DOI: 10.21917/ijsc.2018.0240

AN UNSUPERVISED HEADER INDEPENDENT APPROACH TOWARDS SUBJECT COLUMN DETECTION IN TABLES

K. Karpaga Priyaa, A. Meena Kabilan and C. Saranya
Department of Computer Science and Engineering, Sri Sai Ram Engineering College, India

Abstract
Subject columns are the important columns that help infer the correct subject matter of the table. The main challenging problem is detecting appropriate subject columns in tables with more than the same. Existing approaches restricted to identification of only one subject column in tables with more than one subject column. With this, it is not possible to infer the correct subject matter of the table. In case of subject column detection, the existing approaches requires table information such as table headers, additional evidences about the table from web pages and also training in prior with a labeled set of tables. To solve these issues, in this paper, we proposed a simple header independent semantic based Concept-Voting Subject Column Detection (CVSCD) algorithm. The proposed algorithm identifies possible subject columns in table with more than one subject column, which provides a way to infer table's correct subject matter. Moreover, CVSCD is unsupervised and works for tables without any table information such as table caption, table headers etc. Experimental results have shown that our approach achieved better accuracy compared to the existing approaches on a corpus of tables extracted from web.

Keywords:
Concept-Voting Subject Column Detection (CVSCD), Subject Column, Subject Matter, Table Headers

1. INTRODUCTION

The use of tables is pervasive throughout all communication, research and data analysis [1] since tables are easy to interpret than paragraph. The main source of Web is, a rich set of millions of tables. These tables either have information such as caption, headers, textual description or do not have few or any of the information. In such cases, it is necessary to infer the correct subject matter of the table. In other words, to know what the table is actually dealing with. A key driving factor towards achieving this goal are the subject columns of table. Identifying subject columns and inferring subject matter of table is essential to retrieve the matched tables to user queries in web tables search and in Research tools [3], to find the related tables in web [6] and in Google Fusion tables [4].

Detecting subject columns in tables without table caption or textual description is a major task. Therefore, different methods have been proposed and adopted to detect the subject columns in table. One such method is the Corpus based supervised approach [13], while some unsupervised approaches [4], [14] have also been proposed. In addition to these approaches, an iterative approach [4] was also adopted to solve the problem. Apart from this, researchers have also developed rule-based and intuitive idea [13] of predicting the subject column. The serious problem faced by the existing work is, the confined detection of only one subject column even if the given table has more than one subject column. In addition to this, the unsupervised approaches fail when the table is not provided with table headers.

Hence, we approach this problem with our proposed simple header independent semantic based approach that is able to detect one and more than one subject columns in tables, if applicable. Our approach is purely concept based majority voting, where the concepts are extracted from a knowledge base, rich in concept-entity sets such as Probase [12], [18]. Being an unsupervised algorithm, it does not require any training in prior and is able to detect the subject columns in table with or without headers.

The rest of the paper is organized as follows. In section 2, we describe the related work in Concept Mapping, Table Extraction and Subject Column Detection. In section 3, we explain the proposed algorithm. In section 4, we discuss the experimental evaluation of proposed approach. In section 5, we conclude the paper with discussions and future directions.

2. BACKGROUND

The main focus of this paper is on proposing a simple, effective concept-based unsupervised algorithm for the detection of the subject column using tables extracted from web by crawl and by using a large knowledge base, Probase for concept mapping [9]. We describe some of the related works for concept mapping, table extraction and subject column detection in this section.

2.1 CONCEPT MAPPING USING PROBASE

Probase is a large web-source probabilistic taxonomy with a rich set of millions of hyponym-hypernym relationships that is constructed from a huge corpus of millions of web pages [16]. Using Probase, the task of concept mapping, called conceptualization [7] is achieved by mapping each entity to its corresponding concepts. Concept mapping finds its application in understanding web tables [14], query [15] and question answering [17]. This rich concept nature of Probase allows us to map the given entity (data cell of table) to a number of concepts.

