Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Evaluating Web Data for Data Mining

Zhenyu Huang
University of Memphis, zhuang@memphis.edu

Lei-da Chen
University of Memphis, ldchen@memphis.edu

Mark Frolick
University of Memphis, mfrolick@memphis.edu

Evaluating Web Data for Data Mining

Zhenyu Huang, Fogelman College of Business and Economics, University of Memphis,
Zhuang@memphis.edu
Lei-da Chen, Fogelman College of Business and Economics, University of Memphis,
Ldchen@memphis.edu
Mark Frolick, Fogelman College of Business and Economics, University of Memphis,
Mfrolick@memphis.edu

Abstract

Organizations' operational data constructs the major data source for their data warehouse. The exponential development of WWW has made Internet an immense database containing all kinds of information with various types of data structures. Organizations are increasingly interested in capturing web data into their data warehousing systems to enlarge their data source for decision supporting, therefore improving accuracy and effectiveness of their decision making. This research thoroughly analyzes the data value of web data to data warehousing as well as business decision making, discusses the feasibility and potential problem of loading web data into data warehouse system, and provides a framework for evaluating web data for data warehousing purpose. Web data analysis and evaluation is regarded as a prerequisite for Web Integration - a breakthrough approach in furnishing data warehouse input: extracting, scrubbing, transforming web data and loading it into data warehouse systems to support organization decision making.

Keywords: Data Warehousing, Data Mining, Web, Framework, Data Extraction.

Introduction

Since the first book of Inman - *How to Build a Data Warehouse* in 1992, people have begun to pay attention to Data Warehousing this new area of information systems. Since then, the data warehousing evolved so rapidly that it is now one of the hottest topics in information system technologies. A data warehouse is "a subject-oriented, integrated, time-variant, nonvolatile collection of data used in support of management decision making processes" (*Inmon 1996*). Data warehouse supports informational processing by providing a solid platform of integrated, historical data from which to do analysis. According to Inmon, the data entering the data warehouse comes from the operational environment in almost every case. This situation is true till present.

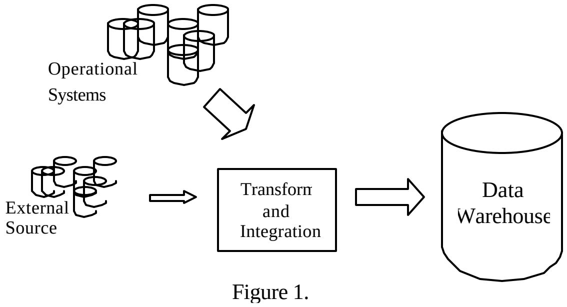

Figure 1.

Figure 1 is a conceptual model that describes how a data warehouse system operates. In a typical data warehousing system, data is extracted from various system files and databases, and is transformed and integrated before being loaded into the data warehouse. By default, there are two kinds of resources provide data input: operational systems and external source. The former forms a main stream of information needed for the data warehouse, the latter often refers to random, mostly paper based external data such as news paper, statistic data, etc.

Besides data warehouse, there exist another hot topic: Internet. Internet has been developing with an exponential speed in recent years. The data of Internet Domain Survey (Internet Software Consortium, 1999), shows that Internet has an amazing growth speed. In January 1993, there were only 1.3 million web hosts, after 6 years, the Internet contains nearly 60 million sites now.

People have begun to think about the possibility and benefits of combining two technologies. Some researchers and practitioners advocated that the web and data warehousing could form a powerful combination and have proven to be highly powerful and successful (Hackathorn 1997, Miley 1998, Gardner 1997). Currently, publishing warehouse data via the Intranet is a highly productive approach that combines Web delivery mechanisms with decision support capability of data warehousing (Figure 2). Lots of discussion primarily focuses on this aspect of the marriage between data warehouse and web (e.g. Anonymity 1999 B, Booker 1999, Corbin 1997, Wilson 1997): data flowing from data warehouse to the web. Data warehouse is the source of information, Internet provides a publication and presentation tool of the data in front of customers.

