Engineering Applications of Artificial Intelligence 35 (2014) 114–130

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Semantic analysis of web documents for the generation of optimal content

Themistoklis Mavridis [a], Andreas L. Symeonidis [a,b,*]

[a] Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, GR 54 124, Greece
[b] Intelligent Systems & Software Engineering Laboratory, Information Technologies Institute, Center for Research and Technology Hellas, GR570 01, Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 10 September 2013
Received in revised form
27 March 2014
Accepted 9 June 2014
Available online 16 July 2014

Keywords:
Latent Dirichlet allocation
Semantic analysis of content
Search engines semantic analysis
SERPs evaluation

ABSTRACT

The Web has been under major evolution over the last decade and search engines have been trying to incorporate the changes of the web and provide the user with improved – in terms of quality – content. In order to evaluate the quality of a document there has been a plethora of attempts, some of which have considered the use of semantic analysis for extracting conclusions upon documents around the web. In turn, Search Engine Optimization (SEO) has been under development in order to cope with the changes of search engines and the web. SEO's aim has been the creation of effective strategies for optimal ranking of websites and webpages in search engines. Current work probes on semantic analysis of web content. We further elaborate on LDArank, a mechanism that employs Latent Dirichlet Allocation (LDA) for the semantic analysis of web content and the generation of optimal content for given queries. We apply the new proposed mechanism, *LSHrank*, and explore the effect of generating web content against various SEO factors. We demonstrate *LSHrank* robustness to produce semantically prominent content in comparison to different semantic analysis based SEO approaches.

1. Introduction

There have been major changes on the ranking schemas of search engines over the last two years.[1] Google, Bing and Yahoo!, have started focusing on users and provide them with high quality content in order to get valuable user feedback. In parallel, Search Engine Optimization (SEO) has been under constant changes in order to capture the up-to-date search engine ranking strategies. This way SEO aspires to assist websites in achieving higher and better related rankings.

Search engines (SE) inherently are not able to produce popular content, only to promote it. Considering this fact, SEO strategies will always be effective, given that they recognize the factors that search engines value for their rankings and exploit them as much as possible.

Up until recently these factors were machine-related, i.e. metrics calculated by machines, not users. Lately, major SE players have been trying to promote the quality of the projected content to the users as the major characteristic for website evaluation.

Koningstein (2012) mentions a number of techniques followed by Google in order to identify documents that have been modified (in terms of content) for achieving higher rankings in a spamming fashion. Additionally, Lamping and Pearson (2011) discuss methods that define the quality of documents according to their semantic relation with others and with given queries.

The major search engines have published their guides with some basic SEO suggestions. The guides of Google,[2] Bing[3] and Yahoo![4] portray technical differences, but they all conclude that content should be the focus of SEO. Content should be unique, of high quality and user oriented.

Nevertheless, due to the fact that search engines are not transparent regarding their ranking schemas, conclusions about them are only extracted through their result pages analysis and by exploration of their patents. Several organizations and companies have defined their own metrics for the evaluation of web documents from various perspectives. Moz[5] (former SEOmoz) metrics

* Corresponding author at: Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, GR 54 124, Greece. Tel./fax: +30 2310 99 4344.
E-mail addresses: themis.mavridis@issel.ee.auth.gr (T. Mavridis), asymeon@eng.auth.gr, asymeon@iti.gr (A.L. Symeonidis).

[1] http://moz.com/google-algorithm-change (last accessed 10/03/2014)

[2] https://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf (last accessed 10/03/2014)

[3] http://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a (last accessed 10/03/2014)

[4] http://web.archive.org/web/20130501164432/http://styleguide.yahoo.com/resources/optimize-search-engines/seo-basics (last accessed 10/03/2014)

[5] http://moz.com/ (last accessed 10/03/2014)

http://dx.doi.org/10.1016/j.engappai.2014.06.008
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

are very popular and are employed by a plethora of users in order to analyze their web presence and evaluate the popularity of their websites.

For the generation of semantically optimal content, we have previously concluded that Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has proven effective (Mavridis and Symeonidis, 2012). LDA's added-value in SEO techniques is also strengthened by Patterson's (Patterson, 2012) analysis on the mechanism that Google employs in order to relate query phrases to documents for indexing, retrieval, description and analysis. The mechanism focuses on the extraction of top phrases out of documents and groups of documents, and calculates a score for the importance of each phrase and a total score for the top phrases; this is, practically, LDA's core objective.

