2C-02

データの構文的かつ意味的解釈によるスキーママッチング手法の提案
(A Syntax-Semantics Approach for Schema Attribute Identification)

姜　逸越　　Jiang Yiyue

株式会社富士通研究所　Fujitsu Laboratories Ltd.

Abstract

Before data analysis, data preparation takes around 80 percent of the whole time, which is expected to be cut down its related workloads. In this study, a supportive way is introduced to improve integrating schemata, which will further help accelerating the data preparation, by employing a set of proposed syntax-semantics-combined features. A comparative evaluation with representative feature sets representing those used by existing methodology is also provided to help validating its feasibility.

1. Introduction

Nowadays, the handling and analysis of heterogeneous information within data lakes and dilated databases is becoming more and more important for various industries to extract their insights and value. Data preparation or data pre-processing takes around 80 percent of the whole time, prior to further data analysis. Hence, it becomes indispensable and has been expected to be hugely cut down the related workloads. As one of the essential methods for data preparation, data integration has been studied [1]. More specifically, integration of the schemata of processed data which are defined with different names, called *schema matching*, has been employed as a necessary supportive way to systematically get the data well organized by machine and help achieving the above goals. Under the situations when processing structured datasets like table data, schemata refer to the sets of the attribute names (column names), and thus schema matching is done by properly identifying and matching the attributes from one set to another.

Many current existing tools by machine investigate the actual values under each attribute, by extracting the syntactic significance for almost all words or phrases as the counted units [1][2]. But such measurements can only aggregate completely identical sub-sequences as same units, and hence have no identification metrics for term definitions or semantic relationships. So they will probably result in low recall when many different terms or expressions belong to a same meaning or entity. Some tools consider the attribute names themselves instead to extract the inner relationships by referring to some pre-determined definitive or semantic dictionaries or ontology databases [3][4]. Those methods are able to indirectly get semantically related information by referring to metadata for matching, but the referral and comparative processing is heavily restricted by computation complexity to handle the attribute names only [5], which makes matching results poor without considering actual values.

Here in this study, we aim to achieve another value-based method for schema matching which is expected to improve the matching of other value-based methods that only focus on syntactic significance. The improvement should be obvious even when different terms or expressions with the same meaning appear, and we achieved it by also taking the potential semantic relationships into consideration. Also, the syntactic features are not lost but fully made use of, to get rid of complexity restriction to achieve the value-based manner.

2. Proposed Method

This method investigates the actual values belonging to each attribute in a syntax-semantics-combined fashion. We integrated distributed representation features for syntactically significant words or phrases, which were measured and extracted by tf-idf vectorization.

Specifically, two table data files containing two schemata (attribute sets) to be matched their attributes, consist of columns that were regarded as virtual documents in a similar way implemented in [2]. All the documents containing Japanese terms or phrases that had been previously parsed by the cell boundaries of the tables and a morphological parser, were firstly analyzed and extracted properly for index words with a tf-idf vectorizer. As the output, a list of tf-idf syntactic index words (or tf-idf coordinate names) $w_1, \dots, w_n$ and document-specific sets of their corresponding tf-idf values (or tf-idf coordinates for virtual document A) $t_A(1), \dots, t_A(n)$ could be obtained for further uses. Afterwards, semantically-featured word vectors were generated for all the index words, based on a Word2Vec distributed representation model [6]. In this work, a 200-dimensional word vector was used to represent each tf-idf index word w_i as $\boldsymbol{v}_i = (v_i(1), \dots, v_i(200))$. And by summing them with their respective tf-idf values as the boosting weights, a combined feature vector which is similar to the one employed for sentence meaning by getting mean of word vectors [7], was further generated for each virtual document (table column) as

$$\boldsymbol{FV}_A = \sum_{i=1}^{n} [t_A(i) \cdot \boldsymbol{v}_i].$$

Cosine similarities between such vector pairs were used to decide the ranked list of matching candidates.

3. Experiments

We used two facility ledger dataset A and B from two factories (A and B, respectively) of Fujitsu Ltd., made by Fujitsu Facilities Ltd. for validation. The facility ledger datasets contain information such as related repairing history and setting position information.

