IDENTIFYING WEBPAGE SEMANTICS FOR SEARCH ENGINE OPTIMIZATION

Themistoklis Mavridis[1], Andreas L. Symeonidis[1,2]
[1]*Electrical and Computer Engineering, Aristotle University of Thessaloniki*
[2]*Informatics and Telematics Institute, CERTH*
Thessaloniki, Greece
themismavridis@gmail.com, asymeon@eng.auth.gr

Keywords: Search Engine Optimization: LDArank: Semantic Analysis: Latent Dirichlet Allocation: LDA Gibbs Sampling: LDArank Java Application: Webpage Semantics: Semantic Analysis SEO

Abstract: The added-value of search engines is, apparently, undoubted. Their rapid evolution over the last decade has transformed them into the most important source of information and knowledge. From the end user side, search engine success implies correct results in fast and accurate manner, while also ranking of search results on a given query has to be directly correlated to the user anticipated response. From the content providers' side (i.e. websites), better ranking in a search engine result set implies numerous advantages like visibility, visitability, and profit. This is the main reason for the flourishing of Search Engine Optimization (SEO) techniques, which aim towards restructuring or enriching website content, so that optimal ranking of websites in relation to search engine results is feasible. SEO techniques are becoming more and more sophisticated. Given that internet marketing is extensively applied, prior quality factors prove insufficient, by themselves, to boost ranking and the improvement of the quality of website content is also introduced. Current paper discusses such a SEO mechanism. Having identified that semantic analysis has not been widely applied in the field of SEO, a semantic approach is adopted, which employs Latent Dirichlet Allocation techniques coupled with Gibbs Sampling in order to analyze the results of search engines based on given keywords. Within the context of the paper, the developed SEO mechanism LDArank is presented, which evaluates query results through state-of-the-art SEO metrics, analyzes results' content and extracts new, optimized content.

1 INTRODUCTION

Over the last decade, search engines have evolved from mere indexing tools to a necessity, to all types of web users. Apart from elaborate architectures and computing power, search engines have incorporated a plethora of metrics in order to provide satisfactory results. On the other side of the coin, Search engine Optimization (SEO) techniques appeared. Gradually, with the incorporation of personalized results from the engines, the explosion of Social Media and the creation of real time search engines, SEO has become a field with a multitude of approaches and a great added-value, since it directly affects the Search Engine Result Pages (SERPs). Currently, the majority of web traffic is driven by the search engines of Google, Bing and Yahoo!, which compete towards returning the most relevant results to a user query through the improvement of their web crawling technology and the addition of sophisticated quality factors. SEO works the other way round, trying to optimize websites in order to increase the traffic they receive from search engines and, thus, achieve better rankings. Due to the nature of the web, though, there will be always some technique producing better results and some wrong SEO "action" that may harm website ranking.

Search engines (SEs) explore the web in order to find all the content they can access. Content is in the form of webpages, containing text and links to files, images etc, as well as javascripts and flash content. SEs have specially designed mechanisms called "crawlers" and use the web's link structure in order to perform the crawling. Based on the content retrieved, information is stored and indexed for later reference. When a query is performed, the search engine returns the most relevant results and ranks them according to their importance, which is interpreted as popularity. The popularity of a document containing some content is determined through complex algorithms comprising hundreds of components called ranking factors.

The major search engines provide guidelines to web content developers, in order to be SE-friendly. They advise them to: a) create webpages targeting the users and not search engines, b) include

keywords that are possible queries related to the context of the webpages, c) follow a clear hierarchy and architecture in the website and include links in the text of the webpages and, d) keep a reasonable number of outgoing links from the page.

Nevertheless, search engines do not have an inherent metric for the evaluation of quality. They can promote popularity but they are not able to generate it.

2 RELATED WORK

2.1 Search Engine Ranking Factors

Within the context of the 2011 SMX Advanced Conference, the correlation of critical SE ranking factors (excluding social media related factors) with Google rankings was presented (Figure 1).

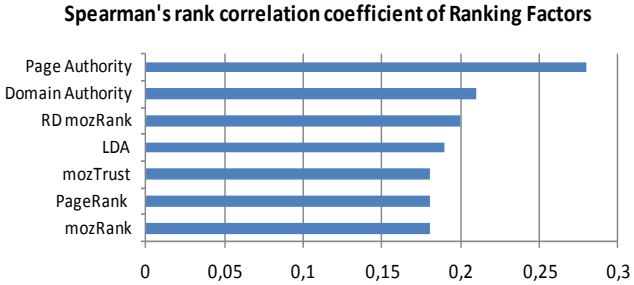

Figure 1: Correlation coefficient of various ranking factors.

