Knowledge-Based Systems 71 (2014) 211–226

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Effect of thesaurus size on schema matching quality ☆

Thabit Sabbah [a], Ali Selamat [b,*], Mahmood Ashraf [c], Tutut Herawan [d]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b] UTM-IRDA-COE & Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[c] Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
[d] Department of Information System, University of Malaya, 50603 Pantai Valley, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 26 June 2014
Accepted 4 August 2014
Available online 17 August 2014

Keywords:
Schema matching
Thesaurus
Information Retrieval
Searching
Performance
Text similarity
Structured vocabulary

ABSTRACT

Thesaurus is used in many Information Retrieval (IR) applications such as data integration, data warehousing, semantic query processing and schema matching. Schema matching or mapping is one of the most important basic steps in data integration. It is the process of identifying the semantic correspondence or equivalent between two or more schemas. Considering the fact of the existence of many thesauri for identical knowledge domain, the quality and the change in the results of schema matching when using different thesauri in specific knowledge field are not predictable. In this research, we studied the effect of thesaurus size on schema matching quality by conducting many experiments using different thesauri. In addition, a new method in calculating the similarity between vectors extracted from thesaurus database is proposed. The method is based on the ratio of individual shared elements to the elements in the compound set of the vectors. Moreover, we explained in details the efficient algorithm used in searching thesaurus database. After describing the experiments, results that show enhancement in the average of the similarity is presented. The completeness, effectiveness, and their harmonic mean measures were calculated to quantify the quality of matching. Experiments on two different thesauri show positive results with average Precision of 35% and a less value in the average of Recall. The effect of thesaurus size on the quality of matching was statically insignificant; however, other factors affecting the output and the exact value of change are still in the focus of our future study.

1. Introduction

For more than two decades, thesauri were exploited in many IR applications. For example, it were used in web document classification [1], summarization [2], indexing [3], and in calculating the semantic similarity of documents written in the same or in different languages [4]. Thesaurus was also utilized to solve the problem of schema matching [5–7]. Recently, thesaurus is used to predict query difficulty in medical domain. It was concluded that the performance of the predictor is influencing with many factors such as the coverage of thesaurus or query mapping quality [8]. Earlier studies assumed that there are no general thesauri such that sufficient coverage are available, so that the use and impact of thesaurus was not studied widely [8]. However, a high quality thesaurus is available for some specific domains, also many thesauri with different coverage abilities and sizes are found in the same domain.

Such as any other controlled vocabularies, thesaurus is reusable and replaceable (i.e. can be reused in many different applications and can be replaced by another compatible thesaurus). However, the quality of the thesaurus is crucially to be assessed before reuse or replacement. According to [9] the size of the vocabulary is one of the main quality issues considered in measuring the quality of the controlled vocabulary. This research is discuss the effect of the thesaurus size on the quality of schema matching, thus, measuring and assessing of the thesaurus quality is out of this research's scope, details on thesaurus quality assessment can be found in [9,10].

Domain specific thesaurus are preferred to the common thesaurus such as WordNet in this research because of the common thesaurus are already used in this field as shown in the next

☆ This is an extended paper that has been presented in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference by Sabbah, Thabit; Selamat, Ali, "Thesaurus Performance with Information Retrieval: Schema Matching as a Case Study," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference, pp. 4494,4498, 13–16 October 2013.

* Corresponding author.
E-mail address: aselamat@utm.my (A. Selamat).

http://dx.doi.org/10.1016/j.knosys.2014.08.002
0950-7051/

paragraphs, moreover this research is studying the effect of the size of domain specific thesauri for single domain.

In information and database systems, schema is stands as the set of formulas (collection of meta-data) imposed on the data in the database. These formulas (also called integrity constraints) are applied to ensure the compatibility and describe the organization and the relations between database's parts and entities [11].

The importance of studying the effect of thesaurus size is coming from the vital need of effective and complete automatic solutions, because of the rapid expansion of application areas in which thesaurus and other vocabulary tools can be utilized such as natural language processing and Information Retrieval. For instance, schema matching forms the first and the crucial step toward data integration, however, the multiplicity of the obtainable common and domain specific vocabulary and linguistics tools that can be used, makes it hard to prefer one tool over others since the influences of tool's features such as size and coverage are not predetermined.

