Exploiting Popularity to Improve Blog Search*

Luiz Guilherme P. Santos[1], **Orientador: Marcos André Gonçalves**[1]
Co-Orientador: Alberto H. F. Laender[1]

[1]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG)
31270-901 - Belo Horizonte - MG - Brasil

{luizgps, mgoncalv, laender}@dcc.ufmg.br

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. 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, 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. We also provide evidence that explicitly incorporating popularity into the search engine algorithm has the potential to significantly improve the final rankings.*

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 [Macdonald and Ounis 2006]. 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]).

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 [Duarte et al. 2007]. 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 [Mislove et al. 2006] as a means to improve the quality of Web search experience.

*Endereço do texto completo da dissertação: *http://www.dcc.ufmg.br/pos/cursos/defesas/1104M.PDF*

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

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 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 [Bao et al. 2007]. For blogs, specifically, other popularity indicators include number of individuals who have subscribed to them, relative click-through ratio [Baehni et al. 2007] and, as considered here, number of times the blog appeared in top lists.

In this dissertation, we focus on blog search considering the blogosphere as a social network where popularity is an important aspect [Ali-Hasan and Adamic 2007]. We start by analyzing the quality of blog search in actual general Web search engines (restricted to a given blog domain). We would expect that a successful search in the blogosphere should return not only relevant blogs, but, desirably, 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 of links, 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.

In sum, the main contributions of this master thesis are: (1) a comprehensive investigation of the ability of current search engines to exploit blog popularity and (2) the proposal and evalution of a simple method to sort the result of queries in blog search engine which exploits popularity, thus demonstrating the potential of exploring this property. These contributions were formally published in [Goncalves et al. 2010] (Journal Qualis A2).

2. Analysis of Blog Popularity

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[2]. 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

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

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.

2.1. PageRank Analysis

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[3] 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.

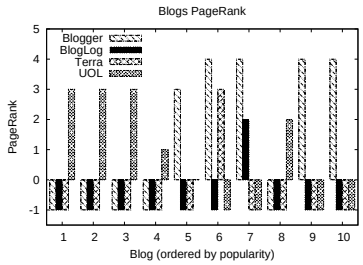


Figure 1. PageRank values for the most popular blogs

Figure 1 shows the values for the analyzed blogs, ordered, in the x-axis, by their popularity within the respective domain. We notice 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.

2.2. 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

[3]toolbar.google.com

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 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.

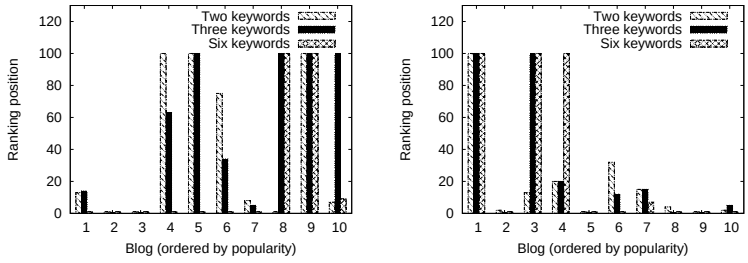
In our first set of experiments, we used the specialized blog search engine provided by one of the blog domains, namely UOL. The other domains use general search engines explored in our second set of experiments. In this experiment, more than 70% of the most popular blogs were not retrieved in the first result page with any of the query types. In the 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. 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.

Figure 2 shows the ranking position of some popular blog as they appeared in the Google and Yahoo! rankings. Note that there are some cases (Figure 2(b), 2(c) and 2(d)) 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, and even so for only one type of query.

	Google	Yahoo!
2 Keywords	52%	42%
3 Keywords	42%	37%
6 Keywords	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



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

(c) Yahoo!, Terra domain (d) Yahoo!, UOL domain

Figure 2. Ranks by Google and Yahoo! for blogs of each domain.

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.

3. Using Popularity to Rank Blogs

In this section, we propose a new search strategy for blogs based on their popularity. The idea here is to incorporate popularity as a factor in the ranking formula. 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.

3.1. Experimental Setup

For these experiments, we relied again on our set of collected popular blogs and the keywords assigned to them and also collected a sample of blogs from the UOL domain. UOL was chosen mainly because its strategy to estimate blog popularity, described previously, 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[4]. 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

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

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).

