International Journal of Engineering &Technology, 7 (3.12) (2018) 157-160

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



EOBAA: Enhanced Ontology Based Alignment Algorithm for Mining Frequent Patterns

[1]D. SrinivasaRao, [2]V. Sucharitha, K.V.V Satyanarayana

[1]Research Scholar, Department of Computer Science & Engineering, KL University, AP
[2]Professor, Department of Computer Science & Engineering, Narayana Engineering College, Gudur, ChittorDist, AP
[2]Professor, Department of Computer Science & Engineering, KL University, AP
**Corresponding Author E-mail: [1]srinimtechcse2007@gmailcom, [2]jesuchi78@yahoo.com*

Abstract

Mining frequent patterns are most widely used in many applications such as supermarkets, diagnostics, and other real-time applications. Performance of the algorithm is calculated based on the computation of the algorithm. It is very tedious to compute the frequent patterns in mining. Many algorithms and techniques are implemented and studied to generate the high-performance algorithms such as Prepost+ which employees the N-list to represent itemsets and directly discovers frequent itemsets using a set-enumeration search tree. But due to its pruning strategy, it is known that the computation time is more for processing the search space. It enumerates all item sets from datasets by the principle of exhaustion and they don't sort them based on utility, but only a statistical proof of most recurring itemset. In this paper, the proposed Enhanced Ontologies based Alignment Algorithm (EOBAA) to identify, extract, sort out the HUI's from FI's. To improve the similarity measure the proposed system adopted Cosine similarity. The experiments conducted on 1 real datasets and show the performance of the EOBAA based on the computation time and accuracy of the proposed EOBAA.

Keywords: *Pruning, frequent patterns, interestingness, Prepost+, ontologies.*

1. Introduction

Similarity measures are most widely done in many domains for finding the similarity items or itemsets based on the dataset. Mostly the similarity measures in psychology, biology, and statistics and other domains. The other name for the similarity referred as semantic similarity or also as ontological similarity. The issue in the semantic web has further transferred the data into related heterogeneous ontology I,e data interoperability. In this paper, the issue focuses on ontology alignment (OA) process. The basic objective of this proposed system is to find the similarities between items and itemsets. For the better performance of the ontology alignment the similarity measure Cosine similarity is utilized for the improved output and this is to find the similar and dissimilar items. This paper investigates the use of semantic similarity measures in this process and their application in a proposed semantic alignment quality (SAQ) measure to evaluate the alignment results of OA systems. The SAQ measure does not require a reference alignment as does the standard OA performance measures of similar, dissimilar items [1].

In chapter 2 describes the literature survey, chapter-3 describes the ontology alignment overview, chapter-4 describes the existing system, and chapter-5 describes the proposed system, the next are results and conclusion.

2. Literature Survey

The authors explained in [4], which utilizes metaphysics to progress bolster in administer mining, is a case of an approach that considers semantic data amid the pre-handling step. In that work, information are raised to more summed up ideas as indicated by the metaphysics, and after that the mining procedure is performed by a traditional affiliation control mining calculation, as Apriori [8]. The creators contend that past information speculation makes it conceivable to consider subcategories in help estimation, creating rules with higher help. Besides, got principles can be simpler to decipher, since they contain abnormal state ideas that speak to wealthier data than particular terms in the database.

Then again, an applicable work has concentrated on the postprocessing step. In [3], for instance, space learning is utilized to sum up low level standards found by normal control mining calculations, with a specific end goal to get less and clearer abnormal state rules. The creators utilized ontologies to sum up the items or ideas in rules in the wake of applying the calculation of information mining, and afterward they connected the information mining calculation again to find the abnormal state in theory runs the show.

Another illustration is depicted in [12], where ontologies are utilized to decide lead intriguing quality. This is finished by checking whether found tenets affirm, repudiate or uncover new data when contrasted with the information accessible in the cosmology.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Besides, the creator moreover proposed input systems to refresh area learning from created rules, in light of the fact that new and fascinating bits of knowledge can be found from the consequences of the mining procedure.