2.2 TABLES EXTRACTION

The manual conversion of text to a structured design like tables is hard for a vast amount of web data. Hence, researches have used the idea of crawling to extract numerous tables from web pages. [10] Crawled over millions of HTML tables from web pages for annotating tables. [11] Developed a table search engine, Table Seer that extracts tables by crawling digital libraries. [5] Used the general web crawl to extract tables from the raw crawled web pages. [2] Measured the semantic distance between the string values using the crawled table corpus. Hence, in this paper, we

adopt the crawling technique to build our table corpus, an input feed to our proposed algorithm.

2.3 SUBJECT COLUMN DETECTION

We brief existing works and their issues for the subject column detection in each of the following three categories: Intuitive Prediction, Supervised approach and unsupervised approaches.

2.3.1 Intuitive Prediction:

- Rule-based assigned the first non-numerical column as subject column, which is encountered on scanning the table from left.
- Feature-based [4] framed the features that a subject column satisfies and detected the subject column using the features. The issue with the intuitive way of prediction is they fail to detect more than one subject columns in tables if it has.

2.3.2 Supervised Approach:

Most work utilized SVM classifier for detecting the subject columns in tables [3], [6], [4], [13]. The serious problems with the supervised approach are:

- Require training in prior with a labeled set of table corpus.
- Results are training corpus dependent.

2.3.3 Unsupervised Approach:

Using Web [4]: tried to predict the possibility of a column to be the subject column by using additional information about the table like headers, caption and description from web pages. Some of the drawbacks are:

- Not able to work if fail to gather information about the table.
- Ignores column with preposition headers like “for”, “against” etc. Sometimes such columns act as subject columns for inferring the correct subject matter of the table.
- Fails if the given table is without headers.

Using Concepts [14]: proposed entity column detector algorithm using confidence score of concepts. We found two issues tied with this approach. One is, though it was able to detect the entity column in case of tables with headers, it always tends to return the column with maximum confidence score as entity column (subject column in our paper) for tables with more than one subject columns. This characteristic resulted in another issue, this algorithm does not work when we want to infer the correct subject matter for tables with headers. In case of tables without headers, it always returns 0 confidence score for all columns. Since, this algorithm takes the headers into account for calculating confidence score. Hence, for tables without headers, it returns no subject columns. To solve this, there should be a way of handling these issues.

Hence, we attempt to propose an unsupervised header independent semantic based CVSCD, which overcomes all the above stated issues. Our algorithm differs from the existing methods in that we do not use any additional information about tables from web pages. Being unsupervised, it is able to detect one or more than one subject column if it is applicable for tables with or without headers in more efficient way. By the unification of these solutions, it pays a way to infer the correct subject matter of tables with headers.

3. CONCEPT-VOTING SUBJECT COLUMN DETECTION ALGORITHM

Normally, the problem of detecting the subject columns is viewed in two-fold: 1) Table with headers: Detecting subject columns in table with headers play a key role in inferring the subject matter of the table. 2) Table without headers: Detecting subject columns in table without headers would help one to know the important columns of the table. These factors necessitated the task of identifying the subject column in tables with no evidence such as table caption, description text etc. However, the major challenging problem lies in detecting the subject columns in tables with more than one subject column. To meet these challenges, we proposed a header independent, semantic-based unsupervised CVSCD that is designed in such a way that it is suitable for any given table with or without headers. In this section, we describe our method of detecting the subject columns with few running examples.

