www.irbjonline.in

International Research Based Journal

Vol(4)-Issue(2), July,29, 2016,ISSN 2348-1943

Review of Machine Learning Based NLP For Conceptual Similarities In Trademark

Manpreet Kaur*
Department of Computer Science & Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib

Abstract—This paper provides a review of Data Mining Techniques used in Different Operations of Natural Language Processing such as Semantic Analysis. The paper presents the review of different techniques used for similarity measures and their contrasting factors.

Keywords—Data Mining, Semantic Analysis, Natural Language Processing, Similarity Measures, Review

I. INTRODUCTION

Information mining is an increasingly important department of laptop science that examines knowledge with the intention to in finding and describe patterns. Due to the fact we reside in a world the place we can be overwhelmed with information, it is important that we discover approaches to categorise this input, to search out the understanding we want, to illuminate constructions, and to be equipped to draw conclusions. Data mining is a very realistic discipline with many purposes in trade, science, and government, similar to exact advertising, web evaluation, sickness prognosis and effect prediction, climate forecasting, credit score danger and mortgage approval, purchaser relationship modeling, fraud detection, and terrorism risk detection. It's centered on methods a couple of fields, but in most cases computer studying, facts, databases, and knowledge visualization.

Data mining is a method of exploration and evaluation, by automatic or semiautomatic manner, of ancient knowledge to be able to become aware of patterns and ideas, which can be utilized in a while new information for predictions and forecasting. With information mining, you deduce some hidden knowledge via examining, or training, the information. The unit of examination is referred to as a case, which can also be interpreted as one appearance of an entity, or a row, in a desk. The talents is patterns and rules. In the approach, you utilize attributes of a case, which might be referred to as variables in information mining terminology. For better working out, that you would be able to examine information mining to on-line analytical processing (olap), which is a model-driven analysis the place you build the model prematurely. Information mining is a knowledge-pushed evaluation, where you seek for the model. You compare the information with information mining algorithms. Special knowledge Mining methods are their

A. Association principles

The algorithm used for market basket evaluation, this defines an item set as a combination of items in a single transaction. It then scans the info and counts the quantity of instances the itemsets appear together in transactions. Market basket evaluation is valuable to detect go-selling possibilities.

B. Clustering

This groups cases from a dataset into clusters containing similar characteristics. You should utilize the clustering system to team your shoppers in your crm utility to find distinguishable corporations of your buyers. Moreover, you should utilize it for finding anomalies for your data. If a case does not fit well to any cluster, it is type of an exception. For example, this perhaps a fraudulent transaction.

C. Naïve Bayes

This calculates possibilities for each viable state of the enter attribute for each single state of predictable variable. These probabilities predict the goal attribute based on the identified enter attributes of latest circumstances. The naïve bayes algorithm is fairly simple; it builds the items swiftly. Consequently, it is very suitable as a starting factor to your predictive analytics project.

D. Decision trees

the most wellknown dm algorithm, it predicts discrete and continuous variables. It uses the discrete enter variables to split the tree into nodes in any such means that each and every node is more pure in terms of goal variable, i.E. Each and every split leads to nodes the place a single state of a goal variable is represented better than other states.

E. Regression trees

For steady predictable variables, you get a piecemeal a couple of linear regression formulation with a separate formula in every node of a tree. Discrete input variables are used to separate the tree into nodes. A tree that predicts steady variables is a regression tree. Use regression timber for estimation of a steady variable; for illustration, a financial institution might use this method to estimate the family sales for a loan applicant.

II. REVIEWS OF EXISTING DATA MINING TECHNIQUE

Anuar et.Al in [1] trademarks are signs of excessive reputational value. As a result, they require safeguard. This

paper reviews conceptual similarities between logos, which occurs when two or more trademarks evoke identical or analogous semantic content material. This paper advances the state-of-the-art by using proposing a computational approach based on semantics that can be used to evaluate logos for conceptual similarity. A trademark retrieval algorithm is developed that employs average language processing approaches and an external skills source within the type of a lexical ontology. The quest and indexing manner developed makes use of similarity distance, which is derived utilising Tversky's concept of similarity. The proposed retrieval algorithm is validated utilising two resources: a trademark database of 1400 disputed instances and a database of 378,943 organization names. The accuracy of the algorithm is estimated using measures from two extraordinary domains: the R-precision ranking, which is almost always utilized in understanding retrieval and human judgment/collective human opinion, which is used in human-computing device techniques.

Zhou et.Al in [2] proposed a novel procedure to address the semantic extension inside the framework of language modeling. Their procedure extracts explicit subject signatures from documents and then statistically maps them into single-word points. The incorporation of semantic knowledge then reduces to the smoothing of unigram language units making use of semantic potential. The dragon toolkit displays our method and its effectiveness are established by using three duties, textual content retrieval, textual content classification, and textual content clustering.

