DESIGN AND ANALYSIS OF APPROXIMATE STRING MEMBERSHIP CHECKING WITHIN WEB BASED FRAMEWORK

S.Balan[1], Dr. P.Ponmuthuramalingam[2]

[1]Ph.D. Research Scholar, PG & Research Department of Computer Science,
Government Arts College (Autonomous), Coimbatore, Tamil Nadu, (India)
[2]Associate Professor& Head, PG & Research Department of Computer Science,
Government Arts College (Autonomous), Coimbatore, Tamil Nadu, (India)

ABSTRACT

The amount of information stored in the web contains structured and unstructured text data. The main problem of users to find the exact information they are looking for. In existence, there are various techniques provides a good performance for web content mining. Approximate string membership checking is used to retrieve the true matches of the substring of text. For example the application areas like computational biology, text retrieval and signal processing carries many problems on spell check data querying to exact string matching. This paper focuses on overview of string matching technique research progress on the approximate matching in the World Wide Web and an improved approximate membership extraction to achieve its objective in web content mining. Which results in the proposed research combines approximate membership extraction vs. approximate membership localization technique and potential redundancy pruning techniques to retrieve the exact data and reduce the redundancy among the retrieval of information.

Keywords: Entity Recognition, String Matching, Tokens, Approximate Membership Extraction (AME), Potential Redundancy Pruning.

I. INTRODUCTION

The major application of approximate String Matching (ASM) is text retrieval to find relevant information in the large text of a given document. Recent years, enumerate text data in database and information systems to identify the exact string similar to a given query string. when user giving a query in web may mislead to unsatisfied the results of the given query. The drawback of the existing technique finds the result based on prefix or substring matches. In order to solve this problem to overcome that semantic relationships are identified for the given string and produce the accuracy of the result which helps the user or author to find the exact information. In a structured database, a web document contains various number of class entities, each entity represents some information stored in the database. Considered a word named as journal that it matches the query term only against the information of its own database. Web search engine searches the given input string and retrieves the result but not exact one, it may produce journal site links, international and national journal names as well as any discussions. Entity Recognition (ER) problem is to identify the exact entity of research paper titles of specific journal or person names in the give document. The string finds the substring of the given name and produces the result, but the accuracy of the substring searching is low and this problem is said as dictionary based membership checking. The main objective of the approximate string membership checking is to find the substring of the given text by using semantic relationships. There is various membership extraction techniques are available for Approximate Membership Checking (AMC), some of the literature reviews are as follows: Most of the approximate string matching techniques are identified by set of tokens that means similarity of the word searching is based on tokens [Chong Sun et al., 2011]. Approximate String Membership Checking (AMSC) is collection of string or document which matches another string n the document [Chakrabarti et al., 2008, Chandel et al., 2006, Li et al., 2008]. The information retrieval is search string carries two phenomena are namely precision and recall [Salt 1968, Paice 1977]. Proportion of retrieved records that are relevant is precision and proportion of relevant records actually retrieved is recall. There are four indexing solutions for approximate string matching namely first one is similarity function, there are two types of string similarities, they are set based and edit based. Second one is string tokenization, processing a string into set of primitive components called tokens. Third one is query type, it is classified in to selection query and joint query. Fourth one is indexed structure it is classified in to indexed schemes such as inverted index and trees [Patrick et al., 1980]. In wide range of filed keyword is important to search an operation for retrieving string match is required for the exact match [Ramya, 2013, Koudas et al., 2006, Lee H et al., 2007]. This paper categorized into four sections. Section-1 contain the introduction to named entity recognition and literature review of approximate membership checking, Section- contain materials and methods of membership extraction, section-3 contains the result and discussion of different data sets tested in approximate membership extraction and Section-4 includes conclusion while references mentioned in the last section.

II. MATERIAL AND METHODOLOGY

2.1 Web Extraction

The web data is extracted and stored in the database depends upon the user requirements and the extraction technique is borrowed from the areas of database grammars and machine learning etc. user retrieve the web data by browsing and keyword searching. These methods contain some limitations are not accurate data, link is lost, so the author focus various data extraction tools compared to few case results [Muslea et al., 2001, Soderland S 1999, Califf et al., 1999, Freitag D 2000, Kushmerick 2000]. There are various characteristics of web data extraction tools namely HTML (Hyper Text Markup language)- aware ttol, wrapper induction tool, Natural Language Processing (NLP) – based tool, modeling based tool and ontology based tool. The page contents are basically classified into semi-structured data and semi-structured text. The main motivation of the tool is used to finish the task easier on the other hand, languages required the user to execute the quries manually. HTML tags used for extracting data from documents those explicit formats XML (Extensible Markup language). The table1 shows that the summary of tool analysis of web extraction data, processing way of data extraction and return formats of the web to store the data in database and the type of data is identified significantly.

