Chapter VIII
User and Context–Aware Quality Filters Based on Web Metadata Retrieval

Ricardo Barros
Federal University of Rio de Janeiro, Brazil

Geraldo Xexéo
Federal University of Rio de Janeiro, Brazil

Wallace A. Pinheiro
Federal University of Rio de Janeiro, Brazil

Jano de Souza
Federal University of Rio de Janeiro, Brazil

ABSTRACT

Due to the amount of information on the Web being so large and being of varying levels of quality, it is becoming increasingly difficult to find precisely what is required on the Web, particularly if the information consumer does not have precise knowledge of his or her information needs. On the Web, while searching for information, users can find data that is old, imprecise, invalid, intentionally wrong, or biased, due to this large amount of available data and comparative ease of access. In this environment users constantly receive useless, outdated, or false data, which they have no means to assess. This chapter addresses the issues regarding the large amount and low quality of Web information by proposing a methodology that adopts user and context-aware quality filters based on Web metadata retrieval. This starts with an initial evaluation and adjusts it to consider context characteristics and user perspectives to obtain aggregated evaluation values.
INTRODUCTION

Years ago, data storage was scarce and had to be cleaned periodically. In those times, a database project would include rules for longevity of data, for migration from secondary memory (disks) to archival memory (tapes), and for deletion of unused data. Current practice, however, is to leave data lingering around, or at most, transfer them to data warehouses to be used opportunistically.

Specifically, data available on the Internet is overwhelming. Lyman and Hal (2003) estimate that there are 167 terabytes of data in fixed Web pages, and also that Web pages created on demand use more than 91.850 terabytes of data stored in databases. To that we can add 440.606 terabytes of new e-mails per year, approximately one-third of which is spam.

Due to the amount of information being so large and being of varying levels of quality, it is becoming increasingly difficult to find precisely what is required on the Web, particularly if the information consumer does not have precise knowledge of his or her information needs (Burgess, Gray, & Fiddian, 2004). On the Web, while searching for information, users can find data that is old, imprecise, invalid, intentionally wrong, or biased, due to this large amount of available data and comparative ease of access. Web search engines are a good example of this situation: in a reply from these mechanisms, one can usually find links to replicated or conflicting information. In this environment, users constantly receive useless, outdated, or false data, which they have no means to assess.

Moreover, the range of suppliers also results in a diverse variety of formats in which information is stored and presented. It is possible to search for information on an unlimited number of contexts and categories across a wide range of information environments, such as databases, application systems, electronic library systems, corporate intranets, as well as the Internet. This information presents different levels of quality, with original sources ranging from multi-national corporations to individuals with limited knowledge. With so

Figure 1. Producer-consumer schema and selection filters
Related Content

Text Mining in Bioinformatics: Research and Application
[www.igi-global.com/article/text-mining-in-bioinformatics/100039?camid=4v1a](www.igi-global.com/article/text-mining-in-bioinformatics/100039?camid=4v1a)

Method of Lexical Enrichment in Information Retrieval System in Arabic

Development of the Leader Integrity Assessment
[www.igi-global.com/chapter/development-leader-integrity-assessment/69743?camid=4v1a](www.igi-global.com/chapter/development-leader-integrity-assessment/69743?camid=4v1a)

A Knowledge Based Approach for Accident Analysis: Application to the Safety of Automated Rail Transport Systems
[www.igi-global.com/article/a-knowledge-based-approach-for-accident-analysis/210062?camid=4v1a](www.igi-global.com/article/a-knowledge-based-approach-for-accident-analysis/210062?camid=4v1a)