Multi-document Summarization for Query Answering E-learning System

Dr.S.Saraswathi
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

S.Janani
Department of Information Technology
Pondicherry Engineering College
Puducherry605014, India

M.Hemamalini
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

V.Priyadharshini
Department of Information Technology
Pondicherry Engineering College
Puducherry605014, India

Abstract:

The proposed E-learning system aims at providing a multi-document summarization for documents retrieved from Google and providing the user a precise answer for his/her query under the domain of "Operating Systems". This E-learning system also provides documents display for various topics, authentication facility for it to be used by certain users. Authenticity of users is maintained by Symmetric Encryption key Algorithm. The system is designed in a manner that the query entered by the user is passed to the tree tagger to obtain the keywords which are then passed to Google to retrieve multiple documents. Multiple levels of summarization are performed on the documents to retrieve exact answer for the user. Document ranking is performed by using tf*idf weights followed by Individual document summarization using keywords and concept words. The similarity of documents is then compared to produce a non-redundant output. Finally the resultant data is submitted to the user.

Keywords - Tagger, Multi-document Summarization, Information Extraction and Retrieval, Ontological tree, Concept-wise retrieval.

I.INTRODUCTION

E-learning can be defined as technology-based learning in which learning material is delivered electronically to remote learners via a computer network.

The availability of e-learning systems makes it possible for users around the world to directly access previously unimagined sources of information. Security in any e-learning system is a major concern. It is necessary to authenticate the users of the system. However conventional e-learning systems provide only the course content to the users. This requires the user to read the complete content for better understanding.

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic.

The query is processed by a Parts Of Speech tagger [1] which detects the keywords for deciding the type of search. This leads to Concept wise search or the Keyword search based on the keywords obtained [2].

Locality based similarity heuristic method [3] is used to extract the answers in which every word location in each document is scored. The quality of this approach depends on the location of the keyword.

Document retrieval based on query answering system [4] focus on solving majority problems to process the natural language query: approaches of syntax analysis and syntax model, semantic model, transformation mechanism from semantic model into database queries.

Clustering algorithms [5] such as K-Means have popularly been used as semantic summarization methods where cluster centers become the summarized set. The goal of semantic summarization is to provide a summarized view of the original dataset such that the summarization ratio is maximized while the error (i.e., information loss) is minimized.

In Discovery Net [6], each distributed dataset is locally summarized by the K-Means algorithm. Then, the summarized sets are sent to a central site for global clustering. The quality of this approach is largely dependent on the performance of K-means.

A scalable clustering algorithm [7] is proposed to deal with very large datasets. In this approach, the datasets are divided into several equally sized and disjoint segments. Then, hard K-Means or Fuzzy K-Means algorithm are used to summarize each data segment. Similar to Discovery Net, a clustering algorithm is then run on the union of summarized sets.

A database system consists of millions of data items that could be picked out as their priorities by proper approach. At first, all the data is classified into different categories. Each category is assigned with a predefined priority. The higher the priority information has more possibility to be chosen. Secondly, each data can only be visited once. In addition, writing a program to perform the task can be very straightforward [8]. However, it is not very easy to design an algorithm that is most efficient for all scenarios.

The technique used to search keyword query dynamically generates new pages, called composed pages [9], which contain all query keywords. The composed pages are generated by extracting and stitching together relevant pieces from hyper-linked Web pages, and retaining links to the original Web pages. To rank the composed pages, the authors consider both the hyper link structure of the original pages, as well as the associations between the keywords within each page.

Naïve algorithm and Page ranking algorithm [15] is used for searching and ranking the documents respectively, used Ontology tree for multiple purposes like inter language conversion, keyword language identification and sub keywords extraction.

Glenisson P.and Mathys J [16] have showed how the bag-of-words representation can be used successfully to represent genetic annotation and free-text information coming from different databases. They evaluated the VSM by testing and quantifying its performance on a fairly simple biological problem. They found that it can establish a powerful statistical text representation as a foundation for knowledge-based gene expression clustering[17].

