2009

How Consistent is Web Information - A Case Study on Online Real Estate Databases

Ningning Wu
University of Arkansas at Little Rock, nxwu@ualr.edu

Irit Askira Gelman
University of Arizona, askirai@email.arizona.edu

Isaac O. Osesina
University of Arkansas at Little Rock, oiosesina@ualr.edu

How Consistent is Web Information– A case study on Online Real Estate Databases

Ningning Wu
Information Science
University of Arkansas at Little Rock
nxwu@ualr.edu

Irit Askira Gelman
University of Arizona
askirai@email.arizona.edu

Isaac O. Osesina
Applied Science, University of Arkansas at Little Rock
oiosesina@ualr.edu

ABSTRACT

Inconsistent information among different websites indicates potential data quality problems such as accuracy, completeness, timeliness, etc. Unless the user is able to tell which information is accurate, it can lead to the user's concern about the believability of the information and will prevent the effective use of information. This paper attempts to study how consistent the information from different websites will be. A case study is conducted based on two widely used real-estate databases, Zillow.com and mls.com. The preliminary results show a large discrepancy in information between the two.

INTRODUCTION

The World Wide Web (Web) is the largest distributed information system that is ever built. It contains interlinked hypertext documents that are managed by autonomous systems and accessed via Internet. With advances in Internet and web technology, information management, and search engine, the Web is quickly becoming the most popular information resource. It enables the spread of information over the Internet and makes the access to information much easy and convenient. In addition it greatly changes the way that information is disseminated, acquired, shared, and utilized.

The Web is evolving from a simple information management system for the purpose of sharing and viewing information in the form of hypertext over a dedicated computer network to a gigantic repository of information created over the Internet, and it keeps growing and evolving. Driven by the need to find, share, and combine web information in a better way so that it can be utilized more effectively, the Web is evolving to embed semantics into the information and services. The semantic web is an extension of the Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content (Lerners-Lee *et al.,* 2001). With the help of blogs and wikis, people are not only the consumers of web information but also the active creators of information. In addition, people are able to interact on the Web and create a new form of social circle, so-called on-line communities through the social network services like Facebook, MySpace, Twitter, etc. This brings the concept of social semantic web, which can be seen as a web of collective knowledge systems, which are able to provide useful information based on human contributions and which get better as more people participate (Gruber 2006). While the semantic web provides ways for business to interoperate across different domains, the social semantic web enables users to share knowledge. Supported by the technologies from semantics web, social software and Web 2.0, the social semantic web is fast booming and becoming a mega-trend of the Web today.

On one hand, the Web is often the first resource people tend to use to search for the information they need because it is convenient, fast, and informative. On the other hand, there is little control over the quality of information as the building blocks of the Web -- hypertext documents –are managed by autonomous systems, each of which has different requirements for quality and quality assuring procedures. For example, timeliness and accuracy of information is crucial for a stock exchanges system, while little is required for a social networking site except accessibility.

There are many issues in Web information quality. First, the quality of information in independent documents might change over time due to many reasons. Second, Information in Web documents is often related in many ways. For

example, a document may contain the original information that is linked by other documents, the information that is extracted or derived from other documents, or the same information that is kept in other documents. Since the source document and referencing document reside on different systems, the discrepancy between the two may exist when their updates are not synchronized. Third, even for the documents that intend to have the same information, they may differ in the content and quality as the same piece of information may be collected from different sources or from the same source but at different times.

Problems arise when a query is submitted to an Internet search engine and returns a number of hits containing different information about the subject, then which information would the user accept? The problem of identifying high quality documents has been well studied in the field of information retrieval. In the context of web information retrieval, several approaches have been proposed to incorporate quality metrics in information retrieval on the Web. The research by Zhu *et al.* aims to improve search effectiveness by incorporating the currency, popularity, information-to-noise ratio, and page cohesiveness measures in both centralized and distributed search (Zhu and Gauch, 2000). The research by Pun *et al.* proposed an ontology-based cohesiveness measure to find high quality web-pages based on how close the concepts in a page are related to each other (Pun and Lochovsky, 2005). Lachica *et al.* proposed a framework for ranking information based on quality, relevance and importance (Lachia *et al*., 2008). In addition, several frameworks have been developed to measure information quality on the Web (Alexander and Tate, 1999; Eppler and Muenzenmayer, 2002; Katerattanakul and Siau, 1999; Knight and Burn, 2005). Most of these researches focus on the information quality assessment of a single web page or web site, and the quality measures are calculated based on both content-based features and non-content based features such as authority, reputation, popularity of the website. Hence, when a query results in multiple hits with similar quality measures, the user will have to decide which one to use. A potential problem arises when the top quality ranking websites are inconsistent on the data of great importance to the user's decision making. This will cause the user's concern about the reliability of the information and would potentially impair their decision making.