However, organizations rarely consider the combination of data warehousing and Internet in the opposite way: “integrating” Internet data into data warehousing systems. The major reasons are that web contents are considered very unreliable, data external to the organization are often considered to have little business value, and external data are hard to manage in complex systems such as a data warehouse. The majority of companies are focusing on capturing data from their internal operational system and limited external data source. However, external data are highly useful to organization and readily available on the web. Organizations are facing internal and external pressure for newer, better, or more timely information beyond the range that internal data can offer. The first step for an organization to integrate these data from the web is to realize the business potential of doing and learning how to evaluate web data before they are included in the decision-support data warehouse.

The emerging area that is concerned with this challenge is called Web Integration (WI). Web integration is a systematic process of drawing valuable Web content as input to the data warehouse (Hackathorn 1997). By integrating web (as major external data source) information, WI lets data warehouses have broader information resources and will be able to provide better decision support than those only utilizing internal operation data. Hackathorn (1997) had similar concept of putting web content into the data warehouse, but he named it “Web Farming” because he thought “catching information from Internet is not dealing with the back yard garden but face acres of Internet farm”. Some other researchers and practitioners have also reported the benefits and significance of putting highly business related web contents into organizations’ information systems (Marchetti 1999, Zorn et al 1999). In the past, data warehouse was based upon the database contents of internal operational systems. However, Web Integration shifts the focus to external and more global perspective of the enterprises (figure 3). Web Integration technology offers a powerful tool to manage and take advantage of Web data. The challenge is to wade through the web, discover and acquire pieces that do have an impact on the business and reorganize them into a format compatible to data warehouse structure to allow future retrieval.

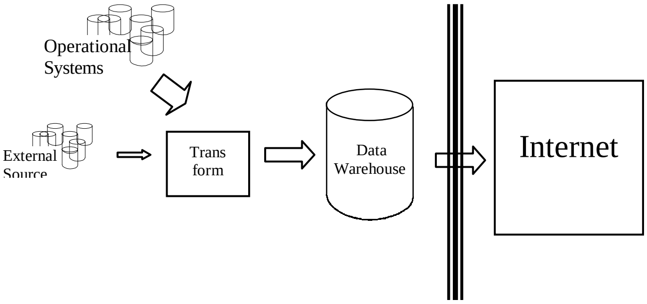

Figure 2. Traditional Data warehouse and Web Combination model

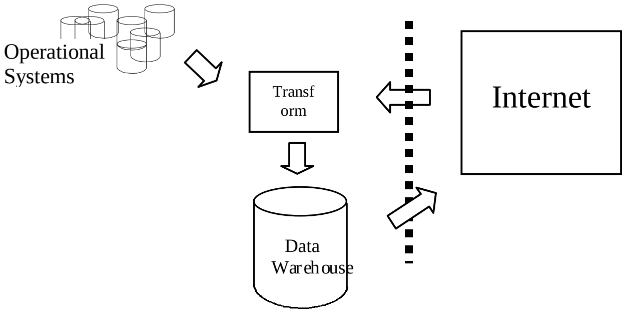

Figure 3 Web Integration model

Web Integration and its value

Internet breaks the time and space limitation, and brings infinite information. Web can be considered as a massive database though the format and structure of this database is not uniform. Koehler (1999) compared Internet to World Brain. As for data warehouse, enlarged and valuable data input could help realize the function of data warehouse to maximize. For many reasons discussed as following, it is necessary for companies to share their eyes on integrating the web resources into their data warehouse.

Abundant information on the web

The WWW has experienced exponential growth in the last few years. According to NetWizard, the number of domains on the web increased from 21,000 in January 1993 to 4.3 million in July 1997, and reached 7.1 million in August 1999. The number of hosts on the web grew from 36.734 million in 1998 to 60.147 million in 1999. Among them, 22.5 million are .com sites, 5.7 million are .edu sites, 0.8 million are .org and 0.7 million are .gov sites.

Information that has strategic implication to organizations is abundant on the web. Electronic Commerce wave tremendously increases organizations’ involvement in the Internet. This will inevitably lead to an dramatic increase in business relevant data, thus make the Internet a more valuable resource for businesses.

Web data are valuable to business issues

Drucker (1997) admonishes IT executives to look outside their enterprises for information. He remarked that the single biggest challenge is to organize outside data because change occurs from the outside. He predicted that the obsession with internal data would lead to organizations being blindsided by external forces.

