STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LIX**, Number 1, 2014

A STUDY REGARDING INTER DOMAIN LINKED DOCUMENTS SIMILARITY AND THEIR CONSEQUENT BOUNCE RATE

DIANA HALIŢĂ AND DARIUS BUFNEA

Abstract. Then main objective of linking inter domain documents is to offer to the reader access to supplementary, semantic related information. However, linking web domains is sometimes artificially used, especially when the goal is to abusively increase the page rank of the destination domain. This paper presents a study regarding inter domain linked documents similarity and their consequent bounce rate. For that, we have advanced a series of experiments which outlines how similarity functions' behavior correlates with a website bounce rate. The method presented here could be used to identify within a web site improper placed outgoing links such as ads or spam links. Based on that, a search engine could fine the results in SERP by downgrading any website that fall in the above presented category.

1. Introduction

In a certain document context, one of the goals of referring a document within another one is to give to the reader access to more semantic related information to the information being currently read. Two common examples in this direction are a scientific article citing one of its references or a web page linking naturally to another web page.

When it comes to linking inter domain web pages (i.e. HTML links that points to documents hosted on a different domain), often the linking is performed in an abusive manner in order to artificially increase the page rank of the destination document or domain and not to lead the visitor to a more semantic related information to the one he is currently interested in.

Received by the editors: April 18, 2014.

2010 *Mathematics Subject Classification.* 68U15, 68M11, 68U35.

1998 *CR Categories and Descriptors.* H.5.3 [**Group and Organization Interfaces**] - Web-based interaction; H.5.4 [**Hypertext/Hypermedia**] - Navigation; I.7.5 [**Document Capture**] - Document analysis.

Key words and phrases. bounce rate, document similarity, page ranking, identify improper placed links.

In the majority of cases, such abusive links are either site wide (for example ads), these one being the most easy to detect, or it might be automatically placed inside the absolute content of a web page (i.e. the article/post's effective content) by different advertising modules integrated in the Content Management System of the source web site.

This paper analyzes the possible relation between the similarity of the source and the external linked document and the bounce rate generated within the destination domain by this link. Such a possible relation could lead to bounce rate estimation and disclosure for any external link between any two third party web sites. A high estimated bounce rate for the destination site via such a link may indicate an improper link, i.e. an abusive, spam or ad link.

2. Analysis of bounce rate in relation with inter linked documents' similarity

Bounce rate represents the percentage of the visitors who enter a site and leave it rather than continue visiting other pages. Generally this have two meanings. First of all, one may find the exactly desired information. For example, a user search the definition of bounce rate and this information is provided accurately by the first web page in SERP (Search Engine Results Page). He is satisfied with the provided definition and he leaves the site without accessing any other result page. However, in most cases the user may not be satisfied with information shown on a particular web page in SERP and he may leave it immediately in order to access the next one returned in SERP. It is widely accepted, that a greater bounce rate, mainly due to the second case, means something negative and that value is directly associated with the quality of the content.

The higher similarity between the outgoing web page and the external referred page is obtained, the more similar content is provided to the user. This leads to a smaller bounce rate for the destination domain, the visitor having much more to read about what he is interested in.

Example:

- An Internet based forum dedicated to pets lover contains a topic with a link that points to a dog raising website. Such a link offers a plus of information to a visitor being susceptible of generating a smaller bounce rate for the destination web site.
- The admission web site of our university contains basic information about the admission process. More detailed information about the admission process is available via a web link on each faculty's web site

which is hosted on a different domain. Such a link will also offer additional semantic related information to the reader as the information he is interested in, generating a smaller bounce rate for a certain faculty's web site.

Counterexample:

- An abusive link, such as a spam or an ad link, in many cases points to a site / external page having a content that is not semantic related with the source. Visitors that follow such a link will consequently generate a higher bounce rate for the destination site: i.e. will click the link, access the destination web page and then close it or return back to the previously accessed web page. In fact, targeted advertising was introduced in order to increase conversion from the advertised site point of view [3].

