



A SEMANTIC IMPROVED APPROXIMATE MEMBERSHIP EXTRACTION WITH EFFECTIVE SIMILARITY SEARCH TECHNIQUES

S.Balan¹, Dr. P.Ponmuthuramalingam²

¹Ph.D. Research Scholar, ²Associate Professor & Head

Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Taminadu, India.

Abstract: Dictionary-based Entity Recognition Problem is mainly used to find entities which are present in the documents. Those entities are locations, person names, company, etc. The aim of large Dictionary-based Entity checking is to find all possible substrings from documents which matches any reference in the given dictionary. Existing research work (Approximate Membership Localization (AML) used by extending Approximate Membership Extraction (AME). AML target is to find non overlapped substrings. The reason for locating non overlapped substrings is to get an approximation of true matched substring by without generating overlapped redundancies. In order to give efficiency of existing research, a proposed semantic based substring matching of documents developed. This returns the semantically obtained result documents for a queried substring. Then the similarity among those returned documents measured by two similarity measurement methods such as Inverted file –base similarity and Cluster based similarity. Next the most important term with respect to queried substring in the document found. These are done by using Term Frequency-Inverse Document Frequency.

Keywords: Semantic based Information retrieval, Inverted-file Based Similarity, Cluster Based Similarity, Term Frequency-Inverse Document Frequency.

I. Introduction

The basic information retrieval process categorized into three steps named as a user, query and information retrieval system. It is a technology providing the information based on the user request. The retrieval system responds the relevant document based on the query. It does not search for specific data and return the real documents but with the information. There are two aspects, one is processing the collection and another is processing the queries (searching). The model consists of documents, queries and matching function which compares queries to documents. It's classified as classical model, non-classical model and alternative model. The classical information retrieval model based on mathematical knowledge, such as boolean, vector and probabilistic models. A non classical model based on principles such as information logic, situation and interaction model. An alternative model based on specific technique such as cluster model and fuzzy model.

In retrieval process, the actual text of the document is not queried. Instead, document representation used to do retrieval process. For that a set of index term or keywords are used to extract the information. The World Wide Web has a huge amount of data that has unstructured and understandable. The semantic web aims to display the information closer to the user. It defines different levels of abstraction to the web. The various applications of data mining techniques to the content, structure and usage of web resources is known as web mining. The main difference is web mining can learn implicit structures and semantic web mining offers to add structure of the web. This paper categorized into six sections. Section-1 contain the introduction to information retrieval and semantic web, Section-2 contain the literature survey, Section-3 contain Existing methods and Drawbacks, Section-4 contain proposed methods and its advantages, Section-5 contain Result and analysis Section-6 includes conclusion while references mentioned in the last section.

II. Literature Survey

Named entity recognition (NER) aims at finding named entities in unstructured text. It is a task of information extraction and integration, and serves many applications, including identifying geographical locations for geotagging [3], identifying gene and protein names from MEDLINE abstracts for text mining [9], identifying names and their categories to improve Web search [8]. In search engine integrated approach the queried data extracted by using xml and RDF formats. According to web search result the most relevant entity result finds in a structured

database is high [1]. Efficient exact similarity joins given two inputs as collection of sets and find the similar sets and produce the efficiency is high it on string based similarity search [3].

A large collection of information tested based on the sparse vector data in a high dimensional space to find all pairs of vectors whose similarity score is above a given threshold to find the all pairs of similar. To improve the performance they develop a new method using the inverted base index and meet the threshold value is high [4]. The web has a huge amount of information in that a particular information need's based on text search various methods are available. The fast similarity search method is for time sensitive applications, to improve the data quality a new algorithm as proposed with an edit distance model of string similarity [5].

Named Entity Recognition is the subject that determines the relationship between two entities. In recent years, they proposed different algorithms to solve this problem and produce the efficiency high in different fields [6]. A text is given as input to find the string based on tokens in a large dictionary. There are various approaches for best matches with queried substring with its similarity to string available in dictionary [7]. Approximate Membership Extraction is provided in a given document true matched substrings are process and find the efficiency of a string is low. To overcome that a new algorithm named as P-Prune that prunes a redundant match substring based on boundary constraint and interval pruning before generating them [10].

III. Existing Methods

To avoid redundancy from a large set of paired results constraints are in Appropriate Membership Extraction and named as boundary constraint and Non-overlapped constraint. To identify unqualified redundant matches in P-Prune algorithm and compared the efficiency over AME. The relationship between web search results and structured entity database techniques has little space and the execution time is low. Using entity extraction there are various methods involved to find quality of search and each entity ranked. Compared with different search methods and produce the accuracy of the entity.

This research aims study of existing entity extraction methods. The two components of extraction methods named as Look-up driven extraction and entity classification method. This method shows the recall and precision based on context, document, frequency features is 54% to 97%. To compute exact set similarity joins using jaccard similarity of two sets the weighted performance value id and similarity results [2]. In many applications a queried text in a search engine is the problem to find the exact information to solve this basic, inverted index approach technique to solve and cut the time space.

A. Drawbacks

- AML provides full text search but the semantic of queries substring may know.
- Information retrieval related to queried text may know.
- Full document search needs to extract the entity information.
- It takes time to retrieve a document and redundancy is high.

IV. Proposed Methods

The proposed research carries implementing a semantic technique for the queried text and returns the semantically obtained result documents for a queried substring. The similarity among those returned documents measured by two similarity measurement methods such as Inverted file –base similarity and Cluster based similarity. The most important term with respect to queried substring in the document is found .This may done by using term frequency and inverse document frequency.

