Faculty of Informatics, Masaryk University, Czech Republic

Habilitation Thesis

Data Quality Management for Recommender Systems

Mouzhi Ge, Ph.D.

October 2017

Abstract

Recommender systems are developed to help users find relevant products that may interest them. The goal of recommender systems is to reduce information overload and provide personalized recommendations for users. Over the last decade, recommender systems have been widely applied in e-commerce, for example, book recommendation on Amazon or movie recommendation on Netflix. Moreover, a number of case studies have stated that the use of recommender systems can increase user satisfaction and produce added value to business. In order to produce high-quality prediction in recommender system, one of the important factors is to use high-quality data. Therefore, data quality management is critical for the predictive analytics systems such as recommender systems. The quality of data used for applications is directly related to quality of predictions in recommender systems. Besides, data quality management has been found as an efficient and effective way to use data for predictions and business decision makings.

The main goal of this habilitation thesis is to provide an overview of my research achievements in the areas of recommender systems and data quality management. For recommender systems, it covers state-of-the-art topics of explanation component in recommender system, diversity in recommendation list, and the new application domains such multimedia and food recommender systems. On the other hand, the data quality research is focused on the data quality assessment, data quality and data integration as well the effects of data quality in enterprises. The thesis will provide a comprehensive discussion on the recent research findings regarding the recommender systems and data quality management.

This thesis is written as a commentary to a collection of 10 peer-reviewed papers published in international journals such as International Journal of Human-Computer Studies, Journal of Computer Information Systems and International Journal of Semantic Computing, also in international conferences such as European Conference on Information Systems and International Conference on Electronic Commerce and Web Technologies, as well as Springer Book Chapters. Each paper will be summarized and briefly discussed, which tends to provide an outline for my research over last 8 years after my Ph.D. In the selected papers, my personal contribution for these papers is between 30% to 90% with an average of approximately 60%.

Table of Content

Chapter 1: Introduction

The extensive growth of data in our daily life has created complexity in decision-making (Ricci et al. 2011, Jannach et al. 2011). Making choices by exploring the entire product catalog manually can be very time-consuming and usually frustrating. Therefore, Recommender systems are developed to reduce the information overload and provide personalized suggestions to assist users' decision making (Adomavicius and Tuzhilin 2005).

Recommender systems are information search and filtering tools (Ricci et al, 2011; Konstan and Riedl, 2012) that help users to make better choices while searching for products such as movies, restaurants, vacations, and electronic products. As Recommender systems are playing an important role throughout the Internet, they have been applied in a large number of Internet applications such as Amazon, YouTube, Netflix, Yahoo, Tripadvisor, Last.fm, and IMDB etc (Davidson 2010, Linden 2003). Moreover, social networks such as LinkedIn and Facebook have also introduced recommendation technology to suggest groups to join and people to follow (Baghaei, 2011).

In order to produce high-quality prediction in recommender system, one of the important factors is to use high-quality data. Numerous business initiatives have been delayed or even cancelled, citing poor data quality as the main reason. The problem of poor data quality has caused various organizational losses, such as losing customers and making incorrect decisions. Case studies of these data quality problems can be found in a plethora of reports, journals and books. Many of the data quality problems are pervasive, costly and even disastrous. For example, more than 60% of 500 medium-size firms were found to suffer from data quality problems (Wand and Wang 1996). It is estimated that an industrial information error rate up to 30% is considered typical and it is often reported that the error rate rises to 75% (Redman 1996). In recognition of the criticality of data quality, organizations have become increasingly aware of its importance for Business Intelligence.

1.1 Development of Recommender Systems

Over the last decade, various techniques in the areas of information retrieval and information filtering have been developed to help users find items that match their information needs and filter out unrelated information items. In contrast to information filtering techniques implemented in search engines, whose aim

is to retrieve the desired information from a large amount of information based on a user query, Recommender Systems are developed to reduce the information overload and provide personalized suggestions to assist users' decision making \cite{Adomavicius2005}.

Rooted in the fields of information retrieval (IR), machine learning (ML) and decision support systems (DSS), from the mid-1990s recommender systems have become an independent research area of its own (Jannach et al. 2011). Recommender systems propose ranked lists of items (that are subsets of a larger collection) according to their presumed relevance to individual users. Relevance is determined from explicit and implicit user feedback such as ratings on items, commercial transactions or explicitly stated requirements. With the rapid growth of electronic commerce, the ubiquity of mobile information access and the advent of the Social Web, the interest in RS research has grown enormously during the past years. This is for example documented by the rapidly growing ACM Recommender Systems conference series as well as by the publication of various focused journal special issues and books. The reasons for this high attractiveness of the field are manifold and include highly visible competitions such as the Netflix prize, increased industrial interest or the new application opportunities for recommendation techniques in the Mobile and Social Web.

Commonly, recommender systems are classified into four categories: collaborative filtering, content-based filtering, knowledge-based systems and hybrid recommendation approaches (Jannach et al. 2011). Collaborative filtering (CF) approaches exploit the wisdom of the crowd and recommend items based on the similarity of tastes or preferences of a larger user community. Content-based approaches, on the other hand, recommend items by analyzing their features to identify those items that are similar to the ones that the user preferred in the past. Knowledge-based recommender systems, finally, rely on explicit user requirements and some form of means-ends knowledge to match the user's needs with item characteristics. In order to benefit from the advantages of the different main approaches, hybrid recommendation systems try to combine different algorithms and exploit information from various knowledge sources. Studies have shown that for example hybrids which combine content-based and collaborative filtering can lead to more accurate predictions than pure CF or content-based recommenders (Gedikli et al. 2011).