3.1 CVSCD ALGORITHM

For the given table, the algorithm runs for each column to predict the possibility of it to be the subject column. The algorithm first mines over the sub-components of the current column, called as data cells. It then employs two-way mapping by using Probase. Finally, by using Concept-Voting, it elects the winning column as the subject column. We present the pseudo code of proposed unsupervised CVSCD Algorithm in Algorithm 1. It is described in two-phases as follows:

Algorithm 1: Proposed CVSCD Algorithm

Input:
Web tables with column set T
Output:
One or more subject columns

1: **for** each col $t_j \in T$ **do**
2: Initialize related concept set $C \leftarrow null$
3: **for** each data cell $td_{ji} \in t_j$ **do**
4: Map td_{ji} to a set of concepts $< c >_{td_{ji}}$
5: Update $C \leftarrow < c >_{td_{ji}}$
6: **for** each $c \in C$
7: Initialize entity set E of c, $E(c) \leftarrow null$
8: Initialize column $subj\_col \leftarrow null$
9: Pass c as concept into Probase
10: Update $E(c) \leftarrow \{< e >_c | e \in E\}$
11: if $E(c)$ covers $(3/4)Vote(t_j)$
12: $subj\_col \leftarrow t_j$
13: **break**
14: **if** $subj\_col \leftarrow null$
15: $t_j \neq subj\_col$
16: **Goto** Step 1

3.1.1 Two-Way Mapping Phase:

For each column $\{t_j | j=col\_no\}$ of table, we use Probase to perform data cell concept mapping that maps each data cell $\{td_{ji}|$

$td_{ji} \in t_j$ *i=data_col_no*} to a set of related concepts $< c >_{td_{ji}}$ as shown in lines 1-5. The mapped set of concepts represents the semantic space. We call this kind of mapping as iterative data cell to concept mapping as this process is carried out for each td_{ji} iteratively. The resultant set of all the related concepts are in *C*. While, the lines 6-10 identifies the other way of mapping called the iterative concept to entity set mapping. In this mapping, we try to map each $td'_{ji}s$ concept $c \in < c >_{td_{ji}}$ to a set of entities *E*, obtained for *c* from Probase. However, line 7 is repeated for each $c \in < c >_{td_{ji}} \in C$. Using two-way mapping, we first collected related concepts *C* to the data cells of the given column and then we collected the entity set *E*(*c*) for each of the related concept. However, at this stage, we cannot decide the correct subject column of the table by using only the related concept and entity sets. Hence, phase-2 of CVSCD brings out a clear idea of deciding the subject column using a voting strategy.

3.1.2 Concept-Voting Phase:

With the related concept-entity sets *E*(*c*) and the data cells $<td_{ji}>|td_{ji} \in t_j$, we first decide on whether the column is a partial semantic column or fully semantic column. Before we define what these terms are, we need to know the number of votes contributed by $<td_{ji}>$ to $c \in C$ through *E*(*c*). To understand this, we use tables in Fig.1. For each column of tables, we carried out the two-way mapping phase and obtained

1) $< c >_{td_{ji}} | \, td_{ji} \in t_j$,

2) $E(c) \, | \, c \in < c >_{td_{ji}} \in C$.

Then, we conducted the concept voting process by using *E*(*c*) and t_j. In this process, we counted the number of $t_j s$ that actually occurred in *E*(*c*). In other words, the number of $td_{ji} \in t_j$ who truly elected for the candidate concept c is counted. This can be dealt the other way in terms of shared concepts. We know that each $td'_{ji}s$ semantic space is a collection of hundreds of concepts and also each *E*(*c*) is a collection of thousands of entities. But we do not consider all the concepts of $td'_{ji}s$ semantic space and not all entities of *E*(*c*). We want only the related concepts that are common/shared among $td_{ji}s$ and entities of *E*(*c*) that are part of t_j. Since we need to identify the correct subject column of the table, only the common/shared concepts among the data cells are considered. Hence, we discard the uncommon concepts and confine to an intersected semantic space of $td_{ji}s$. This intersected semantic space semantically represents the entire column. However, since we are interested only in shared concepts *C* among $td'_{ji}s$, we present only them in Table.1. From Table.1, we observe that either all $td'_{ji}s$ or only few $td'_{ji}s$ alone share concepts. If only few $td'_{ji}s$ of the columns share concepts, we call those columns as partial semantic column. On other hand, if all $td'_{ji}s$ of the columns share concepts, then we call them as fully semantic column.