The columns inside one dataset were matched to the columns inside the other, being given a list of matching candidate columns that were sorted according to the cosine similarity as scores in descending order. For the evaluation upon the matching results, we compared the obtained similarity scores with the tf-idf-only method, as well as receiver operating characteristic (ROC) curves with those using tf-idf-only and Word2Vec-only methods.

4. Results

4.1 Example of Additional Semantic Similarity

For simplicity, one example is the column named "ビル管呼称" containing the term "警報" among its values, and it was to be matched by using our proposed method and tf-idf-only method which represents commonly-used existing methods. By using tf-idf-only method, this column was given no matching candidate, since all the columns inside the other dataset had no exactly the same term, and the matching failed with zero similarity scores only. However, with the proposed method, since there was some column named "機器名称" in the other dataset containing terms such as "警備", "非常", and "放送" which are semantically related to and all have non-zero similarities with "警報" in the Word2Vec model ("警備, 0.301", "非常, 0.207", and "放送, 0.157" in this study), such column was given as a reasonable matching candidate with a non-zero score as well. And some of such matched column pairs with different attribute names (like column "ビル管呼称" and column "機器名称" here) were among the ground truth ones. This indicates that our proposed method has successfully added the semantic related scores with more potential matching candidates to help getting better matching results through recall etc.

4.2 Quantitative Evaluation with ROC Curves

A wider range of investigation within the sorted matching candidate list until the k-th position in descending order of scores, was performed with the analysis on sensitivity and specificity values along with their corresponding variable k. The compared methods all used cosine similarity as the matching scores. Their feature sets included those related to syntax-semantics aspects (tf-idf-weighted Word2Vec), as well as the syntactic tf-idf-only vectors and the semantic Word2Vec-only results. And for a more intuitive comparison purpose, a ROC curve made by using such analytical metrics is given as shown in Figure 1.

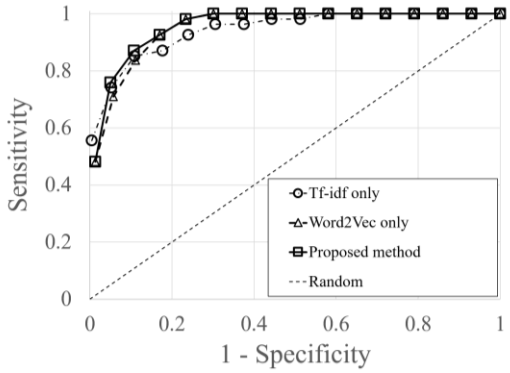

Figure 1. Receiver operating characteristic (ROC) curves of the results obtained with cosine similarity upon tf-idf-only features, Word2Vec features, and our proposed syntax-semantics (combining tf-idf and Word2Vec) features, respectively.

It can be obviously noticed that while given comparison by drawing the ROC curves (the ones closer to the upper left corner are better), for the method proposed in this work (squares), it is outperformed within the narrow range on the very left ($k = 1$) by tf-idf-only method. However, for the most cases where users usually prefer higher recall or sensitivity, even by sacrificing more cost for a wider investigation range over several more schema attributes, the investigated range would be further extended until top k ($k > 1$), and it indicates that our proposed manner has a better overall trade-off between sensitivity and specificity.

5. Conclusions

We have successfully integrated additional semantic features, for another value-based schema matching method by retaining syntactic significance as the boosting weight. It has been demonstrated with improvements in evaluation metrics like recall and the trade-off between sensitivity and specificity, especially when semantic related pairs have different expressions.

References

[1] Stonebraker, Michael, et al. "Data Curation at Scale: The Data Tamer System." CIDR. 2013.
[2] Aumueller, David, et al. "Schema and ontology matching with COMA++." ACM, 2005.
[3] Zhang, Yinuo, et al. "System and method for fuzzy ontology matching and search across ontologies." U.S. Patent Application No. 14/678,943.
[4] Do, Hong-Hai, et al. "COMA: a system for flexible combination of schema matching approaches." VLDB Endowment, 2002.
[5] 佐藤彰洋, et al.“スキーマ構成文字列と主キー制約情報に基づく外部参照関係の推定.”一般社団法人 人工知能学会, 2014.
[6] Mikolov, Tomas, et al. URL https://code. google. com/p/word2vec (2013).
[7] White, Lyndon, et al. "How well sentence embeddings capture meaning." ACM, 2015.