Within Figure 1 the following metrics are identified: a) *Page Authority (PA)*, a calculated metric on how high a given webpage is likely to rank in search results regardless of its content, b) *Domain Authority (DA)*, a metric similar to *PA*, but applicable on a domain level, c) *mozRank* (*mR*), a logarithmic scaled 10-point measure of global link authority/popularity (and *RD* means root domain), d) *mozTrust (mT)*, which quantifies the trustworthiness of a webpage to all the other webpages, e) the well known *PageRank (PR)* and f) *Latent Dirichlet Analysis* (*LDA*). In fact, *LDA* (Blei, Ng & Jordan, 2003) appears to have drawn a lot of attention as far as SE ranking is concerned.

Based on this observation, authors argue that content analysis should also be considered significant for webpage ranking and its potential is explored within the context of this work.

2.2 Semantic analysis of web content

Tf-idf (Salton & McGill, 1983), LSI (Deerwester S. et al, 1988), and pLSI (Hofmann, 1999) have been widely applied for performing text processing and analysis. Based on their primitives, Latent Dirichlet Analysis (LDA) was proposed for the probabilistic modeling of collections of discrete data such as text corpora. Each collection item is modeled as a finite mixture of topics, which, in turn, are modeled as an infinite mixture over an underlying set of bayesian probabilities. The parameters of the model can either be defined empirically, or can be identified by employing Gibbs sampling (Griffiths & Steyvers, 2004).

LDA was first reported as a possible factor in SEO by Bishop (Bishop, 2004) and then by Grubber in his GoogleTechTalks (2007). Since *SEOmoz* experiments have indicated a satisfying correlation between LDA and search engine results, we have developed *LDArank*, a mechanism that employs LDA in order to identify the most important topics related to a query. Incorporating these topics into a website corpus would lead to search-engine-optimized content and, thus, higher rankings of the webpage/website in the SERPs of queries related to its topic. Discussion on the mechanism is provided next.

3 THE *LDARANK* MECHANISM

The developed mechanism provides a generic framework for collecting query results from the top search engines and employs all state-of-the-art metrics in order to perform webpage evaluation and select the top results to perform LDA analysis upon. The facets of the *LDArank* mechanism are depicted in Figure 2:

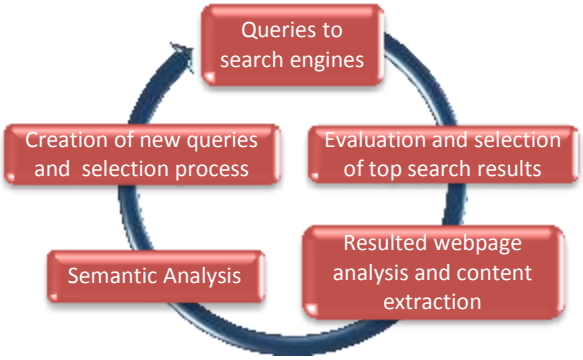

Figure 2: The *LDArank* mechanism

During the first iteration the queries are defined by the user and are submitted to Bing, Yahoo! and Google search engines through their APIs. The results are extracted in JSON format and are analyzed in order to extract the returned URLs. Consequently, evaluation is performed against *PA*, *DA*, *mR*, *sR* (simple Ranking), and the *Visibility Score – VS* (combined ranking), and the top λ results are retained. Next, webpage analysis is performed;

the body of the content along with the anchor text and metadata are extracted, stemming and cleaning is performed, while regular expressions and stop words are removed.

Semantic analysis via *LDA* is performed on the text of the retained results for a given query, in order to recognize the dominant words that compose the dominant content for this query. The output of the semantic analysis is a list that contains the most probable words for the given queries. Based on the most dominant words of the list, a set of queries is designed. All the possible combinations of words are formed, but only the *l* most powerful combinations are retained. This is based on the observation that a user query typically contains a finite number of words. The similarity of the new queries with the original query is calculated in means of NGD-Normalized google distance (Cilibrasi and Vitanyi, 2007), and the top *K* queries are selected.

The above process (of creating queries and identifying dominant words) is repeated until list of words from the current round of semantic analysis contains at least *β%* of common words with the previous round.

All the mechanism parameters are defined through a configuration file, which is parsed during the initiation phase of the mechanism. Table 1 depicts the configuration parameters:

Table 1: *LDArank* configuration parmeters.

Input Parameters
- User query ($q_1, q_2, \ldots q_n$)
- Search engine results threshold (λ)
- *LDArank* topics of analysis threshold (τ)
- *LDArank* beta parameter (β)
- *LDArank n*umber of iterations *(M)*
- *LDArank* number of top words/topic (*α*)
- *LDArank* probability threshold (*ξ*)
- *NGD* threshold (*K*)
- Maximum words itemset (*l*)
- Convergence limit (*cl*)
- Performance limit (*pl*)
- Type of SE employed (Google, Bing, Yahoo! , all)
- *SEOmoz* metric, (*mR, external mR, PA, DA, VS,* all*)*

4 EXPERIMENTS AND RESULTS

In order to provide evidence on the applicability of our model, we discuss an indicative test case. Let's assume that a web content provider would like to set up a website on Software Engineering practices. In order to increase website visibility, and given the preceding analysis on the importance of website content in SE ranking, he/she would like to identify the dominant keywords that he/she should use, in order to achieve his/her goal. To this end, *LDArank* is employed. The following analysis provides a set of experiments and conclusions; nevertheless one may perform an even wider range of experiments, by tuning any of the *LDArank* mechanism parameters.