Information quality is a multi-dimensional concept. A number of frameworks have been proposed to present information quality in different sets of dimensions from different perspectives; however, they share a number of characteristics regarding their classification of the dimensions of quality (e.g., Alexander and Tate, 1991; Naumann and Rolker, 2000; Kahn *et al*., 2002; Katerattanakul *et al.,* 1999; Wang and Strong, 1996).The most common dimensions identified in these frameworks include accuracy, consistency, security, timeliness, completeness, concise, reliability, accessibility, availability, objectivity, relevancy, usability, and understandability.

This paper attempts to study the consistency issue of the related information on the Web. More specifically, we want to study how information about a same entity but from different websites would differ. Here by entity we mean anything that is identifiable and can be described by a number of features. So regardless of how the representations of an entity differ in format and structure, we can use a set of features to represent the entity and evaluate the discrepancies among its various representations. Obviously, inconsistent information not only causes confusion, but also discourages the effective use of data. Methods should be developed to identify the root cause of inconsistent data so as to improve the overall quality of web information.

LITERATURE REVIEW

Consistency has been long known as one of the most important dimensions of data quality and is identified in the majority of data quality frameworks (Dedeke 2000; Naumann and Rolker, 2000; Kahn *et al*., 2002; Katerattanakul *et al.*, 1999; Wang and Strong, 1996). Several studies have been conducted that aim to improve data consistency in different systems. Cong *et al*. (2007) proposed algorithms for computing and repairing data inconsistencies in a traditional database by employing a class of conditional functional dependencies (CFD) to specify the consistency of the data. Angeles and MacKinnon (2004) proposed a Data Quality Manager for the database systems to select best data sources and query execution plans, detect and resolve inconsistent data, and to integrate and rank the query results based on user-defined quality criteria. Svensson *et al.* (2004) used use cases to describe how data modeling concepts and a service-oriented architecture can be used to resolve data inconsistencies in an IT landscape.

Klein (2002) conducted an empirical research on when users detect information quality problem on the World Wide Web and proposed a theoretical model of factors influencing the phenomenon. A number of frameworks have been proposed for assessing information quality on the Web. Knight and Burn (2005) presented the IQIP (Identify, Quantify, Implement, and Perfect) model to assess the quality, refine and perfect the quality retrieval process based on relevancy. Bizer *et al.* (2007) developed a Web Information Quality Assessment Framework (WIQA) and a tool to assess the quality of a website according to the task specific criteria. Preece *et al.* developed a framework for

managing information quality in an e-Science context using semantic web technology. It uses ontology to capture generic and domain dependent quality descriptors, and binding annotation models to associate concepts in the IQ ontology with data and service entities. Moustakis *et al.* (2004) presented a hierarchical framework that uses the Analytical Hierarchy Process to assess website quality.

In the context of measuring web information quality, Eppler and Muenzenmayer (2002) presented an IQM methodology to match information quality criteria with adequate measurement tools so that specific IQ-criteria can be measured in a systematic and planned way. Zhang *et al* (2000) conducted an explorative analysis on information quality of commercial website home pages based on user perceptions of presentation, navigation and quality of web home pages. The study showed empirical evidence of relationships between companies at different positions in the supply chain and the information quality of their Web home pages. Gelman and Barletta (2008) proposed a simple metric, based on the reported hit counts of search engine queries on a pre-defined set of commonly misspelled words, for assisting in the evaluation of the quality of websites.

