www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Semantic based sentence ordering with ontology-mining in heterogeneous data using explicit semantic analysis

[1]Mr. S.Edison, [2]Mrs. K.E. Eswari, M.C.A., M.Phil., M.E.,

Final MCA., Department of MCA, Nandha Engineering College (Autonomous), Erode-52
Associate Professor/MCA Department of MCA, Nandha Engineering College (Autonomous), Erode-52.

ABSTRACT

In this analysis an ontology-oriented architecture where core ontology has been used as knowledge base (KB) and allows data integration of different heterogeneous sources. In existing model used to Natural Language Processing and Artificial Intelligence methods to process and mine data in the health sector to uncover knowledge hidden in diverse data sources. The approach has been applied in the field of personalized medicine (study, diagnosis, and treatment of diseases customized for each patient). AI methods have been used with the objective to mine data in the healthcare sector to uncover knowledge hidden in heterogeneous data sources. A set of learned rules (using Data Mining techniques on structured data, DM rules) and their improvements (applying NLP techniques on data from the Web) are obtained. In additionally proposed system, to apply three phase Ontology, first stop word removal, stemming and semantic (Synonym word) replacement is used for preprocessing. Next phase Naïve Bayes classification is used. Next phase Rules Extraction is processed and final phase Explicit Semantic analysis is made. In this method automatically construct and incorporate document and word constraints to support unsupervised constrained clustering. The result of the evaluation demonstrates the superiority of our approaches against a number of existing approaches.

Keywords: Data Mining, Ontology Mining, Classification Model, Clustering, Automatic Analysis

INTRODUCTION

Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. (Note, however, that reporting is not always considered to be data mining.)

DATA MINING TASKS

- Data mining commonly involves four classes of tasks:
- Clustering - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree

Author for correspondence:
Department of MCA, Nandha Engineering College (Autonomous), Erode-52

learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.

- Regression - Attempts to find a function which models the data with the least error.
- Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.
- Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

OBJECTIVE

Document clustering is being studied from many decades but still it is far from a trivial and solved problem. The challenges are:

- A systematic approach called multi-document summarization is required to generate a summary about particular topic.
- It should address the semantic relationship between the sentences in the summary.
- Since data mining based on health records and diseases has become a hot research topic, Explicit Semantic Analysis should be carried out.
- Rules should be extraction such that Co-occurrence of symptom/medicine/treatment details.
- Should improve the health industry in diagnosis thereby.

RELATED WORKS

David Gil And Antonio Fernandez [1] describe the Internet of Things (IoT) has made it possible for devices around the world to acquire information and store it, in order to be able to use it at a later stage. However, this potential opportunity is often not exploited because of the excessively big interval between the data collection and the capability to process and analyze it. In this paper, we review the current IoT technologies, approaches and models in order to discover what challenges need to be met to make more sense of data. The main goal of this paper is to review the surveys related to IoT in order to provide well integrated and context aware intelligent services for IoT. Moreover, we present a state-of-the-art of IoT from the context aware perspective that allows the integration of IoT and social networks in the emerging Social Internet of Things (SIoT) term.

M. Fazio and a. Celesti [2] describes monitoring activities detect changes in the environment and can be used for several purposes. To develop new advanced services for smart environments, data gathered during the monitoring need to be stored, processed and correlated to different pieces of information that characterize or influence the environment itself. In this paper we propose a Cloud storage solution able to store huge amount of heterogeneous data, and provide them in a uniform way. To this aim, we adopt an hybrid architecture that couple Document and Object oriented strategies, in order to optimize data storage, querying and retrieval. In this paper, we present the architecture design and discuss some implementation details in the development of the architecture within a specific use case.

Javier Medina And Macarena Espinilla [3] describe a data streams generated from devices collecting data from patients, which are monitored within both clinical and home environments, provide useful information for decision making processes. Nevertheless, medical personnel are still required to review and process the data and therefore spend a lot of time and effort to detect situations of concern such as exacerbations with conditions or the occurrence of anomalies in the measurements. In this paper, we propose a methodology for the real-time linguistic analysis of data streams generated from medical monitoring devices based on a rule-based inference engine exploiting a fuzzy linguistic approach. A case study based on health data provided by the Physiological Data Modeling Contest is used to illustrate the proposed methodology and to demonstrate the flexibility to interpret, in a linguistic manner, data streams and the detection of risk situations of interest based on linguistic expressions.

