Educational Administration: Theory and Practice
2024,30(4), 1484-1490
ISSN:2148-2403
https://kuey.net/

Research Article

Optimizing Semi-Automatic Semantic Matched Searching Concept

Dr. D. Elangovan[1*], Dr. Radha.C[2], Dr. M.T. Raghuraman[3],

[1*]Department of Computer Science, Faculty of Science and Humanity, SRM IST, Kattankulathur, Chennai, INDIA
[2]Department of Computer Science Guru Nanak College (AUTONOMOUS), Chennai, INDIA
[3]Department of Computer Science Guru Nanak College (AUTONOMOUS), Chennai, INDIA

Citation: Dr. D. Elangovan et al. (2024), Optimizing Semi-Automatic Semantic Matched Searching Concept, *Educational Administration: Theory and Practice*, *30*(4), 1484-1490,
Doi: 10.53555/kuey.v30i4.1698

ARTICLE INFO

ABSTRACT

The Semantic web presents an opportunity to convey the meanings of web documents in a format understandable by machines. However, the majority of web content remains in a format intended for human consumption, and it's anticipated that creators and developers will continue to prefer this format due to its simplicity. To bridge this gap and realize the vision of the Semantic Web, two main approaches have emerged: annotating information sources with machine-accessible semantics or developing programs to extract semantics from web sources. This proposed research aims to identify documents retrieved from web servers based on knowledge extraction using a Semi-automated semantic matching concept. This matching concept aids users in selecting the appropriate document categorized into factual, procedural, and conceptual based on Bloom's taxonomy. The analysis involves iterating through grammatical rules to apply those relevant and determining if a valid stem is found. The SAS algorithm entails complex grammatical rules, such as removing multiple suffixes and prefixes, which can lead to variations in results. One factor influencing the outcome is whether the algorithm requires the output word to be a real word in the given language. Some approaches don't mandate the word's existence in a lexicon database, while others maintain a database of known word roots that are actual words. Optimization is provided by the SAS algorithm through its efficient memory utilization and swift execution, enhancing overall performance.

Keywords: Semantic, Matched, extraction, web.

1. INTRODUCTION

The primary aim of this proposed research is to identify and retrieve documents from a web server based on their content using Semi-automated Semantic Matching. This matching process assists users in locating the appropriate documents categorized as factual, procedural, or conceptual, according to Bloom's Taxonomy [2][3]. Access to client functionalities is granted via authenticated usernames and passwords, ensuring confidentiality during document retrieval and download based on the user's knowledge level [7].

Security measures have been established to limit access to authorized users, preventing unauthorized entry into any web application services. Files are downloaded to default locations swiftly, with a focus on optimizing memory utilization. Stemming is employed to enhance garbage collection efficiency by identifying and removing redundant words with similar base meanings, thus reducing morphological clutter within documents. In this process, if two words in a document share the same base meaning, they are mapped to the appropriate stem, effectively removing one of the words from the document to streamline its content. However, words with dual meanings are kept separate from this stemming process. Stemming implementations with higher accuracy probabilities are utilized to achieve optimal results from the provided documents. The effective management of total records and information extracted from documents is achieved through probability-based trimming, resulting in over 90% accuracy of the given documents. Implement at both ends such as suffix and prefix of the given documents. Extraction of suffix and prefix based on tree structures [6].The probability of Suffix and prefix removal from the given word of the document leads to the success rate of implementation. Implement knowledge of extraction are categorised as procedural, factual and conceptual based on Semi-automatic semantic matched concept. Semi-Automatic

Semantic Matched concept provides the probability, which an average of 10% difference from factual and conceptual knowledge [1].
Optimization of knowledge extraction based on Semi-automatic Semantic Matching surpasses Naive Bayes and TF-IDF algorithms in terms of memory utilization [5]. Factual, Procedural and conceptual knowledge of extraction may differ from documents to documents. Effective use of Memory utilization in Semi-automatic semantic matched concept is comparatively more than Naive Bayes algorithm and TF- IDF Algorithms [5]. Less Comparison is calculated in Semi-automatic semantic matched concept to compare to Naive Bayes algorithm and TF-IDF Algorithms [8]. Time taken to run the proposed algorithms are optimized. The probability of running time of algorithm provides lesser time compare to proposed algorithms. Non-functional testing ought to increase usability, efficiency, maintainability, and movability. Optimize the proposed work in setup, executes, managed and monitored.