In addition, a number of approaches have been proposed to find high quality web pages in the domain of web information retrieval and Internet search by 1) incorporating quality metrics (Zhu and Gauch, 2000); 2) defining a distance metric for measuring web page cohesiveness (Pun and Lochovsky 2005); 3) using a document quality language model approach (Zhou and Croft, 2005); 4) employing a socio-semantic contextual approach that extends topicality to enhance precision in information retrieval (Lachica *et al.*, 2008); 5) introducing the Portal Data Quality Assessment tool for measuring the data quality based on suitability for the task at hand(Corritore *et al.*, 2003).

OUR METHODOLOGY AND DATA

The study selects the data provided by two real estate services, mls.com and zillow.com as the target data for two reasons. First, it is easy to obtain a common set of houses listed by both services. Second, both services are well-established and considered to have the data of relatively high quality. As the data of two databases may have been collected from different sources, we would expect some discrepancies may exist; however, it would be interesting to see how big the discrepancy is. In addition, the results of this study will offer an insight into the consistency issue of the related information on the Web.

MLS Data

MLS.com is a free Multiple Listing Service search engine. It provides real estate listings for sale by Realtors® and other realty professionals that are members of MLS. It consists of listings of all 50 states in US.

The data quality of MLS varies for each listing agent as MLS does not have direct control over the data quality. Generally the information of a listed property either comes from the county tax records or from the reassessment of the listing agent, and this information is usually verified by the property owner. Thus it is reasonable to assume MLS data are generally current and of relatively high quality. Although measures of data quality of MLS data are not available, anecdotal evidence indicates that errors are not rare (as well as missing values.

Zillow Data

Zillow is created by Rich Barton and Lloyd Frink, founders of pioneering online travel service Expedia.com, with the goal of "helping people make smarter real estate decisions." It was launched in February, 2006. Zillow aims to satisfy the needs of three classes of users in the US market: home buyers, home sellers, and owners. For home buyers, it provides access to data about houses in neighborhoods of interest. A buyer can also review recently sold homes to get a sense of neighborhood trends. Sellers and home owners can find out the value of their home, neighborhood trends, and how their home is valued compared to others in the same ZIP code.

Major components of Zillow include:

(1) A database that currently stores data on 65 million homes in the US. The database includes attributes that describe the legal identification of a property and its physical characteristics, such as the square footage, number of bedrooms, number of bathrooms, number of stories, age, and lot size. Additional attributes inform about the sales history of the property, its assessed value for tax purposes, and tax rates. Data also include aerial images, which, in some areas, enable a bird's eye view with a 3D-like quality, and maps.

(2) The "Zestimate," a home market valuation that is calculated using a proprietary statistical model. Zillow offers both current and historic Zestimates, which facilitate the identification of market trends along time.

(3) The "Zindex," which is the median Zestimate valuation for a given geographic area at a given point in time.

Data quality is of major importance to Zillow. Since the Zestimate and Zindex are both calculated from the home data, the availability and accuracy of these estimates depend on the completeness and accuracy of such data. Due to deficiencies in completeness, the database is currently limited to 65% of the homes in the US. In addition, the Zestimate and Zindex are available for just 45 million homes—less than 75% of the homes reported by the database.

Data are also often inaccurate. The service provider acknowledges the fact that users frequently find inconsistencies between the actual values (e.g., number of bedrooms or the square footage) and the data that Zillow shows.

Zillow draws its data from a multitude of public sources. These include county assessor offices, county recorder offices or similar city or local government offices that are responsible for recording real estate transactions, and other public sources. The quality of the data supplied by these sources varies. For example, some counties provide all the needed data, while others are lacking key attributes such as the number of bedrooms and bathrooms, or, in some cases, the square footage of the home.

Methodology

The initial aim of this study is to compare information about several features of a house e.g. number of bedrooms, number of bathrooms, number of floors, size of the house, year-built, stories, and fireplace. However, due to unavailability in structured form of all the information on MLS's website, only four features are included in the study. They are number of bedrooms, number of bathrooms, size of house, and year-built. In addition, we extract the last-date-sold information for each property from Zillow. We consider last-date-sold could potentially affect timeliness of a house information. For instance, a house that has not changed the owner for long time is likely to remodel overtime compared to a recently built or bought house; however, its information is often not updated until it is reassessed or put for sale. This would cause discrepancy between the data of the two databases if either one contains outdated data.

