Using Distributed Meta-information Systems to Maintain Web Data Quality

Dawn G. Gregg
dawn.gregg@asu.edu
School of Accountancy and Information Management, College of Business
Arizona State University, Tempe, AZ 85287-3606
voice: (602) 965-3631 fax: (602) 965-8392

Introduction

Maintaining the quality of data resources has been a continuing concern for information systems professionals. Over time techniques have been developed for maintaining the appropriate level of quality for individual databases, for data warehouses and for transaction processing systems. However, web-based systems lack the tools and procedures for data quality to be properly maintained. My dissertation seeks to develop methods that can be used to improve web-based data quality. It maps data quality dimensions, as identified in prior research [Wand et. al., 1996, and Wang et. al., 1995], to the web domain and then proposes methods that can help maintain each of these data quality dimensions. These six primary data quality dimensions are presented in Table 1.

Table 1: Data Quality Dimensions [Wand and Wang, 1996, and Wang, Reddy and Kon, 1995]

Data Quality Dimension	Characteristic
Accessibility	Data is available to (easily found by) the user,
Completeness	Every meaningful state of the specified real world system can be represented.
Believability	The extent to which data can be counted on to be correct.
Currency	Data are current if they do not reflect outdated information
Accuracy	Data agrees with an identified source to a desired precision.
Consistency	Two sets of data are consistent if they do not conflict with one another. Referential integrity in databases is one type of consistency constraint.

For the quality of Web-based information to improve, it needs to be constructed according to commonly known standards. Otherwise, it will be difficult for businesses and consumers to locate appropriate information, assess the informational content, and judge its fitness [Goul et. al., 1997]. As yet, however, there are few standards for the capture, description, distribution, and correlation of web material.

My dissertation proposes using distributed meta-information to allow classification of web content and to describe the relationships between content and other web resources. The meta-information would consist of a set of specialized <tags> that could be read by intelligent agents designed for web site search and/or maintenance. Two classes of meta-information are being proposed, content information and maintenance information

Content Information

Meta-information labels can be used to identify the content and quality of Web pages [Resnick, 1997; Resinick & Miller 1996]. Content Web page labeling systems need to be designed to improve end-users' ability to locate specific information or resources that are available on the WWW. Some common types of Web pages that can be labeled include home pages, product pages, sales sites information sites, white papers, software download sites etc. In addition to developing labels based on page type, labels could also include information on the page's subject area. The general requirements for web page meta-information systems have been identified in the literature [Goul et. al., 1997]:

1. The meta-information must provide specific data related to the page type and subject area,
2. The meta-information should be designed to meet the specific needs of the target user population,
3. The meta-information must have a consistent format so autonomous agents can readily use it when processing user search requests.

These labels would improve the accessibility data quality dimension because they help to pin point the exact topic the Web page covers. This allows individual users or automated intelligent search agents to more readily find specific information about web page content, providing end-users with improved discovery of existing web-resources [e.g., Acerman, et. al., 1997; Etzioni, 1997].

Maintenance Information

Meta-information can also be used to ensure that web data remains accurate, consistent and up to date. Several studies have sited broken hyperlinks as "one of the most serious problems facing the WWW today" [Ingham, Caughey and Little, 1996]. However, broken hyperlinks represent only one type of consistency failure that can occur on the web. Often information on a web page can be related to information found on other web pages, yet there is currently no mechanism for ensuring that updates to web based information are automatically propagated to related web sites. This need for content quality on the web begins with hyperlinks but extends to all other types of web information from individual numbers, to textual strings, to graphics, to audio etc. This requires meta-information concepts to be extended to include information about the complex relationships among web-based data as well as to retrieve and update related data efficiently.

To enforce *accuracy, consistency* and *currency* it is necessary that web meta-information systems meet the following two requirements, as derived from database theory [Codd, 1970; Elmasri, & Navathe, 1994]:

1. Maintenance meta-information must provide information related to the data type and appropriate values for specific web data or resources,

2. Maintenance meta-information must also specify how a resource in one file is related to resources in other files.

In addition:

3. Maintenance meta-information must have a consistent format so autonomous agents can readily user it when processing retrieval and update activities.

The third requirement is necessary because it allows intelligent agents to be used to monitor and update web content as long as it contains appropriate maintenance meta-information [e.g., Acerman, et. al., 1997; Etzioni, 1997].

Research Methodology

The research is being conducted in two phases. The first phase focuses on content labels for a specific domain, decision support systems (DSS). It proposes a protocol suite that utilizes meta-information labeling to fully describe DSS such that an intelligent agent can easily discover them. The protocol is being validated using an experiment that assesses a subject's understanding of specific DSS capabilities based solely on the meta-information.