2. PROPOSED WORK- SAS

This proposed work is to download or view the content of the file based on knowledge extraction such as Factual knowledge, procedural knowledge or conceptual knowledge. This proposed work mainly used for the massive user, who needs to download the appropriate file from the intranet web servers [7]. This proposed work may differentiate the category of people who may be interested in technologies such as procedural knowledge, theoretical knowledge such as conceptual knowledge or analytical knowledge such as factual knowledge. The following Figure 2.1 explains the proposed work.
Initially, when the user searches the file, the appropriate documents will be retrieved from the database. The Proposed works categorized into four parts are:
a) Stemming
b) Trimming
c) Knowledge extraction and
d) Memory utilization and Running Time.
Stemming is the process which is used to remove the unwanted words from the documents. A certain list of words from the document. This is not necessary to determine the knowledge of extraction. Those words are considered as a stopword [7].

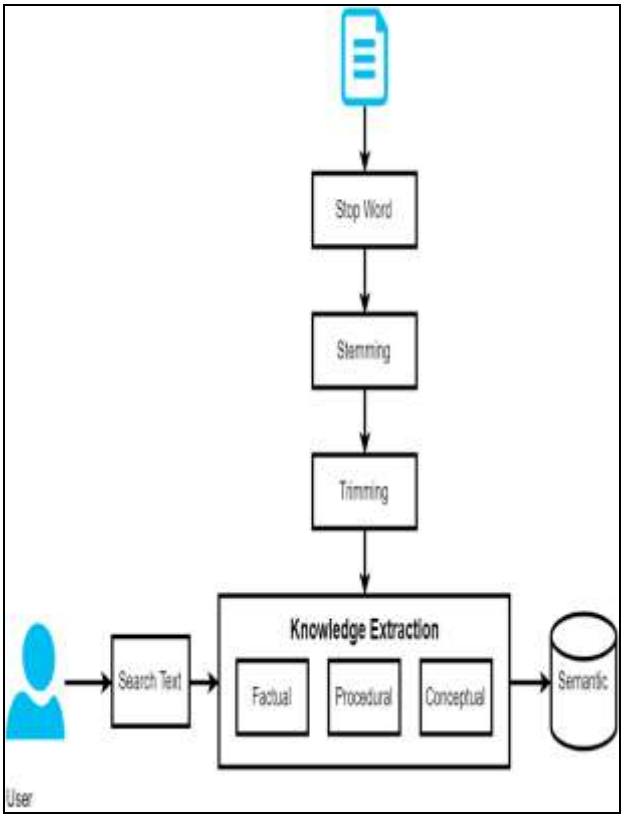

Fig. 2.1 Proposed Work – SAS

The selection of stop words is based on Bloom's Taxonomy, originally proposed by Benjamin Bloom. His work primarily revolves around different types of learning and understanding user behavior. Bloom categorizes learning into three domains: Cognitive, Affective, and Psychomotor. The Cognitive domain emphasizes critical thinking within a specific subject area. The Affective domain centers on understanding

and influencing user behavior, which includes emotional responses, attitudes, and feelings. The third domain, Psychomotor, focuses on skill-based learning and development. [1][4].

2.1 STEMMING

Based on the Blooms Taxonomy, this reflects phase is called Stemming. In other word, a garbage collector who is used to collect the garbage word from the given document, which is not necessary for knowledge extraction.