1.2 Development of Data Quality Management

The initial research on data quality is from the 1980s to the early 1990s. In this phase, data quality research is widespread but not yet systematic. Researchers begin to focus on exploring data quality dimensions, assessment methodologies and improvement strategies. Different sets of data quality dimensions were explored. For example, Brodie (1980) proposed that data quality contained three distinct components: data

reliability, logical integrity and physical integrity. Olson and Lucas (1982) used appearance and accuracy to measure data quality in office automation information systems. Morey (1982) considered information quality to be data accuracy and proposed three data accuracy measures in the context of information systems. O' Reilly (1982) investigated the effects of information quality on the use of information sources. In his study, information quality is measured by accessibility, accuracy, specificity, timeliness, relevance and amount of data. Ballou and Pazer (1985) considered accuracy, completeness, timeliness and consistency in the measurement of data quality in multi-input and multi-output information systems. Laudon (1986) identified completeness, accuracy and ambiguity as data quality dimensions for criminal-record systems. Observing the works above, it is found that different data quality dimensions can be derived from different contexts.

Around 1990s, researchers found that data quality is a key determinant for information system success (DeLone and McLean 1992). Since different sets of dimensions are developed according to different contexts, such as reporting system (Ahituv 1980) and office automation information system (Olson and Lucas 1982), there is no comprehensive set of information quality dimensions in this phase. Furthermore, there appears to be no single accepted definition of data quality. Some researchers (e.g. Morey 1982) consider data quality to be data accuracy, while other researchers (e.g. Keller and Staelin 1987) define data quality in terms of usefulness to consumers. Finally, most of the works in this phase are not validated by practical application.

The second development phase of data quality research can be identified from the 2000 to 2010. In this decade, data quality research becomes intensive, systematic and empirical. Therefore, the amount of information quality papers significantly increases, across a wide range of journals and conferences. From 2000 to 2010, more than 15 data quality books were published. These books have addressed different aspects of data quality research. Two information quality journals have been launched in this period: *International Journal of Information Quality* and the *ACM Journal of Data and Information Quality*. In addition, many leading database and information system conferences such as Special Interest Group on Management of Data (SIGMOD), Very Large Data Bases (VLDB) and Conference on Advanced Information Systems (CAiSE) have included data quality as one of their topics. Furthermore, since 1996, the International Conference on Information Quality is held annually to provide a forum for researchers and practitioners to present research findings and exchange knowledge in the field of data quality. Beyond research developments in academia, industry and government have also begun to pay attention to data quality issues. For example, in 2001, the US president signed data quality legislation into law (Batini and Scannapieco 2006). These newly founded companies and government operations clearly indicate the empirical application of data quality research.

With the advent of Big Data era, after around 2010, organizations are dealing with tremendous amount of data. These data are fast moving and can be originated from various sources such as social networks, unstructured data from different websites or raw feeds from sensors. Big Data practitioners are however experiencing a huge number of data quality problems, which can be time-consuming to solve or even lead to incorrect data analytics. As (Warden 2011) stated "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined". Therefore, Big Data Quality should be one of the critical issues related to Big Data research and its applications. Big Data creates not only value in financial terms but also in terms of operational and strategic advantages (Haug and Arlbjørn 2010). Thus exploring the value of Big Data and its quality management is crucial to the success of world-leading organizations.

Big Data is typically characterized by volume, velocity and variety (Laney 2001). As a consequence, Big Data Quality can possibly be affected by the three characteristics. Let us illustrate the challenge with an example within a Smart City context, in which many sensor data are used for decision making. Smart cities applications are excellent examples, as they are characterized by big data of high volume, velocity and variety. In this environment, for example, higher data velocity can result in frequent changes in data specification. In a traffic surveillance information system, the traffic camera is taking a photo every 5 minutes. The data specification of photo quality is set to be 300 dpi. The traffic photo whose resolution is lower than 300 dpi will be considered as low quality data. When time interval between taking two photos becomes 2 minutes, the data specification of photo quality may be lowered because of flow of the traffic photos turns to be fluent. In this case, data specification can be affected by the data velocity, in turn Big Data Quality problems can be caused by using the obsolete data specifications.

1.3 Goal and Outline of the Thesis

The goal of this habilitation thesis is to demonstrate my research findings in the recommender systems and data quality management. The main results of recommender systems are summarized in Chapter 2 and the results of data quality management are listed in Chapter 3. Each of two chapters is divided by research topic, which is then supported by my selected papers. A brief overview of the paper and my own contributions are described in each paper. Finally, Chapter 4 concludes the thesis and outlines the future research.

1.4 Paper Collections

I have selected 10 papers which are published after my Ph.D. as a collection for this thesis, in which there are 4 journal papers, 4 conference papers and 2 Springer book chapters. The list of the paper including ranking[1] and type is shown as follows:

1. (**CORE A | Journal**) Fatih Gedikli, Dietmar Jannach, Mouzhi Ge, How should I explain? A comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies, Volume 72, Issue 4, pp 367–382, 2014
2. **(CORE B | Conference)** Dietmar Jannach, Markus Zanker, Mouzhi Ge, Marian Gröning, Recommender Systems in Computer Science and Information Systems - a Landscape of Research, 13th International Conference on Electronic Commerce and Web Technologies, Vienna, Austria, 2012.
3. (**CORE B | Journal**) Mouzhi Ge, Fabio Persia, A Survey of Multimedia Recommender Systems: Challenges and Opportunities, International Journal of Semantic Computing, 11(3), 2017
4. (**Springer | Book Chapter**) Mouzhi Ge, Dietmar Jannach, Fatih Gedikli, Bringing Diversity to Recommendation Lists – An Analysis of the Placement of Diverse Items, Enterprise Information Systems, Springer LNBIP, Volume 141, pp 293-305.
5. (**Journal**) Fatih Gedikli, Mouzhi Ge, Dietmar Jannach, Explaining Online Recommendations Using Personalized Tag Clouds, Journal of Interactive Media, Vol. 10, No. 1, ISSN: 1618-162X, 2011.
6. (**ACM | Conference**) Mouzhi Ge, Mehdi Elahi, Ignacio Fernández-Tobías, Francesco Ricci, David Massimo, Using Tags and Latent Factors in a Food Recommender System, 5th ACM International Conference on Digital Health, Florence, Italy 2015
7. (**CORE A | Journal**) Mouzhi Ge, Markus Helfert, Impact of Information Quality on Supply Chain Decisions, Journal of Computer Information Systems, Vol. 53, No. 4, 2013.
8. (**CORE A | Conference**) Mouzhi Ge, Markus Helfert, Dietmar Jannach, Information Quality Assessment: Validating Measurement Dimensions and Process, 19th European Conference on Information Systems, Helsinki, Finland, 2011.
9. (**Springer | Book Chapter**) Qishan Yang, Mouzhi Ge, Markus Helfert, Data Quality Problems in TPC-DI based Data Integration Processes, Enterprise Information Systems, Lecture Notes in Business Information Processing, 2017
10. (**CORE A | Conference**) Markus Helfert, Owen Foley, Mouzhi Ge and Cinzia Cappiello, Analyzing the Effect of Security on Information Quality Dimensions, 17th European Conference on Information Systems, Italy, 2009.

Paper 1 to 6 are in the recommender systems domain, which are focused on the overview of the recommender system in computer science and information system (Paper 2), explanation feature of the recommender systems (Paper 1 and Paper 5), diversity in the recommendation list (Paper 4), and domain specific recommender system in Multimedia (Paper 3) and Food (Paper 6). In order to provide a high quality

[1] The conference and journal ranking are based on CORE http://portal.core.edu.au/conf-ranks/

data for the data usage, data quality management research includes 4 of my papers, which are from paper 7 to 10. Paper 8 and Paper 10 are focused on the data quality assessment, Paper 9 discusses the data quality and data integration. Paper 10 tests the effect of the data quality in decision making.

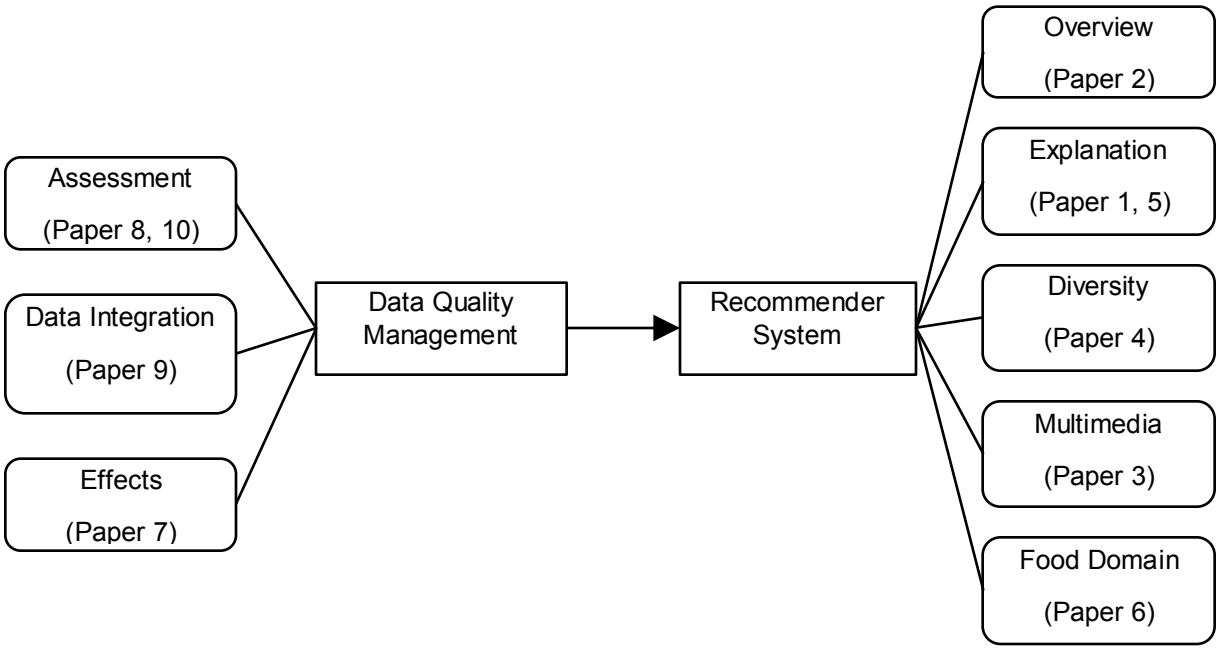

Figure 1: Overview of the selected papers in this thesis

Chapter 2: Recommender Systems

This chapter describes a total of 4 novel research topics in recommender systems, which are (1) overview of recommender system research in computer science and information system, (2) explanation feature in recommender system, (3) diversity in recommendation list, and recommender system in the context of (4) multimedia and (5) food domain. Each research topic will be arranged into one section. In each section, the research motivation, objective and contribution will be discussed.

2.1 Overview of Recommender Systems

Over the last decade, recommender systems have been widely applied in e-commerce, for example, book recommendation on Amazon or movie recommendation on Netflix. Recommender systems are developed to help users find relevant products that may interest them. The goal of recommender systems is to reduce the information overload and provide personalized recommendations for users. There is a growing popularity of using recommender systems in different domains. Given this diversity of research perspectives, the goal of this work is to review and classify recent research in recommender systems in order to quantify the research interests and identify opportunities for future research. We systematically evaluated all publications of a pre-defined set of high-impact journals and conferences in the fields of Computer Science and Information Systems during the period from January 2006 to July 2011. We included both journal articles as well as full papers appearing in conference proceedings. In particular, we considered those journals, where special issues on recommender systems have appeared. The analysis in this work should serve as a basis to understand limitations of current research practice in this field. As RS are IT applications we naturally limit our analysis to publications in the neighboring fields of Computer Science and Information Systems.