Different methodologies, as ExCIS [13], utilize area learning in both pre-handling and post-preparing steps. In this work, the pre-preparing step utilizes metaphysics to manage the development of particular informational indexes for specific mining assignments. The following stage is the utilization of a standard mining calculation which removes designs from these datasets. At long last, in the post processing step, mined standards might be deciphered and additionally separated, as their terms are summed up as per a philosophy. In this way, semantic data utilized as a part of ExCIS bolsters dataset readiness and permits decreasing the volume of separated examples.

Agrawal et al (1994) thought about the issue of finding affiliation governs between things in a huge database of offers exchanges. They displayed two new calculations for taking care of this issue is on a very basic level unique from the known calculations. Another system for finding worldly designs in successive information is the mining of incessant scenes proposed by Mannila et al (1994). In this system, the information are given in a solitary long succession of occasions, and the undertaking is to find visit scenes in the succession, which is later used to produce scene rules. Agrawal et al (1995) presented the issue of mining summed up affiliation rules. Given an expansive database of exchanges, where every exchange comprises of an arrangement of things, and a scientific classification (is-a chain of importance) on the things, they discovered relationship between things at any level. Savasereet a; (1995) exhibited a productive calculation for mining affiliation decides that is on a very basic level not quite the same as known calculations. Since, their calculation not just lessens the I/O overhead essentially yet in addition has bring down CPU overhead for generally cases.

Bifet et al (2008) proposed another approach for mining shut unlabeled established trees adaptively from information streams that change after some time. Their approach depends on a proficient portrayal of trees and a low intricacy thought of loose shut trees, and prompts an online system and a versatile sliding window system for managing changes after some time. Liu Ken-Hao et al (2008) proposed a general post-sifting system to improve strength and exactness of semantic idea recognition utilizing affiliation and fleeting examination for idea learning revelation. Cooccurrence of a few semantic ideas could infer the nearness of different ideas. Lee et al (2009) proposed an effective calculation, to mine close examples in a period arrangement database where each record in the database, moreover called an exchange, contains various time arrangement groupings.

Song et al (2006) exhibited a novel calculation for mining complete visit itemsets. This calculation is alluded to as the Transaction Mapping (TM) calculation. In this calculation, exchange IDs of each itemset are mapped what's more, compacted to nonstop exchange interims in an alternate space and the tallying of itemsets is performed by crossing these interim records in a profundity first request along the lexicographic tree. Wang Jianyong et al (2007) proposed an effective calculation, for mining regular shut arrangements without applicant upkeep. It receives a novel succession conclusion checking plan called BI-Directional Extension (BIDE) and prunes the inquiry space more profoundly contrasted with the past calculations by utilizing the BackScan pruning strategy. An effective calculation named Apriori - development in view of Apriori calculation and the FP-Tree structure is introduced by Wu Bo et al (2008) to mine regular examples.

Zaki and Gouda (2003) introduced a calculation that uses the basic properties of continuous itemsets to encourage quick disclosure. In their work, related database things are gathered together into bunches speaking to the potential expanded continuous itemsets in the database. Cross section traversal methods are utilized by them to rapidly recognize all the genuine expanded visit itemsets. Lucchese et al (2006) exhibited another versatile calculation for finding shut incessant itemsets. They built up a lossless and dense portrayal of all the successive itemsets that are to be mined from an exchange database. Their calculation abuses a partition and-overcome approach and a bitwise vertical portrayal of the database. Wang Yunlan et al (2006) displayed a general incremental calculation for mining successive scenes. Also, it proposes and uses the window-hash-based system to prune competitor scenes.

3. Ontology Alignment Overview

In this paper, the ontology alignment is used for the research to get the formatted output is described by [1]. There are many inputs taken by the ontology alignment such as source ontology Os to find the entities es based on the parameters such as (topic, relation, instance) and all the entity parameters have the same meaning. The entities with the same meaning and the alignment process easily initialize a one-to-one similarity or finding the relation between the source and target entity. Based on the similarity the other relations behind the equality such as predictions may also be initialized as two entities.

