Towards the "Green Web": Fighting Pollution and Promoting High Quality Content on the Web

Jussara M. Almeida Marcos André Gonçalves Raquel O. Prates

Computer Science Department, Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil 31275-010

{jussara, mgoncalv, rprates}@dcc.ufmg.br

ABSTRACT

We here present *GreenWeb*, an ongoing research project which aims at investigating how to estimate the quality of information and how to detect and treat low quality content, more broadly referred to as *pollution*, on Web 2.0 applications.

Categories and Subject Descriptors

H.3.1 **[Information Storage and Retrieval]**: Content Analysis and Indexing **;** H.5.2 **[Information Interfaces and Presentation]**: User Interfaces

General Terms

Measurement, Performance, Human Factors

Keywords

Web 2.0, Content pollution, Information quality, Tag Recommendation

1. INTRODUCTION

The advent and rapid growth of a variety of Web 2.0 applications, enabling and fostering the establishment of online communities and social networks, have contributed to the creation and dissemination of a massive amount of content. This content is often generated by the *end users*, who have different backgrounds and expertise, with no *quality* guarantees. Moreover, all this freedom opens the opportunity for malicious or opportunistic behavior including the introduction of polluted (i.e., *low quality*) content, or simply pollution, into the system. By pollution, we mean *undesired, irrelevant, or redundant content, which aggregates low value to Web services, thus causing discomfort, disorder, or some threat to the user or to the system*. Examples of attempts to introduce pollution into the system include vandalism in Wikipedia [7] and different forms of spamming [2,4].

Content pollution may affect the effectiveness of various services such as searching, content recommendation, and advertising, as well as compromise user patience and satisfaction with the system. This is because users cannot easily identify the pollution before having contact with it, which leads to the consumption of more system resources (e.g., bandwidth in case of video pollution [2]).

We here introduce *GreenWeb*, a project of the Brazilian National Institute for Science and Technology for the Web (INWeb), which aims at developing techniques and tools to identify and fight pollution on the Web, contributing to aggregate value to various information services. GreenWeb focuses on three main axes: (1) reducing pollution in the system; (2) increasing the quality of the available information; while (3) keeping a good cost-benefit tradeoff. As such, it has to address multiple challenges, such as: (1) to collect, store and process huge amounts of data, (2) to address very heterogeneous and time-evolving patterns, and (3) to deal with a high degree of subjectivity, as quality must be assessed from the perspective of specific services and of specific (groups of) users[1]. Given such challenges, GreenWeb has an inherently multi-disciplinary nature, involving various areas from computer science, such as information retrieval, machine learning, user behavior and social network analysis, human-computer interaction, performance measurement, as well as from other areas such as information science and semiotics.

2. CURRENT RESULTS

Selected results of the GreenWeb project include:

- **Automatic Detection of Polluters on YouTube:** In [1], we developed supervised classification algorithms to automatically detect spammers and promoters of polluted video content on YouTube, with high rates of detection. In [5], we improved our solutions using multi-view learning techniques to greatly reduce the amount of labeled data required to effectively identify polluters. We also developed analytic cost models to study the tradeoffs between alternative detection strategies,

[1] On one hand, the same content may have high quality for supporting advertising and classification services but not for supporting searching [9]. On the other, the same information may be considered valuable by some users but not by others.

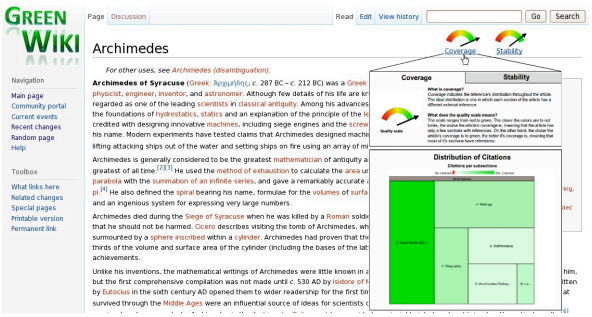

Figure 1: GreenWiki: Assessing the Quality of Wikipedia Articles.