A web crawler written in Java was developed for this study. The program first extracted from MLS the following information about properties: State, City, Address, bedrooms, bathrooms, size, and year-built. Properties whose address contains a house number were then searched for on Zillow. Information about each property's bedrooms, bathrooms, size, and year-built were then extracted from Zillow and compared with MLS's data. Inconsistent rate, defined as the percentage of properties that have inconsistent features with the two databases, is calculated.

OUR FINDINGS

The study selected a random sample of 1067 properties that cover 41 states and 446 cities in the United States. The quality analysis based on these properties was conducted for each target feature in terms of house age and last-date-sold.

Quality Analysis in Terms of House Age

The charts below summarize the findings of the study. Properties are partitioned into 6 age groups. If a house has different values for year-built from two databases, then the smaller value is used for calculating its age. Figures 1 to 4 show inconsistency rates for the properties of different ages in terms of number of bedrooms, number of bathrooms, house size in square feet, and year-built respectively. For example, Figure 1 shows, for each age group, the number of properties that have the same information on number of bedrooms as well as the number of properties that do not. The percent under each age group is the inconsistency rate expressed as the percentage of properties that Zillow and MLS have different values for the target feature.

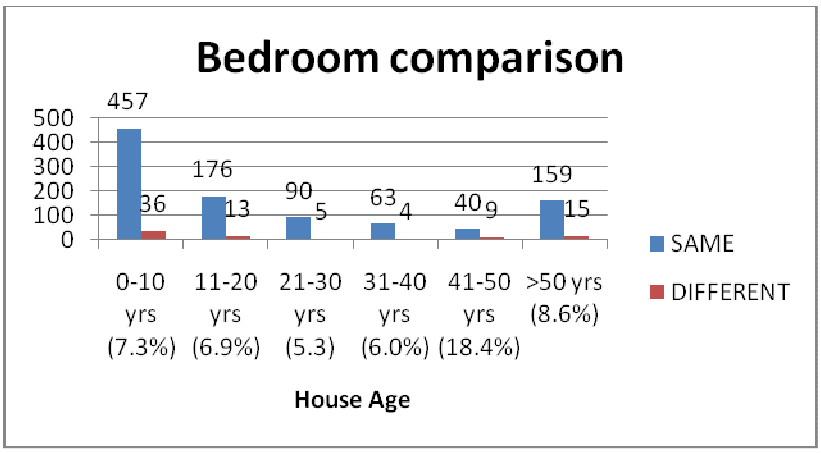

Figure 1: Inconsistency rate of Bedroom

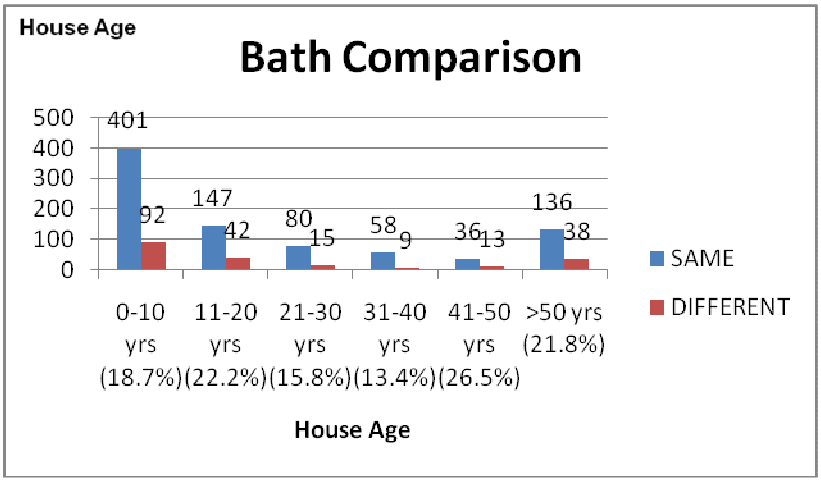

Figure 2: Inconsistent rate of Bathroom

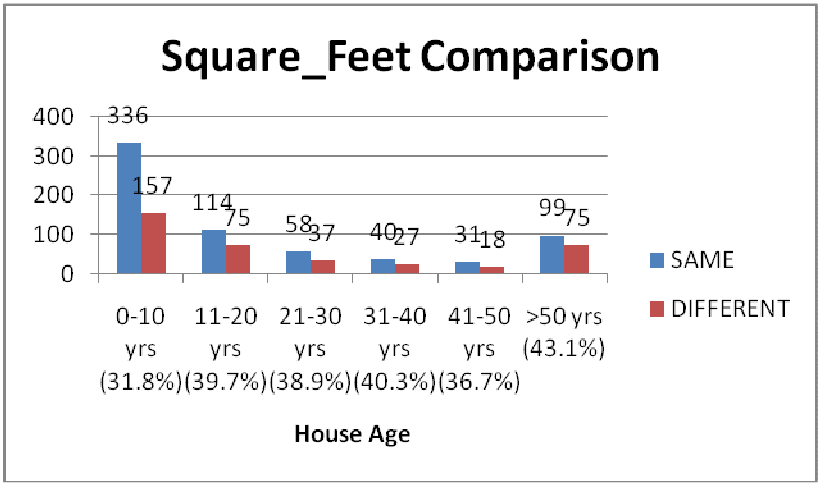

Figure 3: Inconsistent rate of house size

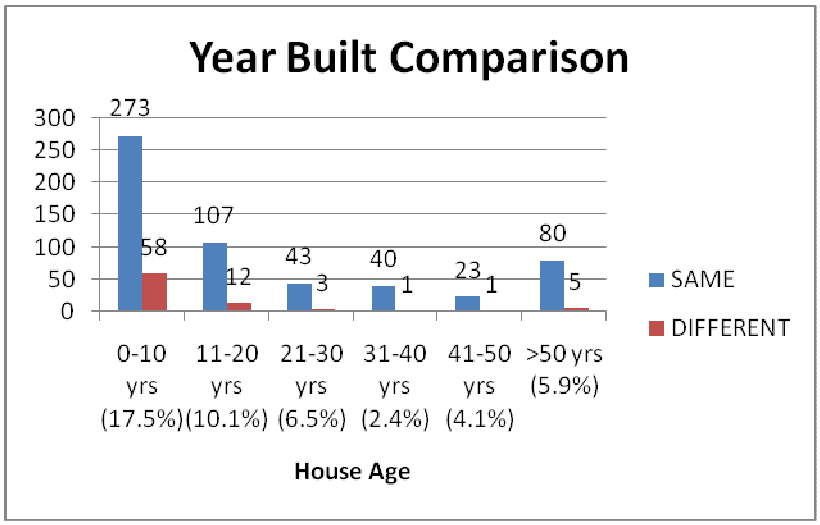

Figure 4: Inconsistent rate of year-built

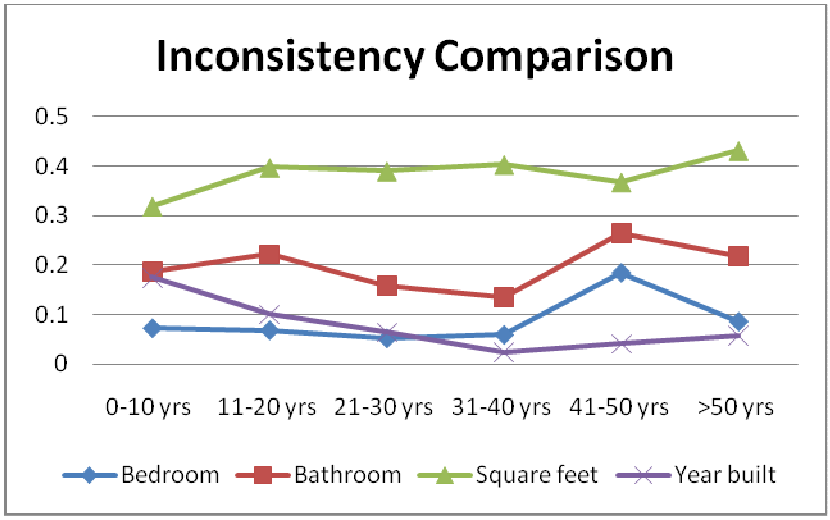

Figure 5: Inconsistency comparison of different features in terms of year built