During software maintenance, developers often cannot read and understand the entire source code of a system and rely on partial comprehension, focusing on the parts strictly related to their task at hand[18].

II.PROPOSED WORK

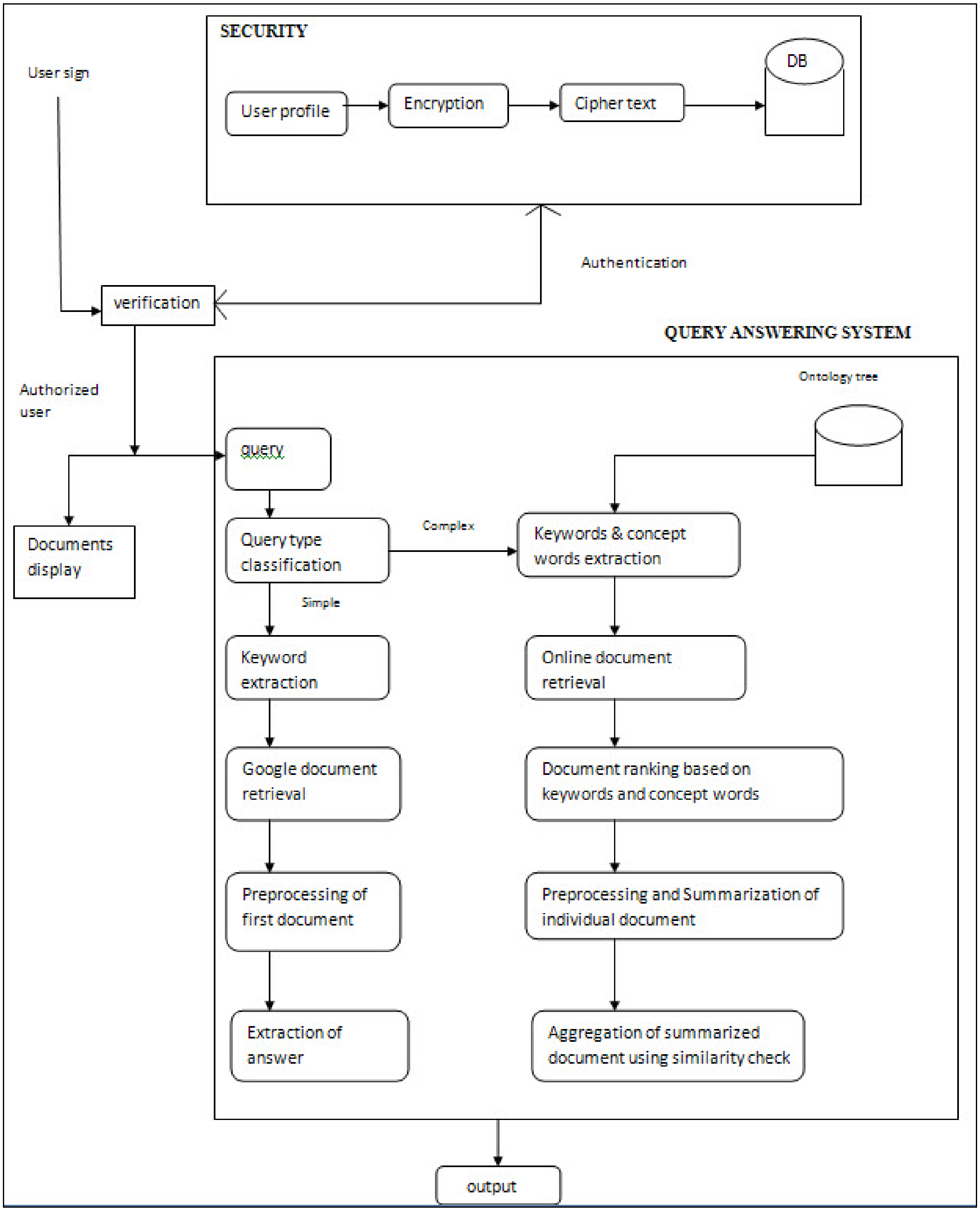
SECURITY
User sign
User profile
Encryption
Cipher text
DB
Authentication
verification
QUERY ANSWERING SYSTEM
Authorized user
Ontology tree
query
Documents display
Query type classification
Complex
Keywords & concept words extraction
Simple
Keyword extraction
Online document retrieval
Google document retrieval
Document ranking based on keywords and concept words
Preprocessing of first document
Preprocessing and Summarization of individual document
Extraction of answer
Aggregation of summarized document using similarity check
output

