Proc. European Conference on Information Retrieval*, pages 289–301, London, UK, 2006.

[17] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting social networks for internet search. In *Proc. Hot Topics in Networks workshops*, pages 79–84, California, USA, 2006.

[18] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *Proc. Text REtrieval Conference*, pages 15–27, Gaithersburg, USA, 2006.

[19] A. Stewart, L. Chen, R. Paiu, and W. Nejdl. Discovering information diffusion paths from blogosphere for online advertising. In *Proc. Int'l Workshop on Data Mining and Audience Intelligence for Advertising*, pages 46–54, San Jose, California, 2007.

[20] R. Willrich, R. de Moura Speroni, C. V. Lima, A. L. de Oliveira Diaz, and S. M. Penedo. Adaptive information retrieval system applied to digital libraries. In *Proc. of the 12th Brazilian Symposium on Multimedia and the web*, pages 165–173, Natal, Rio Grande do Norte, Brazil, 2006.