In order to analyze the possible connection between linked documents similarity and their consequently generated bounce rate we will test in the respect of this some basic similarity functions such as: Cosine, Jaccard, Sorensen and Jaro-Winkler. In a working paper, we will also address the same topic using semantic similarities.

3. Previous Work

As the Internet have evolved and surfing the Web, as its main application, become a common and a daily based activity for the society, the spam linking spread as a negative phenomenon that affect mainly the quality of search results return by a search engine. The biggest companies involved in the search industry and equally the research community have step up their efforts in order to identify solutions to detect and limit this negative phenomenon's impact.

Ever since link analysis was used in building search engines optimization algorithms, corresponding spamming techniques have been developed [9]. As a consequence, multiple negative effects were caused by spamdexing. This effects lead to the appearance of new challenges in this area research. In a comprehensive survey on the main principles on web spam detection [8], authors had categorize all existing techniques into three categories based on the type of information they use: content-based methods, link-based methods, and methods based on non-traditional data analysis such as user behavior, clicks and HTTP sessions. These techniques are able to detect up to 80% of spam pages [2] and they should be applied together, at least by combining link based and content based analysis [1]. Previous studies related to web spam detection were using automatic supervised or unsupervised classification [1], power-law distribution, algorithms for collusion detection [4] or confusion

matrix and precision-recall matrix [7]. That was a key challenge for search engine industry, so spamdexing was cast into a machine learning problem of classification on directed graphs [9].

4. Predicting bounce rate using external linked documents' similarity

4.1. **Ideal similarity.** Based on the above observations, that high content related web pages are less susceptible for generating a high bounce rate and poor content related pages can lead to a higher percent of visitors that bounce, we intend in this section to study the possible correlation between bounce rate and several similarities functions. In the case of an ideal similarity function, depicted in figure 1, the bounce rate will follow a linear regression path.

By using a properly chosen similarity function, one could estimate the bounce rate of a third party external link based only on content of linked documents' similarity (currently, bounce rate may only be obtain by a site's owner using web analytics tools such a Google Analytics). Such a method could be used by a third party (for e.g. the search engine) to identify improper placed and abusive links such as ads or spam links. Consequently, a web site that relies on a large number of links that fall in the above categories could be fined by a search engine.

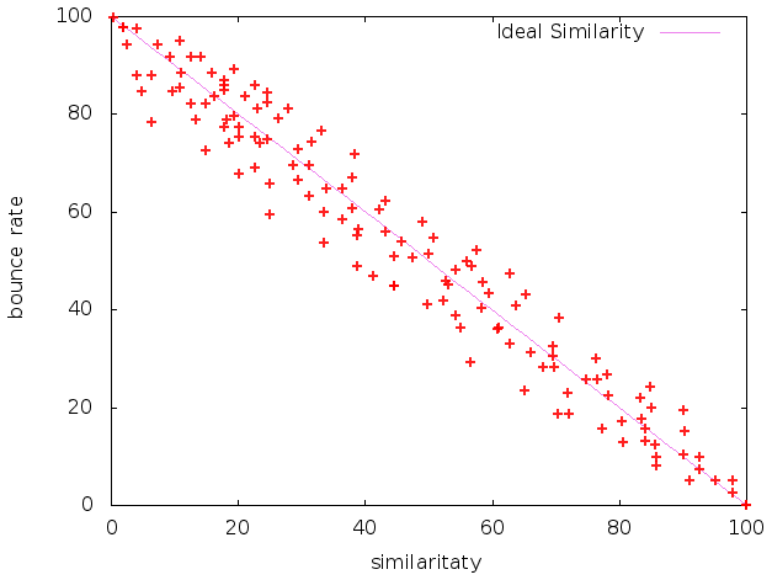

Figure 1. Ideal similarity

The perfect correlation between bounce rate and similarity of inter domain linked documents, shown in an ideal case [fig:1], is not always found, especially considering usual mathematical similarity functions. We choose to take into consideration more similarity functions in order to determine which one fits better our goal. Intuitively, we say that a similarity function is better than other if the pairs (similarity, bounce rate) obtained from each link between two inter domain linked pages are closer to the diagonal (as we can see in the an ideal case [fig:1]). In our case, similarity represents the horizontal coordinate (abscissa) of the point in a two-dimensional rectangular Cartesian coordinate system and bounce rate represents the vertical coordinate (ordinate) of the above considered system. Taking into consideration the line which is parallel with the secondary bisector of the Cartesian coordinate system and passes through the points $(100, 0)$ and $(0, 100)$, a similarity function is the best if the sum of all distances, from the graphical representation points, represented by the above named pairs, to that line is minimal, i.e. if