Rimma pivovarov and Adler j. Perotte [4] presents the Unsupervised Phenome Model

(UPhenome), a probabilistic graphical model for large-scale discovery of computational models of disease, or phenotypes. We tackle this challenge through the joint modeling of a large set of diseases and a large set of clinical observations. The observations are drawn directly from heterogeneous patient record data (notes, laboratory tests, medications, and diagnosis codes), and the diseases are modeled in an unsupervised fashion. We apply UPhenome to two qualitatively different mixtures of patients and diseases: records of extremely sick patients in the intensive care unit with constant monitoring, and records of outpatients regularly followed by care providers over multiple years. We demonstrate that the UPhenome model can learn from these different care settings, without any additional adaptation. The results shows that (i) the learned phenotypes combine the heterogeneous data types more coherently than baseline LDA-based phenotypes; (ii) they each represent single diseases rather than a mix of diseases more often than the baseline ones; and (iii) when applied to unseen patient records, they are correlated with the patients' ground-truth disorders. Code for training, inference, and quantitative evaluation is made available to the research community.

Cristina Soguero-Ruiz And Kristian Hindberg [5] developed a learning system capable of exploiting information conveyed by longitudinal Electronic Health Records (EHRs) for the prediction of a common postoperative complication, Anastomosis Leakage (AL), in a data-driven way and by fusing temporal population data from different and heterogeneous sources in the EHRs. We used linear and non-linear kernel methods individually for each data source, and leveraging the powerful multiple kernels for their effective combination. To validate the system, we used data from the EHR of the gastrointestinal department at a university hospital. We first investigated the early prediction performance from each data source separately, by computing Area Under the Curve values for processed free text (0.83), blood tests (0.74), and vital signs (0.65), respectively.

When exploiting the heterogeneous data sources combined using the composite kernel framework, the prediction capabilities increased considerably (0.92). Finally, posterior probabilities were evaluated for risk assessment of patients as an aid for clinicians to raise alertness at an early stage, in order to act promptly for avoiding AL complications. Machine-learning statistical model from EHR data can be useful to predict surgical complications. The combination of EHR extracted free text, blood samples values, and patient vital signs, improves the model performance. These results can be used as a framework for preoperative clinical decision support.

Didier Dubois And Weiru Liu [6] propose and advocate basic principles for the fusion of incomplete or uncertain information items that should apply regardless of the formalism adopted for representing pieces of information coming from several sources. This formalism can be based on sets, logic, partial orders, possibility theory, belief functions or imprecise probabilities. We propose a general notion of information item representing incomplete or uncertain information about the values of an entity of interest. It is supposed to rank such values in terms of relative plausibility, and explicitly point out impossible values. Basic issues affecting the results of the fusion process, such as relative information content and consistency of information items, as well as their mutual consistency, are discussed.

Diogo Machado and Tiago Paiva [7] describe a Diabetes management is a complex and a sensible problem as each diabetic is a unique case with particular needs. The optimal solution would be a constant monitoring of the diabetic's values and automatically acting accordingly. We propose an approach that guides the user and analyses the data gathered to give individual advice. By using data mining algorithms and methods, we uncover hidden behavior patterns that may lead to crisis situations. These patterns can then be transformed into logical rules, able to trigger in a particular context, and advise the user. We believe that this solution is not only beneficial for the diabetic, but also for the doctor accompanying the situation. The advice and rules are useful input that the medical expert can use while prescribing a particular treatment. During the data gathering phase, when the number of records is not enough to attain useful conclusions, a base set of logical rules, defined from medical protocols, directives

and/or advice, is responsible for advise and guiding the user. The proposed system will accompany the user at start with generic advice, and with constant learning, advise the user more specifically. We discuss this approach describing the architecture of the system, its base rules and data mining component. The system is to be incorporated in a currently developed diabetes management application for Android.

In this work two different methods were used: association rules and Bayesian networks.

- Association rules: reveal links, and the weight of these links, between variables. By applying this algorithm to our users we were able to conclude rules such as e.g. "At Wednesday in the afternoon your usual meal results in high glycemic values.". This connection of days and times of the week with crisis occurrences is fundamental to avoid or correct incorrect behaviors.
- Bayesian networks: show variable probabilistic dependencies. In contrast to the last example, where we found relations between variables, with Bayesian networks it's possible to approach the problem of crisis prediction in a different manner. After creating a network for a user we can now ask e.g. "what is the probability of having hypoglycemia given that today is Thursday."