Figure 2: GreenMeter: Assessing the Quality and Recommending Tags on LastFM.

considering system resource usage (e.g., resources wasted with pollution inserted by false negatives) and the cost associated with human manual effort.

- **Automatic Quality Assessment of Content:** We developed machine-learning methods that exploit various quantitative metrics to provide indicators of quality of *Wikipedia articles* [3]. Moreover, we also proposed new strategies for the communication of these indicators to users [6] and implemented these new interfaces in a tool called GreenWiki[2] (Figure 1).

  We also developed quantitative metrics to drive the characterization of the quality of four *textual features* – title, tags, description, user comments – associated with multimedia objects on four Web 2.0 applications. Feature quality was analyzed from the perspective of its potential as supporting data for various information services [8,9].

- **Novel Tag Recommendation Methods:** Motivated by the results of our feature quality characterization, we designed novel methods for suggesting *high quality tags*, aiming at improving both the quality of this feature and the results produced by various services that often rely on it as data source. By high quality, we mean tags that describe the content of the target object reasonably well (e.g., for supporting content recommendation) and/or that discriminate it from other objects (e.g., for searching and object classification). Our methods jointly exploit tag co-occurrence patterns, quality metrics and the contents of multiple textual features [1]. We also investigated the use of learning-to-rank techniques (notably Genetic Programming) for tag recommendation purposes.

  Some of the proposed feature quality metrics [8,9] were used to build a *tag quality estimator*, which was implemented, along with one of our tag recommendation methods, in the GreenMeter tool[3], with a current prototype for the popular LastFM application (Figure 2).

3. CONCLUSIONS AND FUTURE WORK

We have presented the GreenWeb project, which focuses on detecting and fighting content pollution and on promoting high quality content, ultimately contributing to create a "cleaner" Web for all. Several promising results have already been obtained on the identification of content polluters, on the assessment of quality of user generated content, and on the recommendation of high quality content. Future work includes extending our solutions to consider other pieces of evidence of information quality (e.g., user and social behavior aspects, dynamic patterns), other scenarios and applications, and by performing a broader evaluation based on user experiments.

4. ACKNOWLEDGMENTS

This work is supported by the INWeb (MCT/CNPq grant 57.3871/2008-6), and by the authors' grants from CNPq and FAPEMIG.

5. REFERENCES

[1] F. Belém, E. Martins, T. Pontes, J. M. Almeida, M. A. Gonçalves, "Associative Tag Recommendation Exploiting Multiple textual Features", Proc. ACM SIGIR, 2011.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting Spammers and Content Promoters in Online Video Social Networks", Proc. ACM SIGIR, 2009.

[3] D. Dalip, M. Gonçalves, M. Cristo, P. Calado, "Automatic Quality Assessment of Content Created Collaboratively by Web Communities. A Case Study of Wikipedia", Proc. JCDL 2009.

[2] http://www2.dcc.ufmg.br/projetos/greenwiki/mediawiki/.

[3] http://sites.google.com/site/greenmeterdemo/

[4] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, H. Garcia-Molina, "Combating spam in tagging systems: An evaluation", ACM TWEB 2(4), 2008

[5] H. Langbehn, S. Ricci, M. Gonçalves, J. Almeida, G. Pappa, F. Benevenuto, "A Multi-view Approach for Detecting Non-Cooperative Users in Online Video Sharing Systems", Journal of Information and Data Management, v.1, 2010.

[6] R. Pereira, "Quality of Wikipedia Articles for Users – Analysis and Interaction Proposal", Master's Dissertation, DCC/UFMG, 2011.

[7] M. Potthast, M. Stein, T. Holfeld, Overview of the 1st International Competition on Wikipedia Vandalism Detection. CLEF (Notebook Papers/LABs/Workshops),2010.

[8] F. Figueiredo et al., "Evidence of Quality of Textual Features on the Web 2.0". Proc CIKM 2009.

[9] J. Almeida, M. Gonçalves, F. Figueiredo, F. Belém, H. Pinto, "On the Quality of Information for Web 2.0 Services", IEEE Internet Computing 14(6), 2010.