Figure 1 Architecture Diagram

The entire system is classified such that it consists of the following major divisions as refered from Fig. 1

1. Verification
   1.1 User profile information
   1.2 Encryption
   1.3 Cipher text

2. Documents display
3. Query Answering System
   3.1 Input Query
   3.2 Query type classification
      3.2.1 Simple queries
         3.2.1.1 Keyword extraction
         3.2.1.2 Extraction of online documents
         3.2.1.3Extraction of answer
         3.2.1.4Result
      3.2.2 Complex queries
         3.2.2.1 Keywords and concept words extraction
         3.2.2.2Online document retrieval
         3.2.2.3 Document ranking based on keywords and concept words
         3.2.2.4 Summarization of individual document
         3.2.2.5 Aggregation of summarized document using similarity check
         3.2.2.6 Result

1. Verification

Security is must in e-learning system. There may be a chance of stealing user’s personal information. So there is a need to encrypt the user’s information.

1.1 User profile information

Authorized users only can access the system. So users need to register. Here users will give their personal information.

1.2 Encryption

The user details are encrypted using simple cost effective encryption technique[10]. Each character of the input string is converted into a stream of 8 bits from its ASCII value. The binary stream is then reversed. A 4 bit secret key is generated one for each character of the string. The reversed bits are divided by the key and the resulting remainder and quotient are allotted 4 and 5 bits respectively.

1.3 Cipher text

The 4 bit remainder and the 5 bit quotient are then appended together to form the cipher text which is then stored in the database.

During each login by the user, username and password are encrypted and they are checked with the encrypted data in the database. When the username and password match, user is allowed to proceed. This module prevents from the unauthorized user by misusing the system.

2. Documents display

The course content on various topics are presented to the user.

3. Query Answering System

3.1 Input Query

The input to the system is query. The user will give the query in search engine.

3.2 Query type classification

The query can be classified broadly as

- Simple queries
- Complex queries

3.2.1 *Simple queries*

Simple queries are identified by the keywords what, when, where, define, which.

3.2.1.1Keyword extraction

The Query given by the user is fed as an input to the Tree Tagger. It tags the input sentence and returns the parts of the sentence. From the output of the tagger, verbs and nouns are identified and are set as main keywords to perform the document search.

3.2.1.2 Extraction of online documents

The keywords are passed to the Google search engine and Google result page consisting of links to various documents are obtained. The URL for the links are identified and they are used to extract the exact documents. The Html documents obtained are now converted into notepad documents by calling the “HtmlToText” converter.

3.2.1.3Answer extraction

For queries like ‘what is thrashing’, ‘when deadlock occurs’ the first document is taken for extracting brief answers. Preprocessing of the documents removes the irrelevant content. Using weighted means algorithm the passages in the filtered document are ranked according to the frequency of keywords. A pattern matching is performed on each passage for the keywords. The passage with the maximum weight is identified and it is taken as the result.

For queries like ‘how to prevent thrashing’, ‘how to detect deadlocks’ the keyword pattern of verb i.e. prevent, thrashing is passed on to the ontology tree to consider the concept words for the extraction of answer. Concept words also will be consider for the weight calculation.

3.2.1.4Result

The passage with the maximum weight is identified and it is taken as the result.The result will be displayed to the user.

3.2.2 Complex queries

Complex queries are identified by the keywords explain, describe, detail.