Brand	Product	Characteristic
Nissan	Sunglass	innovative
Volkswagen	Watch	Easy-to-use
Hyundai	Mp3 player	Affordable
Nike	Handbags	Backpacks

(a)

Scientist	Life	Quantity	SI Unit
Issac Newton	1643-1727	Force	newton
James Watt	1796-1819	Power	watt
Michael Faraday	1791-1867	capacitance	farad
Joseph Henry	1797-1878	inductance	henry

(b)

Milk	Semi skimmed	1.7
Cheese	Salty	1.39
Butter	Unsalted	2.29
sausages	Reduced fat	3.49

(c)

Fig.1(a) and (b) Tables with Headers, (c) Table Without Headers

Table.1. Common Voters and their Shared Concepts

Table	Col no. *j*	Common Voters ($td'_{ji}s$ Sharing Concepts)	Some Shared Concepts *C*
(a)	1	Nissan, Volkswagen, Hyundai, Nike	Company, organization, **brand,** manufacturer, international company
	2	Sunglasses, Watch, Mp3 player	**Product**, item, gift, consumer product, valuable item
	3	Innovative, affordable	Term, word, descriptive term
(b)	1	Isaac Newton, James Watt, Michael Faraday, Joseph Henry	**Scientist**, inventor, engineer, person, man
	2	-	-
	3	Force, power, capacitance, inductance	Parameter, **quantity**, energy source, pressure, physical control
	4	newton, watt, farad, henry	Quantity, specification, unit, **SI Unit**, power unit.
(c)	1	Milk, cheese, butter	diary product, Food, ingredient, commodity, product
	2	Salty, Reduced fat	Item, term, word
	3	-	-

This variation in column semantic representation is because, the knowledge base which we use in this paper, may not include all table related information. But it is enriched with all possible sets of instance-concept pairs, that is enough and helpful for machines to interpret what the table is mentioned [14]. However, at this stage, we still cannot decide the correct subject column of the table. Since, even a partial semantic column may sometimes be a subject column. Hence, the decision making is based on the principle of the Boyer-Moore majority voting algorithm [2], which normally chooses the winning candidate i.e. subject column if it is voted by at least $\lfloor n/2 \rfloor$ $td'_{ji}s$. But in our task we elect a column to be the subject column only if it is voted by more than 75% of $td'_{ji}s$. To justify this, we use Fig.1(a) and Fig.1(c). For Fig.1(a), consider the characteristic column, we can get the number of common voters from Table.1 as 2 though the total number of $td'_{ji}s$ is 4. Similarly, for Fig.1(b), consider column 2, for which the number of common voters are found to be 2 out of 4. From this, we find only 50% votes are contributed to characteristic column and column 2. Irrespective of the vote count, we can intuitively say that these columns are not of much importance to know what the table is about. While the other columns, in Fig.1(a), the number of common voters to columns Brand and Product are 4 and 3 respectively. Here, column1 achieves 100% voting and column 2 achieves 75% voting. Hence, we can declare column 1 and column 2 as subject columns of Fig.1(a). Similarly, in Fig.1(b), columns Scientist, Quantity and SI Unit, achieves 100% voting. Hence, we can declare columns 1, 3 and 4 as subject columns. It may be surprising about why all 3 columns are declared to be subject columns. The reason behind this is dealt in the following part of inferring subject matter. In case of table without headers, consider Fig.1(c), column 1 achieves 75% votes whereas column 2 achieves only 50% vote. Hence, we declare column 1 as subject column.

Thus, our proposed two-phase algorithm helps in detecting more than one subject columns of table, which in turn pays a way to infer the correct subject matter of the table.