4.1 Experiment setup

Various alternatives have been explored in order to illustrate *LDArank* versatility and ease-of-use (some omitted due to space limitations). The aims of the experiments were to: a) identify whether the size of the resulting word cloud is related to SE ranking of webpages, b) identify whether the type of words in residing in a webpage (generic or more specialized) affects SE ranking and, c) to evaluate the convergence capabilities of all the metrics considered.

To this end, two sets of terms are considered for the analysis: a) a set comprising 15, more generic terms on Software Engineering and b) a set comprising 40 terms, more focused on software engineering processes.

Parameters *M, K, l, cl,* and *pl* were kept constant in the performed *LDArank* experiments. The experiments performed had varying values with respect to *α, λ, ξ* and *τ*, and were evaluated against the core SE metrics identified: *sR*, *mR*, *PA*, *DA*, *VS*, *mR* with merged engine (*mRm*), *PA* with merged engine (*PAm*), and *DA* with merged engine (*DAm*).

4.2 Results

Experiments run on the first set of terms (generic) resulted into three groups, according to the size of word cloud generated, with respect to the values of the number of topics, number of top words and the probability threshold. These groups are: a) Group A – a small scale group, b) Group B – a medium scale group and, c) Group C – a large scale group.

Group A produced a total of 44 words, group B 554 words, and group C 921 words. Comparing the top words of group A against the top words of the other two groups (Figure 3) it can be argued that the groups are well separated.

Group C produced more content than the group B, which produced more content than group A. Moreover, group C's produced content is characterized by more variety in contrast to the other two groups and the top words of the content produced, in terms of occurrences, in the small-scale case are top words in both the medium and large-scale cases.

Figure 3: Group A, B and C word clouds.

Group D was built from the second set of terms (specialized) and produced a total of 143 words. Comparing group D to group C, 39, 14, and 12 words are the same, out of 100, 40, and 20 top words, respectively. It is, thus, obvious specialized content leads to different rankings in search engines.

Figure 4: Group D word cloud.

From the mean value and the standard deviation of convergence of each evaluation metric per group, it could be stated that using merged results led to lower convergence percentages. *PAm* presented the highest convergence percentage and *PA* led to high convergence percentages in the medium scale cases. It should be mentoned that *sR* led to high convergence percentages in small-scale and large scale cases, *mR* had high convergence percentages in the medium-scale cases and *DA* had high convergence percentages in the medium and large scale cases. Therefore, *Pam* and *sR* seem to be the most efficient evaluation metrics that confirm the high value of Spearman's rank correlation coefficient of them and the new search engines' trends and updates.

5 CONCLUSION

In this paper, a new mechanism for the optimization of website ranking in search engines based on Latent Dirichlet Allocation with Gibbs sampling. LDA is used in a different approach in our model and the results of the experiments run using the proposed model confirm the search engines' latest trends regarding the Google Panda Updates towards a more content based ranking and a focus on domains by considering domain-level metrics to be equally important to page-level ones. Furthermore, the model reveals a detail about the engines' algorithm about the top results of their search engine results pages.

The next step for the evaluation of the proposed mechanism is the application of it on a website in order to confirm the level of benefit it provides to the production of optimized content and the effect of it on the website's rankings in the SE results pages.

REFERENCES

Grubber A., Rosen-Zvi M., Weiss Y. 2007. Hidden Topic Markov Models, *Artificial Intelligence and Statistics (AISTATS)*..

Blei D.M., Ng A.Y., Jordan M.I. 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research* vol.3, pp. 993-1022.

Griffiths T.L., Steyvers M. 2004. Finding scientific topics, *In Proceedings of the National Academy of Sciences of U.S.*,101(1), pp. 5228–5235, 2004.

Bishop C. M. 2004. Recent Advances in Bayesian Inference Techniques, *Proceedings of SIAM Conference on Data Mining (keynote speech)*.

Salton G., McGill M. J. 1983. *Introduction to modern information retrieval*, McGraw-Hill.

Deerwester S. et al 1988. Improving Information Retrieval with Latent Semantic Indexing, *In Proceedings of the 51st Annual Meeting of the American Society for Information Science*, 25, pp. 36–40.

Hofmann T. 1999. Probabilistic Latent Semantic Indexing, *In Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57.

Cilibrasi, R., Vitanyi, P. 2007. The Google Similarity Distance, *IEEE Trans. Knowledge and Data Engineering*, 19(3), pp. 370-383.