Understand Short Text by Harvesting and Analyzing Semantic Knowledge

S. Balu[1]
[1] Assistant Professor CSE,
K.S.Rangasamy College of Technology,
Tiruchengode-637215

R. Karthika[2], S. Karthikeyan[3], B. Mohammed Yacub[4]
[2,3,4] IV B.E CSE,
K.S.Rangasamy College of Technology,
Tiruchengode-637215.

Abstract: **Trademarks have a sign on high reputational value. So, they require protection. Conceptual similarities occurs when two or more trademarks come out with identical or analogous semantic content. Here the state-of-the-art by proposing a computational approach based on semantics that can be used to compare trademarks for conceptual similarity. A trademark retrieval algorithm is developed that employs natural language processing techniques and an external knowledge source in the form of a lexical ontology. The search and indexing technique developed uses similarity distance, which is derived using Tversky's theory of similarity. The accuracy of the algorithm is estimated using measures from two different domains: the R precision score, which is commonly used in information retrieval and human judgment/collective human opinion, which is used in human machine systems. The proposed algorithm employs Web Service Model Ontology search with Support Vector Machine (SVM) techniques and the word similarity distance method, which was derived from the WordNet ontology, together with a new trademark comparison measure. WordNet is employed in this algorithm due to its lexical relationships, which mirror human semantic organization, and because it has also been proven successful in many previously developed works. The trademark comparison measure is derived from the Tversky contrast model a well-known model in theory of similarity search.**

I.INTRODUCTION

A trademark is used to identify the brand owner of a particular product or service. Trademarks can be licensed to others; for example, Bully land obtained a license to produce Smurf figurines. The trademark owner may pursue legal action against trademark invasion. Most of the

countries recognize common law trademark rights, which means action can be taken to protect an unregistered trademark if it is in use.

Trademarks, as determined by the European Office of Harmonization in the Internal Market (OHIM). They do insignificant intellectual property (IP) goods that permit well or service to be well validated to clients. Each year many trademarks registered and used that outlet. Trademarks are exclusive words or figures with advance reputational significance, used in commerce to comparison between products and services. They allow products or tasks to be goods tenable and compared by traders. Searching for conceptually similar trademarks is a text retrieval problem. However, traditional text retrieval systems based on keywords are not capable of retrieving conceptually related text. This limitation motivates research into semantic technology, which addresses this problem by using additional knowledge sources.

Invasion may occur when invasion party, uses a trademark which is similar to a trademark owned by another party, in relation to products or services which are identical or similar to the products or services which the registration covers having existence trademark look for systems as a general rule use text-based acts to get back technology. These searches look for trademark that matches some or all words in a question line wording. As indicated in their latest printing on trademark knowledge-bases and look for systems. Two trademarks are necessary not same to make an infringement. The conceptual different of text files that part of same domain, utilization same notations, or demonstration same consideration has been used broadly.

II.EXISTING SYSTEM

Existing Trademark Search Systems: The underlying technology embedded in existing trademark search systems is primarily based on text-based retrieval. Such systems search for trademarks that match some or all words in a string text query. In a recently launched search system, the OHIM provides an option that allows users to search for trademarks in different languages.
This newly upgraded system also provides advanced search options that offer three search types: word prefix, full phrase, and exact match. The word prefix mode returns trademarks with a prefix that matches the query. The full phrase mode finds trademarks with terms that include the query input, and the exact match returns trademarks that match the query input exactly. In the United Kingdom, the Intellectual Property Office (IPO) provides search options that are similar to the OHIM search service, with an additional option that searches for similar query strings.

The system employs an approximate string-matching technique, along with several pre-defined criteria, such as the number of similar and dissimilar characters in the words and the word lengths, to retrieve similar trademarks.

