Analyzing the relation between bounce rate and document similarity

Diana-Florina Halițǎ

Faculty of Matemathics and Computer Science
Babeș-Bolyai University

June 4th, 2014

Table of Contents

1. Abstract
2. Why?
3. Bouncerate
4. Bounce rate and similarity
5. Previous work
6. Results

Abstract

- Linking inter domain documents:
  - main objective: offering access to supplementary, semantic related information
  - is sometimes artificially used
- The study:
  - experiments which outlines how similarity functions' behavior correlates with a website bounce rate
  - identify improper placed outgoing links
  - downgrading website in SERP

Why?

- the linking is performed in an abusive manner
  - increasing page rank of the destination document
  - not leading the visitor to a more semantic related information to the one he is currently interested in
- abusive links are
  - located site wide (i.e. ads); easy to detect
  - automatically placed inside the absolute content of a web page

About bouncerate

Definition

Bounce rate represents the percentage of the visitors who enter a site and leave it rather than continue visiting other pages.

Meaning:

1. one may find the exactly desired information, so he leaves the site without accessing any other result page.

Fact

The definition of bounce rate is provided accurately by the first web page in SERP

About bouncerate

Definition

Bounce rate represents the percentage of the visitors who enter a site and leave it rather than continue visiting other pages.

Meaning:

2. finding unsatisfactory information; one may leave the site immediately in order to access the next result in SERP.

Fact

A greater bounce rate means something negative and that value is directly associated with the quality of the content.

About bouncerate - example

Links that generates small bouncerate:

- a forum dedicated to pets lover linked to a dog raising website.
- the admission web site of our university linked to all faculty's web site

About bouncerate - counterexample

Links that generates a higher bouncerate:

- An abusive link which points to a site $\Rightarrow$ users will click the link, access the destination web page and then close it or return back to the previously accessed web page.

Bounce rate and similarity

Testing all gathered data against the following similarity functions:

1. Cosine Similarity
   - measures the angle between two vectors
   - it does not take into consideration the order of the strings
2. Jaro-Winkler Similarity
   - is a lot more accurate especially because
   - it takes into consideration the order of the strings
   - it is designed and best suited for short strings

Bounce rate and similarity

Testing all gathered data against the following similarity functions:

3. Jaccard Similarity
   - is defined as the length of the intersection divided by the length of the union of the two sets of strings
   - is designed to find textually similar documents in a large corpus, such as the Web
4. Sorensen Similarity
   - is similar to Jaccard coefficient
   - it has some different properties, including retaining sensitivity in more heterogeneous data sets and gives less weight to outliers.

Previous work

- surfing the Web has become a common and a daily based activity for the society
- the spam linking spread as a negative phenomenon that affects mainly the quality of search results return by a search engine
- corresponding spamming techniques have been developed

Categories based on the type of information they use:

- content-based methods
- link-based methods
- methods based on non-traditional data analysis such as user behavior, clicks and HTTP sessions

Previous work

Fact

Spamming techiques were able to detect up to 80% of spam pages and they should be applied together, at least by combining link based and content based analysis