2.1.1 ALGORITHM FOR STEMMING

Step 1: Let D be the Document, d be the words in the D document, act as an input.
Step 2: Let n be the number of word count in the given document D.
Step 3: Process to calculate the word count.

```
//to read the file
FileReader fr = new FileReader ("http://localhost/filename.pdf");
BufferedReaderbr = new BufferedReader(fr);
String line = "", str = "";
int n = 0;
int b = 0;
while ((line = br.readLine()) != null) {
str += line + " ";
b++;
}
// to find the word count
StringTokenizerst = new StringTokenizer(str);
while (st.hasMoreTokens()) {
String s = st.nextToken();
n++;
```

Step 4: Process of Stemming as follows

```
for(i=0;i<=n;i++)
{
if(d[i].equals(stopword[i])){
d[i].replace(d[i],d[i++]);
}
```

Step 5: Stop the process.

2.2 TRIMMING

Next Phase is called trimming, which is used to extract the particular characters. For example, an adjective word ends with “ing” format. The words like “studying”, the suffix wording is removed from the word of the given document. This process includes both prefix and suffix words. Remove of the adjective word either in prefix or suffix is called trimming.

Classification of trimming are categorized into two levels, the First level is prefixed, which means front part of the characters are removed from the word of the given documents is called prefix trimming. The second level is a suffix, which means the tail parts of the words are removed from the word of the given document is called suffix trimming [6].

2.2.1 ALGORITHM FOR TRIMMING

Trimming the **suffix** part of the word d, from the given document D

```
for(i=0;i<=n;i++){
String s1=d[i];
String s2=suffix[i];
d[i]=removeSuffix(s1,s2);
}
String removeSuffix(String s1, String s2){
if(s1 !=null &&s1.startsWith(s2)){
returns.split(s2)[1];
}
return s;
```

From the above algorithm, which is used to split the word into characters and remove the suffix from the given characters.

Trimming the **prefix** part of the word d, from the given document D

```
for(i=0;i<=n;i++){
String s=d[i];
String p=prefix[i];
```

```
d[i]=removePrefix(s,p);
}
String removePrefix(String s, String p){
if(s1 !=null && p !=null && s1.endsWith(p)){
returns.substring(p.length)
}
return s;
```

From the above algorithm, which is used to split the word into characters and remove the prefix from the given characters.

Third Phase is of Knowledge Extraction. This knowledge extraction are classified into three categories are a) Factual Knowledge b) Procedural Knowledge and c) Conceptual Knowledge as shown in Figure 2.1

2.2.2 Experimental Result

a) Factual Knowledge may be an even affirmation of one thing. Factual data is Associate in nursing affirmation. After taking two ideas and add them along, then one thing is affirmed. As an example, "run" and "boys" square measure joined along to provide the affirmation "boys run". All affirmations square measure either true or false b.

b) Procedural Knowledge additionally called imperative data is that the data exercised within the performance of some task. See below for the particular that means of this term in psychology and property law.

c) Conceptual Knowledge it's a connected network of data, a network within which the linking relationships square measure as outstanding because of the distinct bits of data.

In **Semi-automated semantic matched concept**, the Probability of having both the Outcome O and Evidence E is: (Probability of O occurring) multiplied by the (Prob of E given that O happened)[7].

The evidence, P (Outcome or Evidence) = P (Evidence given that the Outcome) times Prob (Outcome), scaled by the P (Evidence)

Based on SAS, the probability to find out the given document is Factual knowledge, concept knowledge and procedural knowledge.