3.2.2.1 Keyword and concept words extraction

If the query contains the mentioned keywords the entire information related to the queries will be grouped based on the predetermined concepts for the particular query. To group the related data tree structure[12] is used for better and efficient search reasults. This increases both the vastness of the subject covered by the process and also the quality[10] of the retrieved documents as all the aspects of the related domains are addressed.

The obtained keywords from the POS Tagger are passed on to a pre-built tree structure[13][14] inorder to obtain the relevant concept words for carrying out the required concept-wise search.

3.2.2.2 Online document retrieval

The keywords and concept words are passed to the Google search engine and Google result page consisting of links to various documents are obtained. The URL for the links are identified and they are used to extract the exact documents. The Html documents obtained are now converted into notepad documents by calling the “HtmlToText” converter.

3.2.2.3 Document ranking based on keywords and concept words

For document ranking, tf*idf weighted approach[tf] is used. The term frequency of each keyword and concept word is found out for every document in the document collection. The inverse document frequency of each keyword and concept word is calculated using the formula
$IDF = \log((N-n)/n)$ where,

N- number of documents in the collection,
n- number of documents containing the term

The weight of each term with respect to each document is then calculated.

$$W=tf*IDF$$

The document is then ranked using the average of the weights of the terms (keywords + concept words) taken over the document. The concept words from the ontology help identifying the more relevant documents.

3.2.2.4Summarization of individual document

Using weighted means algorithm the passages in the documents are ranked according to the frequency of keywords and concept words. The summarized size of the document is decided by the actual size of the document. 1/4th of the actual document size is considered for summarization. Primary keyword followed by concept wise keyword summarization is carried out separately to retrieve the summarized documents.

3.2.2.5Aggregation of summarized document using similarity check

When passages are represented as term vectors, the similarity of two passages corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity.

Given two passages a and b, their cosine similarity is

$$\text{cossim}(a,b) = \frac{a * b}{|a| * |b|}$$

3.2.2.6Result

Now a set of solutions extracted . All the possibilities are now consolidated; avoiding repetition and apt solution is given to the user.

III. PARAMETERS

1.MOS(Mean Opinion Score)

Mean of the remarks obtained from the users of the system.

$$MOS = (OS_1 + OS_2 + OS_3 + \text{..........} OS_n) / n$$

Where OS is Opinion Score.
MOS and OS range from 0 to 10.

2.Summarization Ratio

It is the ratio of the size of the Summarized text to the size of the original document.

Summarization ratio = (number of lines in summarized text) / (number of lines in original document)

Value ranges from 0 to 1.

The above one is calculated for each single document.
Effective summarization ratio is calculated for multiple documents by taking average of individual documents.

3.Precision

No of relevant documents retrieved divided by the total no of documents retrieved by a search.

Precision= (number of relevant documents retrieved) / (total number of documents retrieved)
It ranges from 0 to 1.

IV.RESULTS

A.SIMPLE QUERIES

TABLE 1
SIMPLE QUERY RESULTS BASED ON QUERY TYPE

Question type identifying word	Total no of questions	No of questions answered	No of unanswered questions	% of retrieval	Mean opinion score
Simple what	65	53	15	81.53	8.3
Complex what	35	26	9	74.28	7.9
Define	53	47	6	88.67	8.7
when	27	17	10	62.96	7.4
how	23	15	8	65.27	7.1
Different types of	26	22	4	84.61	8.1

Example

Simple what – 1) What is thrashing? 2) What is deadlock?

Complex what – 1) What are the necessary conditions for deadlock to occur?

2) What is thrashing in paging?

Different types - 1) What are the different CPU scheduling algorithms?

B.COMPLEX QUERIES

TABLE II
COMPLEX QUERY RESULTS

No of queries	Results obtained queries	Precision	Summarization ratio of the system	Mean opinion score
20	13	0.9432	0.35	8.42

V.CONCLUSION

Multi document summarization is important in e-learning system which can be utilized for improving the effectiveness of retrieval and accessibility of learning objects in e-learning.