This literature review work has indicated the importance of recommender systems in the fields of Information Systems (IS) and Computer Science (CS). Given the different roots of the fields, CS researchers focus more on algorithms, whereas IS researchers are more interested in the systems-perspective and the effects of RS on the users. Correspondingly, different research designs and methods dominate in the two communities as documented by this work.

As an outlook, we see evidence that increased mutual exchange of results from the two communities can help further advance the research of recommender systems.

Paper: Dietmar Jannach, Markus Zanker, Mouzhi Ge, Marian Gröning, Recommender Systems in Computer Science and Information Systems - a Landscape of Research, 13th International Conference on Electronic Commerce and Web Technologies, Vienna, Austria, 2012.

Contribution (30%): This work is mainly done by my Master student Marian Gröning at the Technical University of Dortmund in Germany. I have initiated and refined the idea, guide the student to carry out all the analysis of the reviewed papers. I have mainly written the Section 3 in the paper.

2.2 Explanations in Recommender Systems

In recommender systems, one possible approach to support the end user in the decision making process and to increase the trust in the system is to provide an explanation for why a specific item has been recommended (Friedrich and Zanker, 2011). In general, there are many approaches of explaining recommendations, including non-personalized as well as personalized ones. An example of a non-personalized explanation would be Amazon.com's "Customers who bought this item also bought..." label for a recommendation list, which also carries explanatory information.

This work mainly investigates the effects different explanation types for recommendations on users. we aim at evaluating different explanation types in a comprehensive manner and consider the desired effects and quality dimensions such as efficiency, effectiveness, persuasiveness, perceived transparency, and satisfaction (Tintarev and Masthoff 2011) in parallel. To that purpose, we conducted a laboratory study involving 105 subjects in which we compare several existing explanation types from the literature (Herlocker et al., 2000) with a tag-based explanation approach.

In this work, we aim at detecting interdependencies between more than two quality dimensions. In particular, the goal is to analyze the influence of efficiency, effectiveness, and perceived transparency on user satisfaction. Based on the dependencies between the different effects of explanation types, we aim to derive a first set of possible guidelines for the design of effective and transparent explanations for recommender systems and sketch potential implications of choosing one over the other. These guidelines were validated through a qualitative interview-based study involving 20 participants.

We aim to obtain a deeper understanding of the value of the recently proposed tag and preference-based explanation types proposed in (Gedikli et al., 2011). We included two variants of this explanation method in our experimental study and compare their performance with the other explanation types in the different quality dimensions. Since acquiring explicit tag preferences is costly and can be cumbersome for the user, one of the two tag-based explanations incorporates a new method to automatically estimate the user's detailed preferences from the item's overall ratings.

> **Paper**: Fatih Gedikli, Dietmar Jannach, Mouzhi Ge, How should I explain? A comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies, Volume 72, Issue 4, pp 367–382, 2014
>
> **Contribution (30%)**: initiate the idea of the paper, conduct all the data analysis in the paper and write up the data analysis in the paper.

As mentioned in last work, one of the explanation interfaces we evaluated is the Personalized Tag Clouds, which is proposed as an extension to the basic tag cloud interface presented above. It provides more information by using additional "tag rating data" which was reported in Gedikli and Jannach (2010) as an additional knowledge source for recommender systems. In Gedikli and Jannach (2010) the authors present a recommendation approach, in which users rate items by rating their attached tags. While the general idea of "tag preferences" was also reported in Vig et al. (2009) the novel idea consists in allowing users to rate tags in the context of an item. The intuition behind this idea is that the same tag may have a positive connotation for the user in one context and a negative in another. For example, a user might like action movies featuring the actor Bruce Willis, but at the same time this user might dislike the performance of Bruce Willis in romantic movies. In (Gedikli and Jannach, 2010) the authors show that the predictive accuracy of recommender algorithms can be improved when incorporating such user- and item-specific tag rating data. In the Personalized Tag Clouds explanation interface, we pick up on this idea but aim to use the tag rating data to improve the quality of explanations for recommendations. An example of the Personalized Tag Clouds interface for a comedy movie is shown in Figure 2.

Figure 2: Personalized Tag Clouds

Paper: Fatih Gedikli, Mouzhi Ge, Dietmar Jannach, Explaining Online Recommendations Using Personalized Tag Clouds, Journal of Interactive Media, Vol. 10, No. 1, ISSN: 1618-162X, 2011.

Contribution (50%): initiate the idea of the paper, conduct the experiment and write up most of the paper.

2.3 Diversity in Recommendation List

As there is a growing popularity of using recommender systems in e-commerce, a variety of recommender algorithms have been proposed over the last fifteen years. Most of these algorithms focus on improving recommendation accuracy. Accordingly, the performance of recommender systems was evaluated by accuracy metrics such as Mean Absolute Error (MAE) or Precision and Recall. A recent literature survey shows that still today both the Information Systems and Computer Science community very strongly rely on these measures. However, some researchers have proposed that being accurate alone is not enough (McNee et al. 2006). Additional and complementary metrics, including diversity, novelty and serendipity as well as transparency could be used to evaluate the quality of recommender systems (Castells et al. 2011). Among the proposed metrics, diversity has been widely discussed and considered to be a factor that is equally important as accuracy (Fleder and Hosanagar 2007). The concept of diversity in recommender system research can be generally divided into inherent diversity and perceived diversity. Inherent diversity considers diversity from an objective view and is often measured by the dissimilarity among the recommended items. Perceived diversity, in contrast, defines diversity from a subjective perspective and can only be determined through a user evaluation.

how to place diverse items in a recommendation list has not been explored so far in recommender system research. Considering the possible effects of differently positioning the diverse items, we believe that the question of how to arrange the diverse items is an important research topic in recommender systems. Therefore, this work has investigated how to place the diverse items in a recommendation list and analyze

the effects of different item placements on the perceived diversity, on serendipity, and on user satisfaction. also, we have developed a set of guidelines of how to arrange diverse items so as to improve recommender's overall perceived quality.