Figure 5 shows the inconsistent rates for all four features. The house size, represented by square feet, has the highest average inconsistency rate. Over one-third of properties have inconsistent information on house size, and there is no exception for the newly built houses. The possible reason for such a high rate of inconsistency is that there is no common standard for calculating square footage of a house and the calculation varies for different geographic areas. Some real estate agents tend to use the square footage listed in the county tax records in their marketing materials. Unfortunately, this information is often outdated especially with older homes because over time basements get finished and additions are constructed and these changes will increase the chances that tax records will be outdated and inaccurate.

The feature with the second highest average inconsistency rate is bathroom. This is also understandable because bathroom along with kitchen are the two most popular areas for home remodeling. Although addition of a new bathroom is not rare for older homes, it is interesting to see that close to 20% of properties built within last ten years has inconsistent bathroom information.

The feature year-built has the lowest average rate of inconsistency; however houses of less than 10 years old have the highest inconsistency rate, about 17.5%. This is interesting because once a house is built, its built year will not

change unless it is rebuilt, which is rare let alone for the houses of less than 10 years old. The rate decreases as the house age increases until it reaches the lowest point when the house age is between 31 to 40 years, then it starts to increase slightly for the houses of 41 years and older.

Overall, Figure 5 shows that the older homes do not necessarily have a higher inconsistency rate than the newer homes. Further study is needed to investigate the causes for inconsistent information between the two sites.

Quality Analysis in Terms of Last-sold-date