As markets become turbulent, the old way of doing business with data only from internal operational systems becomes less effective. A company must know more about its customers, suppliers, competitors, and other external factors than ever before. It must enhance the information from internal systems with information about external factors. The synergism of the combination creates the greatest business benefit for the enterprise. Much of this external data is readily available on the web.

Online public libraries (e.g. library of conference, ACM digital libraries) provide examples, and so are online public services (e.g. www.weather.com and www.map.com etc). Hackthorn (1997) studied web information and categorized web data into several types. His categories are mainly based on the web data functions and most of them are from commercial sites. Another category that is more widely accepted bases on publisher type (Koehler 1996, 1999). Web documents contain different types of information and behave differently according to publisher type. Publishers can be identified by TLD (Top Level Domain), sometimes 2LD (Second Level Domain). For example, commercial sites can be identified by .com, of which the purpose is to sell and promote products and services. Companies can locate competitors' production information, suppliers' price strategy, supplementary products development information easily from this kind of web sites.

Reduce the cost of acquiring data for data warehouse

Building a data warehousing system requires heavy monetary as well as human resource investment. According to Meta Group (http://www.metagroup.com), the average 1997 data warehouse project (with staffing costs) requires $1.9 million to implement. In his book, W. H. Inmon (1998), estimates that, on average, 80% of the time building a data warehouse will be spent on extracting, cleaning, and loading data. Upon the completion of data warehouse, the work of data collection will be continued and sometime expanded to keep up with the business needs. Tremendous labor and relevant cost are involved with data warehousing operations.

Web Integration can help reduce the operating cost of data warehouse for several reasons. Web data are free to use, individuals can use most information that they see on the web without paying for the usage, which will save a lot on operating a data warehouse that includes web data as a large percent source. And these free data are valuable to the business as discussed previously and may not be provided by internal systems. Meanwhile, in the following part about WI operation, WI won't increase labor and financial investment tremendously.

Toolkits are available for web integration.

WI will become a labor-intensive operation without corresponding technologies and software support. Software products that aim and capture data from web and load data directly into organizations' data store are commercially available. These tools systematically utilize the web as one massive database and query the relevant data from the web (Wilent 1997). They retrieve data from web page, and incorporate the relevant data into new HTML documents or business applications. Besides Web Automation Toolkit, there exist a large number of software applications suitable for accomplishing WI task. For example, FlashSite is a product of Incontext (http://www.incontext.ca). FlashSite's two primary functions can well meet WI's requirements. First, it will download web sites, web pages, or it will prepare a map (site map) diagramming the web sites. Second, it will periodically check the selected downloaded web site or web page against the then-current counterpart. Combining the technologies and the effect of web analysts, WI is able to effectively reduce the labor and other costs associated with data acquisition.

Possible problems with Web Integration

Although people increasingly adopt Internet as business information source, there are still a lot of potential problems existing in web information that need to be alerted. Web data are not always reliable. Various data formats become a barrier for organizations to systematically utilize web information because data is not easy to be managed without same format. Other pitfalls of web data also deserve organizations to pay attention to.

Rich formats of web data make it difficult to organize web information flow

Internet web is a multimedia environment, it includes all types of media such as artwork or static pictures, full or partial motion video, sound (midi, voice), text, graphics, animation, tables etc. Multiple data types make Internet an interesting and attractive world but they also bring problems when people try to combine data from different web sites into one file. Multi-format data appear to be disorder, random and irregular, therefore is hard to be utilized.

Data warehouse has rigid format requirements on data. Though data on the web is computer recognized, but it is not data warehouse compatible. Transformation of hypertext into a structured database is tough. The process of loading web data into data warehouse looks like putting picture, text, and video into same column of a relational database table.

Web data are of little discipline and volatile and constantly changing

One of characteristics of data warehouse is that data warehousing deals with nonvolatile data collection (Inmon 1996). On the contrary, the characteristics of web data could be volatile and ephemeral. The paradigm of the Web is radically different from the paradigm for the data warehouse. The web's diversity often challenges individual's imagination and appreciation for new forms of creative expression.