2.2 Text Extraction

Web data extract the information and stored as HTML documents those documents contains various accesses of pages. To extract the needed information, text extraction technique is used by text to tag on line basis and cluster the results in different areas. The size of the data is huge so the researchers are forced to crawl the data, analyze and store the content entirely.

Table1. Summary of the Tool Analysis

Sno	Tools	Methods	Processing way	XML	HTML	Type of Page
1	Languages	Minerva Web-OQL	Manual Manual	Yes	Partial	Semi Data
2	HTML-Aware	XWRAP Road Runner	Automatic Automatic	Yes Yes	None None	Semi Data Semi Data
3	NLP - Based	WHISK RAPER	Semi-Automatic Semi-Automatic	No No	Full Full	Semi Text Semi Text
4	Induction	WEIN STALKER	Semi-Automatic Semi-Automatic	No No	Partial Partial	Semi Data Semi Data
5	Ontology	BYU	Manual	No	Full	Semi data / Semi Text

Table2. Text Extraction from various Techniques [Krupl'05]

Sno	Techniques	Mean	Median	Standard Deviation
1	Threshold	56.21	61.63	31.89
2	Expectation Maximization	48.77	48.98	30.66
3	K-Means	57.44	61.17	32.96
4	Farthest-First	62.53	77.03	33.75
5	Prediction	52.40	55.30	30.01

Extract text may contains text, image etc., so there is no necessity to download and index the entire pages to save the space and time. Natural Language Processing (NLP) techniques are used to extract the specific information while removing banners, advertisements and other texts [Soderland 1997, Mooney 2005]. The users method based on the content of the page structure in a given website [Krupl et al., 2005]. The table2 shows that the extraction of text using various techniques calculated the values of mean, median and standard deviation.

2.3 Advanced Web Search

It aims to retrieve the web page files increasing amount of data. Some of the observations are length of the sentence is relevant or irrelevant, number of links is higher, number of irrelevant is lower. The author [Jericho 2006] identified HTML Parser technique is used in java library able to extract the information about tag trees. Some of the HTML tags are font, center, s, b, I etc. it scored 88.3% terms of coverage. Zipf's law (1949) inspired the word extraction from number of sentences. There are three steps to extract the relevant text namely find good sentence, find a sequence with start and end sentences, find sequence quantity and then reformatting text.

2.4 Approximate Membership Extraction

Sometimes there is a need to locate data without exact information about the subject. In those situations approximate membership extraction technique is valuable, it is also known as Approximate String Membership (ASM). Exact string matching for instance searching a word river may return the relevance of liver, rover. So the error of the margin is too large to improve the irrelevant data. To identify the exact value [Hall & Bowling, 1980] named a new technique is called precision and recall. Edit distance is used to identify the difference between two strings named as A and B, there are different types of edit functions most common is hamming distance and subsequence distance. It can be defined as B, A, L and S ( ), finds the set of all substrings in B. Such that S (A, B[I, J]) < L, where L represents maximum number of errors allowed and S ( ) represents distance function [Navarro, 1998].

2.5 Potential Redundancy Pruning

The goal of this technique is to predict the accuracy of classifier based on noisy or error data. To avoid the problem of Approximate Membership Extraction (AME) there are three strategies: - first one, weight pruning is used to sum weight of all segments. Second one, Interval pruning is used for weight removal of the domains. Third one, boundary pruning is used to combine the segments of the domain. Pre-Pruning is needed to decide whether a substring is extracted or not. Figure1 shows the various issues of pruning techniques are identified to achieve the overall performance.