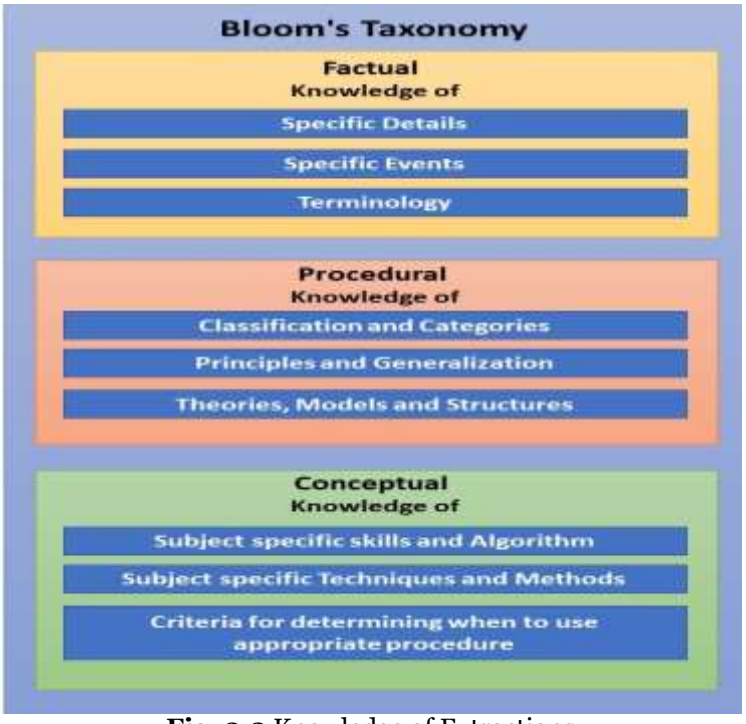

Fig. 2.2 Knowledge of Extractions

Consider D as given document, N be the number of word in the given document. Initially the given document D is used for Stemming process. The Process which is used to eliminate the unwanted word based on Blooms taxonomy. Next process is of trimming which is used to remove the prefix and suffix of the adjective word from the given document D. Consider the Sample data for Knowledge extraction such as Factual knowledge, Procedural Knowledge and Conceptual knowledge. The probabilities of outcomes are calculated and compare with the existing algorithms as follows. The Proposed system downsizing the running time and memory utilization.

2.2.3 Determination of Knowledge Extraction

To determine the optimum performance of knowledge extraction techniques of e-content are as follows
Initially, Stemming process to be determines as follows.
Assume that stopwords=array("a", "about", "above", " across", "alter", "afterwards", "again", "against", "all", "almost", "along", "already", "also", "although", "am", "amongst", "amount", "an", "another"...) // more than hundred words based on blooms taxonomy.
If the words available in the stop words then removing the word from the document, this process called Stemming.

- Next process is called trimming, which is used to remove the suffix and prefix of the word. Mainly trimming is used for adjective words.
- Knowledge of classification are categorized into three.

- List of sample word for **Factual knowledge are**:

What, list, define, tell, name, locate, identify, distinguish, acquire, write, underline, relate, state, recall, select, repeat, recognize, reproduce, measure, memorize, etc,
List of sample word for **Procedural knowledge are**:
Demonstrate, summarize, illustrate, interpret, contrast, predict, associate, distinguish, identify, show, label, collect, experiments, recite, classify, discuss, select, compare, translate, prepare, change, rephrase, interpret, differentiate, draw, explain, estimate, fill in, choose, operate, perform, organize, apply, calculate, develop, solve, make use of, predict, design, construct, access, practices, classify, solve. Etc.,
List of sample word for **Conceptual knowledge are**:
Analyse, resolve, justify, infer, combine, integrate, plan, create, generalize, assess, decide, rank, grade, test, recommend, select, explain, judge, contrast, survey, examine, differentiate, investigate, compose, invent, improve, imagine, hypothesis, decide, judge, prove, predict, evaluate, rate.
As SAS proposed algorithm discussed The Probability of having both the Outcome O and Evidence E is: (Probability of O occurring) multiplied by the (Probability of E given that O happened).
Probability of having both the Outcome O and Evidence E is: (Probability of O occurring) multiplied by the (Probability of E given that O happened). →(1)
The evidence, P (Outcome or Evidence) = P(Evidence given that the Outcome) times P(Outcome), scaled by the P(Evidence) →(2)
Based on the equation, to determine knowledge of extractions,
Assume that,
Let D be the document contains 256 words
Let k_1 be the knowledge word of factual contains 21 words
Let k_2 be the knowledge word of Procedural contains 48 words
Let k_3 be the knowledge word of Conceptual contains 35 words
Let wd_1 be the total number of factual words (81 words) in the document.
Let wd_2 be the total number of Procedural words (98 words) in the document.
Let wd_3 be the total number of Conceptual words (69 words) in the document.

