Review report on the doctoral dissertation

Ordered by: Ghent University,
Council of the Faculty of Engineering and Architecture.

Cooperation: Polish Academy of Sciences, Systems Research Institute, Warsaw.

Title of the PhD-thesis: Handling metadata in the scope of coreference detection in data collections.
(in Polish: Zastosowanie metadanych przy wykrywaniu podobieństwa w kolekcjach danych)

Author: Marcin Szymczak, MSc

Supervisors: Guy de Tre, Prof. Dr. and Sławomir Zadrożyń, MSc, PhD, DSc

1. Introductory remarks and scope of the research

The dissertation falls within the field of technical sciences, the area of competence of Ghent University and Systems Research Institute of the Polish Academy of Sciences, where the dissertation was submitted. The discipline of research is information technology.

The compact view into the goals and content of the dissertation is given in the summaries written in three languages: Dutch, Polish and English. However, the Author has forgotten to write the title of the work in Dutch – it is given on the cover only in two languages, namely English and Polish.

The main goal of the dissertation was formulated in Section 1.3 by posing four main research questions, namely:

1) How does one establish detection of coreferent elements in XML schemas based on a comparison of the tags and their sequences – paths?
2) Is it possible to establish schema matching based on content data?
3) What kind of information can be used to improve the detection of coreferent tuple and data fusion in homogeneous or heterogeneous collections of data and how can it be used?

4) Is it possible to establish practically useful semantical mappings between the values of attributes and to what extent a partial order relation can help in the data fusion of heterogeneous data collections?

Thorough research into those four areas was reported exhaustively in chapters 2, 3, 4, 5 and 6.

The answers to those four main research questions were presented in a concise and readable way in chapter 7, titled “Overall conclusions”. Thus, chapters 1 and 7 bind the work into a cohesive whole.

2. Overall impression

The Author skillfully guides the reader through his work (which comprises seven chapters and 192 pages of text). The sequence of chapters is correct.

In my opinion, the scientific content is of the level expected from a modern dissertation in the field of IT. The need for research effort is well motivated by literature resources and the Author’s own ideas. The presented material is original and innovative. To some extent, the work has its source in the previous original achievements already published. Those publications are listed before the start of sections dealing with the relevant problems and their solutions. The originality of the research is well documented and related to the state of the art in the explored field and to the specific tasks set by the author.

Moreover, the author verified sufficiently the proposed algorithms by performing computations with the use of nontrivial examples, which lends credibility to the practical efficiency of his approach. Thus, the practical aspect of the research results is clearly presented.

3. Substantive aspects

As regards the contents, the chapters can be grouped as follows:

- Chapter titled “Introduction”, which contains the fundamental definitions (coreferent data, metadata, data fusion, schema matching, object mapping, and others), as well as problem statement and outline of the thesis.

Creative and original chapters, aimed at solving the problems posed in Section 1.3. Those include the following chapters (the titles are given below according to the “Table of contents” of the dissertation – see the remark in the Section “Minor remarks” of this review report):
- Chapter 2 “Coreference detection in schema based on schema alone”. Coreference is considered in a hierarchical way, and the method is based on paths in XML schemas. A binary notion is applied (the existence or non-existence of coreference); however, the uncertain results are allowed. The accuracy of the method is evaluated by the comparison of its results with alternative approaches. The material has been presented in a clear and precise manner.

- Chapter 3 “Coreference detection in database schemas based on coreferent content data”. Here, a novel content data based approach has been presented in which a proper matching between corresponding schema elements of heterogeneous datasets is constructed. The method uses: possibilistic truth values to describe certainty of matchings and similarity, as well as fuzzy integers to express cardinalities of sets. Coreference in the content of the data sources is used to obtain the meaning of the considered schema elements. The necessity of existence or construction of a knowledge base is discussed. However, the construction of such a database is the subject of the next chapter.

- Chapter 4 “Dynamically constructed order relation as metadata”. Here, the original algorithm, called dynamical order construction (DOC), is specified. The results presented are based on two publications where Mr. Szymczak is a co-author. Necessary definitions (A-partition, generative multirelation and multiproximation) have been formulated and the general strategy for the inference of order relation from the generative multirelation has been described. Evaluation and discussion of the results is satisfactory.