$$\sum \frac{|x_i + y_i - 100|}{\sqrt{2}}$$

is minimal, where x_i and y_i are the coordinates of a point on the graphical representation.

4.2. **Experimental results.** Various similarity functions can be used for determining whether two strings are similar. In this paper, we have tested all gathered data against the following similarity functions: Cosine Similarity, Jaro-Winkler Similarity, Sorensen Similarity and Jaccard Similarity. The Cosine Similarity measures the angle between two vectors, but it does not take into consideration the order of the strings. Jaro-Winkler similarity is a lot more accurate especially because it takes into consideration the order of the strings, but it is designed and best suited for short strings. The Jaccard coefficient is defined as the length of the intersection divided by the length of the union of the two sets of strings. An important class of problems, that Jaccard similarity addresses well, is finding textually similar documents in a large corpus, such as the Web. Sorensen similarity coefficient is similar to Jaccard coefficient, but it has some different properties, including retaining sensitivity in more heterogeneous data sets and gives less weight to outliers.

In order to have the best possible accuracy of the results and for reducing the noise induced in the similarity algorithm by the master pages' HTML code, we have tested all the similarity functions for the absolute content of the documents (i.e. we ignore all the content that falls within the footer, header or menu).

The absolute content of a web page was completely determined using a Java library named *boilerpipe*, which is able to remove or extract full text from HTML pages. This library provides algorithms which detects and removes the master page template around the main content of a web page. This library was released under Apache License 2.0.

In order to follow our ideas we took for experimental purpose, an educational website $http://www.cs.ubbcluj.ro$ (Computer Science Faculty website), and we generated all the triplets $(landingpage, refferer, bounce\ rate)$ through two methods:

- a client side tool provided by Google, named Google Analytics;
- a server side tool, integrated in the website's master page template, developed by us.

We have chosen to experiment on the above website because we have administrator rights on the faculty website and we may properly measure for each external incoming link the generated bounce rate.