2.2.4 Experiment

To find the probability of the given document based on Semi –automatic semantic matched concepts are as follows:
To calculate probability of evidence for factual, consider the following formula,
To find the probability of Evidence for Factual, using **equation (1)& (2)**

a) Find the Probability (Likelihood of Evidence)

Probability (Likelihood of Evidence) = 1 / Number of words in factual
Probability (Likelihood of Evidence) = 1 / 21 = 0.05 →(3)
Let k_1 (Factual words) are 21 then Probability (Likelihood of Evidence) is 1 divided by 21 is 0.05.

b) Prior Probability of Outcome

Prior Probability of outcome = No of words in the document / Total No of words in the document.
Let wd_1contains 81 words are found in the document and the total no of words in the document D contains 256 words, now calculate the Prior Probability.
Prior Probability of outcome = 81 / 256 = 0.32 →(4)

c) The Probability of Evidence

Probability of Evidence = Total No of Knowledge Words / Total No of words in the document.
Let k_1 contains 21 words are knowledge words in factual and the total no of words (256 words) in the D, now calculate the Probability of Evidence.
Probability of Evidence = 21 / 256 = 0.08 →(5)

Finally calculate the probability of outcomes using SAS algorithm,
Probability of outcomes = probability of factual * prior probability of outcome / probability of evidence.
Probability of outcomes = 0.05 * 0.32 / 0.08 = 0.18
From the result of equation (4), (5) and (6), determine the following table 2.2.3 as follows

Knowledge Representation	Total No. of Knowledge Words (k)	Total.No.of Words in the Document (wd)	Probability (Likelihood of Evidence)	Prior Probability of Outcome	Probability of Evidence	Probability of Outcome (SAS)
FACTUAL	21	81	0.05	0.32	0.08	0.18
PROCEDURAL	48	98	0.02	0.38	0.19	0.04
CONCEPTUAL	35	69	0.03	0.27	0.14	0.06

Table 2.2.2 Probability of Outcome using SAS

2.2.5 Evaluation of Knowledge Extraction

Following graph fig. 2.3 had been generated based on the tabulations of table 2.2.2 from derived by working as needed with the sample data. From the following graph, it is understood that the probability of factual becomes high comparing with procedural and conceptual. Consider there are 21 Factual words, 81 words in the document and 256 is total number of words in the document. Based on that calculation can be done to find the Probability (Likelihood of Evidence), Prior Probability of Outcome, Probability of Evidence and finally find the probability of outcome. Based on the table 2.2.3, it generate probability of outcomes that shows factual knowledge has a highest probability.

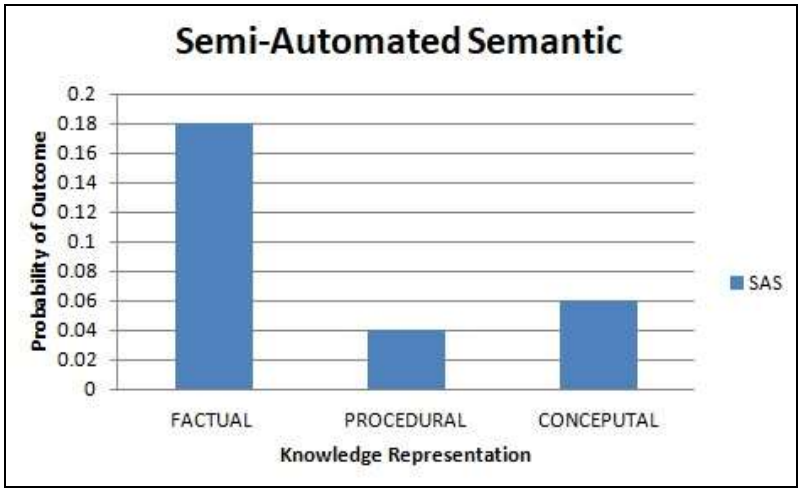

Fig.2.3 Comparison of Factual Knowledge on SAS

Similarly it generate probability of outcomes that shows procedural and conceptual knowledge has a highest probability.