Thus the complete e-learning system, whose design is proposed aims at obtaining the apt solution using summarization technique in multiple documents, uses Ontology tree for the purpose of concept keywords extraction. Also, security in the system and documents display for complete reading also provided effectively to the users. Thus generic platform for incorporating summarization in e-learning system is also obtained.

References

[1] Charniak, Eugene. 2007. "Statistical Techniques for Natural Language Parsing". *AI Magazine* 18(4):33-44.

[2] Hans van Halteren, JakubZavrel, WalterDaelemans. 2004. Improving Accuracy in NLP Through Combination of Machine Learning Systems. *Computational Linguistics*. 27(2): 199-229.

[3] Praveen Kumar, Shrikant Kashyap, Ankush Mittal and Sumit Gupta, "A Query Answering System for E-Learning Hindi Documents", SOUTH ASIAN LANGUAGE REVIEW VOL.XIII, Nos 1&2, January-June,2003.

[4] Nguyen Tuan Dang, Do Thi Thanh Tuyen, "Document Retrieval Based on Question Answering System", Second International Conference on Information and Computing Science, IEEE2009.

[5] Ha-Thuc, Duc-Cuong Nguyen, PadminiSrinivasan "A Quality-Threshold Data Summarization Algorithm" from ieee international conference on research, innovation and vision for the future in computing & communication technologies, rivf 2008, ho chi minh city, vietnam, 13-17 july 2008. ieee 2008.

[6] [5] Wendel, P., Ghanem, M., Guo, Y., "Scalable clustering on the data grid", In Proceedings of 5th IEEE International Symposium Cluster Computing and the Grid (CCGrid), 2005.

[7] Hore, P., Hall, L. O. "Scalable clustering: a distributed approach", In Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2004.

[8] Ping Cai, Liang He "Weighted Information Retrieval Algorithms for Onsite Object Service" Proceedings of the International Multi-Conference On Computing in the Global Information Technology (ICCGI'07), 2007.

[9] Ramakrishna Varadarajan and VagelisHristidis, "A system for query-specific document summarization," in CIKM '06: Proceedings of the ACM conference on Information and knowledge management, 2006, pp. 622–631.

[10] Sarker, M.Z.H.; Parvez, M.S., "A Cost Effective Symmetric Key Cryptographic Algorithm for Small Amount of Data"**,** 9th International Multitopic Conference, IEEE INMIC2005, pages:1-6.

[11] R.Satheesh Kumar, E.Pradeep, K.Naveen, R.Gunasekaran, "Enhanced cost Effective Symmetric Key Cryptographic Algorithm for Small Amount of Data", International Conference on Signal Acquisition and Processing, IEEE2010.

[12] Gilberg, R.; Forouzan, B. (2005), "8", "Data Structures: A Pseudocode Approach With C++", Pacific Grove, CA: Brooks/Cole, ISBN 0-534-95216-X

[13] Heger, Dominique A. (2004), "A Disquisition on The Performance Behavior of Binary Search Tree Data Structures", *European Journal for the Informatics Professional***5** (5)

[14] C.R.Aragon and R.G.Seidel, "Randomized search trees", Proc. 30th IEEE FOCS (2000), 540-545

[15] Dr.Saraswathy, Asma siddhiqaa, Kalaimagal, Kalaiyarasi, "Bilingual Information Retrieval System for English and Tamil" Journal of Computing, Volume 2, Issue 4, Aprial 2010.

[16] 16. P. Glenisson, P. Antal, J. Mathys, Y. Moreau, B. De Moor, "*Evaluation of the Vector Space Representation in Text-Based Gene Clustering*", Pacific Symposium on Biocomputing 8 pp391-402, 2003.

[17] 17. P. Glenisson, P. Antal, J. Mathys, Y. Moreau, B. De Moor, "*Evaluation of the Vector Space Representation in Text-Based Gene Clustering*", Pacific Symposium on Biocomputing 8 pp391-402, 2003.

[18] A. Lakhotia, "Understanding Someone Else's Code: An Analsis of Experience," *The Journal of Systems and Software*, vol. 23, pp. 269-275, 1993.