Habilitation Thesis Reviewer's Report

Masaryk University	
Faculty	Faculty of Informatics
Procedure field	Informatics
Applicant	Mouzhi Ge, Ph.D.
Applicant's home unit, institution	Faculty of Informatics, Masaryk University
Habilitation thesis	Data Quality Management for Recommender Systems
Reviewer	Prof. Alan Said
Reviewer's home unit, institution	School of Informatics, University of Skövde

The habilitation thesis focuses on two separate, but still related, research topics, namely recommender systems, and data quality. The research presented is supported by a number of research publications ranging from conference and journals as well as a few book chapters.

Given the topics covered in the thesis, it would appear as if the order of the topics could very well have been reversed in order to increase readability. Intuitively, issues relating to data quality provide a basis on which various information systems can be built on top. In this case, the basis would have been provided for the work concerning recommender systems. It appears as if this intuitive notion has been considered given that it is how Figure 1 (p 6) presents the topics and connections in the thesis. However, the thesis itself (and the order of papers) starts by presenting aspects of recommendation and flows into data quality aspects, i.e. in reverse order to Figure 1. If this is not an oversight, it would have been interesting to understand the motivation behind this order or presentation. A short discussion and motivation on this would have sufficed to make the thesis stand stronger and more on point.

As for the two specific topics, recommender systems and information quality, they are complementary and have a clear sense of belonging together. However, looking at the works which have been included in each part, the feeling of belonging is not as clear. In the recommender systems part, it appears as if the work, although more than enough for a habilitation thesis, does not carry a well-defined common thread throughout. All works do cover some aspect of recommender systems, there is however a large variety in specific topics, i.e. diversity, explanations, multimedia, and food/health. Generally, the papers do not connect to each other in an intuitive manner. It would have been reasonable to tie the works together under a common theme, even if under a heterogenous moniker such as "complex recommendations in rich data" or similar through in an introduction or discussion section. The part of the thesis which covers information quality contains works which are similarly disconnected, i.e. covering topics from data quality dimensions, through data integration, and aspects of data quality in decision making. Again, what is missing is a well-defined common thread, or an

answer to why these papers were put together into a thesis. A short text which would tie everything together under a common umbrella would have been great.

Both the recommender systems section and information quality section would be more understandable if the bridges (Chapter 2 and Chapter 3) respectively would have provided a common theme for the individual sections as well as one for habilitation in general. Chapter could have served as a better bridge between the topics, in its current form it sets the stage for recommender systems and data quality management in a historical context but does not provide a common theme for the rest of the thesis.

The papers included in the thesis are all of good quality, a fair portion of them have had a good impact in the community given the number of citations they have received. However, given the focus of this thesis, it is not entirely clear how the selection of publication outlets has taken place. Granted, there are many outlets for research papers, however, given the specific topic of this thesis it would have been expected that some of the papers had been published in the premier venues in the field.

Reviewer's questions for the habilitation thesis defence (number of questions up to the reviewer)

1. From your perspective, which community has most use from papers in your thesis?
2. Your thesis focuses very much on recommender systems, your papers on recommender systems cover a broad range of topics. Given this, it is interesting to see that none of the paper have been published in the traditional venues for recommender systems, e.g. ACM RecSys, UMAP, SIGIR. Why is this so?
3. Similarly, you mentioned ACM JDIQ and IJIQ, together with SIGMOD, VLDB, and CAiSE as prominent outlets for research on data quality. Why is it that the works included in the thesis are published at other venues?
4. On page 1, in the second paragraph of the introduction you say that recommender systems have had an important role on the Internet with many applications. What about the offline, real world, have these systems played an important role outside of the Internet? How?
5. In the third paragraph of the introduction you mention the importance of high-quality data and speak about many businesses going out of business due to lacking data quality. Given that user-generated information (such as explicit ratings, and implicit interaction patterns) often are very noisy due to poor data collection processes, incomplete data recordings, and/or false statements by the end users. How can and should these systems be applied on this data in the first place?
6. There have been a number of works stating that the underlying data of recommender and personalization systems is flawed, yet large multinationals still make very large profits using this data. How does the poor data quality problem mentioned in the introduction reflect on that?
7. In section 1.1 you mention that the difference between recommendation and information filtering is the fact that recommender systems are not query-driven. However, in information retrieval, there is a research topic focusing on "query-less search". These query-less search systems use information retrieval technologies for a

use case that resembles recommender systems very much. Can you elaborate a little on the differences (if any) between query-les search and recommendation?

8. In Section 1.2, on page 3, you state that "there appears to be no single accepted definition of data quality". Can you reason on whether there should one? Given the various application scenarios and use cases, is it even feasible to have a unified expression for constitutes good or bad data?

9. In paper 4 you develop a set of guidelines for arranging of diverse items. These guidelines, although reasonable, are based on users interacting with static list of items which are referred to as recommendations in the paper when they are in fact not. Can you reason about how the guidelines can have been affected by the static lists of items (instead of proper recommendations)?

10. Some of the papers that serve as the basis for the thesis develop guidelines (paper 4) and challenges (paper 3), can you reason about how these guidelines and challenges are reflected in the other works which are included in the thesis?

11. Paper 6 focuses on using tags for recommendation and does so with good results. Tag-based and Tag-aware recommender systems have been researched heavily in the past. Apart from the application on foods, in what way does this work expand on earlier work – if we consider the system agnostic to what type of items are recommended?

12. The included papers span 8 years. How well would you say that e.g. paper 10 (2009) aligns with the findings and recommendations in papers 3 (2017) and paper 9 (2017)?

13. Looking back at the work conducted in the thesis, what would you have changed?

Conclusion

The habilitation thesis entitled *"Data Quality Management for Recommender Systems"*
by Mouzhi Ge *fulfils* requirements expected of a habilitation thesis in the field of Informatics.

In Gothenburg on June 3, 2018

....
signature

3