A similar quality analysis was performed for the same four features in terms of last-date-sold. The result is shown in Figure 6. For the features bedrooms and year-built the inconsistency rate of the properties sold within last 5 years is no better than those sold less recently. This is interesting for two reasons. First, bedroom and year-built are the least likely changed features of a property sold recently, so they are supposed to have a low inconsistency rate. Second, the properties with the recent last-sold-date are likely to contain the updated information due to its last transaction, so the chances for them to have inconsistent information from two databases should be lower. However, the result is different from our expectation. More than 10% of properties sold within last 15 years have different built-year information and the rate drops to zero for the properties whose last-sold-dates are more than 20 years ago. The feature square feet has the highest inconsistent rate --over 30% of properties have inconsistent sizes regardless of when they were sold last time except for the properties that were not sold within last 36 years. The feature bathroom has the second highest inconsistent rate. Figure 6 shows that the most recently sold properties do not have a lower inconsistency rate compared to other properties.

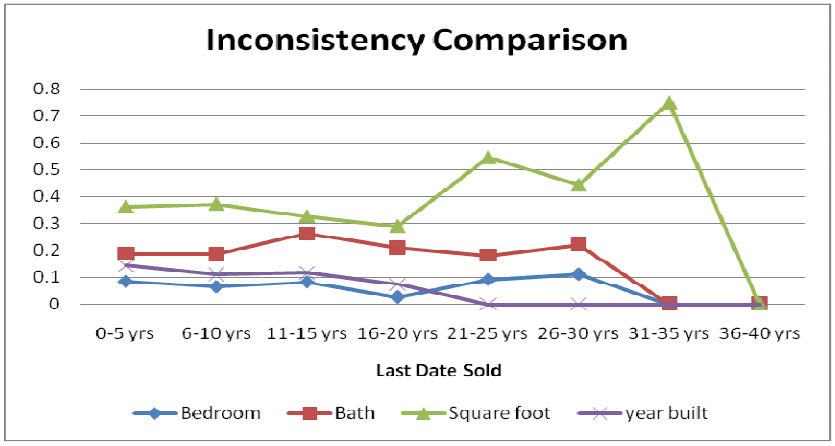

Figure 6: Inconsistency comparison of different features in terms of last sold date

Summary

Out of 1067 properties, only 35% of properties have consistent information on all four features, namely bedroom, bath, square feet, and year-built. The rest 65% of properties have inconsistent information on either one or more features. The detailed result is shown in Table 1. For example, 32.9% of properties are inconsistent on one feature, 22.1% on two features, 5.4% on three features, and 4.3% on all four features. The result shows a large discrepancy between zillow.com and mls.com. Such a discrepancy points to potential quality problems of at least one database if the other one is assumed to more accurate.