Paper: Mouzhi Ge, Dietmar Jannach, Fatih Gedikli, Bringing Diversity to Recommendation Lists – An Analysis of the Placement of Diverse Items, Enterprise Information Systems, Springer LNBIP, Volume 141, pp 293-305.

Contribution (80%): refine the idea of the paper, conduct the whole experiment and write up most of the paper.

2.4 Multimedia Recommender Systems

The widespread availability of media technologies (e.g., digital and video cameras, MP3 players, and smartphones) dramatically increased the availability of multimedia data. Multimedia data allow fast and effective communication and sharing of information about people's lives, their behaviors, work, interests, but they are also the digital testimony of facts, objects, and locations. Usually images and videos are used by media companies as well as the public to record daily events, to report local, national, and international news, to enrich and emphasize web content. However, the extensive growth of multimedia information in our daily life has created information overload and increased complexity in decision making (Albanese 2011). People are facing the problem of dealing with the huge amount of multimedia data in a limited time, and they are also facing the challenge of quickly find the multimedia data that are interesting for them. It is thus usually time-consuming to manually select a preferred multimedia object. Therefore, more recently recommender systems are used to help users select the suitable multimedia objects.

Multimedia recommender system is emerging as a new research topic. With the advent of Big Data era, the multimedia data are fast moving and can be originated from various sources such as social networks, unstructured data from different websites or raw feeds from sensors. Big multimedia data have created a huge number of problems for users to choose the suitable multimedia objects. Therefore, multimedia recommender system is developed to compute customized recommendations for users accessing multimedia collections, using semantic contents and low-level features of multimedia objects, past behavior of individual users, and social behavior of the users' community as a whole.

In this work, we have conducted a survey on the research papers across multimedia information systems and recommender systems. We have then further focused on the papers that cross both research communities

and especially papers on multimedia recommender systems. The selected multimedia recommender system papers are reviewed and summarized by three features, which are recommender algorithm, multimedia object and application domain. We have discussed each feature and possible research opportunities. Based on the review, we have proposed a set of research challenges for multimedia recommender systems that can help both researchers and practitioners further explore this domain, and provide insights of how to perform the follow-up research in the field of multimedia recommender systems.

Paper: Mouzhi Ge, Fabio Persia, A Survey of Multimedia Recommender Systems: Challenges and Opportunities, International Journal of Semantic Computing, 11(3), 2017

Contribution (80%): Lead and write most section of the paper, conduct the paper analysis and derive the research agenda.

2.5 Health-Aware Food Recommender Systems

Among the recommender system application domains, food recommendation is emerging as a new research topic. Nowadays with the increasing changes in the food sector and lifestyles, many people are facing the problem of making better, i.e., healthier food choices (Freyne and Berkovsky 2010), especially in urban living–areas. Food, for many people, has become a black box that may lead to bad eating habits and poor healthy conditions. Therefore, the research has addressed the design of food recommender systems that suggest valuable food options to the user (high utility), by taking user preference, diet constrains, nutrition factors, food costs, etc. into account.

However, most of the existing applications provide just generic food advices that are not tailored to user's specific tastes or poorly match them. One of the recent food recommender systems is described in (Freyne and Berkovsky 2013). This is a meal planner system that provides personalized recommendations using a content-based recommendation technology. We have conjectured that the accuracy of the system predicted user's preferences (ratings) can be improved by better modelling and acquiring user preferences. for example, by letting the users, with tags, to signal what are in their opinion, the most important ingredients and features of the recipes. We also focused on the overall system usability, and tried to improve it by designing an effective human-computer interaction. This paper therefore is filling a research gap by proposing a mobile food recommender system that is easy to use high-quality personalized food recommendations.

In this work, we showed how to incorporate user selected tags, which describe important attributes of food, in a recommender algorithm based on matrix factorization, i.e., which uses latent factors for modelling both users and recipes. Also, we described how to exploit this algorithm, which is able to predict the rating that a user will give to a not yet rated item, by designing a complete human computer interaction in a tablet-based application. Our proposed solution is with high prediction accuracy, i.e., the recommended food recipes receive high evaluations from the users. In fact, our proposed algorithm significantly outperforms state-of-the-art algorithms, in terms of rating prediction error (MAE and RMSE).

Paper: Mouzhi Ge, Mehdi Elahi, Ignacio Fernández-Tobías, Francesco Ricci, David Massimo, Using Tags and Latent Factors in a Food Recommender System, 5th ACM International Conference on Digital Health, Florence, Italy 2015

Contribution (70%): proposed the key model in the paper and write up 90% of the paper.

Chapter 3: Data Quality Management

In order to produce high-quality prediction in recommender system, one of the important factors is to use high-quality data. Therefore, data quality management is critical for recommender systems and other predictive analytics systems. In this chapter, I will discuss three main aspects in data quality management, which are (1) data quality dimensions and assessment, (2) data quality in data integration and (3) effects of data quality in decision making

3.1 Data Quality Dimensions and Assessment

Most influential data quality research originated from information system research. Information system researchers initially identify and employ a set of dimensions to address the information quality problems within information systems. As information quality awareness and requirements have increased, researchers have begun to focus on data quality frameworks (Ballou and Pazer 1985), data quality dimensions (Wang and Strong 1996; Pipino 2002), data quality assessment (Wand and Wang 1996) and data quality management (Wang 1998).