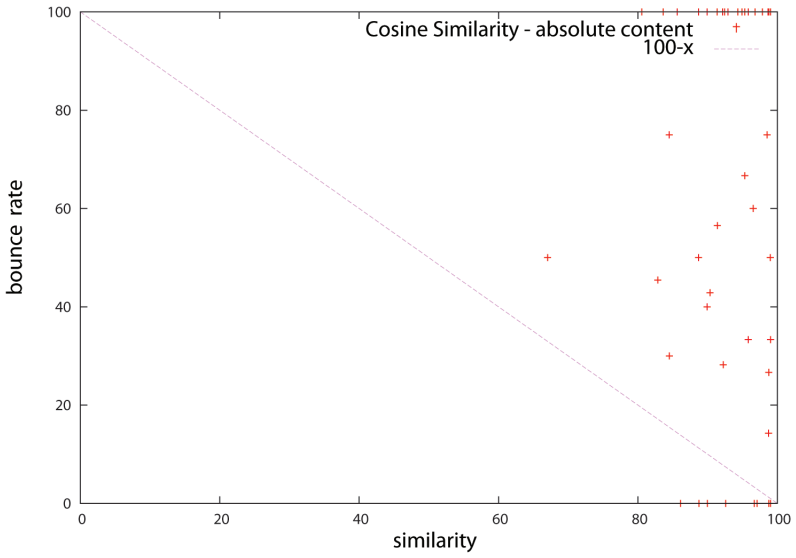

FIGURE 2. Cosine similarity

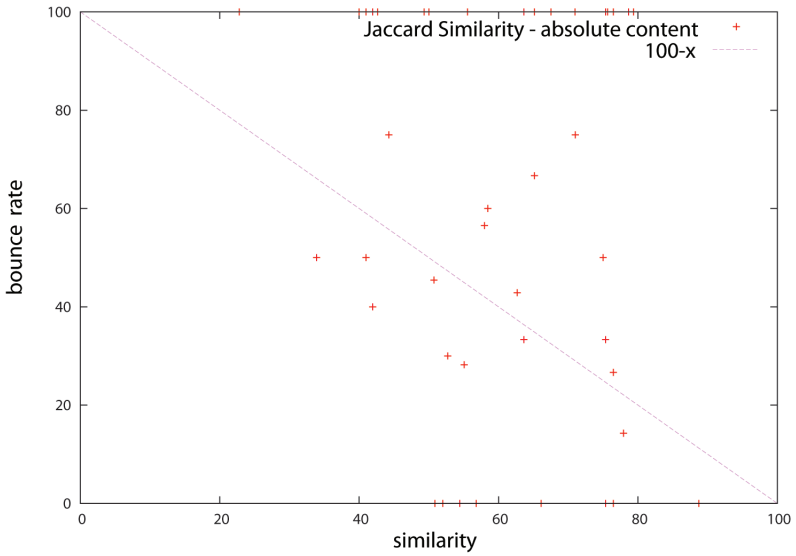

FIGURE 3. Jaccard Similarity

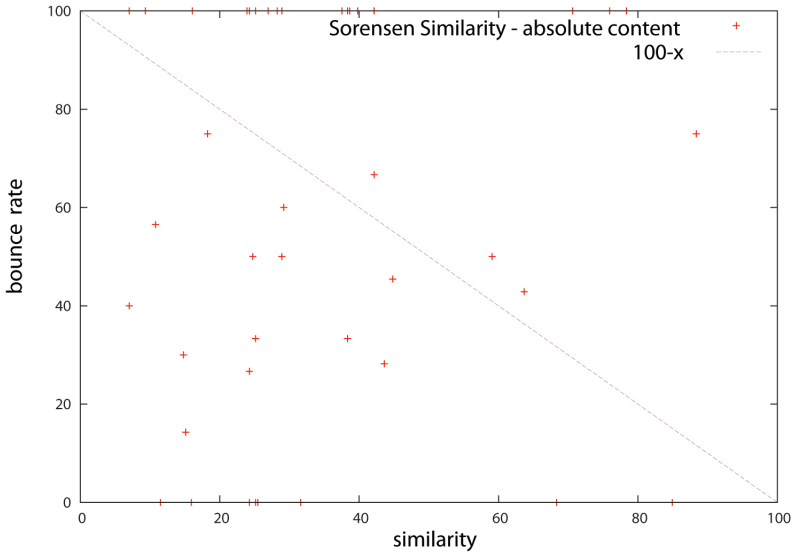

FIGURE 4. Sorensen similarity

As we can see in the above figures, the best similarity function which gives us the expected results is Jaccard distance applied only to the absolute