The second phase will develop a Web data model and an XML based protocol that will allow Web based information to be maintained in a manner consistent with organizational goals for data quality. First, a formal definition for the data model will be created using a conceptual-formal software development methodology. Then a proposal for a protocol suite that will facilitate the maintenance of Web-based information will be defined.

Conclusions

The development effort represents a new approach to managing and maintaining web content. The meta-information labels developed as a part of this research can improve the six quality dimensions presented in Table 1. Content information labels help to improve the *accessibility* of web information by improving an end-user's ability to locate specific content or resources that are available on the WWW. It is also possible for content labels to provide a measure of the *completeness* of a specific web resource [Goul et. al., 1997]. Meta-information maintenance labels can be used to insure the *currency*, *accuracy* and *consistency* of web content. Both types of meta-information systems should help end-users better assess the *believability* of web content because end-users will have access to the meta-information labels and will be able to use those labels to judge the likelihood that a specific piece of web content is reliable.

The power of the distributed meta-information approach to managing web data quality is that it distributed with actual web pages. This provides universal access to the content and maintenance information and contributes to the disintermediation of the WWW. These meta-information labels allow individuals or automated intelligent agents to more readily find specific information about web page content or to update the distributed web content. The distributed nature of the approach can provide greater flexibility to local webmasters and web page designers increasing local autonomy, as well as increasing expandability of web systems [Ozsu and Valdurez, 1991].

References

Acerman, Billsus, Gaffney, Hettich, Khoo, Kim, Klefstad, Lowe, Ludeman, Muramatsu, Omori, Pazzani, Semler, Starr, and Yap, Learning Probabilistic User Profiles - Applications for finding interesting websites, notifying users of relevant changes to web pages, and locating grant opportunities, *AI Magazine,* v18n2, Summer, 1997, 47-56

Bond, A.H., and Gasser L., Eds. *Readings in Distributed Artificial Intelligence*, Morgan Kaufmann, 1982.

Bowman, C. Mic, Danzig, P. B., Manber, U. and Schwartz, M. F. "Scaleable Internet Resource Discovery," *Communications of the ACM,* (37:8), August 1994, pp. 98-107+.

Codd, E. F., "A Relational Model for Large Shared Data Banks," *Communications of the ACM*, v13 n6, June 1970.

Elmasri, Ramez and Navathe, Shamkant, *Fundamentals of Database Systems,* Redwood City, California:The Benjamin/Cummings Publishing Company, 1994

Etzioni, O. "Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web," *AI Magazine,* (18:2), Summer 1997, pp. 11-18

Goul, M.; Philippakis, A., Kiang, M.; Fernandes, D.; and Otondo, R., "Requirements for the Design of a Protocol Suite to Automate DSS Deployment on the World Wide Web: A Client/Server Approach," *Decision Support Systems*, (19:3), March 1997, pp. 151-170.

Ingham, D., Caughey . and Little, M. " Fixing the Broken Link Problem: the W3Objects Approach," *Computer Networks and ISDN Systems,* (28:7-11), May 1996, pp. 1255-1268.

Isakowitz, T., Bieber M., and Vitali F., "Web Information Systems," *Communications of the ACM*, (41:7), July 1998, pp78-80.

Nielsen, J. "Impact of Data Quality on the Web User Experience." *Jakob Nielsen's Alertbox,* July 12, 1998, http://www.useit.com/alertbox/980712.html.

Ozsu, M. Tamer and Patrick Valduriez, *Principles of Distributed Database Systems,* Englewood Cliffs, New Jersey: Prentice Hall, 1991.

Resnick, P. "Filtering Information on the Internet," *Scientific American,* , March 1997, pp. 26-32

Resnick, P. and Miller, J. "PICS: Internet Access Control without Censorship," *Communications of the ACM*, (39:10), October, 1996, pp. 87-93

Wand, Y., and Wang R. Y. "Anchoring data quality dimensions in ontological foundations;" *Communications of the ACM,* (39:11) Nov. 1996, pp. 86-95

Wang, R. Y., Reddy, M. P., and Kon, H. B. "Toward Quality Data: An Attribute-based Approach," *Decision Support Systems,* (13), 1995, pp. 349-372

Wang, R. Y. and Strong, D. M. "Beyond Accuracy: What data quality means to data consumers; *Journal of Management Information Systems,* (12:4), 1996, pp. 5-34