Another symptom of web data is inconsistent. The most common "failure to respond" reasons were "No DNS" (no domain name server) and the 400 errors (page not found or access denied). Koehler (1999) conducted a thorough survey of web site/page permanency and consistency. In his survey, after a 6-month period, 12.2% of the web sites and 20.5% of web pages collected for the study failed to respond when queried, as did 17.7% and 31.8% respectively after 1 year. Also, it was found that more than 97% of web sites underwent some kind of change over the 6-month period they were followed, after 1 year, more than 99% had changed. But different types of web pages, web sites and domains behave differently.

Copyright of WebPages

Copyrights offer an important topic that couldn't be ignored when digging Web Data. When Web increasingly becomes an important resource, the needs to protect their copyrights of their Web "products" become increasingly impetus for companies. Although those web pages or web sites under copyright currently only owe a small part of the whole Internet resources, individuals need to pay for these kinds of information protected by copyrights and may pay more in the future. It is believed to be a trend for Internet Web pages to have copyrights. But, In another aspect, only when organizations can have the protection of copyrights for their web resources, they are willing to invest more on developing high quality web pages, which will benefit other organizations who utilize these high quality web pages. Copyright law on the Internet is evolving. When using information from the Internet, Kirshenberg (*1998*) reminds users to assume that the work, whether an article or a photograph, is protected under copyright law unless it explicitly states otherwise.

Information overload

A study released by Pitney Bowes Inc., showed that the average businessperson in the United States, Canada and the United Kingdom sends or receives 190 messages a day (McCure 1998). It is discussed above that abundant information is one of benefit of Web data, whereas it is also their drawback when companies want to systematically use this data. Too much information is confusing, time-consuming and run out of processing ability of humans. McCure (1998) claimed "Information overload can have detrimental effects on both individuals and companies". Everything seems to fit into their requirement while it is often found to be the contrary. It is a common experience for people that when they use the search engine by keywords, the searching results would be thousands and millions items, which make it difficult to locate what is exactly wanted. Another situation is that individuals face tremendous similar or identical data, which also takes them too much time to select the appropriate one.

Credibility and Reliability of Web data

The reliability of Web content is an important issue that companies must manage carefully. Most people have the 'flake free' image of Web content. In reality, the Web is a global bulletin board where the wise and the foolish have equal space. Acquiring content from the web should not reflect positively or negatively on its quality. Correctly evaluating web data is basic and critical to use it efficiently.

Evaluating business value of Web data

Data quality models for Web Integration

Data quality has been intensively studied (Klobas 1995, O'Brien 1996, Wang and Strong 1996, Wang 1998, Orr 1998, Tayi et al 1998, Redman 1998, Rieh and Belkin 1998, Hackathorn 1998). Researchers have established various frameworks or standards to evaluate information quality from various perspectives. For example, Wang (1998) identified four roles of "information product": information suppliers, information manufacturers, information consumers and information product managers. Then he contended that just like other products, Information product also has quality. Four categories and associated dimensions are identified to assess information (product) quality: intrinsic IQ (accuracy, objectivity, believability, and reputation), accessibility IQ (Access, security), contextual IQ (relevancy, value-added, timeliness, completeness, amount of data), and representational IQ (interpretability, ease of understanding, concise representation, and consistent representation). According to Wang (1998), accuracy is merely one of the four dimensions of intrinsic IQ category. He argued that representational and accessibility IQ emphasize the importance of the role of information.

O'Brien's model is one of most comprehensive data quality assessment model. According to the model provided by O'Brien (1996), all kinds of information can be evaluated at these three aspects: time, form and content, and each aspect is a multi-dimensional construct.

Rieh and Belkin (1998) research demonstrated that people assessed information quality based on source credibility and authority at either institutional or individual level. After scrutinizing former information quality studies, as well as summarizing characteristics of information problems and strategies in the WWW, Rieh and Belkin (1998) identified seven criteria for judgement of web information quality: source, content, format, presentation, currency, accuracy, and speed of loading. Further, they pointed out that quality and authority are indeed important to people searching in the Web.

Hackathorn (1998) contended that "acquiring content from the Web should not reflect positively or negatively on its quality". He suggested thinking of web resources in terms of quality and coverage (figure 4). In the two-dimension framework, commercial data, government data, and corporate data are placed relatively according to their relative quality and relevance. Hackathorn's data model is superior to previous ones on its consideration of characteristics of both Web and data warehousing. Issues of quality and coverage are measurable and of importance to data warehouse. "Quality" dimension makes sure that information downloaded from Web is business relevant and valuable for decision making, "coverage" dimension silts those data not compatible for data warehousing structure or those not fit organizational needs. However, he did not further his discussion on this data quality model such as items for these two dimensions etc, therefore model itself is not operationizable.