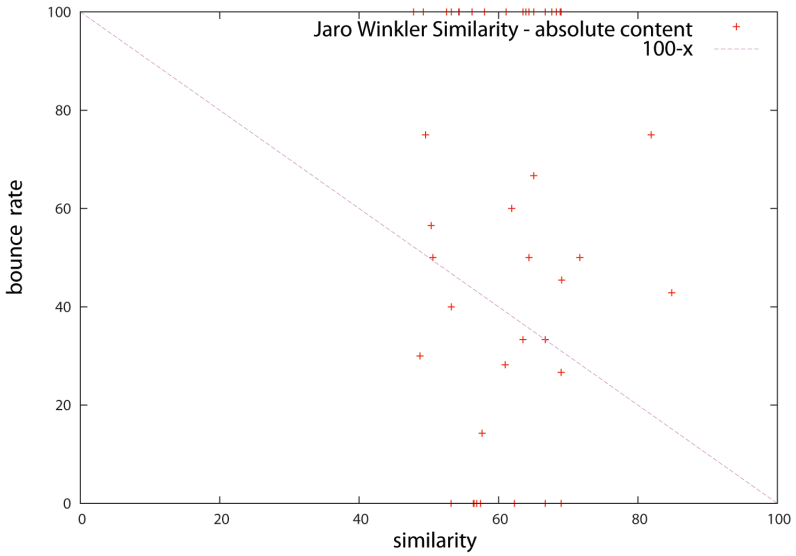

FIGURE 5. Jaro-Winkler Similarity

contents of the web pages. Given this, we will use this similarity function for our future graphical representations.

We also can prove that Jaccard is the best similarity function in our case, by calculating for each similarity function used and for both absolute content and full content comparison, the distance from all points which are represented on the figure to the line of equation $y = x - 100$.

Similarity function	Method	Value of the Sum
Cosine	absolute content	1804.476000579978
Jaccard	**absolute content**	**1414.03085464388**
Sorensen	absolute content	1769.3699189543242
Jaro-Winkler	absolute content	1528.5359097516346

TABLE 1. Sum of the distances from all points to the line of equation $y = x - 100$

5. Conclusions and Future work

In this paper we have advanced a series of experiments in order to see whether the bounce rate correlates with different similarity functions. A high content similarity between linked and source content may imply future visitor

requests for pages hosted on the destination domain, these action reducing the bounce rate implied by such links. Starting with this idea it is possible, as future work, to obtain better results if we analyze the similarity of the referrer's page content with the content of each internal link found in the corresponding landing page.

Future work may also imply testing this ideas on semantic similarity functions. Other ideas would be giving different weights to similarity functions or choosing a similarity function and weighting and fine tuning various specific properties of the content, such as URL, page headings, page title or keywords.

A work in progress paper of the same authors is studying scrapper sites identification based on their content similarity with the content they automatically retrieved and usually link to.

References

[1] L. Becchetti, C. Castillo, D. Donato, *Link-Based Characterization and Detection of Web Spam*, 2nd International Workshop on Adversarial Information Retrieval on the Web, AIRWeb, Seattle, USA, August 2006, pp. 1-8

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi,R. Baeza-Yates, *Link Analysis for Web Spam Detection: Link-based and Content Based Techniques*, ACM Transactions on the Web (TWEB), Volume 2, Issue 1, New York, USA, February 2008, pp. 1-41

[3] A. Farahat, M. Bailey, *How Effective is Targeted Advertising?*, Proceedings of the 21st World Wide Web Conference 2012, Lyon, France, April 16-20, 2012, pp. 111-120

[4] Z. Gyongy, H. Garcia-Molina, P. Berkhin, J. Pedersen, *Link Spam Detection Based on Mass Estimation*, 32nd International Conference in Very Large Data Bases (VLDB), Seoul, Korea, 2006, pp. 439-450

[5] A. Huang, *Similarity Measures for Text Document Clustering*, Proceedings od the New Zealand Computer Science Research Student Conference, Hamilton, New Zealand, 2008, pp. 49-56

[6] J. Leskovec, A. Rajaraman, J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2010

[7] M. Najork, *Detecting Spam Web Pages through Content Analysis*, International World Wide Web Conference Comittee, Edinburgh, Scotland, 2006, pp. 83-92

[8] N. Spirin, J. Han, *Survey on Web Spam Detection: Principles and Algorithms*, ACM SIGKDD Explorations Newsletter, Volume 13, Issue 2, December 2011, pp. 50-64

[9] D. Zhou, C. Burges, T. Tao, *Transductive Link Spam Detection*, Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web, AIRWeb, New York, USA, ACM Press, 2007, pp. 21-28

Babeş-Bolyai University, Faculty of Mathematics and Computer Science, 1 M. Kogălniceanu St., 400084 Cluj-Napoca, Romania

E-mail address: diana.halita@ubbcluj.ro, bufny@cs.ubbcluj.ro