The assessment of data quality is a key determinant of data quality management, as one cannot manage data quality without measuring it appropriately (Stvilia et al. 2007). By adapting a general definition of assessment (Gertz et al. 2004), IQ assessment can be defined as the process of assigning numerical or categorical values to IQ dimensions in a given setting. Over the last decade, a number of IQ assessment frameworks have been proposed (e.g. Pipino et al. 2002, Lee et al. 2002); however, in practice, organisations are facing still difficulties when implementing these assessment frameworks (Batini et al. 2009). One major difficulty is to understand and coordinate the quality assessment process for raw data and information products. Typical questions in that context are for example the following: which dimensions are suitable for measuring the quality of raw data in contrast to the quality of information products? How to coordinate the different assessment processes? Examining some of these issues, we conduct a literature review which reveals that most proposed frameworks are too generic to be used for assessment purposes or merely remain at a theoretical stage. Subsequently, in this work we aim to address the limitations of some data quality frameworks, and develop a practical data quality model on the basis of valid and reliable measurements.

Paper: Mouzhi Ge, Markus Helfert, Dietmar Jannach, Information Quality Assessment: Validating Measurement Dimensions and Process, 19th European Conference on Information Systems, Helsinki, Finland, 2011.

Contribution (80%): proposed the key model, conducted all the data analysis in the paper and write up 80% of the paper.

Among the data quality dimensions mentioned in the work above, security and accessibility have attracted research attentions in the data quality management. Therefore, we focus on these aspects and their implications on other data quality dimensions. Fehrenbacher and Helfert (2008) show that the importance of security and accessibility as IQ criteria has increased. This is accompanied with an increase in security requirements and complexity of information systems. Due to the increasing complexity and variety of access methods, question about its impact arises. What are implications of security measures on other data quality criteria? Does architecture have a significant (moderating) effect on the relationship between data quality criteria? What is the difference in the impact of accessibility from a workstation compared to a mobile device?

In order to address current limitations, this research focuses on the security and accessibility dimension of data quality. Review of related research shows that most data quality frameworks consider accessibility and security; however, researchers classify or consider these data quality dimensions diversely among various data quality frameworks. Furthermore, our research indicates an impact of security and accessibility on other data quality dimensions. An experiment is conducted to evaluate the effect on data quality dimensions of varying levels of security to an Information System. It allows for a thorough analysis of accessibility as a dimension of data quality. We propose a research model and illustrate results of an experiment, which support our research hypothesizes.

Paper: Markus Helfert, Owen Foley, Mouzhi Ge and Cinzia Cappiello, Analyzing the Effect of Security on Information Quality Dimensions, 17th European Conference on Information Systems, Italy, 2009.

Contribution (30%): conducted all the data analysis in the paper and wrote section 2 and part of section 3 of the paper.

3.2 Data Quality and Data Integration

Data quality problems appear frequently in the stage of data integration when extracting, migrating and populating data into data repositories. Data quality is considered an important aspect that influences the data integration process (Kimball and Caserta 2011). Previous research indicates that understanding the effects of data quality is critical to the success of organizations. Numerous business initiatives have been delayed or even cancelled, citing poor-quality data as the main reason. Most initial data quality frameworks consider that data quality dimensions are equally important (Knight and Burn 2005). More recently, as (Fehrenbacher and Helfert 2012) states, it is necessary to prioritize certain data quality dimensions for data management. However, as far as we know, there is limited research on prioritizing data quality dimensions and guiding the data quality management in the data integration process. As far as we know, there is still no study that focuses on the data quality problems aligning with the TPC- data integration benchmark.

Therefore, in this work, we intend to find out which data quality dimensions are crucial to DI and also attempt to derive the guidelines for proactive data quality management in data integration. The contributions of this paper are shown in three parts. The TPC-DI processes are investigated based on the data flow from different data sources to a data warehouse. Then we demonstrate some typical data quality problems which should be considered in the data integration process. We specify these data quality problems and classify them into different data quality dimensions. Finally, in order to proactively manage data quality in DI, we derive a set of data quality guidelines that can be used to avoid data quality pitfalls and problems when using the TPC-DI Benchmark.

> **Paper**: Qishan Yang, Mouzhi Ge, Markus Helfert, Data Quality Problems in TPC-DI based Data Integration Processes, Enterprise Information Systems, Lecture Notes in Business Information Processing, 2017
>
> **Contribution (50%)**: initiate the idea of the paper, wrote at least 60% of the paper.

3.3 Effect of Data Quality in Decision-Making

To further investigate data quality, some researchers studied the effect of data quality on decision-making; they demonstrate that increasing information quality level fosters decision effectiveness, decision performance, and decision quality. These findings suggest that decision-making depends on high-quality information but further extent of this relationship has not been investigated thoroughly (Fisher et al. 2003).

Although extant research confirms that increasing data quality increases decision-making quality, it does not detail the relationship regarding individual data quality dimensions. To further refine data quality effects on decision-making, it is crucial and valuable to investigate how different data quality dimensions affect decision-making.

Given this research gap, we studied the effects of data quality dimensions on decision-making, focusing on three frequently cited data quality dimensions: (1) accuracy, (2) completeness, and (3) consistency. Since the three dimensions are cited often and studied in numerous data quality literatures, measurement of the dimensions is mature (Elizabeth 2004). We find that different terms have been used in the decision-making literature as dependent variables including decision quality, decision effectiveness, and decision performance. Although these terms are different in a literal sense, their primary measurement is usually decision correctness. In this work, we use the term decision quality, found in a wide range of literature. Therefore, our research objective is detailed as studying the effects of data accuracy, completeness, and consistency on decision quality.

Paper: Mouzhi Ge, Markus Helfert, Impact of Information Quality on Supply Chain Decisions, Journal of Computer Information Systems, Vol. 53, No. 4, 2013.

Contribution (90%): initiate the idea of the paper, conducted all the experiments and data analysis, wrote at least 80% of the paper.