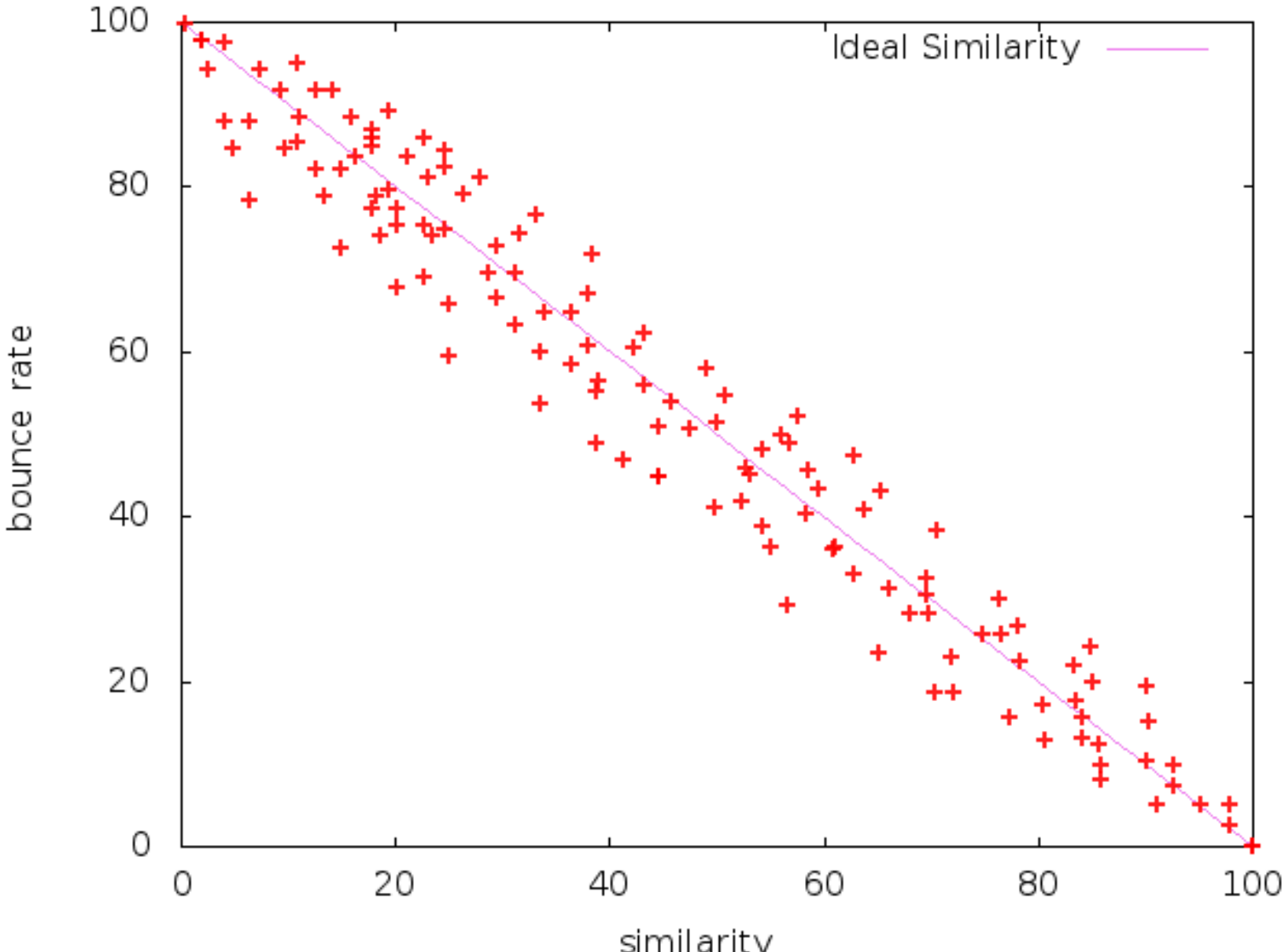
Fact

Other spamming techniques:

- *automatic supervised or unsupervised classification*
- *power-law distribution*
- *algorithms for collusion detection*
- *confusion matrix and precision-recall matrix*

Spamdexing was cast into a machine learning problem of classification on directed graphs.

Ideal similarity



Ideal similarity

- The perfect correlation between bounce rate and similarity is not always found
- take into consideration more similarity functions
- Intuitively, a similarity function is better than other if the pairs (similarity, bounce rate) are closer to the diagonal
- A similarity function is the best if the sum of all distances, from the points (similarity, bounce rate) to the $y = 100 - x$ line is minimal, i.e. if