No. of features different	0	1	2	3	4
No. of properties	376	351	236	58	46
Percentage of properties	35.2%	32.9%	22.1%	5.4%	4.3%

Table 1: Inconsistency rate analysis

CONCLUSION AND FUTURE WORK

In this project, we used two well-established real-estate databases as the case study to evaluate information consistency among different sites on the Web. The preliminary results, based on a random sample of 1067 properties, show that about 65% of the properties has inconsistent information with two databases. Such a high inconsistent rate is beyond our expectation because the two sources have been widely used by the real-estate professionals, perspective home buyers, and sellers and are considered to have relatively high quality. The study also reveals the extent of inconsistent information on the Web might be greater than what we thought.

Inconsistent information among different websites indicates potential data quality problems such as accuracy, timeliness, believability, reliability, etc. It can impair the user's decision making and prevent the effective use of information. Our future research question will be whether it is possible to produce accuracy measures for each database.

A recently popular approach in the data quality literature to objective accuracy assessments applies data mining techniques. This approach (e.g., Luebbers *et al*., 2003) relies on the assumption that error rates are low, such that patterns in the data with strong support reflect correct links and instances that disagree with these patterns are probably incorrect. A second approach directs the use of a sample of clean data (e.g., Rakov, 1998) in database accuracy measurement. However, as is often the case, a sample of clean data is out of reach.

Our initial findings imply that the approach that applies data mining techniques for the assessment of data accuracy may be inappropriate with the databases of Zillow and MLS at this stage. If subsets of the data are infected by high error rates, application of a data mining algorithm indiscriminantly may yield invalid patterns. In contrast, the method that is proposed by (Rakov, 1998) may be particularly suitable for accuracy measurement due to its special sensitivity to variations in accuracy. This approach identifies data subsets whose quality is homogeneous in data sets where the quality of the data as a whole is heterogeneous. Furthermore, assuming that error rates of one database, say MLS database, do not exceed few percents, such data may prove useful in the detection of high error rates, and in measurement of the accuracy of data that are characterized by high error rates, in place of the sample of clean data that is assumed in the application of the data analysis algorithm in (Rakov, 1998).

We plan to examine the value of the method described by (Rakov, 1998) in the measurement of Zillow database accuracy, using MLS data as a clean sample. The objective of such application will be limited to the detection of data subsets with high error rates, and corresponding accuracy measurement. The success of our approach may suggest its potential usefulness in other scenarios that involve multiple, overlapping sources. Specifically, data that can be assumed to have low error rates may be useful in initial assessments of the accuracy of overlapping data in other sources.

REFERENCES

1. Alexander, J. E. & Tate, M. A. (1999) Web Wisdom: How to Evaluate and Create Information Quality on the Web. Mahwah, NJ: Erlbaum.
2. Bizer, C., Cyganiak, R., Maresch, O., and Gauss, T. (2006) The WIQA - Web Information Quality Assessment Framework, *http://www4.wiwiss.fu-berlin.de/bizer/WIQA/index.htm*
3. Calero, C., Caro, A., and Piattini, M. (2008) An Applicable Data Quality Model for Web Portal
4. Data Consumers, *World Wide Web, Vol 11*, 465-484.
5. Cong, G., Fan, W., Geerts, F., Jia, X., and Ma, S. (2007) Improving Data Quality: Consistency and Accuracy, *32nd International Conference on Very Large Data Bases,* 2006, 315-326.
6. Dedeke, A. (2000) A Conceptual Framework for Developing Quality Measures for Information Systems. *Proceedings of 5th International Conference on Information Quality,* 126-128.
7. Gelman, I. A. and Barletta, A. (2008) Initial Study of A "Quick and Dirty" Website Data Quality Index. *Proceedings of the 13th International Conference on Information Quality*, Nov. 2008, MIT, Cambridge, MA.
8. Gruber, T. (2006) Where the Social Web Meets the Semantic Web. *Keynote presentation as the 5th Internatinoal Semantics Web Conference*, Nov. 2007, Athens, GA.
9. Eppler, M. and Muenzenmayer, P. (2002) Measuring Information Quality in the Web Context: a Survey of State-of-the-Art Instruments and an application Methodology, *Proceedings of 7th International Conference on Information Quality,* November 2002, MIT, Cambridge, USA, 187-196.
10. Klein, B.D. (2002) When Do Users Detect Information Quality Problems On The World Wide Web? *8th American Conference in Information Systems,* 2002, 1101-1103.