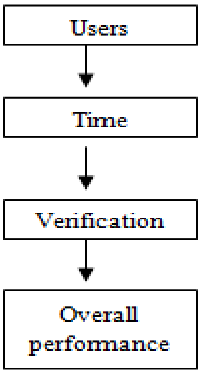
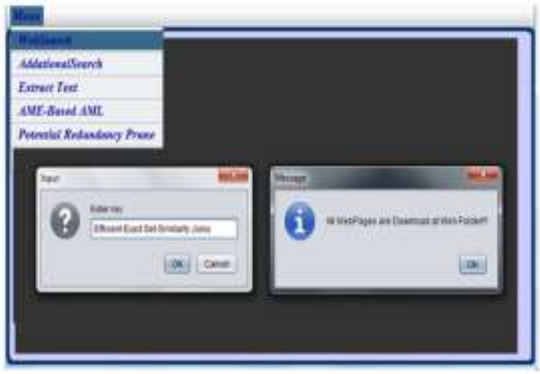


Fig.1: Issues Of Pruning Technique

III. RESULTS AND DISCUSSIONS

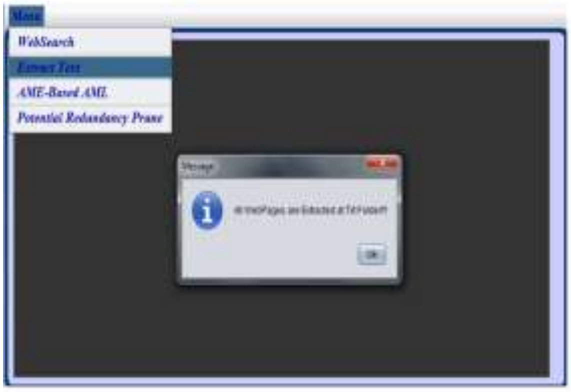
Figure2 shows the web extraction search technique. The panel offers a search mechanism for searching the texts and it allows the user to extract the text information and stored as a HTML file in the folder. The retrieved texts are represented in the following of semantic relationships such as text, tags and the results stored depend upon the user.



Enter key
Efficient Exact Set-Similarity Joins

All WebPages are Download at Web Folder!!

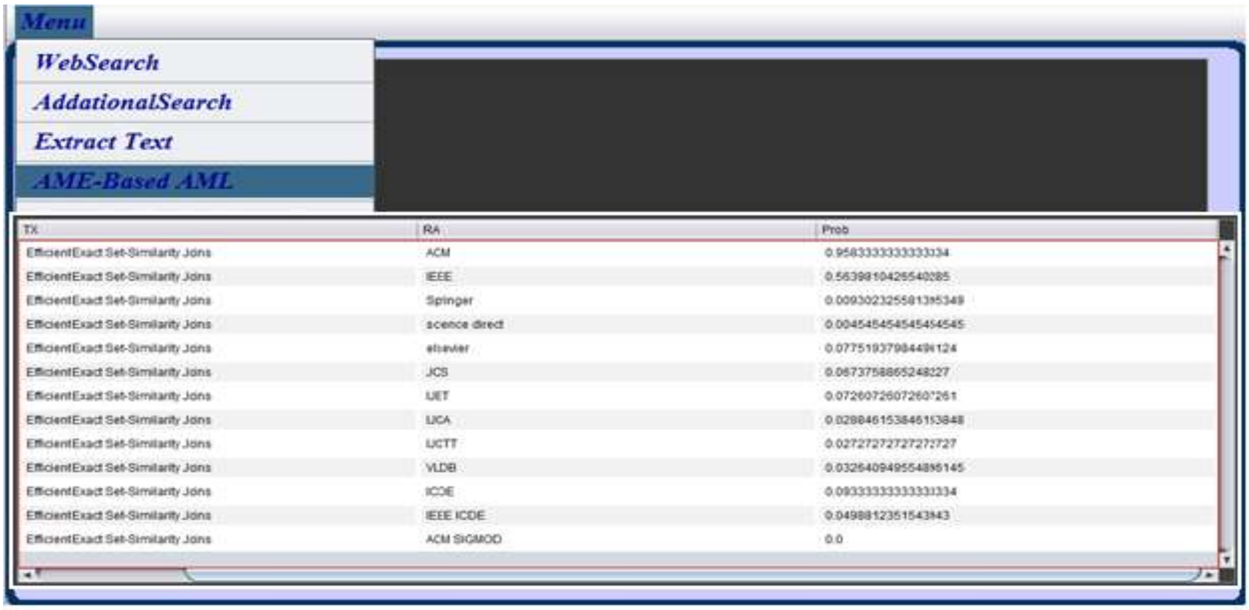
Fig.2: Web Extraction Search



All WebPages are Extracted at Txt Folder!!