Behzad Golshan And Alon Halevy [8] describe the field of data integration has expanded significantly over the years, from providing a uniform query and update interface to structured databases within an enterprise to the ability to search, exchange, and even update, structured or unstructured data that are within or external to the enterprise. This paper describes the evolution in the landscape of data integration since the work on rewriting queries using views in the mid- 1990's. In addition, we describe two important challenges for the field going forward. The first challenge is to develop good open- source tools for different components of data integration pipelines. The second challenge is to provide practitioners with viable solutions for the long-standing problem of systematically combining structured and unstructured data.

Maria Ganzha and Marcin Paprzycki [9] describe The Internet of Things (IoT) idea, explored across the globe, brings about an important issue: how to achieve inter-operability among multiple existing (and constantly created) IoT platforms. In this context, in January 2016, the European omission has funded seven projects that are to deal with various aspects of interoperability in the Internet of Things. Among them, the INTER- IoT project is aiming at the design and implementation of, and experimentation with, an open cross-layer framework and associated methodology to provide voluntary interoperability among heterogeneous IoT platforms. While the project considers interoperability across all layers of the software stack, we are particularly interested in answering the question: how ontologies and semantic data processing can b e harnessed to facilitate interoperability across the IoT landscape. Henceforth, we have engaged in a "fact finding mission" to establish what is currently at our disposal when semantic interoperability is concerned. Since the INTER-IoT project is initially driven by two use cases originating from (i) (e/m) Health and (ii) transportation and logistics, these two application domains were used to provide context for our search. The paper summarizes our findings and provides foundation for developing methods and tools for supporting semantic interoperability in the INTER-IoT project (and beyond).

Antoni Olivé [10] describes a vision of a universal ontology (UO) aiming at solving, or at least greatly alleviating, the semantic integration problem in the field of conceptual modeling and the understandability problem in the field of the semantic web. So far it has been assumed that the UO is not feasible in practice, but we think that it is time to revisit that assumption in the light of the current state-of-the-art. This paper aims to be a step in this direction. We try to make an initial proposal of a feasible UO. We present the scope of the UO, the kinds of its concepts, and the elements that could comprise the specification of each concept. We propose a modular structure for the UO consisting of four levels. We argue that the UO needs a complete set of concept composition operators, and we sketch three of them. We also tackle a few issues related to the feasibility of the UO, which we think that they could be surmountable. Finally, we discuss the desirability of the UO, and we explain why we conjecture that there are already organizations that have the

knowledge and resources needed to develop it, and that might have an interest in its development in the near future [11].

METHODOLOGY

Document clustering has been investigated for use in a number of different diseases of text mining and information retrieval. Initially, Text clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a Text. Text clustering has also been used to automatically generate hierarchical clusters of Texts and then uses these clusters to produce an effective Text classifier for new Texts. Ontology textual data, one of the most important distance measures is Text similarity [12]. Since Text similarity is often determined by word similarity, the semantic relationships between words may affect Text ontology results [13].

The sharing common named entities (NE) among Texts can be a cue for ontology these Texts together. Moreover, the relationships among vocabularies such as synonyms, antonyms, heteronyms, and hyponyms, may also affect the computation of Text similarity. Text ontology has been number of analysis for different Text database model such as HTML document, XML document and SGML document [14]. The existing system only investigated for use in a number of different diseases of text mining but the number of different kinds of Text ontology need information retrieval. So improve the Text ontology method for web mining techniques in this thesis work. The proposed system is developed an application for recommendations of news diseases to the readers of a news portal. The following challenges gave us the motivation to use ontology of the news diseases:

- The number of available diseases was large.
- A large number of diseases were added each day.
- Diseases corresponding to same news were added from different sources.
- The recommendations had to be generated and updated in real time.

The ontology algorithm is reducing and search Texts for recommendations in users have been interest to a few numbers of clusters of Texts. This improved our time efficiency to a great extent and different from sources Texts. The main motivation of this work has been to investigate possibilities for the improvement of the effectiveness of Text ontology by finding out the main reasons of ineffectiveness of the already built algorithms and get their solutions [15].

Initially applied the K-Means and Agglomerative Hierarchical ontology methods on the data and found that the results were not very satisfactory and the main reason for this was the noise in the graph, created for the data. The tried for pre-processing of the graph to remove the extra edges [16].

Heuristic applied for removing the inter cluster edges and then applied the standard graph ontology methods to get much better results. The information tried a completely different approach by first ontology the words of the Texts by using a standard ontology approach and thus reducing the noise and then using this word cluster to cluster the Texts. Their results are found that this also gave better results than the classical K-Means and ontology algorithm methods [17].

COSINE SIMILARITY

In this module, two documents are selected. Then the vector values for two documents are found out. The cosine similarity measure is applied. Then the correlation between two documents is found out using the following formula.