To overcome the existing methods a new algorithm for format style, top most ranking algorithm and text extraction based on images used. In web, each information stored in different format style. To find a text in list or paragraph format style needed. According to the page rank method the highest ranking clustered page results displayed. Semantic relation is to extract the text from the image. The various similarities of these selected documents are found by using two similarity measurement methods are as follows:

A. Inverted-File-Based Similarity Search

It is most common for searching the structured text, namely denoted as vocabulary (set of all words in text) and occurrences (each occurrence of the word in the text). The main aim of this similarity search is to find query elements, retrieval and manipulation. Text search is a technology, it provides a variety of information based on the text in a computer to produce a quick result a good operating system performs to find the file in the computer. Inverted file based similarity classified in two ways, one is a keyword searching and another one is indexing.

B. Cluster Based Similarity Search

The main idea of the cluster based similarity search is cluster the document based on the relevant user query. Documents that have same cluster behave different similarity based on the information. The main use of clustering is to speed up the search based on different methods such as cluster that are close to another cluster value. Clustering

categorized in two ways such as online and offline. While searching in online the efficiency problem is high and in offline various clustering techniques measured to find the relevant document and produce the result based on the term word in the given document.

C. Term Frequency–Inverse Document Frequency

The main purpose of inverse document frequency is to measure word importance. The web has a huge amount of information, there are a lot of words have different meaning. To show a particular word information retrieval method using classification or clustering techniques. Each and every document has a particular keyword based on that keyword word frequency measures and displayed. This concept shows in information retrieval and text mining. All terms are important to get access to relevant information on a query. For example group of documents in the bike industry, in this the term bike is the keyword in every document. To term frequency is identified by using the formula $\log(M/m_i)$. M represents the total number of documents, m_i represents the number of documents containing the term i and (m_i/M) represents the fraction of documents.

D. Advantages

- The semantic meaning of the results of queries substring documents retrieve and display
- Most viewed page retrieved by clustering
- Most important term of concept extracted and identified.
- Automatically detect high accuracy information to improve precision.
- The time complexity is less.

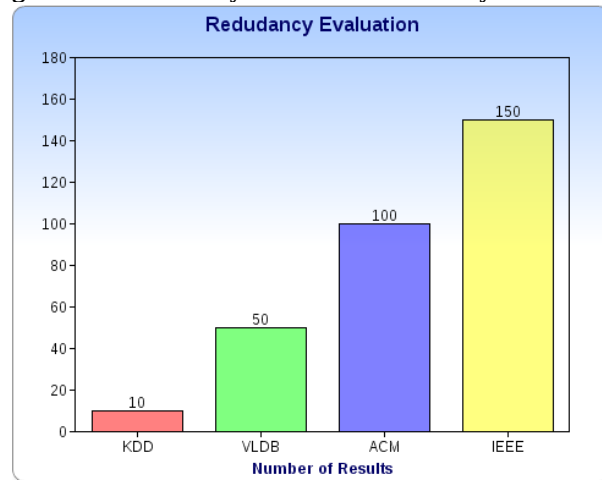
V. Result & Analysis

This survey is mainly focused on Dictionary based entity extraction. There are various methods to extract the collection of entities from the document and its aims for a new method to extract the entity from various documents using two semantic similarity search methods and term frequency inverse document frequency. The above existing methods comparison of entity extraction of a document is low and to improve the efficiency and show a correct document with accuracy a new method is developed. The table 1 shows the entity extraction while searching in google. It displays the description as different information based on the queried text.

Table 1: Example of Entity Extraction

Conference List	Description
KDD conference	19 th ACM SIGKDD conference
VLDB conference	VLDB 2010 conference overview
IEEE conference	It shows conference style and format, website.

Figure 1: Redundancy Evaluation for Entity Extraction



The figure 1 shows the redundancy evaluation for entity extraction using google search engine.

VI. Conclusion

This research aims with the study and analysis of Dictionary based entity recognition problem. The retrieval method focused on semantic similarity search techniques to find the exact true matched documents. The redundancy evaluation of entity extraction verified by using google search engine. To improve the efficiency of entity extraction

techniques named as Relevance feedback with images, Document Template Style, Top most Ranking methods are implemented in future work.

References

- [1]. Agrawal, Sanjay, et al. "Exploiting web search engines to search structured databases." Proceedings of the 18th international conference on World Wide Web. ACM, 2009.
- [2]. Arasu, Arvind, Venkatesh Ganti, and Raghav Kaushik. "Efficient exact set-similarity joins." Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006.
- [3]. Amitay.E, N. Har'El, R. Sivan, and A. Soffer. Web-awhere: geotagging web content. In SIGIR, pages 273–280, 2004.
- [4]. Bayardo, Y. Ma, and Srikant, "Scaling Up All Pairs Similarity Search," Proc. 16th WWW Int'l Conf. World Wide Web, pp. 131-140, 2007.
- [5]. Bocek, Thomas, et al. Fast similarity search in large dictionaries. University, 2007.
- [6]. Borthwick, "A Maximum Entropy Approach to Named Entity Recognition," PhD thesis, New York Univ., 1999.
- [7]. Chakrabarti, Chaudhuri, Ganti, and Xin, "An Efficient Filter for Approximate Membership Checking," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 805-818, 2008.
- [8]. Pasca.M. Acquisition of categorized named entities for web search. In CIKM, pages 137–145, 2004.
- [9]. Tsuruoka.Y and J. ichi Tsujii. Improving the performance of dictionary-based approaches in protein name recognition. Journal of Biomedical Informatics, 37(6):461–470, 2004.
- [10]. Zhixu Li, Laurianne Sitbon, Liwei Wang, Xiaofang Zhou, " AML: Efficient Approximate Membership Localization within a Web-Based join framework", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 2, 298-310, February 2013.