2.2.6 Estimation of Memory Utilization

To Calculate the memory utilization of an algorithm, need to determine the total space available in the memory by using the built – in runtime. Total Memory () method returns the total memory size in megabyte. As same as need to determine the free space available in the memory by using the same built – in runtime. free Memory () method. To determine the memory usage of an algorithm, need to subtract total size of memory with free space available in the memory.
Algorithms to determine the Memory Utilization are as follows

```
<%
//algorithm to determine the memory utilization
//Declare the MEGABYTE as static variable with private access specifier
Private static final long MEGABYTE = 1024L * 1024L;
//Declare the Megabyte method with byte as a parameter to convert bytes to megabyte
Public static long bytes To Mega bytes (long bytes) {
return bytes / MEGABYTE;
}
//Use runtime interface and class to determine the total memory and free space //available in memory and subtract it to find the memory utilized by algorithm
long memory = runtime. total Memory () - runtime. Free Memory ();
out. println("Used memory is bytes: " + memory);
out. println("Used memory is megabytes: "
```

+ bytesToMegabytes (memory)); %>
The algorithm described above is commonly used to determine memory utilization across various methods such as Naïve Bayes, TF-IDF, and the Semi-automatic semantic algorithm. However, SAS provides additional optimization compared to other methods.

3. CONCULSION:

The SAS Algorithm is employed to assess the likelihood of semantic word usage within e-content sourced from a provided document. This methodology involves stemming and trimming words from the document and categorizing them based on their factual, procedural, or conceptual nature. These categorized words are then reconstructed into tree structures to ascertain the probability of outcome and evidence. These efficient techniques primarily aim to enhance time management, optimize memory usage, and improve overall efficiency through the implementation of the SAS (Semi-Automated Semantic) algorithm. The Knowledge Extraction and Memory Utilization algorithm yield more precise results. This proposed approach is primarily focused on reducing the execution time of the algorithm and optimizing memory utilization for loaded documents.

ACKNOWLEDGEMENT

I would like to express my special appreciation and heartfelt thanks to the Pro Vice-Chancellor, Dean Deputy Dean and HOD of Computer Science, FSH, SRM Institute of Science and Technologies, Kattankulathur, for her valuable guidance, scholarly inputs and consistent encouragement, throughout the research work.

REFERENCES:

1. Yassine Gargouri, "Ontology Maintenance using Textual Analysis, Systematic-cybernetics and informatics", Vol-1, Number-5, PP-63-68, 2016.
2. Muqeem Ahmed,"Semantic Based Intelligent Information Retrieval through Data mining and Ontology", IJCSE VOL- 5, ISSUE - 10,PP 210-217,Oct-2017.
3. S. Banerjee,"A Semantic Web Based Ontology in the Financial Domain", International Journal of Computer and Information Engineering, Vol:7, No:6,PP 807-810, 2013.
4. H. Srimathi, "Semantic Web based Personalized eLearning", International Journal of Applied Information Systems (IJAIS), Volume 2– No.1, PP 11-16, May 2012.
5. Akiko Aizawa, "An information-theoretic perspective of tf–idf measures", Elsevier Science Ltd, Information Processing and Management 39,PP 45–65,2003.
6. AyushSinghal and JaideepSrivastava "Data Extract: Mining Context from the Web for Dataset Extraction", International Journal of Machine Learning and Computing, Vol. 3, No. 2, PP 219-223, April 2013.
7. D. Elangovan, “Semi-Automated Semantic Matched Concept Extraction Model for E-Content Development”, International Journal of Applied Engineering Research, PP 2973-2975, November 5, 2016.
8. Dino Isa, Lam Hong, Lee, V. P. Kallimani, R. Rajkumar,” Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model”, computer and information science. CCSE, VOL 1 NO 4, PP 79-90, November 2008.