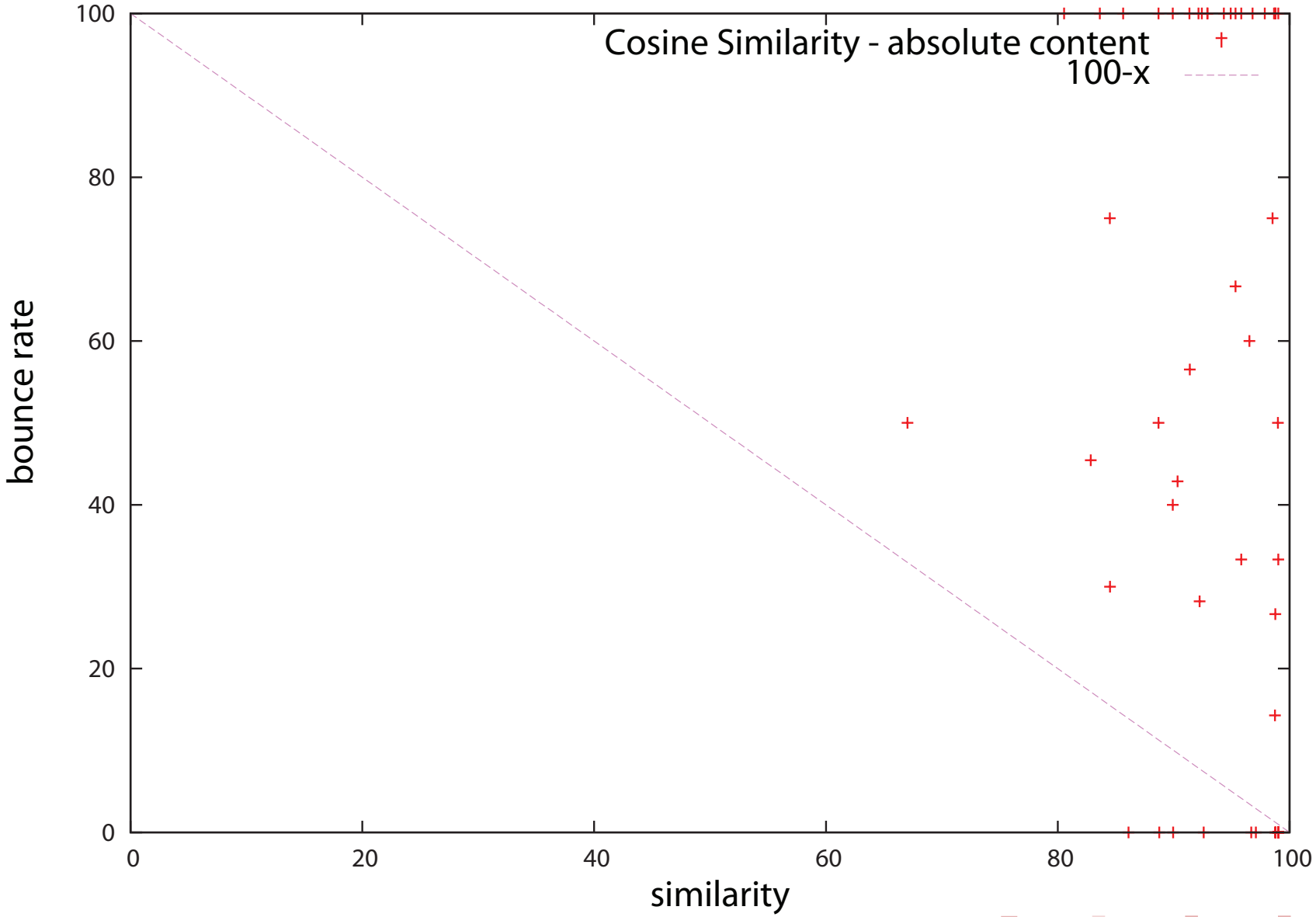
$$\sum \frac{|x_i + y_i - 100|}{\sqrt{2}}$$