Chapter 4: Conclusion

This habilitation thesis is organized by a collection of my published papers after my Ph.D. A total of 10 papers have been selected and related commentary is also provided in this thesis. The thesis has cover two important and related research components that are recommender systems and data quality management. Data quality management can be considered as a pre-step for measuring, cleaning, and managing the data for data predictions in recommender systems. To assure a high-quality data, data quality management contributions are focused on how to measure and improve the data quality, how to avoid data quality pitfalls in data integration and how to reduce the data quality effects on business decision making. The contributions in recommender systems include the overview of recommender system research in different communities, explanation feature of the recommender, diverse elements in the recommendation list and how to apply the recommender system in the multimedia and food domain, e.g. to provide domain specific recommendations such as multimedia recommendations and healthy food recommendations.

As the future work, I will frame the data quality research and recommender system research in the lifecycle of Big Data Analytics, and I plan to conduct research along the lifecycle of Big Data Analytics such data pre-processing, data cleansing, data prediction, data interpretation and visualization. Also, both data quality management and recommender system can be centered by data warehouse or data lake, which can facilitate to promote the research question of how to generate high quality recommendations with Big Data quality management.

References

1. Adomavicius G, Tuzhilin A (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6): 734–749.
2. Ahituv, N. (1980), A systematic approach toward assessing the value of an information system, MIS Quarterly, 4(4), pp. 61-75.
3. Albanese M. , d'Acierno,A., Moscato, V., Persia F. and Picariello A. , A multimedia semantic recommender system for cultural heritage applications, in IEEE Fifth International Conference on Semantic Computing, 2011, pp. 403–410.
4. Baghaei, N.; Kimani S., Freyne J.; Smith G., Berkovsky S., Brindal E. (2011): Engaging Families in Lifestyle Changes through Social Networking. International Journal of Human-Computer Interaction 27(10):971-990.
5. Batini, C. and Scannapieco, M. (2006), Data Quality, Concepts, Methodologies and Techniques, Publisher: Springer, Berlin, Germany.
6. Ballou, D.P. and Pazer, H.L. (1985), Modeling data and process quality in multi-input, multi-output information systems, Management Science, 31(2), pp. 150-162.
7. Baghaei, N.; Kimani S., Freyne J.; Smith G., Berkovsky S., Brindal E. (2011): Engaging Families in Lifestyle Changes through Social Networking. International Journal of Human-Computer Interaction 27(10):971-990.
8. Brodie, M. L. (1980), Data quality in information systems. Information and Management, 3(6), pp. 245-258.
9. Castells, P., Vargas, S., Wang, J.: Novelty and diversity metrics for recommender systems: choice, discovery and relevance. In: Proceedings of International Workshop on Diversity in Document Retrieval, Dublin, Ireland, pp. 29–37 (2011)
10. DeLone, W.H. and McLean, E.R. (1992), Information system success: the quest for dependent variables, Information System Research, 3(1), pp. 60-96.
11. Davidson J., Liebald B., Liu J., Nandy P., Van Vleet T., Gargi U., Gupta S., He Y., Lambert M., Livingston B., Sampath D. (2010): The YouTube video recommendation system.

In Pro- ceedings of the fourth ACM conference on Recommender systems (RecSys '10). ACM, New York, NY, USA, 293-296.

12. Davidson J., Liebald B., Liu J., Nandy P., Van Vleet T., Gargi U., Gupta S., He Y., Lambert M., Livingston B., Sampath D. (2010): The YouTube video recommendation system. In Pro- ceedings of the fourth ACM conference on Recommender systems (RecSys '10). ACM, New York, NY, USA, 293-296.
13. Elizabeth, M.P. (2004) Assessing Data Quality With Control Matrices. Communications of the ACM. 47(2), 2004, 82-86
14. Fisher, C.W., Chengalur-Smith,I., Ballou, D.P., The Impact of Experience and Time on the Use of Data Quality Information in Decision Making, Information Systems Research. 14(2), 2003, 170 – 188
15. Friedrich, G., Zanker, M., 2011. A taxonomy for generating explanations in recommender systems. AI Magazine 32 (3), 90–98.
16. Fehrenbacher, D. and Helfert, M. (2008). An empirical research on the evaluation of data quality dimensions. In Proceedings of the 13th International Conference on Information Quality (Neely, P., Pipino, L. and Slone, S. Eds), pp. 230-245, MIT, USA, Cambridge.
17. Freyne, J. and Berkovsky, S. Evaluating recommender systems for supportive technologies. In User Modeling and Adaptation for Daily Routines, pages 195-217. Springer, 2013.
18. Freyne, J. and Berkovsky, S. Intelligent food planning: personalized recipe recommendation. In Proceedings of the 15th International Conference on Intelligent User Interfaces, pages 321-324. ACM, 2010.
19. Fleder, D., Hosanagar, K.: Recommender systems and their impact on sales diversity. In Proceedings of the 8th ACM Conference on Electronic Commerce, San Diego, CA, USA, pp. 192–199 (2007)
20. Gedikli, F., Bagdat, F., Ge M. and Jannach D. (2011), RF-Rec: Fast and accurate computation of recommendations based on rating frequencies, in 13th IEEE Conference on Commerce and Enterprise Computing, 2011.
21. Gedikli, F.; Jannach, D.: Rating Items by Rating Tags, Proceedings of 2nd Workshop on Recommender Systems and the Social Web at ACM RecSys‘2010, Barcelona, Spain (2010) 25–32.