The drawbacks in existing system are:

- Approximate string matching is possible.
- It requires only one substitution operation.
- Coordinate matching is not possible.
- The fewer operations required to make the strings identical, the more similar they are. The most common

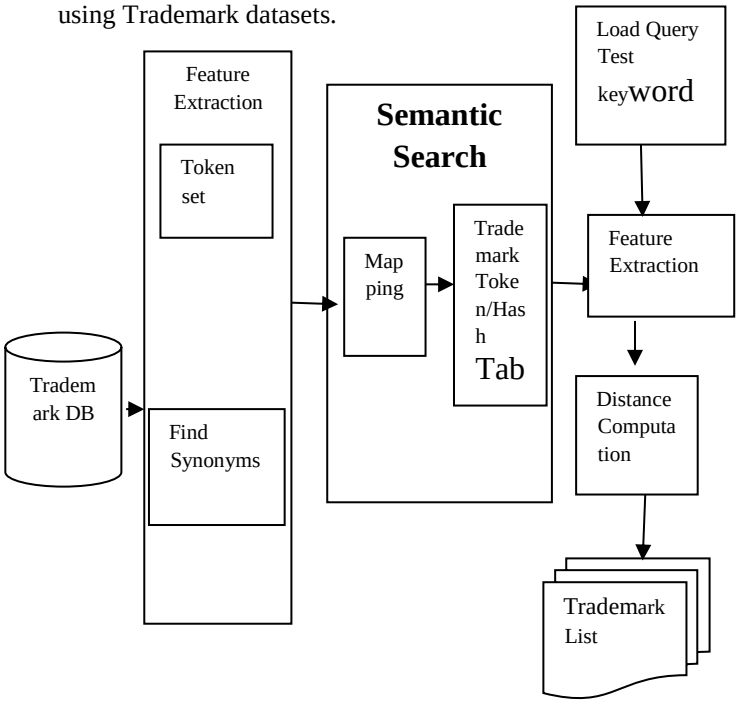
retrieval method employed in the existing trademark search system.

- Simple search tasks may not work well with traditional information systems.
- It can only Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data.
- Risk in decision making

environment.

III.PROPOSED SYSTEM

Relational Keyword search based on WSMO (Web Service Model Ontology) based K-SVM Classification algorithms have been studied for decades, and the literature on the subject is huge. Therefore, it is decided to choose a WSMO K-SVM as representative algorithm in order to show the potential of the proposed approach, namely: the partitioned the cluster semantic word extraction algorithm known as K-Support Vector Machine. Trademark comparison based on conceptual similarities. This work extends the conceptual model by developing and evaluating a semantic algorithm for trademark retrieval based on conceptual similarity. The proposed algorithm employs NLP techniques and the word similarity distance method, which was derived from the WordNet ontology, together with a new trademark comparison measure. WordNet is employed in this algorithm due to its lexical relationships, which mirror human semantic organization, and because it has also been proven successful in many previously developed works. These algorithms were run with different combinations of their parameters, resulting in sixteen different algorithmic instantiations. Thus, as a contribution of this work, compare the relative performances on the studied application domain using Trademark datasets.



In order to make the comparative analysis of the algorithms more realistic, two relative validity indexes related keyword result to the user that have been used to estimate the number of clusters automatically from data.

Each trademark is represented by two kinds of features i.e., the trademark tokens and the synonym list. The feature extraction step begins with a spelling correction process that corrects any spelling mistakes using a spellchecker. Then, frequent words (i.e., 'no,' 'and,' 'the,' etc.) are removed, and the trademarked words are extracted in the form of tokens. The trademark tokens extracted here are sets of English root words. The second feature is defined as the synonym set of the tokens and is extracted from the WordNet database. The synonym set, as defined in the context of this algorithm, includes the synonyms, the direct hypernyms, and the direct hyponyms of the corresponding tokens.

To reduce computational time during the search process, the features are indexed using a hashing technique. The mapping function is designed so that the trademark similarity distance computation is performed only on the set of trademarks that consist of at least one of the terms in fs, i.e., the synonyms set belonging to the trademark query.

The distance computation is based on the similarity concept introduced in Synonym vector learning theory. It defines the similarity between two objects as a function of unique and shared information about the object. Based on this idea, the similarity equation between a trademark query and the token set and the synonyms set of the query, respectively and word_sim is the word similarity measure computation employed in this algorithm. The trademark distance computation is then performed between the trademarks using the trademark similarity equation. A trademark retrieval system using the proposed retrieval algorithm is developed, and the algorithm is tested on two databases. Two experiments are then conducted to evaluate the performance of the proposed algorithm. The first evaluation is conducted using an information retrieval measure (i.e., R-precision score), and the second evaluation is conducted through an open call task (i.e., crowdsourcing). The result will be displayed as tabular format.

The advantages of proposed system are:

- Most importantly, it is observed that Classification algorithms indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner's job.
- This method in applications shows that it has the potential to speed up the computer inspection process.
- Better support with Classification group of data's.
- Highly efficient.
- Provide good result.