- Chapter 5 “Dynamically constructed order relation in data fusion”. The theoretical formal base for considerations is the given definition of fusion function, which – roughly speaking – takes a multiset of tuples and maps it onto one tuple. Next, the DOC-driven fusion is considered as a proposition of a novel method. The approach is described step by step, and evaluated exhaustively in practice.

- Chapter 6 “Semantical mapping of attribute values in data fusion”.

Chapters 2-6 form a logical sequence and include procedures that enable handling metadata in the scope of coreference detection in data collections, which is the general purpose of the work. As shown in those chapters, the Author has successfully completed the research tasks within the scope initially set. This statement is confirmed in Chapter 7 “Overall conclusions” where the compact answers to the research challenges set in Chapter 1 are given.

4. Minor remarks

“Table of contents” is not complete. Subsections are not listed, e.g. 2.1.1, 2.1.3, etc. Moreover, the number ascribed to the “Acknowledgement” is the same as that assigned to “References”, namely “I”.

3
Chapter titles listed in the “Table of contents” of the dissertation differ from those given in the main text.

The titles in the table of contents are as follows:
Chapter 3 “Coreference detection in database schemas based on coreferent content data”.
Chapter 4 “Dynamically constructed order relation as metadata”.

Titles used in the text:
Chapter 3 “Coreference detection in schema based on coreferent content data”.
Chapter 4 “Dynamical construction of a knowledge base: a partial order relation on the domains of attributes”.

Concepts are defined in proper places; the key definitions are also complete.

The Author used a large number of appropriately selected sources. The referencing is very accurate. The References at the end are given in the order of their occurrence in the text. However, there is some inconsistency in the case of source [1] which is placed as the 14th position on pages 1-4, together with positions [15-21]. Thus, perhaps it would be more convenient to list the references in an alphabetical order.

The works cited in chapters 2, 4, 5 and 6 are co-authored by Mr. Szymczak. However, it is not clear in what part and to what extent those chapters present the Author’s original ideas as compared to the co-authored works cited.

To sum up this part of the review, in which I focused on the formal details, I would like to emphasise that the above remarks do not affect or belittle in any respect the scientific results obtained by the Author.

5. Some questions

a) The term “data” is the most basic concept within this work. The term “information” plays also an important role in the context of the problems examined and is used for example on pages 1-3 where the process of data fusion should assure the minimal loss of information; see also the last sentence on pages 1-9 or 6-43 – “onomastic information”. Question: How is “information” defined for the purpose of this dissertation and what is the difference between the term “data” and the term ”information”? Formal definitions and illustrative example are requested.

b) How to define, if the Author recognises such a need, the term “coreferent information” in the context of the definition which emerges as an answer to question a) ?

c) According to Definition 12, page 5-5, “fusion function takes a multiset of tuples and maps that multiset onto one tuple”. How can it affect the preservation of information in the sense of its definition ?
d) A relational database is considered (cf. chapter 5). Is it possible to adapt the presented approaches to coreference detection in other types of databases?

6. Summary and conclusions

To sum up briefly the Author’s findings from the point of view of their novelty, it may be concluded that three methods were proposed and analysed:
- two novel schemas solving the problem of metadata matching and, consequently, the detection of coreference in those metadata; the first technique operates within XML schema and is characterised by intuitiveness and relative simplicity; the second one is based on content data and it allows one to address the attribute granularity, as well as the coverage problems.
- one novel method (called DOC) aimed at the automatic construction of a knowledge base related to the domain of a given attribute.

It has been shown that the methods give better or similar results to the approaches known from the literature.

Additionally, it was necessary to analyse a number of specific problems, each of which was an interesting challenge. Both the central problem and the specific subtasks were formulated, solved and illustrated correctly.

The creative aspect is easily seen. The research process resulted in new findings in the explored field. Successful completion of the dissertation required advanced knowledge and high-level IT skills.

The research activities were conducted and documented property. The scope of research was sufficient and the results were satisfying.

Thus, despite some minor remarks, my general evaluation of this dissertation is very positive. The work presents an independent solution to a scientific problem of practical importance, which fulfills the requirements of a doctoral dissertation. In conclusion, my overall recommendation is to admit the candidate unconditionally to the public defence of his thesis.

[Signature]

5