Measuring Wikipedia: A hands-on tutorial

Luca de Alfaro
Univ. California at Santa Cruz
School of Engineering
1156 High Street MS: SOE3
University of California
Santa Cruz, CA 95064 USA
+1-650-248-2856
luca@cs.ucsc.edu

Felipe Ortega
GSyC/Libresoft
Univ. Rey Juan Carlos
C/ Tulipán s/n
28933. Mostoles.
Madrid. Spain
+34-91-488-8105
jfelipe@libresoft.es

ABSTRACT

This tutorial is an introduction to the best methodologies, tools and practices for Wikipedia research. The tutorial will be led by Luca de Alfaro (Wiki Lab at UCSC, California, USA) and Felipe Ortega (Libresoft, URJC, Madrid, Spain). Both cumulate several years of practical experience exploring and processing Wikipedia data [1], [2], [3]. As well, their respective research groups have led the development of two cutting-edge software tools (WikiTrust and WikiXRay), for analyzing Wikipedia. WikiTrust implements an author reputation system, and a text trust system, for wikis. WikiXRay is a tool automating the quantitative analysis of any language version of Wikipedia (in general, any wiki based on MediaWiki).

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *Computer-supported cooperative work; Web-based interaction.* K.4.3 [**Computers and Society**]: Organizational Impacts – *Computer-supported collaborative work*.

General Terms

Algorithms, Measurement, Documentation, Experimentation.

Keywords

Wikipedia, empirical research, measurements, data mining, WikiTrust, WikiXRay

1. INTRODUCTION

Since its inception back in 2001, Wikipedia has evolved to become a massive collaborative project, involving millions of editors worldwide. It is currently ranked as the 7th most popular website, according to Alexa web ranking, receiving hundreds of millions of visits every month. As a result, Wikipedia is still one of the most attractive targets for researchers of many different disciplines. Nevertheless, many of these researchers must overcome serious difficulties to develop their studies on Wikipedia, especially when it comes to retrieving, analyzing and interpreting the enormous amount of data produced by this project.

Often, researchers wishing to analyze information from Wikipedia (edit logs, content evolution, activity metadata, web graphs, etc.) get frustrated because almost all information is accessible, but it is also very difficult to process it efficiently. For instance, if we consider the English Wikipedia, a researcher interested in looking at the whole information set of activity logs may find more than 270 million changes, performed by more than 1.6 million of registered editors, together with an undetermined number of anonymous editors. That means more than 2.5 TB of plain text. Therefore, it is not a trivial task to store, process and analyze this data set. Without adequate guidance on the best available tools and practices to apply for these purposes, researchers and practitioners will be lost in the *Wikipedia data jungle*.

The tutorial will blend in-depth presentations of state-of-the-art methodologies and tools to analyze Wikipedia, with live demos and practical sessions, to grasp the real problems and challenges that Wikipedia researchers must face off on a daily basis. After attending this session, attendees should be able to:

- Outline a general picture of the different research perspectives currently applied to Wikipedia.
- Easily find previous research works and information sources to contextualize their own work on Wikipedia.
- Store, navigate and process information retrieved from the Wikipedia data jungle.
- Discriminate the optimum set of available tools that fits their own Wikipedia research needs, as well as develop their own set of tools.
- Create and refine their research roadmap, to achieve concrete goals and results.
- Feel comfortable using and extending WikiTrust [1], [2] and WikiXRay [3].

2. REFERENCES

[1] Adler, B.T., K. Chatterjee, de Alfaro, L., Faella, M., I. Pye and V. Raman. 1998. Assigning Trust to Wikipedia Content. In Proceedings of the 2008 International Symposium on Wikis (Porto, Portugal). In press.

[2] Adler, B.T. And de Alfaro, L. 2007. A content-driven reputation system for the Wikipedia. In Proceedings of the 16th Intl. Conference on World Wide Web. WWW '07. (Banff, Alberta, Canada). ACM Press, New York, NY, 261-270. DOI= http://doi.acm.org/10.1145/1242572.1242608

[3] Ortega, F. 2009. Wikipedia: A quantitative analysis. PhD. thesis. ETSIT. Universidad Rey Juan Carlos. Madrid. http://libresoft.es/Members/jfelipe/thesis-wkp-quantanalysis