22. Gertz M., Ozsu T., Saake G., & Sattler K. (2004). Report on Dagstuhl Seminar Data Quality on the Web, SIGMOD Report.
23. Haug, A., Arlbjørn J.S. (2010) Barriers to master data quality, Journal of Enterprise Information Management. Vol. 24 No. 3, pp. 288-303
24. Herlocker, J. L., Konstan, J. A., Riedl, J. T., 2000. Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00). Philadelphia, Pennsylvania, USA, pp. 241–250.
25. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G. (2011): Recommender Systems - An Introduction. Cambridge University Press
26. Keller, K.L. and Staelin, R. (1987), Effects of quality and quantity of information on decision effectiveness, Journal of Consumer Research, 14(2), pp. 200-213.
27. Knight, S.A., Burn, J.M.: Developing a framework for assessing information quality on the World Wide Web. Informing Science: International Journal of an Emerging Transdiscipline, 8(5), pp.159-172 (2005).
28. Kimball, R., Caserta, J.: The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data. Publisher: Wiley (2011).
29. Konstan JA, Riedl J (2012). Recommender systems: from algorithms to user experience. User Model User-Adapt Interact 22(1-2): 101–123.
30. Konstan JA, Riedl J (2012). Recommender systems: from algorithms to user experience. User Model User-Adapt Interact 22(1-2): 101–123.
31. Lee Y., Strong D., Kahn B., & Wang R. Y. (2002). AIMQ: A Methodology for Information Quality Assessment. Information & Management. 40(2) 133-146.
32. Linden, G., Smith, B. & York, J.(2003): Amazon.com recommendations: Item-to-item collaborative filtering. In IEEE Internet Computing, 7, 76-80.
33. Laudon, K.C. (1986), Data quality and due process in large inter-organizational record systems. Communications of the ACM, 29(1), pp. 4-11.
34. Laney, D. (2001), '3D Data Management: Controlling Data Volume, Velocity, and Variety', Technical report, META Group
35. Morey, R.C. (1982), Estimating and improving the quality of information in a MIS, Communications of the ACM, 25(5), pp. 337-342.

36. McNee, S., Riedl, J., Konstan, J.: Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. Montréal, Canada, pp. 1097–1101 (2006)
37. Olson, M.H. and Lucas, H.C. (1982), The impact of office automation on the organization: some implications for research and practice, Communications of The ACM, 25(11), pp. 838-847.
38. O'Reilly III, C.A (1982), Variations in decision Makers' use of information source: the impact of quality and accessibility of information, Academy of Management Journal, 25(4), pp. 756-771.
39. Pipino L., Lee Y.W., & Wang R.Y. (2002). Data Quality Assessment. Communications of the ACM. 45(4) 211-218.
40. Ricci F, Rokach L, Shapira B, Kantor PB (2011). Recommender Systems Handbook. Springer.
41. Redman, T. (1996), Data quality for the information age, Publisher: Artech House, Boston, Massachusetts, USA.
42. Stvilia B., Gasser L., Twidale M. B. & Smith L. C. (2007), A Framework for Information Quality Assessment, Journal of the American Society for Information Science and Technology. 58(12) 1720-1733.
43. Tintarev, N., Masthoff, J., 2011. Designing and evaluating explanations for recommender systems. In: Recommender Systems Handbook. pp. 479–510.
44. Vig, J.; Sen, S.; Riedl, J.: Tagsplanations: Explaining Recommendations Using Tags. Proceedings of the 13th Onternational Conference on Intelligent User Interfaces. Sanibel Island, USA (2009) 47–56.
45. Wand, Y. and Wang, R.Y. (1996), Anchoring data quality dimensions in ontological foundations, Communications of the ACM, 39(11), pp. 86-95.
46. Wang R. Y. (1998). A Product Perspective on Total Data Quality Management, Communications of the ACM. 41(2) 58-65.
47. Wang, R. Y. & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information System 12(4) 5-34.
48. Warden P. (2011) Big Data Glossary, O'Reilly publishing.

A collection of selected publications

1. Fatih Gedikli, Dietmar Jannach, Mouzhi Ge, How should I explain? A comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies, Volume 72, Issue 4, pp 367–382, 2014
2. Dietmar Jannach, Markus Zanker, Mouzhi Ge, Marian Gröning, Recommender Systems in Computer Science and Information Systems - a Landscape of Research, 13th International Conference on Electronic Commerce and Web Technologies, Vienna, Austria, 2012.
3. Mouzhi Ge, Fabio Persia, A Survey of Multimedia Recommender Systems: Challenges and Opportunities, International Journal of Semantic Computing, 11(3), 2017
4. Mouzhi Ge, Dietmar Jannach, Fatih Gedikli, Bringing Diversity to Recommendation Lists – An Analysis of the Placement of Diverse Items, Enterprise Information Systems, Springer LNBIP, Volume 141, pp 293-305.
5. Fatih Gedikli, Mouzhi Ge, Dietmar Jannach, Explaining Online Recommendations Using Personalized Tag Clouds, Journal of Interactive Media, Vol. 10, No. 1, ISSN: 1618-162X, 2011.
6. Mouzhi Ge, Mehdi Elahi, Ignacio Fernández-Tobías, Francesco Ricci, David Massimo, Using Tags and Latent Factors in a Food Recommender System, 5th ACM International Conference on Digital Health, Florence, Italy 2015
7. Mouzhi Ge, Markus Helfert, Impact of Information Quality on Supply Chain Decisions, Journal of Computer Information Systems, Vol. 53, No. 4, 2013.
8. Mouzhi Ge, Markus Helfert, Dietmar Jannach, Information Quality Assessment: Validating Measurement Dimensions and Process, 19th European Conference on Information Systems, Helsinki, Finland, 2011.
9. Qishan Yang, Mouzhi Ge, Markus Helfert, Data Quality Problems in TPC-DI based Data Integration Processes, Enterprise Information Systems, Lecture Notes in Business Information Processing, 2017
10. Markus Helfert, Owen Foley, Mouzhi Ge and Cinzia Cappiello, Analyzing the Effect of Security on Information Quality Dimensions, 17th European Conference on Information Systems, Italy, 2009.