ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (432-435), Month: April - June 2015, Available at: www.researchpublish.com

An Entity Recognition Framework for Data Mining Techniques

[1]Sruthi Varghese, [2]Nagaraj Naik

[1]PG Student, [2]Sr. Assistant Professor, Department of CS&E, MITE, Moodabidri, Mangalore, India,

Abstract: **A large amount of assorted information are posted and retrieved on web by its users and administrators. For the web users, the main issue is to browse through the exact data they are looking for. This web content mining leads for the developing of an efficient technique for retrieving the exact web contents for users query. Nowadays there are many techniques which provide a good performance in web content mining. These existing techniques could be replaced by improved web content techniques which could be utilized for real world applications. Approximate Membership Extraction (AME) was one among them. Approximate Membership Extraction (AME) provides a full coverage to the true matched substrings from a given document, but many redundancies cause a low efficiency of the AME process and decrease the performance of real world applications using the extracted substrings. As a counter measure to this Approximate Membership Localization (AML) can be used to retrieve true matches for clean references. In this project an AML technique is used for extracting those true matches for clean reference which are non-overlapped. Finally a comparison is made between AME and AML.**

Keywords: **Approximate Membership Localization (AML), Approximate Membership Extraction (AME), P-Prune Algorithm**

1. INTRODUCTION

Data Mining is defined as extracting the information from the huge set of data. In other words we can say that data mining is mining the knowledge from data. Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

Text mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

2. LITERATURE SURVEY

In [1] Agarwal et al. suggested a technique that matches the query terms only against the information in its own database. Combination of pre-processing and web search engine adaptations in order to implement entity search functionality at very low space and time overhead. The main tasks are to identify relevant information in a structured database using a web search query very efficiently and effectively. The search in each structured database is soiled in that it exclusively uses the information in the specific structured database to find matching entities. The result from the structured database search is therefore independent of the result from web search. The major drawback is that it takes a high processing time to search from a structured database.

Eleni Mangina et al [2] have proposed a key phrase extraction algorithm along with a tiling process for a document in an e-learning environment. Their technique has applied IUI techniques for an online e-learning environment. For key phrase extraction technique, they have employed user modeling techniques, information retrieval techniques and extraction mechanisms and collaborative filtering methods. The main aim of their system was to recommend documents for the users of e-learning environment exactly based on their requirement and query with minimal inconvenience and non-intrusive manner. For this, the Key Extraction Algorithm was used to automatically extract queries and then those extracted results were filtered and provided to users.

In [3] Arasu et al. suggested a similarity join operation for reconciling representation of an entity. Set similarity join algorithm define that given two record compilation of sets recognize the entire couple of set, individual from every assortment that are extremely related. The information frequently has different contradiction which has to be predetermined earlier than the data can be worn for exact data examination. The conception of resemblance is captured numerically using a string based similarity. Apart from string based similarity semantic relationship flanked by entities can be subjugated to recognize diverse representation of the identical thing. The algorithm is characterized as signature based algorithms that first generate signature for record sets, subsequently find every one of twosome sets whose signature overlie, and finally acquiesce the division of these applicant brace that gratify the set- relationship predicate. The major drawback is that it just compare with minimum amount of database so that it does not give exact similarity.

Yashaswini et al [4] have developed a suffix stripping algorithm to retrieve Kannada information from web. Their work was mainly focused on extracting suffixes from Kannada languages for retrieving Kannada text available in online on Unicode. Using their algorithm, they have stripped fourteen different major classification of Kannada suffix and few other sub classes of Kannada suffix. Also the suffixes associated with nouns, adjectives and stop words were also stripped using their algorithm. This algorithm was used for text extraction, and other text recognition and speech recognition techniques. This algorithm for stripping suffixes was implemented along with a stemming algorithm.

Kaushik et al., Suggested [5] Given two input collections of sets, a set-similarity join (SSJoin) identifies all pairs of sets, one from each collection, that have high similarity. Recent work has identified SSJoin as a useful primitive operator in data cleaning. A large number of different similarity functions such as edit distance, jaccard similarity, and generalized edit distance have been traditionally used in similarity joins. It is well-known that no single similarity function is universally applicable.

3. EXISTING AND PROPOSED SYSTEM

1. Existing System

Approximate Membership Extraction (AME), finding all substrings in a given document that can approximately match any clean references. The objective of AME guarantees a full coverage of all the true matched substrings within the document, where the true matched substring is a true mention of the clean reference. On the other hand, it generates many redundant matched substrings, thus rendering AME unsuitable for real-world tasks based on entity extraction. Indeed, redundant pairs are qualified to be part of AME results, but are unlikely to be true matches in real-world situations.

The major limitations of AME are that causes redundancy and lower the performance efficiency. If the input string is “raj”, the AME retrieve all the substring that match that input string. It will retrieve names such as raj, rajkumar, anusharaj, rajanathan..The AME does not match the exact data it is not suitable for the real world entities.

In the existing system, we are not able to get the exact matched substrings from the documents. The redundant of data will be the outcome of the existing system.

2. Proposed System

The proposed approach is to overcome the problems in approximate membership extraction (AME). Approximate membership localization propose at situating non overlapped substrings references approximately mentioned in a given record, generally in documents each string can be mentioned more than once. This will create data redundancies. Similarly for exactly matched strings there will be always only one true value that means the true mentioned strings should not overlap. In AML by using the score value and the similarity value the data redundancies should be avoided. In order to discover the non-overlapped substring pruning concept is used. It prunes redundant matched substring before generating them.