Further advancing on this context, current work attempts to assess LDA's influence for content production with respect to search engine rankings. We argue that the domain of a website and the metrics employed for ranking web documents is correlated to the effect of LDA on content generation. *LSHrank* extends our previous approach (Mavridis and Symeonidis, 2012) and has been created in order to analyze search engine result pages (SERPs) and draw conclusions regarding the correlation of SE ranking schemas and the semantic analysis of content. Assessment is performed against Moz metrics, and the algorithm is tuned in order to produce the optimal content. The remainder of this paper discusses the state-of-the art regarding semantic analysis in SEs and presents the LDA's relation to SEO. Section 3 describes the architecture of *LSHrank*, while Section 4 discusses the experiments performed in order to assess LDA various parameters. Section 5 explores in depth LDA and *LSHrank* performance with thorough experiments based on the results of Section 4. Furthermore, Section 6 presents a set of experiments run to compare *LSHrank* against other semantic analysis SEO approaches. Finally, Section 7 summarizes work performed, probes on future work and concludes the paper.

2. Related work/background theory

In the following section we discuss state-of-the-art on semantic analysis in web search. There exist multiple approaches, from capturing user input and using ontologies to identifying concepts and performing semantic content analysis. Besides discussing state-of-the-art in all these approaches, we further elaborate on semantic content analysis, the use of LDA as a semantic analysis algorithm and the novelties of our approach.

2.1. User input to define semantics on content

One of the dominant approaches in performing semantic analysis of web content largely exploits user input. Bakir and Kulshreshtha (2012) provide details on Google's intention to capture user-experience through the `Panda Updates`, with the first update[6] being announced by Google early 2011. Authors describe the deployment of human reviewers in order to capture problems that have to do with user-experience on web content and on data that an information system stores and indexes. In the same context, Lawrence et al. (2005) presented a system on capturing user-related signals in order to determine relevance scores to the search results of Google. Going even further, Google attempts to perform hierarchical categorization of search results in topics by capturing user input, as described by Sandler et al. (2012). Moreover, Pedersen et al. (2012) present the way Google uses access history of a user on content in order to calculate semantic similarities using topic vectors for the content of documents.

Lei et al. (2006) follow a different approach; they propose *SemSearch*, a search engine that provides users with precise answers to their queries through a simple interface. *SemSearch* captures user input and creates semantic entities in order to provide them with the desirable result. Still on the context of personalized search, Teevan et al. (2005) propose an automated information retrieval mechanism that exploits user search queries and their browsing history in order to provide efficient client-side algorithms. In a similar fashion, Yin and Shah (2010) of Microsoft have proposed a methodology in order to use search logs into recognizing the user intent and semantics in search queries. The representative terms form a tree built based on directed maximum spanning, hierarchical agglomerative clustering and pachinko allocation model. The use of human raters via Amazon's Mechanical Turk update[7] verified the methodology's success on recognizing significant representative terms.

2.2. Recognition of documents concepts and entities

Another approach widely applied strives to recognize entities and concepts related to documents. Hong et al. (2013) presented a methodology that identifies entities highly correlated to search query entities. It is based on various semantic factors that are not related to keywords nor link structure, indicating that Google can identify the semantic relatedness of results to entities.

Van Durme and Pasca (2008) discuss a method aiming to extract large numbers of concept classes and corresponding instances from documents using distributional similarity metrics based on the logic of Pantel and Ravichandran (2004). Following this methodology, Pasca (2013) analyzed Google's approach to recognize semantic classes/concepts in documents by defining frequency and diversity scores on class-instance pairs, linked through “is-a” relationships.

Lee et al. (2013) described a framework of Google used to recognize candidate terms for an entity by analyzing the documents related to the entity. The proposed framework employs measurement of the frequency of the appearance on various terms in web documents, focuses on the “known for” terms and recognizes the semantics of an entity in order to identify the important terms/words related to it.

In the same context, Zhou et al. (2005) present the ability that Google has in recognizing entity names in search results and in altering these results by providing documents that are related to these entities. On the other hand, Mishne et al. (2009) discuss how Yahoo! can identify the most probable interpretation of a query in order to rank the documents according to their semantic relevance to the interpretation chosen. As Padovitz and Nagarajan (2012) state, Microsoft has also been exploring the field from the entities perspective, attempting to define the complexity of extracting an entity identification from a corpora of documents.

Finally, Federici (2013) discuss how Google can calculate the transition probabilities that represent the relationship of the entities. In his approach, the term “entity” consists of the query, the documents related to the query, the search session related, the time of the query submission, anchor texts in the links of the documents, the advertisements presented and the associated domains of the documents. The calculation of the probabilities considers all the information of the entities in order to rank search results in an accurate fashion, identify semantic relation among queries and to improve the results of vertical search features. In general, the probabilities express the strength of the relationships of entities and the use of user behavior data.

[6] http://googleblog.blogspot.com/2011/01/google-search-and-search-engine-spam.html (last accessed 10/03/2014)

[7] https://www.mturk.com/mturk/ (last accessed 10/03/2014)