3.2 INFERRING SUBJECT MATTER

As discussed already, we can make use of the subject columns that were detected by CVSCD algorithm, to infer the subject matter of the table. To understand this, consider Fig.1(a), where the subject columns detected by CVSCD algorithm is *Brand* and *Product.* With the predicted subject columns in hand, we can intuitively say that table is dealing with "the brand and its products" rather than "product". The motivation of this paper has inferred in two different subject matters. With the subject columns *Scientist* and *Quantity*, the Fig.1(b) is about "Scientist and their inventions". While, with subject columns *Quantity* and *SI Unit*, we can infer that the table is about "Quantities with their standard units". The accurate subject matter of Fig.1(b), with all 3 subject columns is about "Scientist and their inventions with corresponding standard units". One may feel that this much accurate information about table is not necessary. But there are applications for which incurring the accurate subject matter of the table plays a key role. For example, in search engines, inferring accurate subject matter of huge amount of web tables is important because their task is to find and integrate related tables for a given user query. To achieve this, our proposed algorithm CVSCD, forms a strong base.

But such a kind of accurate or correct prediction of subject matter of the table lacks in the existing approaches.

In terms of correct subject matter with proper table information has proved with the Fig.1(b). In case of tables without headers as in Fig.1(c), we cannot infer the subject matter though we are able to detect the subject columns. But it is possible to detect the subject matter using the concepts discovered for the subject column1. In this case, the most general concept *product* can be assigned as the subject matter of the Fig.1(c).

Thus, we have theoretically justified how important the proposed CVSCD algorithm is for the web tables. We prove it experimentally in the following section.

4. EXPERIMENTAL EVALUATION

In this section, we perform the experimental evaluation of the proposed algorithm. The goal of this section is to compare the task of subject column detection using different approaches in terms of accuracy.

4.1 TABLE CORPUS

We extracted tables from millions of crawled web pages [10] and constructed a corpus of 1340 web tables in structured form using CSV files [8] from various domains. We manually labeled the subject columns on the extracted table corpora. Among 1340 web tables, 810 tables are found to have more than 1 subject column, among 810 tables, 270 tables are without headers. While the rest of 530 tables with 1 subject column, among 530 tables, 185 tables have no headers. The characteristics of the web table corpora are:

- Extracted web tables may not have table caption, text description about table.
- There may not be any table headers.

4.2 METHODS FOR COMPARISON

- *Rule-based* [13]: This is an intuitive approach of assigning the first column to be the subject column by scanning the table from left to right.
- *SVM Classifier* [3] [13]: This approach tries to overcome the drawback in rule-based method by a supervised prediction of subject column. Being a binary classification model, we considered another table corpus of 1000 extracted tables which is a collection with and without headers, of which we used 800 to train the model and 200 to test the model. Among 1000, 640 tables are with more than one subject column.
- *Entity Column Detector* [14]: To preserve the time for training which is a major drawback in SVM, [14] developed an unsupervised algorithm that attempts to identify the subject column but always restricted to one in tables with more than one subject column.
- *Proposed CVSCD Algorithm*: The proposed algorithm in Algorithm 1, overcomes the problem stated in the previous approaches.

4.3 EVALUATION

For the purpose of evaluation, we assign the positive cases as tables with more than one subject column and negative cases as tables with one subject column. We execute all the four methods on the constructed table corpus of 1340 tables and observed that *Rule-based* and *Entity Column Detector* approaches did not predict more than one subject column in 810 tables (more than one subject column). The reason behind this prediction of Rule-based approach is, it always attempts to assign the first non-numerical column from left as the subject column. Similarly, with entity column detector algorithm, it always takes the column with maximum confidence score as subject column. In addition to this, it does not return any subject column for tables without headers, as the confidence score is 0. Hence, we did not take them for further evaluation. But we found SVM Classifier and the proposed CVSCD algorithm were able to predict more than one subject column. So we graphically represented our observations for those approaches in Fig.2.