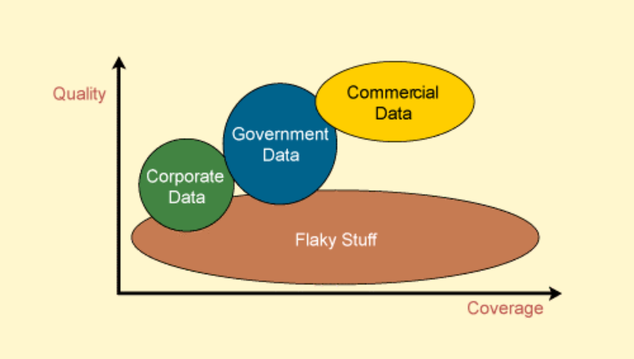

Figure 4. Hackathorn's web data evaluation model

Web data matrix model

We propose a data matrix model for WI web data evaluation. Basically, this is a two-dimension model, quality and coverage are two axis. Quality can be measured with Rieh and Belkin (1998) criteria for web information quality. 1. Source, where a document comes from. For example, the scholoars looked at "edu", "gov", or "com" URLs, and they thought education and government sites have "better quality" (Rieh and Belkin 1998). Also, from source, web users (WI analysts as well) can put some credibility in types of institution. 2. Content, as to what is in the document. Usefulness is a suitable judgement for the web content. 3. Format, formal characteristics of a document. For example, how the pages are presented, and how the information itself is structured are two important aspects of web data format. 4. Presentation, how a document is written/presented. The writing style is a typical symbol which makes this item different from item three. 5. Currency, whether a document is up-to-date. As we commented earlier, currency of web data is a significant value leading us to web integration. 6. Accuracy, whether the information in a document is accurate. And 7. The speed of loading, how long it takes to load a document. Waiting too long for a single page from an old server is waste of resources. WI analysts should avoid select too many such types of web sites as major web resources.

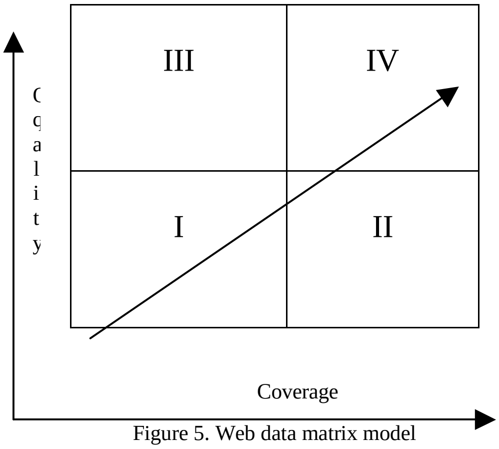

Figure 5. Web data matrix model

Coverage can be measured by following items. 1. Scope, the business relevance degree of data. Is the scope of web information too broad to be useful for one certain business task? 2. Variety, data is presented in which type and how many types of media. Zorn et al (1999) stress text mining technologies to web content is promising for accessing information on the web. Organizations should select pages under their technology process ability. Too complicated web pages may be too time, computer, and human labor consuming.

This matrix offers at least three aspects of functions. First, it helps to understand the development stages of Web information. Second, web data could be categorized into different types according to this matrix. To make it simple, we evaluate data at two levels in each dimension. For example, we have low vs. high quality, and low vs. high coverage data. Therefore, model split web data into four quadrants. Third, web information in different parts of this matrix has different role to play in decision-making supporting.

(1). In general, Web pages always develop from lower stages (area I or II) with a low quality and low coverage or a low quality but higher coverage to higher stages (high quality area as area III or area IV) no matter compare them "horizontally" with other similar "mature" pages or "vertically" with themselves.

(2). The parts of matrix will stand for different types of Web data.

Information resources at part III, IV are of high quality (for example, accuracy, currency, meaningful content etc), at II, IV are of wide coverage (for example, broad scope, multiple media pages). The interesting aspect of the Web is that its information resources occupy all the quadrants in this figure.