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

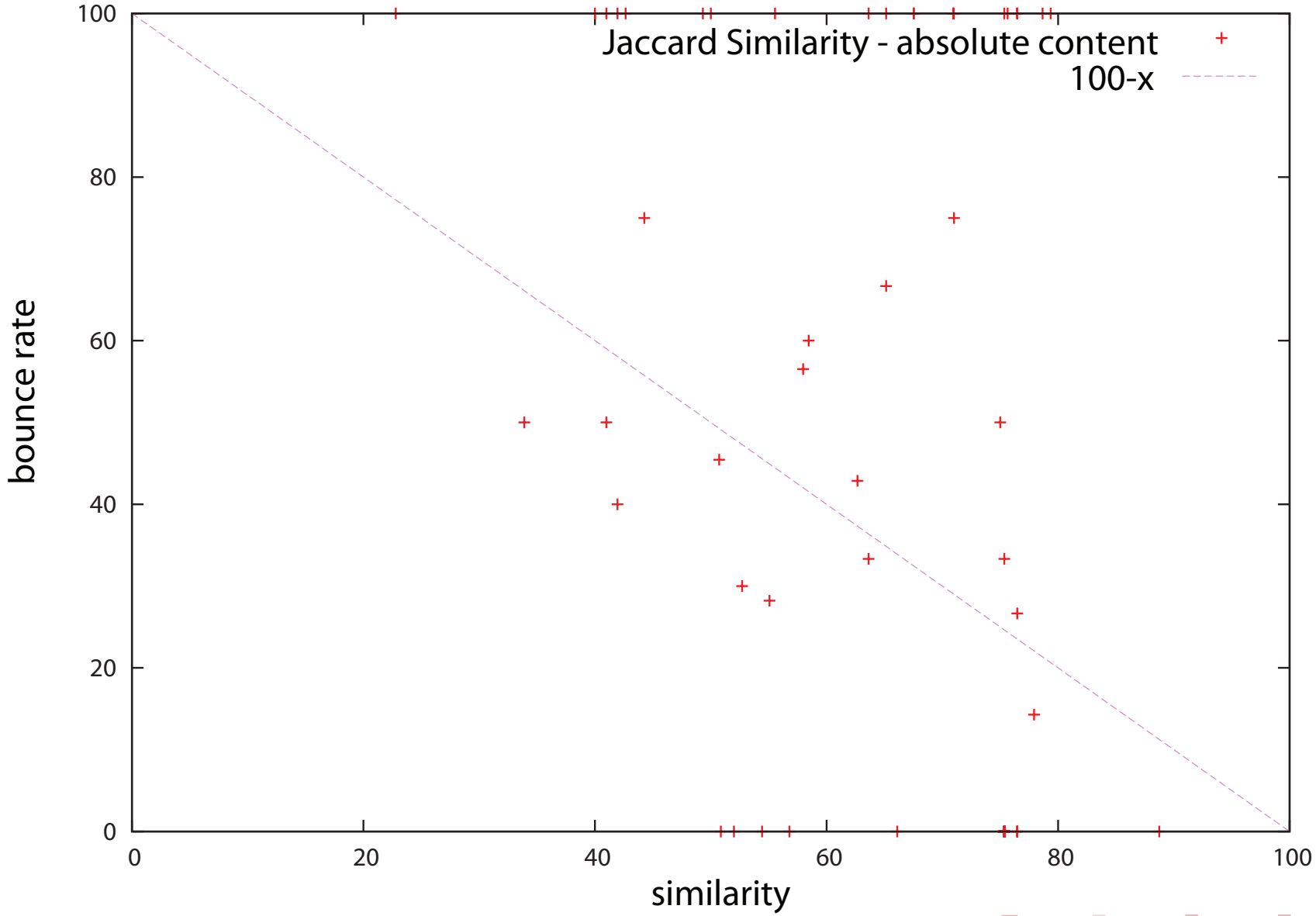
Experimental results

- all similarity functions were tested using the absolute content of a document
- source: an educational website $http://www.cs.ubbcluj.ro$ (Computer Science Faculty website)
- 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.