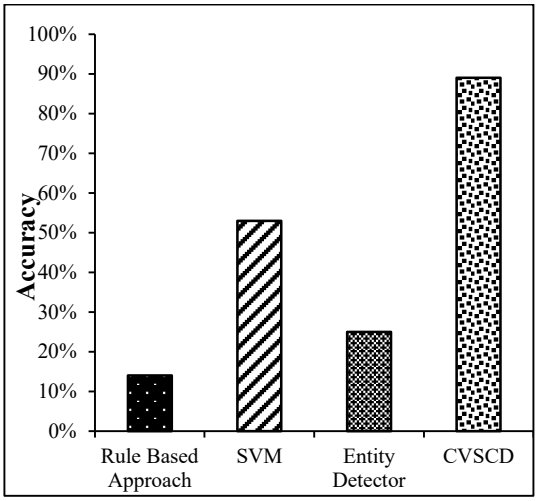

Fig.2(a). Accuracy (Y-axis) achieved by subject column detection approaches

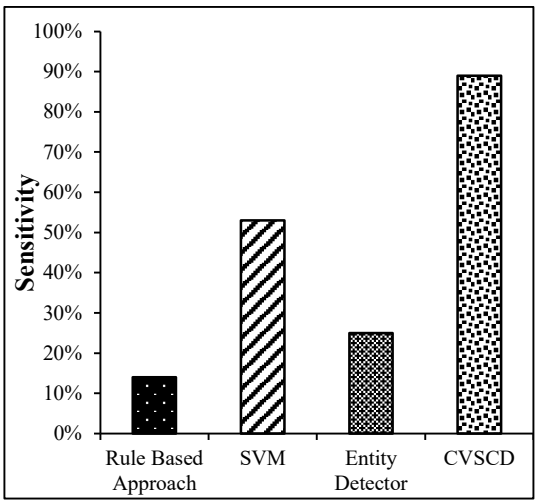

Fig.2(b). Sensitivity (Y-axis) achieved by subject column detection approaches

The Fig.2(a) gives the accuracy of detecting more than one subject columns in tables (more than one subject column with or without headers). Our experimental evaluation shows that for SVM classifier, the accuracy was 53% and the proposed CVSCD algorithm was able to achieve 89% accuracy. SVM classifier was able to detect more than one subject columns in some cases but sometimes it incorrectly predicts the subject columns in most of the cases. But, CVSCD algorithm not only correctly predicts more than one subject columns but also correctly predicts the subject columns in most cases. Similarly, Fig.2(b) shows the comparison of sensitivity among the four approaches. Rule-based and Entity column detection approaches leads to 0% as both fail to detect more than one subject columns in positive cases. The specificity of the proposed approach has been compared to other methods in Fig.2(c), where CVSCD approach still proves to show positive result. Hence, the proposed algorithm gives better results for the collection of tables with more than one subject column with or without headers.

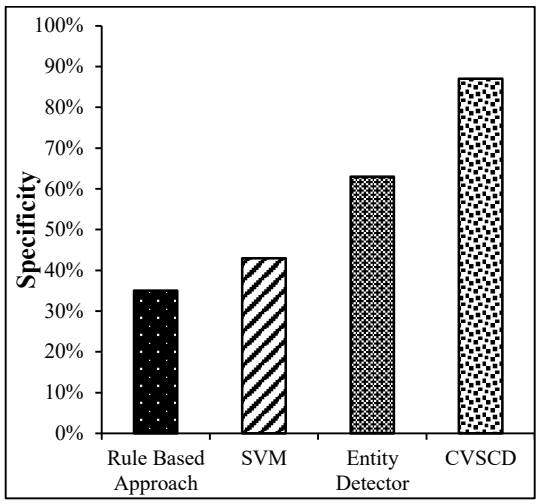

Fig.2(c). Specificity (Y-axis) achieved by subject column detection approaches

5. CONCLUSION

We proposed an unsupervised header independent subject column detection algorithm that detects more than one subject column and acts as a key driving factor for inferring the correct subject matter of the table. It works in both cases of tables, with and without headers. This algorithm provides a strong base for inferring accurate subject matter which is very much useful in applications that involve web search engines to retrieve related tables and at one step further of integrating the semantically related tables in Google Fusion. As a future work, we would like to handle the numerical columns which can also be a subject column to convey the subject matter of the table.