4. TECHNIQUES USED IN THE PROPOSED SYSTEM

The techniques used in the proposed system are:

- Scoring Correlation
- Similarity Function
- Inverse document Frequency
- Pruning Algorithm

4.1 Scoring Correlations

Scoring correlations for each document depends upon three appropriate parameter frequency, distance and document relevance.

Frequency freq: the number of times each reference is mentioned in each document of Docs.

Distance dist: the distance between the mention of each clean reference and the position of T.x.

Document importance imp: documents retrieved on the web are of different magnitude with respect to. their significance to the query, i.

Given a list of elements T with an attribute T.X and a clean reference list R, for each clean reference r in R.A, the probability that r is correlated to a value T.x of T.X can be measured by equation 1.

$$P(r,T.x)=\frac{\sum_{d\in Docs} imp(d).score(r,d)}{\sum_{d\in Docs} imp(d)} \quad (1)$$

Where, imp(d) is the importance of the document and is calculated using equation

$$imp(d)=\frac{\log(2)}{\log(1+\frac{[rank(d)]}{B})} \quad (2)$$

And score(r,d) is the local score of clean reference r in document d and the equation for finding score is given in equation 3.

$$score(r,d)=\omega_a.\frac{freq}{N}+(1-\omega_a)\sum_{1\le i\le freq}\frac{|d|-dist_i}{freq\,|d|} \quad (3)$$

Where is the length of document d, freq is the frequency that r mentioned in d, is the distance between the ith mention of r and, wa is the weight given to the frequency of a reference mentioned in d, and the query entity T.x in d. N is a normalization factor 1-wa is the weight given to the distance between each mention of the reference and the position of T .x in d.

4.2 Similarity Calculation

Approximate membership localization is to find all match substrings for each reference. Similarity calculation gives the similarity value between the strings. There is no overlap between the inputs values means it does not give the similarity value. This similarity process is done using word net dictionary. The word net matches all entries of the database with the given input query. First it performs syntactic checking and then checks for linked synonyms. Using the result of sentence structure checking, linked synonyms it calculates the value of similarity. The one with largest similarity to its matched entity is the best matched substring.

4.3 Inverse Document Frequency

Inverse document frequency (IDF) is a mathematical gauge that indicates how essential a word is to a document in a collection. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases relatively to the number of times a word materialize in the document. Where N is the overall quantity of document .df is defined as document frequency. The inverse document frequency calculated as

IDF=log (N/df)

Where N=Total number of document

df=Document frequency

4.4 Pruning Algorithm

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The twin goal of pruning is minimised complexity of the final classifier as well as better predictive accuracy by the reduction of over fitting and removal of sections of a classifier that may be based on noisy or erroneous data.

By pruning technique it avoid redundancies thus avoid the problem of AME (Approximate membership extraction). **B**asically there are three pruning strategies. Those three are given below.

Prune 1 (Weight Pruning). A domain D of r should be removed, if the sum weight of all segments in D is smaller than δ.ωt(r)

Prune 2 (Interval Pruning). In a domain of r, if there is an interval t whose weight is larger than 1-δ/δ .ωt(r) on the left (right) side of the strong segment, then this interval and other segments and intervals on the left (right) side of t should be removed from the domain.

Prune 3 (Boundary Pruning). The leftmost and rightmost partitions of a domain of r should be two segments of r.

5. CONCLUSION

Approximate Membership Localization (AML) only aims at locating true mentions of clean references. According to experimental results on several real-world data sets, P-Prune is proved to be several times faster, sometimes even tens or hundreds of times faster, than simply adapting formerly existing AME methods. Web-based join Structure is developed which is a search-based approach joining two tables using entity recognition from web documents and it is a typical real-world application greatly relying on membership checking.

REFERENCES

[1] S. Agrawal, K. Chakrabarti, S. Chaudhuri, V. Ganti, A. Konig, and D.Xin, "Exploiting Web Search Engines to Search Structured Databases," Proc. 18th WWW Int'l Conf. World Wide Web, pp. 501-510,

[2] Eleni Mangina and John Kilbride, "Evaluation of keyphrase extraction algorithm and tiling process for a document/resource recommender within e-learning environments", Computers & Education, Vol. 50, pp. 807–820, 2008.

[3] A. Arasu, V. Ganti, and R. Kaushik, "Efficient Exact Set-Similarity Joins," Proc. 32nd VLDB Int'l Conf. Very Large Data Bases, pp. 918-929, 2006.

[4] Yashaswini Hegde, Shubha Kadambe and Prashantha Naduthota, "Suffix Stripping Algorithm for Kannada Information Retrieval", International Conference on Advances in Computing, Communications and Informatics (ICACCl), 2013.

[5] S.Chaudhuri, V.Ganti, and R.Kaushik, "A Primitive Operator for Similarity Joins in Data Cleaning," 2006.

Authors Profile:

Ms. Sruthi Varghese completed the Bachelor's Degree in Information Technolgy from SSCET, Palani, and Pursuing M.Tech in Computer Science Engineering at MITE, Mangalore under VTU, Belgaum.

Mr. Nagaraj Naik Senior Assistant Professor MITE, Mangalore. Completed his M.Tech in Computer Science and Engineering having 8.5 years of academic experience and his areas of interest are java programming, operating systems.