11. Lachica R., Karabeg, D. and Rudan S. (2008) Quality, Relevance and Importance in Information Retrieval with Fuzzy Semantic Networks, *Proceedings of 4th International Conference on Topic Maps Research and Applications*, 2008, Germany.
12. Luebbers D., Grimmer, U., and Jarke, M. (2003) Systematic Development of Data Mining-Based Data Quality Tools, VLDB, 2003.
13. Kahn, B. K., Strong, D. M. and Wang, R. Y. (2002). Information quality benchmarks: Product and service performance, *Communications of the ACM, 45* (4), 84–192.
14. Katerattanakul, P. and Siau, K.(1999) Measuring Information Quality of Web Sites: Development of an Instrument, *Proceedings of the 20th International Conference on Information Systems* December 12-15, Charlotte, NC, USA, 279 – 285.
15. Knight, S.A and Burn, J. (2005) Developing a Framework for Assessing Information Quality on the World Wide Web, *Information Science Journal, vol 8, 2005,* 159-172.
16. Merners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantics Web. *Scientific American Magazine*, May 2001.
17. Moustakis, V.S., Litos, C., Dalivigas, A., and Tsironis, L. (2004). Website Quality Assessment Criteria. *Proceedings of 9th International Conference on Information Quality*, Nov. 2004, MIT, Cambridge, USA, 59-73.
18. Naumann, F. and Rolker, C. (2000) Assessment methods for information quality criteria, *Proceedings of 5th International Conference on Information Quality*, p.148–162.
19. Preece, A., Jin, B., Pignotti, E., Missier, P., Embury, S., Stead, D., and Brown, A. (2006). Managing Information Quality in e-Science Using Semantic Web Technology, *3rd European Semantic Web Conference 2006, LNCS 4011, 472–486*
20. Pun, J. C. C. and Lochovsky, F. (2005) Finding High-Quality Web Pages Using Cohesiveness. *Proceedings 10th International Conference on Information Quality*, November 2005, MIT, Cambridge, USA.
21. Rakov, I. (1998) Data quality and Its Use for Reconciling Inconsistencies in Multidatabase Environments, Ph.D. Dissertation, George Mason University, May 1998.
22. Rudra, A. And Yeo, E. (1999) Key Issues in Acheiving Data Quality and Consistency in Data Warehousing among large Organizations in Australia. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.
23. Svensson, E., Vetter, C., and Werner, T. (2004) Data Consistency in a Heterogeneous IT Landscape: A Service Oriented Architecture Approach, *Proceedings of the 8th IEEE Intl Enterprise Distributed Object Computing Conf , 2004, 1541-7719/04*
24. Wang, R.Y. and Strong, D. M. (1996) Beyond accuracy: What data quality means to data consumers, *Journal of Management Information Systems, Spring,* 5–33.
25. Zhang, X., Keeling, K. B. and Pavur, R. J. (2000) Information Quality of Commercial Web Site Home Pages: An Explorative Analysis, *Proceedings of the Twenty First International Conference on Information Systems*, Brisbane, Queensland, Australia, 164 – 175.
26. Zhou, Y. and Croft, W.B.(2005) Document Quality Models for Web ad ho retrieval. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Oct.-Nov. 2005, Bremen, Germany, 331-332.*
27. Zhu, X. and Gauch, S. (2000) Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web, *Proceedings of the 23rd Annual International ACM Conference on Research and Development in Information Retrieval,* Athens, Greece, 288-295.