The set of entities that are required for the general ontologies are E, the set of real ontologies O, and M-as set of possible alignment relations. The partial method genalign: E x O x O -> E xM. The output of the alignment process is based on the alignment systems is with the confidence level 1.

4. Existing System

1. Proposes a new optimization algorithm called PrePost+ as well as the use of a compact tree like data structure for efficiently discovering FI's sets from chosen datasets.
2. PrePost+ employs N-list to represent itemsets and directly discovers frequent itemsets in an itemset& N-list search tree.
3. For avoiding repetitive search, it also adopts Children–Parent Equivalence pruning to greatly reduce the search space. For evaluating the performance of PrePost+.
4. PrePost+ contains two integral components. They are:

- NL_intersection
- Building_Pattern_Tree

5. NL_intersection are used to generate N-lists of (k+ 1)-itemsets by intersecting N-lists of k-itemsets.
6. Building_Pattern_Tree to search the set-enumeration tree (from level 2) and discovery all frequent k-itemsets extended from each frequent 2-itemset thus reducing the search space.
7. To address prior systems constraints wsearch space iterations are considered.
8. Different types of both real and synthetic data sets can be used in a series of experiments to compare the performance of the proposed algorithms with prior approaches.
9. Experimental results show that proposed methods outperform prior approaches substantially in terms of execution time, especially when data contain lots of long transactions.

Cosine similarity is a common vector based similarity measure similar to dice coefficient. Whereby the input string is transformed into vector space so that the Euclidean cosine rule can be used to determine similarity.

$$x.y = \|x\|\|y\| \cos \emptyset$$

$$\text{Similarity} = \cos(\emptyset) = \frac{X.Y}{\|X\|\|Y\|} = \frac{\sum_{i=0}^{n} X_i Y_i}{\sqrt{\sum_{i=0}^{n} X_i^2} \sqrt{\sum_{i=0}^{2} Y_i^2}}$$

Where X,Y are the components of the x_i, y_i.
By using this siilarty measures.

5. Proposed System

1. Nevertheless, relative importance of each item is not considered in frequent pattern mining.
2. The meaning of item set utility is interestingness, importance, or profitability of an item to users. Utility of items in a transaction database consists of two aspects:
- The importance of distinct items, which is called external utility, and the importance of items in transactions, which is called internal utility.
3. Utility of an item set is defined as the product of its external utility and its internal utility. An item set is called a high utility item set if its utility is no less than a user-specified minimum utility threshold; otherwise, it is called a low-utility item set.
4. Existing approaches enumerate all item sets from datasets by the principle of exhaustion and they don't sort them based on utility, but only a statistical proof of most recurring itemset.
5. So we propose an Enhanced Ontologies based Alignment Algorithm to identify, extract, sort out the similar and dissimilar items. An algorithmic implementation is as follows:
6. So using the above algorithm we will obtain a subset M of total items(I) that contains a priority index to formulate HUI's and the efficiency is highlighted using an application prototype.

Algorithm Steps:
Step-1: Start.
Step-2: Load dataset.
Step-3: Implement PS.
Step-4: Similarity Index.
Step-5: Results.
Step-6: End

6. Experimental Results

The experimental results show the performance of the proposed algorithm and the existing algorithms based on overall computation time and space usage. The existing and proposed algorithms introduced the finding of frequent patterns by adopting the cosin similarity measures.

The one real datasets were used in the existing and proposed system for frequent pattern mining. These dataset were downloaded from FIMI repository (http://fimi.ua.ac.be). The mushroom dataset contains characteristics of various kinds of mushrooms. By utilizing the proposed system the similar and dissimilar items are identify based on the mushroom dataset.

Data Sets / Data sources are used

In this section, we will explain about the dataset used is mushroom dataset. Which is used to find out the poison and edible with features of the mushrooms?