Fig.3: Text Extraction Technique

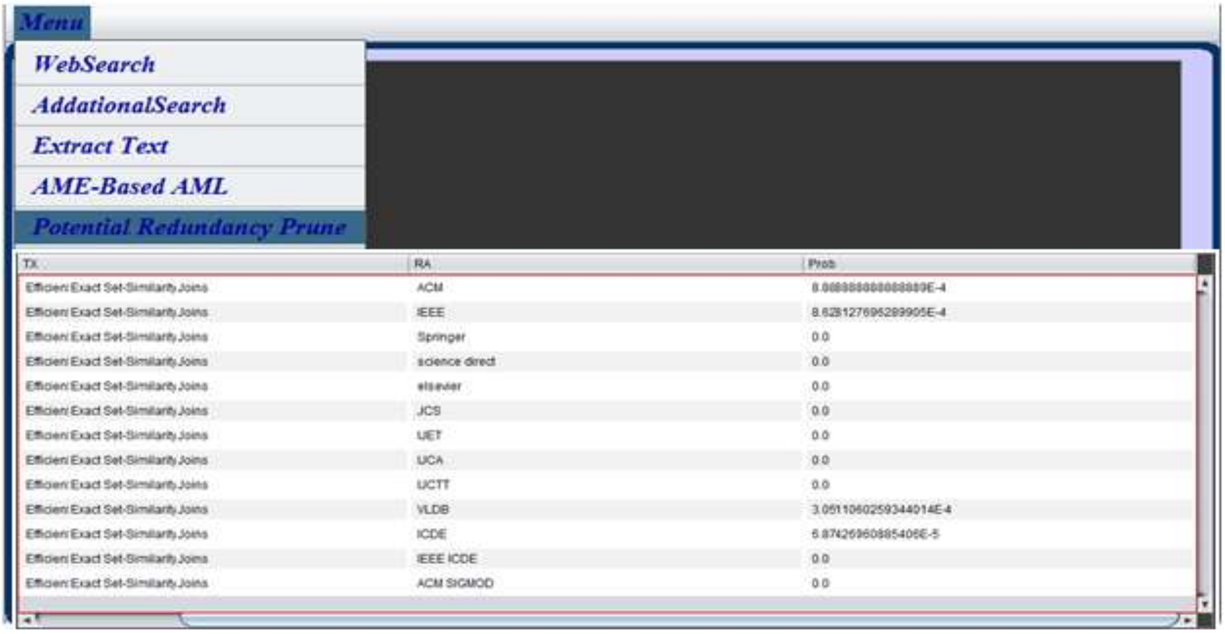
Figure3 shows the text extraction technique. The panel offers an extraction of text mechanism for retrieving the texts in web document and it allows the user to extract the text information in notepad file. Figure4 shows the approximate membership extraction based approximate membership localization. The panel offers the searching text named as efficient exact set similarity joins stored in the database list as journals and find the probability of the text.



TX	RA	Prob
EfficientExact Set-Similarity Joins	ACM	0.9583333333333334
EfficientExact Set-Similarity Joins	IEEE	0.5639810426540285
EfficientExact Set-Similarity Joins	Springer	0.009302325581395349
EfficientExact Set-Similarity Joins	science direct	0.004545454545454545
EfficientExact Set-Similarity Joins	elsevier	0.07751937984496124
EfficientExact Set-Similarity Joins	JCS	0.0673758865248227
EfficientExact Set-Similarity Joins	IJET	0.07260726072607261
EfficientExact Set-Similarity Joins	IJCA	0.028846153846153848
EfficientExact Set-Similarity Joins	IJCTT	0.02727272727272727
EfficientExact Set-Similarity Joins	VLDB	0.032640949554896145
EfficientExact Set-Similarity Joins	ICDE	0.09333333333333334
EfficientExact Set-Similarity Joins	IEEE ICDE	0.0498812351543943
EfficientExact Set-Similarity Joins	ACM SIGMOD	0.0

Fig.4: Approximate Membership Extraction

Figure5 shows the potential redundancy pruning technique and The panel offers the searching text named as efficient exact set similarity joins stored in the database list as journals and find the probability of the text. Compared to the AME based AML method pruning produces the accuracy of text in the searched journal ACM is 8.8 and IEE is 6.6



TX	RA	Prob
Efficient Exact Set-Similarity Joins	ACM	8.888888888888889E-4
Efficient Exact Set-Similarity Joins	IEEE	8.628127696289905E-4
Efficient Exact Set-Similarity Joins	Springer	0.0
Efficient Exact Set-Similarity Joins	science direct	0.0
Efficient Exact Set-Similarity Joins	elsevier	0.0
Efficient Exact Set-Similarity Joins	JCS	0.0
Efficient Exact Set-Similarity Joins	IJET	0.0
Efficient Exact Set-Similarity Joins	IJCA	0.0
Efficient Exact Set-Similarity Joins	IJCTT	0.0
Efficient Exact Set-Similarity Joins	VLDB	3.0511060259344014E-4
Efficient Exact Set-Similarity Joins	ICDE	6.87426960885406E-5
Efficient Exact Set-Similarity Joins	IEEE ICDE	0.0
Efficient Exact Set-Similarity Joins	ACM SIGMOD	0.0