1.1. Schema matching related works

Schema matching, which is the process of identifying the semantic correspondence, or finding the equivalent elements between two or more schemas, is still an open research area since more than two decades. This is not only because schema matching is one of the basic operations [12] in many applications such as data integration, data warehousing, and semantic query processing, but also because it is an increasingly important problem itself [13], and as well as the uncertainty in the results of schema matching techniques [14,15]. Many approaches and tools were used to solve the problem of schema matching such as Cupid [16], LSD [17], and Corpus [18]. In addition, many surveys and classifications were published [19,20]. Few features of matching process were not in the focus of proposed approaches, and aspects such as structural, element, linguistics, and data model were discussed widely. Following is a summarization of the techniques used in schema matching approaches.

Many techniques were employed to carry out matching process; Machine-learning techniques were used in [17], learner-based approaches contains learner modules and specific module to direct learners. These approaches use neural networks advantages to find out the similarity between data sources. In [21] the object-oriented characteristics were exploited to determine the mapping between data sources' attributes. The problem of matching is not solved using this approach as well many proposed works using metadata; however, it is shifted into another problem, which is the problem of ontology mapping. Most of current schema matching tools use rules to carry out the matching, by using information such as elements names and descriptions, data types, hierarchy structure, and constraints. They are employed in determining the similarity at either element level or schema level [16,21,22].

Most effective rule-based schema matching methods usually consist of three phases; linguistic, constraint-based, and structural matching [23]. In linguistic phase, methods depend on string matching in general to find out the similarity between elements names. Current schema matchers usually use WordNet, a large lexical database of English [24] to consider the semantic relationships between elements labels [6]. However, it is common that algorithms in this category use combined methods to get high computed similarity, methods of label normalization to improve schema matching was also by [6,7]. Cupid matcher exploits linguistic matching in a comprehensively and efficiently manner to produce high similarity [16]. Incorrect results that are obtained from linguistic matching phase are usually adjusted in constraint-based matching phase. Data type constraint, data types' compatibility measurement method are usually used as the initial solution of incorrect or ambiguous results of linguistic matching phase [25,16]. Structural matching phase is used to solve the problems of context similarity, these problems are generally appear in XML schema matching where the structure document and the constraints on nodes and edges differs from rational schemas [23] describes such problems in details.

Based on the conclusion of [8], this paper studies the effect of thesaurus size (in aspects of number of terms, number of lead-in terms, and number of cross relations) on the results of schema matching using thesaurus.

1.2. Research contributions

Although there are few exiting works in the thesaurus based schema matching field, the main contributions of this research encompass:

- Presenting an experimental study of the effect of thesaurus size on schema matching quality. Three agricultural thesaurus of different size are utilized and compared, and the results are evaluated through several objective functions.
- A new method to compute the similarity between vectors extracted from the thesaurus is proposed.
- Moreover, this paper explains in detail many of the technical aspects to be considered when using thesaurus.
- The experimental results shows that the effect of thesaurus size in the quality of matching is statistically insignificant. However, an increment in the average of similarity with distinctive values are recorded.

1.3. Research limitations

This research is studying the effect of thesaurus size on the quality of schema matching, by utilizing three thesauri from the Agriculture domain to carry out the matching process on the element level, and the results are analyzed in many different perspectives. Therefore, some other perceptions such as thesaurus construction and evaluation, results (Precision, Recall, and F-measure) optimization, and the method complexity are not in the scope of this research.

In the rest of this paper, Section 2 explains the methodology. Section 3 presents the study setup. Section 4 shows the results as well as a discussion of these results. Finally, this work is concluded in Section 5.

2. Schema matching based on linguistic analysis with thesaurus

This paper studied the impact of thesaurus size on the quality of schema matching. The applied methodology is based on exploiting thesaurus to carry out the matching process. Fig. 1 shows the methodology framework, and the next subsection explains it in details.

The method consists of three main phases as shown in Fig. 1. Numbers is circles 1, 2 and 3 represent these phases. In phase one, two schemas (S_x and S_y) are part of the input of the (Apply Thesaurus) process, thesaurus is the other part of input for this process, and the output of (Apply Thesaurus) process are two sets of vectors of terms (S_x mass and S_y mass). These two sets of vectors will form the input of phase two, which is (Calculating Similarity Matrix) to produce the Similarity Matrix (SM) between the schemas' elements; The third phase is (Extracting the Final Mapping) that uses SM as an input to generate the final mapping list.