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \left\langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \right\rangle$$

Correlation Formula

For example, the string “I have to go to school” is present in one document. the string “I have to go to temple” is present in other document. Then the data is prepared such that

[i , have , to , go, school , temple] = [1,1,2,1,1,0]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]
Formula: cos = 1*1 + 1*1 + 2*2 + 1*1 + 1*0 + 0*1 / sqrt((1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2) * 1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2)

Lexical Unigram, Bigram, Skip-Gram Match

- Two sentences (from text files) are taken. Contents are fetched into strings. Sentences are split into Words.
- For Unigram, the presence of various unigrams in the sentence i is checked against the sentence j.
- For Bigram, Each bigram in the sentence j is searched for their presence in the corresponding sentence i part.
- A skip gram is any combination of words in the order as they appear in a sentence but allowing gap between word occurrences.
- Here, 1-skip bigram is considered where 1-skip bigram allowing one word gap between words in a sentence as they appear.

Lexical Longest Common Subsequence

- The longest common subsequence of sentence i – sentence j pair is the longest sequence of words that is common to both sentences.
- Lexical_LCS_match= LCS (sentence i, sentence j)/length of unigrams in the sentence j.
- If the value of Lexical_LCS_match is 0.80 or more, i.e., the length of the common words in pair of sentence is greater than the length of the sentence j, then the sentence pair is considered as an entailment pair.

PREPROCESSING DOCUMENT

Add Stem Word Document

In this module, enter the given word and stem word using text box control and click save button stem word saved into the table. The details are saved in 'Stemword' table. The stem word details view on grid view controls.

Add Stop Word Document

In this module, enter the stop word using text box control and click save button stop word saved into the table. The details are saved in 'Stopword' table. The stop word details view on grid view controls.

Add Synonym Word Document

In this module, enter the given word and synonym word using text box control and click save button synonym word saved into the table. The details are saved in 'synonym word' table. The synonym word details view on grid view controls.

Rule Extraction Ontology

- Co-occurrence of medicine names with disease in single paragraph is retrieved during rules extraction.
- Co-occurrence of medicine names, disease and treatment in single paragraph is retrieved during rules extraction.

EXPERIMENTAL RESULTS

In two sets of data. This example looks at the strength of the link between the price of a convenience item and distance from the Contemporary Art Museum in El Raval, Arcelona. Table 4.1 shows Spearman's Rank Correlation Coefficient experimental result for existing system. The table contains word pair, word pair one value [W_1], word pair one rank value [R_1], word pair two value [W_2], word pair two rank value.

Table 4.1 Word Pair Relation

S.NO	Word-Pair	W_1	R_1	W_2	R_2	W-Pair [n]
1	cord-smile	38	2	14	12	13
2	monk-oracle	15	13	10	13	8
3	noon-string	22	8	16	11	13
4	glass-magician	25	7	17	10	13
5	monk-slave	15	13	18	9	8

6	coast-forest	27	5	17	10	15
7	crane-implement	18	11	21	7	11
8	car-automobile	30	3	18	9	17
9	brother-lad	19	10	38	1	19
10	bird-crane	29	4	18	9	17
11	bird-cock	29	4	25	4	23
12	coast-hill	27	5	30	2	22
13	car-journey	44	1	23	5	23
14	implement-tool	21	9	26	3	16
15	boy-lad	26	6	38	1	24
16	forest-graveyard	17	12	9	14	8
17	midday-noon	15	13	22	6	15
18	furnace-stove	17	12	20	8	17
19	magician-wizard	17	12	21	7	17
20	lad-wizard	38	2	21	7	21

Spearman's rank correlation coefficient (r_s) is a reliable and fairly simple method of testing both the strength and direction (positive or negative) of any correlation between two word pair or document.

$r_s = 1- [N \Sigma d^2/n^3-n]$

Where $d^2 = [R_2-R_1]^2$, n= Word Pair Count, N= Total number of word pair

Table 4.2 Spearman's Rank Correlation Coefficient

R_2-R_1	d^2	n	n^3	r_s
10	100	13	2197	0.0842
0	0	8	512	1
3	9	13	2197	0.9175
3	9	13	2197	0.9175
-4	16	8	512	0.3650
5	25	15	3375	0.8511
-4	16	11	1331	0.7575
6	36	17	4913	0.8529
-9	81	19	6859	0.7631
5	25	17	4913	0.8978
0	0	23	12167	1
-3	9	22	10648	0.9830
4	16	23	12167	0.9736
-6	36	16	4096	0.8235
-5	25	24	13824	0.9637
2	4	8	512	0.8412
-7	49	15	3375	0.7083
-4	16	17	4913	0.9346
-5	25	17	4913	0.8978
5	25	21	9261	0.9458
Spearman's Rank Correlation Coefficient				0.8239

The above Table 4.2 shows the spearman's rank correlation coefficient between two word pair for existing system. The table contains difference between rank values, square of rank values, word pair count and cube value of word pair values and spearman's rank correlation coefficient for each

word pair (r_s) details are shown. The over all word pair spearman's Rank Correlation Coefficient value is 0.8239.