Liu et.Al in [3] his paper, a novel semantic retrieval approach for faraway sensing photographs utilizing association rules mining is presented. Unlike the common content material-founded snapshot retrieval approaches, organization rules are mined and used to express the semantic understanding of snap shots rather of low-degree aspects. The long-established picture is to begin with segmented into many objects; after which the categorized organization principles between the homes of objects are mined and changed to semantic understanding through semantic annotation approach; sooner or later the semantic retrieval is executed utilizing the similarity dimension method. The experimental outcome point out that the proposed approach can provide better retrieval efficiency than the existing content material-established photograph retrieval approaches.

Costachioiu et.Al in [4] remote sensing platforms accumulate significant amounts of information daily. Accordingly, tremendous archives of knowledge had been created. To be able to provide access to this knowledge effective search and retrieval methods need to be developed, comparable to photo information mining techniques. On this paper we present a framework for an photograph information mining system utilising the Latent Dirichlet Allocation textual content-mining algorithm to provide a excessive-level semantic mannequin of information, the search being performed within the LDA mannequin house.

Hao et.Al in [5] this paper proposes a knowledge retrieval model combining talents search with information mining technologies. On this model, data mining is integrated into the entire retrieval approach of query optimizing, browsing, results inspecting, and assets developing. It realizes skills retrieval through more than a few approaches, distinctive phases, and multi-modes, and vastly improves skills retrieval stage and efficiency. Moreover, they explore related expertise retrieval approaches and algorithms including association evaluation-situated suggestion retrieval and inductive finding out-headquartered classification retrieval, and validate them experimentally.

Shyu et.Al in [6] this paper, they gift our lately developed content material-founded multimodal Geospatial know-how Retrieval and Indexing process (GeoIRIS) which involves computerized function extraction, visual content material mining from colossal-scale picture databases, and high-dimensional database indexing for speedy retrieval. Utilising these underpinnings, they have developed strategies for intricate queries that merge understanding from heterogeneous geospatial databases, retrievals of objects founded on form and visual characteristics, analysis of multiobject relationships for the retrieval of objects in precise spatial configurations, and semantic items to hyperlink low-level photo facets with high-level visible descriptors. GeoIRIS brings this various set of applied sciences collectively into a coherent system with an purpose of enabling snapshot analysts to more speedily establish primary imagery. GeoIRIS is ready to reply analysts' questions in seconds, similar to "given a query snapshot, exhibit me database satellite pix that have identical objects and spatial relationship which can be within a specific radius of a landmark."

Elia et.Al in [7] this paper is excited about Catalog a, a application bundle founded on Lexicon-Grammar theoretical and practical analytical framework and embedding a ling ware module constructed on compressed terminological digital dictionaries. They're going to here exhibit how Catalog a can be used to reap efficient knowledge mining and knowledge retrieval by the use of lexical ontology associated to terminology-based automatic textual evaluation. Also, they will show how accurate knowledge compression is vital to build effective textual analysis program. Hence, they are going to right here discuss the construction and functioning of a program for semantic-headquartered terminological data mining, wherein a imperative role is played by means of Italian simple and compound-phrase digital dictionaries. Lexicon-Grammar is likely one of the most profitable and steady approaches for common language formalization and computerized textual analysis it was installed by means of French linguist Maurice Gross for the period of the '60s, and subsequently developed for and applied to Italian with the aid of Annibale Elia, Emilio D'Agostino and Maurizio Martin Elli. Essentially, Lexicon-Grammar establishes morph syntactic and statistical units of analytic ideas to learn and parse big textual corpora. The analytical method here described will prove itself proper for any type of digitalized textual content, and will symbolize a principal help for the constructing and implementing of Semantic web (SW) interactive structures.

Zhuang et.Al in [8] this paper, they suggest a procedure of transductive finding out to mine the semantic correlations

among media objects of exclusive modalities so that to achieve the go-media retrieval. Pass-media retrieval is a brand new form of looking science through which the question examples and the again outcome may also be of exclusive modalities, e.G., to question photographs by means of an example of audio. First, in keeping with the media objects points and their co-existence information, we construct a uniform cross-media correlation graph, where media objects of one of a kind modalities are represented uniformly. To participate in the pass-media retrieval, a confident score is assigned to the query illustration; the ranking spreads alongside the graph and media objects of goal modality or MMDs with the perfect scores are lower back. To lift the retrieval performance, we also advise distinctive techniques of lengthy-time period and quick-term relevance feedback to mine the information contained in the constructive and poor examples.

Huang et.Al in [9] this paper, they purpose to make use of Linked knowledge to generate semantic annotations for widely wide-spread patterns extracted from textual records. First, they extract semantic relations from textual documents and merge them into a set of semantic graphs. Then, they practice a usual subgraph discovery algorithm on the set of graphs to generate regular patterns. In the end, they annotate the discovered patterns utilizing Linked data. Their process can also be applied in such domains as terrorist network evaluation and organic network evaluation. The efficacy of their strategy is established through an empirical scan that discovers and validates relationships between political figures from significant number of reports on the web.

Wang et.Al in [10] they proposed a novel semantic video retrieval system that integrates web snapshot annotation and concept matching function to bridge photos, ideas and movies. For web photograph annotation, they exploit textual and visible understanding within the web photo to gain strong photograph annotation. For inspiration matching function, they determine the concept members of the family by calculating the similarity between two concepts by way of word net. On the basis of web photograph annotation and thought matching operate, the proposed procedure reaches the pursuits of usability and intelligence on semantic video retrieval. The experimental outcome disclose that their proposed system can effectively seize the person's intention between picture concepts and video standards for semantic video retrieval.