@relation mushroom
@attribute cap-shape {x,b,s,f,k,c}
@attribute cap-surface {s,y,f,g}
@attribute cap-color {n,y,w,g,e,p,b,u,c,r}
@attribute bruises {t,f}
@attribute odor {p,a,l,n,f,c,y,s,m}
@attribute gill-attachment {f,a}
@attribute gill-spacing {c,w}
@attribute gill-size {n,b}
@attributegill-color {k,n,g,p,w,h,u,e,b,r,y,o}
@attribute stalk-shape {e,t}
@attribute stalk-root {e,c,b,r}
@attribute stalk-surface-above-ring {s,f,k,y}
@attributestalk-surface-below-ring {s,f,y,k}
@attributestalk-color-above-ring {w,g,p,n,b,e,o,c,y}
@attributestalk-color-below-ring {w,p,g,b,n,e,y,o,c}
@attribute veil-type {p}
@attribute veil-color {w,n,o,y}
@attribute ring-number {o,t,n}
@attribute ring-type {p,e,l,f,n}
@attributespore-print-color {k,n,u,h,w,r,o,y,b}
@attribute population {s,n,a,v,y,c}
@attribute habitat {u,g,m,d,p,w,l}
@attribute class {p,e}

Table 1: Performance of ES & PS

	Similarity	Dissimilarity	Time	Accuracy
Existing System	**1230**	**983**	**89.431422312**	**10.870987789987**
Proposed System)	**4208**	**3916**	**57.101636705 seconds**	**0.9958148695224027**

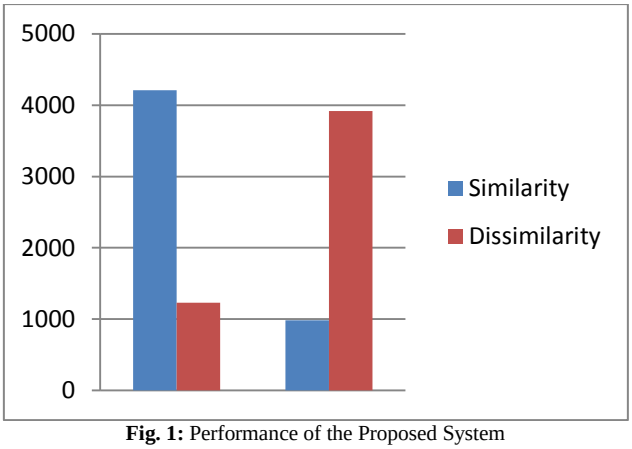

Fig. 1: Performance of the Proposed System

In the above graph, the performance is calculated based on the identifying the similarities and dissimilarities by using the proposed system. The better performance can be seen by using this proposed algorithm.

7. Conclusion

The proposed system focuses on various parameters to find the semantic similarities and dissimilarities with adopted cosine similarity measures for the better performance. The proposed system based on computation time and memory management if the dataset is very huge. The experiments shows the advanced hybrid method ontology alignment with cosine similarity become most performed technique.

References

[1] M. Ehrig, Ontology Alignment: Bridging the Semantic Gap, Springer Science+Business Media, LLC 2007.
[2] J. Euzenat, A. Ferrara, C. Meilicke, A. Nikolov, J. Pane, F. Scharfffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn, "The Results of the Ontology Alignment Evaluation

Initiative 2010. Ontology Matching Workshop, International Semantic Web Conference, Shanghai, China 2010.

[3] J. Euzenat, "An API for ontology alignment," Proc of the 3rd International Semantic Web Conference (ISWC-2004), Hiroshima, Japan, 2004.

[4] N. Goodman, "Seven strictures on similarity" In N. Goodman (Ed.), Problems and projects. Pp. 437-447. Bobbs-Merrill, New York, 1972.

[5] D.L. Medin, R. L. Goldstone, and D. Gentner, "Respects for Similarity," Psychological Review 100(2), pp. 254-278, 1993.

[6] V. Cross and T. Sudkamp, Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications. New York: Physica-Verlag, ISBN 3-7908-1458, 2002.

[7] A.Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web," Proc. 11th Int 'l Conf. World Wide Web (WWW '02), pp. 662-673, 2002.