Measuring Web Quality

Ricardo Baeza-Yates
Yahoo! Labs
Barcelona, Spain
rbaeza@acm.org

ABSTRACT

Measuring the quality of web content, either at page level or website level, is at the heart of several key challenges in the Web. Without doubt, the main one is web search, to be able to rank results. However, there are other important problems such as web reputation or trust, and web spam detection and filtering. However, measuring intrinsic web quality is a hard problem, because of our limited (automatic) understanding of text semantics, which is even worse for other media. Hence, similarly to human trust assessing, where we use past actions, face expressions, body language, etc; in the Web we need to use indirect signals that serve as surrogates for web quality. In this keynote we attempt to present the most important signals as well as new signals that are or can be used to measure quality in the Web. We divide them using the traditional web content, structure, and usage trilogy. We also characterize them according to how easy is to measure these signals, who can measure them, and how well they scale to the whole Web.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: ranking

General Terms

Algorithms, Design, Experimentation

Keywords

Web quality, web search, web spam, web trust, ranking.

1. SUMMARY

We use web content, structure and usage to present our analysis. Of course, all these signals are or can be used to asses the quality of one or more of these elements. That is, although the focus here is to measure the quality of web content, we can also measure the quality of a link or even the quality of a user.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

1.1 Web Content

In addition to standard content signals such as those coming from word usage (e.g. TF-IDF) or the text layout (e.g. font size) we explore new signals such as lexical quality [3] or usage of entities [6]. Here are also important signals coming from social networks such as the relation of friends or followers with web content.

1.2 Web Structure

Beyond signals coming from the link structure (e.g. PageRank) and graph measures used in complex networks (e.g. assortativity), we can use other signals such as where in the structure of the Web a page or website is [2].

1.3 Web Usage

Although signals coming from usage are the most interesting ones, the main problem is that they are available only for the owners of each website. The main players in this case are search engines and click through rate (CTR) seems to be the best signal, if we make sure that it is unbiased [4]. However, new signals are appearing such as mouse tracking [1] and we also explore the relation of user engagement metrics with content quality [5].

2. REFERENCES

[1] Ernesto Arroyo, Ted Selker, and Willy Wei. Usability tool for analysis of web designs using mouse tracks. CHI Extended Abstracts 2006. Montréal, Québec, Canada, 484-489.

[2] Ricardo Baeza-Yates, Carlos Castillo, and Felipe Saint-Jean. Web Dynamics, Structure, and Page Quality. Web Dynamics 2004, New York, USA, 93-112.

[3] Ricardo Baeza-Yates and Luz Rello. On Measuring the Lexical Quality of the Web. Web Quality 2012, Lyon, France.

[4] Ricardo Baeza-Yates and Yoelle Maarek. Usage Data in Web Search: Benefits and Limitations. SSDBM 2012, Chania, Greece, 495-506.

[5] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of User Engagement, 20th conference on User Modeling, Adaptation, and Personalization (UMAP 2012), Montréal, Canada, 164-175.

[6] Pablo Mendes, Peter Mika, Hugo Zaragoza, Roi Blanco. Measuring website similarity using an entity-aware click graph. CIKM 2012, Maui, USA, 1697-1701.