REFERENCES

[1] Table, Available at: https://en.wikipedia.org/wiki/Table

[2] Wim H. Hesselink, "The Boyer-Moore Majority Vote Algorithm", Available at: http://www.cs.rug.nl/~wim/pub/whh348.pdf.

[3] Sreeram Balakrishnan et al., "Applying WebTables in Practice", *Proceedings of the Biennial Conference on Innovative Data Systems Research*, pp. 1-6, 2015.

[4] Chandra Sekhar, Thanapon Noraset, and Doug Downey, "Methods for Exploring and Mining Tables on WIKIPEDIA", *Proceedings of ACM Workshop on Interactive Data Exploration and Analytics*, pp. 18-26, 2013.

[5] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu and Yang Zhang, "Webtables: Exploring the Power of Tables on the Web", *Proceedings of the VLDB Endowment*, pp. 538-549, 2008.

[6] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin and Cong Yu, "Finding Related Tables", *Proceedings of ACM International Conference on Management of Data*, pp. 817-828, 2012.

[7] Dongwoo Kim, Haixun Wang and Alice Oh, "Context-Dependent Conceptualization", Available at: http://uilab.kaist.ac.kr/research/IJCAI13/ijcai13_dongwoo_camera_ready.pdf.

[8] Oktie Hassanzadeh, Michael J. Ward, Mariano Rodriguez-Muro and Kavitha Srinivas, "*Understanding a Large Corpus of Web Tables through Matching with Knowledge Bases-An Empirical Study*", Available at: https://pdfs.semanticscholar.org/f3d7/550fcdf9c284874c05931ced2ffbcb2accc0.pdf.

[9] J. Liang, Y. Xiao, Y. Zhang, S.W. Hwang and H. Wang, "Graph-Based Wrong is a Relation Detection in a Large-Scale Lexical Taxonomy", *Proceedings of 31st AAAI on Artificial Intelligence*, pp. 1178-1184, 2017.

[10] G. Limaye, S. Sarawagi and S. Chakrabarti, "*Annotating and Searching Web Tables Using Entities, Types and Relationship*", *Proceedings of the VLDB Endowment*, pp. 1338-1347, 2010.

[11] Y. Liu, K. Bai, P. Mitra and C.L. Giles, "Tableseer: Automatic Table Metadata Extraction and Searching in Digital Libraries", *Proceedings of 7th ACM/IEEE Joint Conference on Digital Libraries*, pp. 91-100, 2007.

[12] J. Park, H. Cho and S.W. Hwang, "Understanding Relations using Concepts and Semantics", *Proceedings of 3rd International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*, pp. 1-15, 2017.

[13] Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao and Chung Wu, "Recovering Semantics of Tables on the Web", *Proceedings of the VLDB Endowment*, pp. 528-538, 2011.

[14] Jingjing Wang, Haixun Wang, Zhongyuan Wang and Kenny Q. Zhu, "Understanding Tables on the Web", *Proceedings of 31st International Conference on Conceptual Modeling*, pp. 141-155, 2012.

[15] Z. Wang, K. Zhao, H. Wang, H. Meng and J.R. Wen, "Query Understanding through Knowledge-based Conceptualization", Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/paper_msr.pdf, Accessed on 2015.

[16] Wentao Wu, Hongsong Li, Haixun Wang and Kenny Q. Zhu, "Probase: A Probabilistic Taxonomy for Text Understanding", *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 481-492, 2012.

[17] Wen-Tau Yih, Ming-Wei Chang, Christopher Meek and Andrzej Pastusiak, "Question Answering using Enhanced Lexical Semantic Models", *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1744-1753, 2013.

[18] Y. Zhang, Y. Xiao, S.W. Hwang and W. Wang, "Entity Suggestion with Conceptual Explanation", *Proceedings of 26th International Joint Conference on Artificial Intelligence*, pp. 4244-4250, 2017.