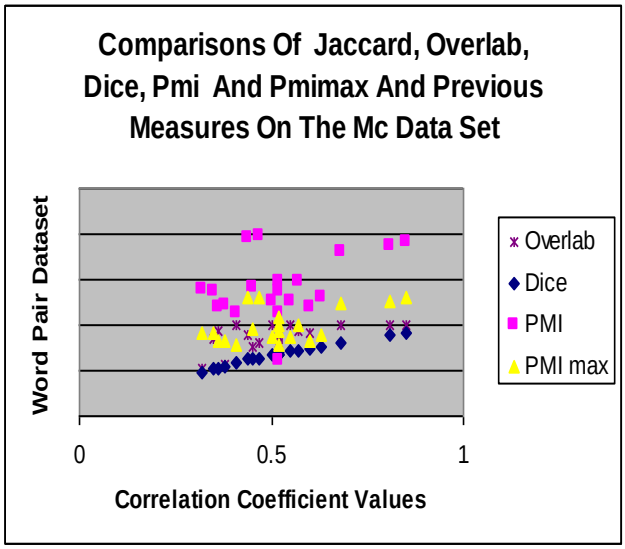

Fig 4.1 Comparisons of Values Mc Data Set

The above Fig 4.1 shows experimental result for existing system analysis. The table contains word pair, word jaccard value, word overlab values, word dice values, word PMI values and its PMI max values details are shown. The word pair count details are measure the cor-relation coefficient score value in each word pair using precision and recall measure. The over all word pair coefficient values are jaccard value is 0.584, overlab value is 0.875, dice values is 0.695, PMI values are 2.846 and PMI max values are 4.025.

CONCLUSION

This paper demonstrated how to construct various text and word constraints and apply them to the constrained co-ontology process. A novel constrained co ontology approach is proposed that automatically incorporates various word and document constraints into information-theoretic co- ontology. It demonstrates the effectiveness of the proposed method for clustering textual documents. There are several directions for future research. The current investigation of unsupervised constraints is still preliminary. Furthermore, the algorithm consistently outperformed all the tested constrained clustering and co-clustering methods under different conditions. The enhanced cosine similarity approach results in better ontology process. The future enhancements can be made for documents of different languages. Investigation for better text features that can be automatically derived by using natural language processing or information extraction tools can be made.

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, 31(3), 264-323, 2014.

[2] M.Fazio Celesti I.S.Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2011, 269-274.

[3] Javier Medina And Macarena Esp., S.Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2013, 89-98.

[4] H.Cho, I.S. Dhillon, Y. Guan, and S.Sra, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," Proc. Fourth SIAM Int'l Conf. Data. Mining (SDM), 2014.

[5] Cristina Soguero-Ruiz And Kristian Hindberg and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," Proc. SIAM Int'l Conf. Data. Mining (SDM), 1-12, 2018.

[6] Didier Dubois And WeiruLiu, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co- Clustering," IEEE Trans. Knowledge and Data Eng, 22(10), 2012, 1459-1474.

[7] Diogo Machado, Tiago Paiva and J.-F.Boulicaut, "Constrained Co-Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM), 2018, 25-36.

[8] Behzad Golshan And Alon Halevy, J. Ghosh, S. Merugu, and D.S.Modha, "A Generalized Maximum Entropy Approach to Bregman Co- Clustering and Matrix Approximation," J. Machine Learning Research, 8, 2017, 1919-1986.

[9] Maria Ganzha And Marcin Paprzyck, and T.M. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, 39(2/3), 2016, 103-134.

[10] Antoni Olivé, K. Wagstaff, C. Cardie, S. Rogers, and S. Schro ¨dl, "Constrained K- Means Clustering with Background Knowledge," Proc. 18th Int'l Conf. Machine Learning (ICML), 2011. 577-584.

[11] I.S.Dhillon and D.S.Modha. Concept decompositions for large sparse text data using clustering. Machine Learning,. Also appears as IBM Research Report RJ 10147, 42(1):143–175, 2001, 2010.