In part IV, for example, the commercial online databases from Dialog Information Services and similar vendors have traditionally supplied businesses with high-quality information about numerous topics (Hackathorn 1998). As aforementioned, Library Services, Digital Libraries, and Meta-Discovery Services, all belong to this quadrant. But, the complexity of using these services and the infrequent update cycles will limit their usefulness (Hackathorn 1998). In part III, government databases have become tremendously useful in recent years. Public information was often available only by spending many hours of manual labor at libraries or government offices before. Corporate Web sites often contain vast amounts of useful information in white papers, product demos, and press releases, eliminating the necessity to attend trade exhibits to learn the 'latest and greatest' in a marketplace. Company-Specific Content Providers, IT Trade Publications, Investment Services, and Government Agencies all locate in this part.

Part II contains "flaky" contents. Its value is not in the quality of any specific item but in its constantly changing diversity. In combination with the other Web resources, the flaky content acts as a wide-angle lens to avoid tunnel vision of one's marketplace. Some search engines and some news agencies belong to this type, such as General Discovery Services, Meta-Discovery Tools and General News Agencies (Hackathorn 1998).

Part I area information can be ignored for Web Integration. Many personal web pages belong to this type.

(3). The function of different parts of information to decision supporting varies. The part IV information can be used to support strategic decision. So, part IV can also be called strategic information area. EIS/ESS systems can take advantages from it. The information in part III has a narrower focus, it is suitable for low-level managers such as production managers or distribution managers. This area is operation/tactic information area. The part II information will perform together with other parts of information. Also, marketing managers can utilize this part of information to find some market trends, this time, the quality is not so vital but the coverage make more senses.

There is not a clear-cut border between information of these four different quadrants for their decision supporting functions. A professional person's homepage may provide critical information for a CEO of a large corporation when he is making important strategic decision. However, this kind of situation is quite occasional and is not our research focus.

Without doubt, high quality information is essential for strategy related decision making. Meanwhile, the coverage of information is another requirement for strategic decision. Because senior can not base their decision for the future several years on a single point or a single line of information, but on a wide covered plane of information. For example, they must synthesize manufacturing, marketing, finance, competition, human resource information to reach a well-fit plan. According to these considerations, part IV will be best selection. However, there is some short of this part web sites. Because, strictly say, there are not many web sites can be categorized into this type, with both high quality and broad coverage. But, a combination of many sources of part III can be in lieu of part IV information, which can be considered compensation. For example, web analyst can choose several leading finance web sites as resources for finance support of strategic decision making, several leading manufacturing web sites as resources for production management support, and so on.

DSS, ESS, EIS are all important decision supporting systems for management. ES is another kind of system that provides support for knowledge workers, management etc. one characteristic of these systems is a "drill down" feature (Lucas 1996). That means narrow and further (deep) instead of broad coverage information will be needed for such kinds of systems. Web sites belonging to part III best fit this requirement. Most high quality web sites are highly specific and so can be categorized into part III. For DSS, ESS, EIS systems, there are a lot of web resources available. There are two ways for these systems to utilize web resources. Fist, systems can build hyperlink to these web sites, managers can click these hyperlinks and wade through these sites to locate their wanted information. Some potential pitfalls exist in this method. Because this way require managers to have some training on web using. Another pitfall is that, this kind of utilization will cost senior management lot of time. Which is opposite the design idea of EIS/ESS systems: providing figures at high, summary level. The second method can save these problems, which need web analyst to locate, summarize information from pre-selected part III web sites and load into Data Warehouse. Web integration can provide processed information for DSS, EIS, ESS etc. be merging web resources, DSS, EIS, ESS will provide management broader vision, fresher ideas, which make these systems more attractive.

Besides these systems, middle level and field managers can also directly take advantage of part III information for their daily operation decision making. For instance, production managers can surge ISO web site to find some products standard, competitors' web stations for some manufacturing aspiration etc.

Part II information is outstanding for its coverage and will be allure to marketing managers. For marketing issues, broad and scattered information may be more interesting sometime.

References

Anonymity, (A). "Ardent gets ahead of the XML herd", *Computing*, April 29, 1999, pp. 16.

Anonymity, (B). "Trends and developments in data warehousing", *America's Network* (103:8), 1999, pp. 22-24.