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

3.2. 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, 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. 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 $\times$ 3 types of query $\times$ 2 evaluations. We evaluated them using the Normalized Discounted Cumulative Gain metric [Järvelin and Kekäläinen 2000], a commonly used Information Retrieval metric that considers several levels of relevance and the position in the rank in which relevant documents appear.

In this context, 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. 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.

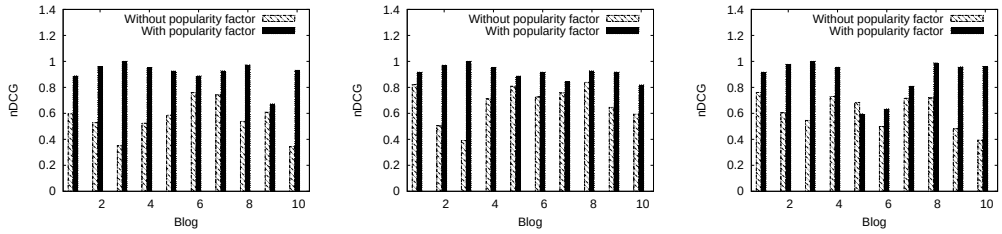


Figure 3. NDCG with and without popularity factor for two (a), three (b) and six (c) keywords

Figure 3 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 2. The overall gains of the strategy that considers popularity are up to 63% for queries with two keywords, 34% for three and 43% for six. All results were statistically significant with 99.9% confidence (t-test).

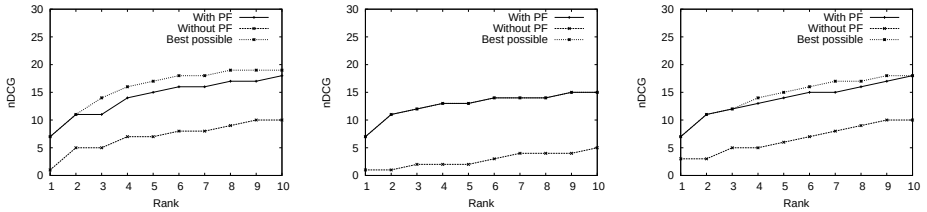
	2 keywords	3 keywords	6 keywords
With PF	0.912	0.915	0.879
Without PF	0.558	0.679	0.613

Table 2. Overall results for NDCG

Figure 4 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 4 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 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 boosting for the popular ones.

4. Conclusions and Future Work

In this dissertation 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



(a) deborahbllom.zip.net, 2 keywords (b) diariodeleticia.zip.net, 3 keywords (c) deborahbllom.zip.net, 6 keywords

Figure 4. Cumulative NDCG for queries with the largest gains

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 blogs' ranking. Our experiments, show that this strategy has the potential to improve the quality of the blog search process and the satisfaction of the users.

As future work, we would like to run additional experiments, with samples of top blogs from other "regions of the blogosphere" e.g., from Western English speaking countries) to check whether our observations would still hold. Additional experiments could also help to better understand when the popularity boosting is more beneficial and when not to use it. In our experiments, this happened in only one out of the 30 queries analyzed (one 6-keyword query: "peace love magic images religion Jesus") and was due to a large number of popular, but not necessarily relevant blogs, promoted in detriment of more relevant (though not as popular) ones. This issue requires a deeper investigation.

References

Ali-Hasan, N. and Adamic, L. A. (2007). Expressing social relationships on the blog through links and comments. In *ICWSM'07*.

Baehni, S., Guerraoui, R., Koldehofe, B., and Monod, M. (2007). Towards fair event dissemination. In *Proc. ICDCSW'07*, page 63.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. In *Proc. WWW'07*, pages 501–510.

Duarte, F., Mattos, B., Bestavros, A., Almeida, V., and Almeida, J. (2007). Traffic characteristics and communication patterns in blogosphere. In *ICWSM'07*.

Goncalves, M. A., Almeida, J. M., dos Santos, L. G., Laender, A. H., and Almeida, V. (2010). On popularity in the blogosphere. *IEEE Internet Computing*, 14:42–49.

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proc. SIGIR'00*, pages 41–48.

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

Mislove, A., Gummadi, K. P., and Druschel, P. (2006). Exploiting social networks for internet search. In *Proc. 5th HotNets-II*, California, USA.