III. Euclidian Similarity

Euclidean distance is a regular metric for geometrical problems. It's the average distance between two features and will also be easily measured with a ruler in two- or three-dimensional house. Euclidean distance is generally used in clustering problems, including clustering textual content. It satisfies all the above four conditions and thus is a true metric. It's also the default distance measure used with the K-approach algorithm.

IV. Cosine Similarity

When files are represented as time period vectors, the similarity of two documents corresponds to the correlation between the vectors. That is quantified as the cosine of the angle between vectors, that's, the so-called cosine similarity. Cosine similarity is one of the most preferred similarity measure utilized to textual content files, comparable to in countless understanding retrieval functions and clustering too.

V. Analysis and Discussions

Within the present method NLP together with external capabilities supply headquartered on lexical ontology is employed. The looking and indexing process is developed making use of similarity distance (Tversky's theory of similarity). This synopsis proposes a conceptual model of trademark retrieval based on conceptual similarity. The mannequin employs typical language processing systems, competencies sources and a lexical ontology to compute conceptual similarity between textual logos. The proposed mannequin improves on existing trademark search units through offering a means of refining the search to conceptually associated logos.

Future work involves a user gain knowledge of to assess the effectiveness of this mannequin, research on the phonetic similarity of trademark assessment, as good as integrating the developed instruments into a trademark retrieval approach situated on visible, conceptual and phonetic similarity. The goal is to provide more correct retrieval and a better platform for examiners conducting trademark analysis.

To increase the entire method, external capabilities source can also be developed into a laptop finding out process which will also be informed to generate required metric. The shopping and indexing algorithm may also be researched upon and up to date to make the process of shopping and indexing turbo.

Data	Euclidean	Cosine	Jaccard
Classic	0.56	0.85	**0.98**
Hitech	0.29	**0.54**	0.51
Re0	0.53	**0.78**	0.75
Wap	0.32	0.62	**0.63**

Table 1: Comparison of Different Similarity Algorithms on Different Test Dataset

VI. Conclusion

In this review paper, we have looked at the existing techniques of data mining and their different enhancement. A particular article is looked over for further analysis and its problem is defined. Finally the solution is proposed which can be the work for future.

References

[1] Anuar, Fatahiyah Mohd, Rossitza Setchi, and Yu-Kun Lai. "Semantic Retrieval of Trademarks Based on Conceptual Similarity."

[2] Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. "Dragon toolkit: incorporating auto-learned semantic knowledge into large-scale text retrieval and mining." Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on. Vol. 2. IEEE, 2007.

[3] Liu, Jun, and Shuguang Liu. "Semantic retrieval for remote sensing images using association rules mining." Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International. IEEE, 2015.

[4] Costachioiu, Teodor, et al. "A semantic framework for data retrieval in large remote sensing databases." Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, 2012.

[5] Hao, Yan, and Yu-feng Zhang. "Research on knowledge retrieval by leveraging data mining techniques." Future Information Technology and Management Engineering (FITME), 2010 International Conference on. Vol. 1. IEEE, 2010.

[6] Shyu, Chi-Ren, et al. "GeoIRIS: Geospatial information retrieval and indexing system—Content mining, semantics modeling, and complex queries." Geoscience and Remote Sensing, IEEE Transactions on 45.4 (2007): 839-852.

[7] Elia, Annibale, Mario Monteleone, and Alberto Postiglione. "Cataloga: A Software for Semantic-Based Terminological Data Mining." Data Compression, Communications and Processing (CCP), 2011 First International Conference on. IEEE, 2011.

[8] Zhuang, Yue-Ting, Yi Yang, and Fei Wu. "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval." Multimedia, IEEE Transactions on 10.2 (2008): 221-229.

[9] Huang, Zhaohui, et al. "Semantic text mining with linked data." INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. IEEE, 2009.

[10] Wang, Bo-Wen, et al. "Semantic Video Retrieval by Integrating Concept-and Content-Aware Mining." Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on. IEEE, 2011.

[11] Wang, Yong, and Ya-Wei Zhao. "Transplantation of Data Mining Algorithms to Cloud Computing Platform When Dealing Big Data." Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on. IEEE, 2014.

[12] Zhang, Lining, Lipo Wang, and Weisi Lin. "A semantic subspace learning method to exploit relevance feedback log data for image retrieval." Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on. IEEE, 2013.

[13] Kiu, Ching-Chieh, and Chien-Sing Lee. "Learning objects reusability and retrieval through ontological sharing: A hybrid unsupervised data mining approach." Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on. IEEE, 2007.

[14] Costachioiu, Teodor, et al. "A semantic framework for data retrieval in large remote sensing databases." Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, 2012.

[15] Hao, Yan, and Yu-feng Zhang. "Research on knowledge retrieval by leveraging data mining techniques." Future Information Technology and Management Engineering (FITME), 2010 International Conference on. Vol. 1. IEEE, 2010.