Booker, E., "Data Warehousing -- Unleash The Treasure Trove", *Internetweek*, iss. 771, 1999, pp.41-

Corbin, L., "Data warehouses hit the web", *Government executive* (29:2), 1997, pp. 47-48.

Flanagan, P., "The 10 hottest technologies in telecom," *Telecommunications* (30:5), 1997, pp. 29-38.

Gardner, D.M., "Cashing in With Data Warehouses and the Web," *Data Based Advisor*, 1997, pp. 60.

Hackathorn, R., "Farming the Web", *Byte* (22:10), 1997, pp. 43.

Hackathorn, R., "Rouging the Web for Your Data Warehouse", *DBMS* (9:11), 1998, pp. 36.

Inmon, W. H., "The data warehouse and data mining," *Communications of the ACM* (39:11), 1996, pp. 49-50.

Inmon, W.H., "Using the data warehouse", New York, John Wiley, 1994

Inmon, W.H., “What is a Data Warehouse?”, http://www.cait.wustl.edu/cait/papers/prism/vol1_no1/ (Current April, 2000)

Inmon, W.H., (1998). “Data warehouse performance”, New York, John Wiley.

Internet Software Consortium, http://www.isc.org/dsview.cgi?domainsurvey/WWW-9907/report.html, (Current April, 2000)

Kirshenberg, S., “Info on the Internet: user beware!” *Training & Development* (11:52), 1998, pp. 83.

Klobas, J. E. “Beyond information quality: fitness for purpose and electronic information resource use”. *Journal of information science* (21:2), 1995, pp. 95-114.

Koehler, W., "A descriptive analysis of Web document demographics: A fist look at language, domain names, and taxonomy in Latin America. In C. Chen (Ed.), *Proceedings of the 9th International Conference*, 1996, New Information Technology, Pretoria, South Africa, November 11-14, 1996 (pp. 159-170), West Newton, MA: MicroUse Information.

Koehler, W., “An analysis of web page and web site constancy and permanence”, *Journal of the American Society for Information Science* (50:2), 1999, pp. 162-180

Levine, S., “If you build it, will they come? Data warehouses”, *American’s Network* (8:102), 1998, pp. 18.

Lucas, H.C., “Information Technology for Management”, New York, McGraw Hill, 1997

Marchetti, M., “Shifting gears”, *Sales and Marketing Management* (151:12), 1999, pp. 38-48.

McCure, J.C, “Data, Data Everywhere” *Management Review (American Management Association)* (10:87), 1998, pp. 10.

Miley, M., “Snapshots from the Web”, *LAN Times* (15:2), 1998, pp.29.

NetSizer, http://www.netsizer.com (Current April, 2000)

Netwizards, http://www.nw.com (Current April, 2000)

Nua’s Index of Net Surveys, http://www.nua.ie/surveys/ (Current April, 2000)

O'Brien, J., “A. Management Information Systems: Managing Information Technology in the Networked Enterprise”, 3rd. Ed. Richard D. Irwin, 1996.

Orr, K., “Data quality and systems”, *Communications of the ACM* (41:2), 1998.

Redman, T.C., “The impact of poor data quality on the typical enterprise”, *Communications of the ACM* (41:2), 1998.

Rieh, S. Y. & Belkin, N. J., "Understanding judgment of information quality and cognitive authority", *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, 35, 1998, pp. 279-289.

Scientific American, http://www.archive.org/ (Current April, 2000).

Sprague R.H. & Watson, H.J., “Decision Support for Management”, Upper Saddle River, N. J., Prentice-Hall, 1996.

Tayi, G.K. and Ballou, D.P., “Examining data quality”, *Communications of the ACM* (41:2), 1998.

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

Wang, R.Y., (). “A product perspective on Total Data Quality Management”, *Communications of the ACM* (41:2), 1998.

Wilder, C. “Web Data – Tapping The Pipeline – Web sites can offer a wealth of customer data”, *Information Week,* March 15, 1999

Wilent, S., "Pulling packages from the Web", *Databased Web Advisor*, 1997, pp. 32.

Wilson, L. “Weaving a web warehouse supplement”, *Software Magazine*, July 1997, pp. 13-14.

Zorn, P., Emanoil, M., Marshall, L., and Panek, M. “Mining meets the web”, *Online* (23:5), 1999, pp.16-28.