Fig.5: Potential Redundancy Pruning

VI. CONCLUSION AND FUTURE ENHANCEMENT

This research is concerned with the study and analysis of web extraction, text extraction, approximate membership extraction and potential redundancy pruning. The retrieval method is based on text extraction and the software prototype allows user to retrieve the accurate information of the web based on journal database such as ACM, IEEE, Springer, VLDB, etc. the prototype is tested with test data and found successful. It can be further extended in the following directions such as image search and style of documents.

REFERENCES

[1]. Salton G, Automatic information organization and retrieval, McGraw-Hill, New York, 1968.

[2]. Paice C.D., information retrieval and the computer, MacDonald and Jane’s Computer Monographs, London, 1977.

[3]. Chakrabarti K, Chaudhuri S, Ganti V and Xin D., An efficient filter for approximate membership checking, in SIGMOD, 2008.

[4]. Chandel A, Nagesh P.C and Sarawagi S., Efficient batch top-k searches for dictionary-based entity recognition, in ICDE, 2006.

[5]. Li C, Lu J, and Lu Y., Efficient merging and filtering algorithms for approximate string searches in *ICDE*, 2008.

[6]. Chong Sun, Jeffrey Naughton, he Token Distribution Filter for Approximate String Membership Checking, Fourteenth International Workshop on the Web and Databases (Web DB 2011), June 12, 2011 - Athens, Greece.

[7]. Ramya B, Survey of Spatial Approximate String Search, International Journal of Computer Trends and Technology (IJCTT) – volume 6 number 3– Dec 2013

[8]. Koudas N, Marathe A, and Srivastava, Flexible string matching against large databases in practice, in VLDB, Pages 1078-1086, 2006.

[9]. Lee H, Ng R.T and Shim K, Extending q-grams to estimate selectivity of string matching with low edit distance, in VLDB, pages 195-206, 2007.

[10]. Patrick A.V Hall, Geoff R.Dowling, Approximate String Matching, Computing surveys, vol. 12, No.4, December 1980.

[11]. Muslea I. RISE: Repository of online information sources used in information extraction tasks. http://www.isi.edu/muslea/RISE/.

[12]. Muslea I., Minton S., and Knoblock C. Hierarchical wrapper induction for semi structured information Sources. Autonomous agents and multi-agent systems 4, (2001), 93-114.

[13]. Soderland s, learning information extraction rules for semi-structured and free text. Machine Learning 34, 1-3 (1999), 233-272.

[14]. Califf M.E and Mooney R., R.J. Relational Learning of Pattern-Match rules for information extraction. In proceedings of the sixteenth national conference on Artificial Intelligence and Eleventh conference on Innovative applications of artificial intelligence (Orlando, FL, 1999), pp. 328-334.

[15]. Freitag D. Machine Learning for Information Extraction in Informal Domains. Machine Learning 39, 2/3 (2000), 169-202.

[16]. Kushmerick N. Wrapper induction: Efficiency and expressiveness. Artificial Intelligence Journal 118, 1-2

(2000), 15-68.

[17]. Soderland S, “Learning to Extract Text-based Information from the World Wide Web”, in Proc. Of KDD 1997, Newport Beach, California, USA, 1997.

[18]. Mooney R.J, and Bunescu R, “Mining Knowledge from Text Using Information Extraction”, in Proc. of SIGKDD 2005, Chicago, Illinois, USA. Aug. 2005.

[19]. Krüpl B, M. Herzog, and W. Gatterbauer, “Using visual cues for extraction of tabular data from arbitrary HTML documents”, in Proc. of WWW 2005, Chiba, Japan, 2005.

[20]. Jericho HTML Parser, (2006), http://jerichohtml.sourceforge.net/doc/index.html

[21]. Hall, P. A. V. and G. Dowling, "Approximate String Matching," Computing Surveys, Vol. 12, No.4 (December 1980), pp. 381-402.

[22]. Navarro, G., Approximate Text Search, PhD. dissertation, Department of Computer Science- University of Chile, Santiago, 1998.