IV.ALGORITHM USED

Support Vector Machine (SVM) is one of the most attractive and potent classification algorithms and has been successful in recent times. SVM dedicates to find the excellent separating hyperplane between two classes, thus can give excellent generalization ability for it. In order to find the excellent hyperplane, the labelled records as the training set. However, the separating hyperplane is only determined by a few crucial samples (Support Vectors, SVs), no necessity to train SVM model on the whole training set. This paper presents a novel approach based on clustering algorithm, in which only a small subset was selected from the original training set to act as the final training set. The algorithm used here works to select the most informative samples using K-means clustering algorithm, and the SVM classifier is built through training on those selected samples. Experiments show that this approach greatly reduces the scale of training set, thus effectively saves the training and predicting time of SVM, and at the same time guarantees the generalization performance.

V.CONCLUSION

The work presented in this work was motivated by the realization that despite the large number of invasion cases based on conceptual similarity, traditional knowledge recover systems do not handle this particular issue well. It is also motivated by the understanding that trademark analogy, one of the factors that contributes to the likelihood of confusion, may be linked to the semantics of trademarks, i.e., their lexical meanings. This work contributes to the state-of-the-art by proposing a semantic algorithm to compare trademarks in terms of conceptual similarity. The algorithm brings forward an entirely new similarity contrast approach in the domain of trademark recover. It utilizes NLP techniques, together with an external knowledge source in the form of a lexical ontology. The evaluation using both knowledge recover measures and human judgment shows a significant improvement because the algorithm provides better results than the traditional baseline technique.

The algorithm is not limited to the use of a specific word measure. This advantage provides the flexibility to choose any word measure suitable for particular applications or requirements. The results from the experiment performed in this work confirm that the comparison of trademarks based on their conceptual similarity can be conducted using linguistic sources. Future work to improve the accuracy of the proposed semantic algorithm should include a study comparing the use of various lexical resources. In addition, the authors are working on extending the current approach to include retrieving trademarks with phonetic similarities and integrating their previous work on visual similarity with their new algorithms

VI.REFERENCES

[1] B. Furlan, V. Batanovic, and B. Nikolic,"Semantic similarity of short texts in languages with a deficient natural language processing support," Decis. Support Syst., vol. 55, no. 3, pp. 710–719, 2013.

[2] F. M. Anuar, R. Setchi, and Y. K. Lai, "A conceptual model of trademark retrieval based on conceptual similarity," in Proc. 17th Int. Conf. Knowl.Based Intell. Inf. Eng. Syst., Kitakyushu, Japan, 2013, pp. 450–459

[3] Latika Pinjarkar, Manisha Sharma, Content Based Image Retrieval for TrademarkRegistration:A SurveyInternational Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013 ISSN (Print) : 2319-5940 ISSN (Online) : 2278-1021

[4] Zhenhai Wang, Kicheon Hong , A Novel Approach for Trademark Image Retrieval by Combining Global Features and Local Features, Journal of Computational Information Systems 8(4) : 1633–1640,2012

[5] J. Oliva, J. I. Serrano, M. D. del Castillo, and A. Iglesias, "SyMSS: A syntax-based measure for short-text semantic similarity," Data Knowl. Eng., vol. 70, no. 4, pp. 390–405, 2011

[6] Tatsuaki Iwanaga, Hiromitsu Hama, Takashi Toriu, Pyke Tin and Thi Thi Zin, A Modified Histogram Approach to Trademark Image Retrieval, IJCSNS International Journal of Computer Science and Network Security, 11(4):56-62, 2011

[7] Rusiñol Marçal , Aldavert David , Dimosthenis Karatzas , Interactive Trademark Image Retrieval by Fusing Semantic and Visual Content , in Advances in Information Retrieval ,Springer :1-12, 2011

[8] H. Qi, K. Q. Li, Y. M. Shen, and W. Y. Qu, "An effective solution for trademark image retrieval by combining shape description and feature matching," Pattern Recognit., vol. 43, no. 6, pp. 2017–2027, 2010.

[9] J. Schietse, J. P. Eakins, and R. C. Veltkamp, "Practice and challenges in trademark image retrieval," in Proc. 6th ACM Int. Conf. Image Video Retrieval, Amsterdam, The Netherlands, 2007, pp. 518–524.