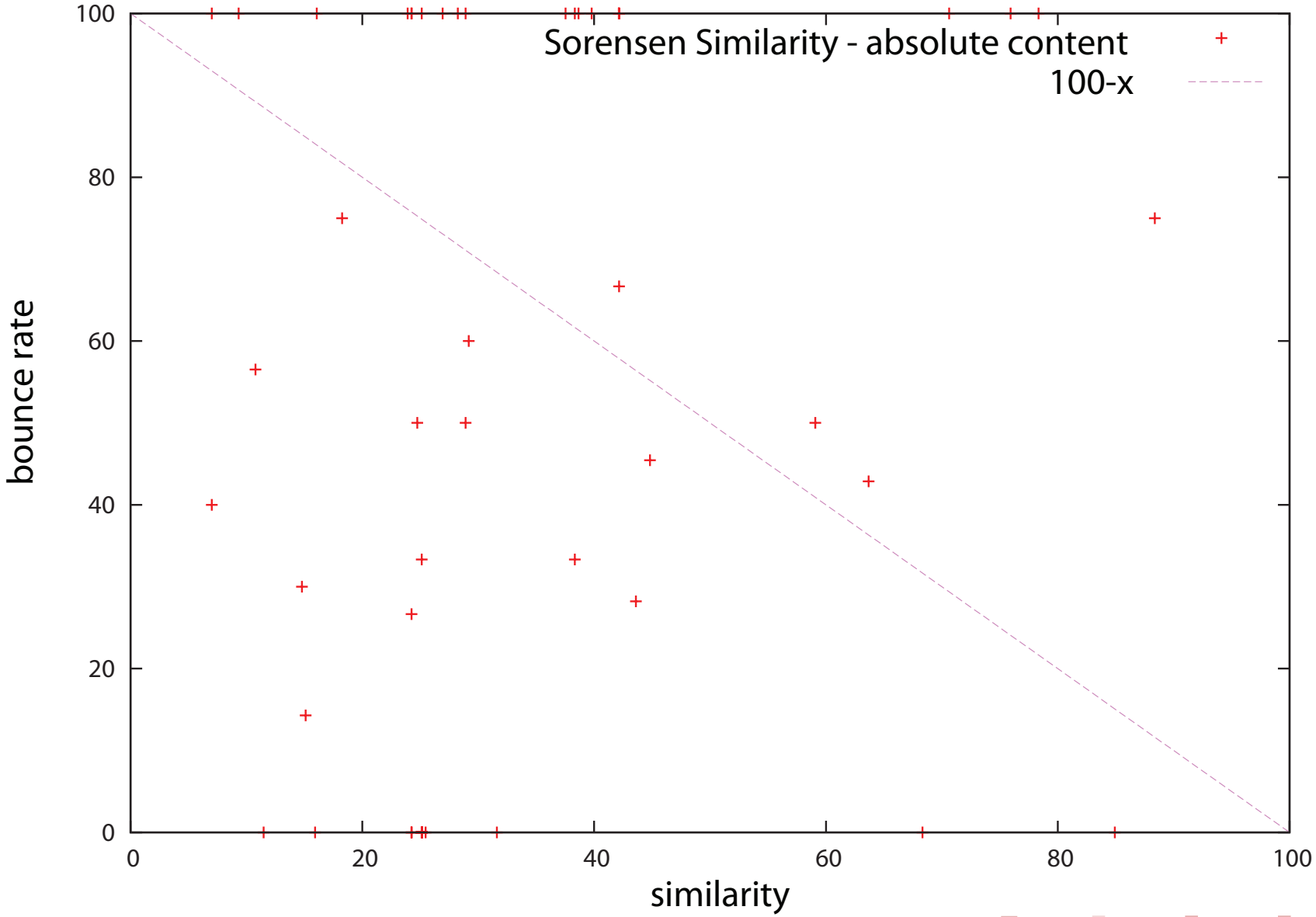
Cosine similarity



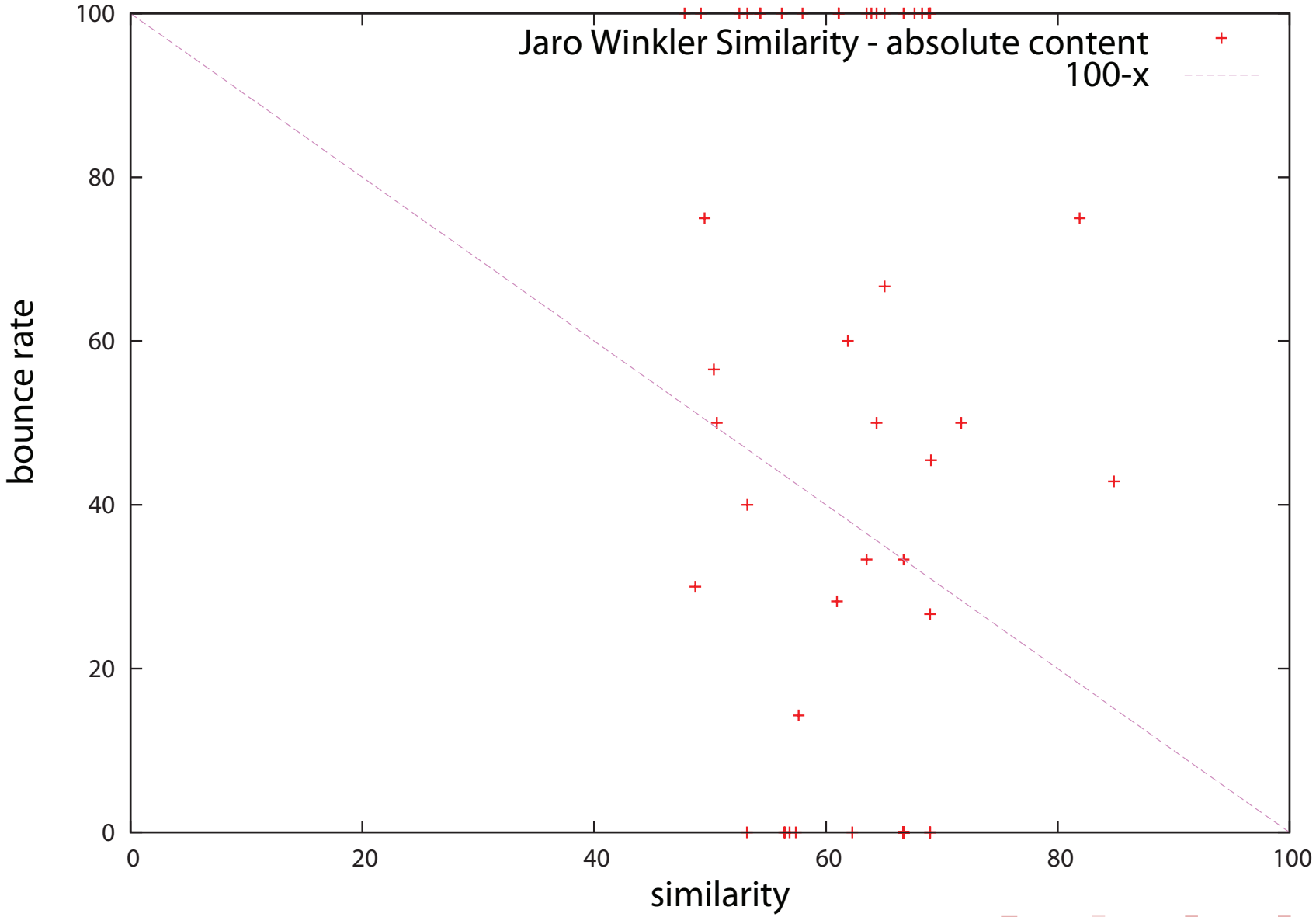
Jaccard similarity



Sorensen similarity



Jaro-Winkler similarity



Experimental results

- the best similarity function : Jaccard

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 $y = x - 100$

Conclusions and Future work

Conclusions

- correlate the bounce rate with different similarity functions

Future work

- analyze the similarity of the referrer's page content with the content of each internal link found in the corresponding landing page.
- testing this ideas on semantic similarity functions
- giving different weights to similarity functions
- choosing a similarity function and weighting and fine tuning various specific properties of the content, such as URL, page headings, page title or keywords.

Thank you!

Q & A