Journal of Computer Science 2 (3): 224-228, 2006
ISSN 1549-3636

A New Indexing Technique for Information Retrieval Systems Using Rhetorical Structure Theory (RST)

Muhammad Shoaib and Abad Ali Shah
Department of Computer Science and Engineering
University of Engineering and Technology, Lahore, Pakistan

Abstract: Effective information retrieval requires an efficient indexing technique. With the availability of huge volume of information, it has become necessary to capture the semantic of the document, which is almost impossible with the existing techniques. Moreover in the existing technique the weights once assign are remains unchangeable through out the cycle. In this paper an indexing technique assignment using Rhetorical Structure Theory with dynamic weight assignment technique has been presented. The nodes of Rhetorical Parsing tree contain relations and text spans which can be used for indexing by indexer. The results are promising for different texts. Enhancing the technique of NLP can improve the proposed algorithm to accommodate more relations and huge documents.

Key words: Indexing, RST, dynamic weight assignment

INTRODUCTION

Organizing the text documents based on their contents is called indexing. Indexing is an important process in an Information Retrieval System. Indexing has three primary purposes in Information Retrieval[1].

* Access to easy location of document by topics
* To relate one document to another by defining topic areas
* To indicate the relevant document for a specific information.

So any index created must be evaluated that up to what level it satisfies the above-mentioned purpose. In the past the indexing has been done by manually by some trained person. These trained persons were considered to be familiar with the topic of text. An uncontrolled indexing language was generally used which permits the indexer more flexibility in document description. The main problem considers for manual indexing are lack of consistency[2-6], Exhaustively[1], Specificity[1], indexer-user-mismatch[7] etc.

With the increase in electronic texts online, the problem with manual indexing has been increased such as it is too slow and expensive. Due to this, need of automatic indexing was considered. It was Luhn[8] who first suggested that certain words could be automatically extracted from texts to represent their content.

Automatic text indexing is much faster and ratio of errors is low. Retrieval effectiveness of automatic indexing is much better than manual[9]. Many automatic-indexing techniques have been developed for retrieval system on the web[1-3]. Besides all these efforts it has been established that Precision is only 30%[10]. Early automated indexing technique were keyword based. These keywords are believed to express that documents. These keywords are usually assigned weights. Usually some IR Model like Extended Boolean, Vector based, probabilistic etc[5-7] are used to assign the weights. This technique suffers drawbacks like return of small amount of relevant information and lacking of semantic information. The weight assigning technique is static which has its own limitations[11]. Thus it is necessary that more semantic information must be captured to increase the performance and weight assignment should be dynamic.

The paper presents an indexing technique with dynamic weight assignment using Rhetorical Structure Theory (RST)[12], the theory of computational and linguistics. The technique presented is keyword as well as relation based. Precision rate has been improved with the help of RST.

RHETORICAL STRUCTURE THEORY

Mann & Thompson developed Rhetorical Structure Theory (RST). They indicate the existence of twenty-five relations. In Table 1, we give some of the RST relationships (other details can be seen in[12]). The relations can relate parts and sub-parts of a text. The text semantics can be captured from these relations. RST is a linguistically useful method for describing text documents and characterizing their structure. It explains a range of possibilities of structure by comparing various kinds of "building blocks" that can be observed in text documents. Using this theory, two spans of text

Corresponding Author: Mr. Muhammad Shoaib, Department of Computer Science & Engineering, University of Engineering and Technology, Lahore, Pakistan, Tel: +092+333-4214364

(adjacent in most cases, but exceptions can be found) are related such that one of them has a specific role relative to the other. For example, an evidence for the claim follows a claim. The claim spans a *nucleus* and the evidence spans a *satellite*. The order of these spans is not constrained, but there are more likely and less likely orders for all of the RST relationships. A general format of a RST relationship and its two spans are shown in Figure 1.

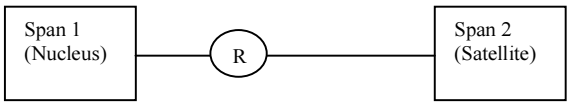


Fig. 1: General view of RST relationships between its two spans

Table 1: Some common RST relationships and their spans

Relationship Name	Nucleus	Satellite
Contrast	One alternative	The other alternative
Elaboration	Basic information	Additional information
Background	Text who's understands is being facilitated.	Text for facilitating understanding
Preparation	Text to be presented	Text, which prepares the reader to expect and interpret the text to be presented.
Antithesis	Ideas favored by the author	Ideas disfavored by the author
Circumstance	Text expressing the events or ideas occurring in the interpretative context	An interpretative context of situation or time
Condition	Action or situation resulting from the occurrence of the conditioning situation	Conditioning situation

Due to the ability of RST to define coherence relations very formally and elaborately makes motivate to develop an algorithm for reorganizations of relations and to use these relations for Indexing. The system developed on the basis of these relations will be able to capture the semantic of the documents.

Previous work: The substantial evidences show that the first automatic indexing system was SMART[13]. SMART was initiated at Harvard University in 1961.The first generation of SMART system was developed in the early 1970.The basic design of SMART was based on the use of various kinds of stored dictionaries, word suffix lists, Phrase tables, and hierarchical term arrangements [3,4]. The relevance feedback methodologies [6] were introduces in SMART along with other retrieval methodologies.

SMART lead to advances in other aspects of automated text manipulation, like new retrieval models generation of new automatic indexing methods, term weight etc.

SMART was unable to retrieve the semantics information of documents. From SMART till now, this time many attempts have been done to improve indexing techniques and overall retrieval effectiveness of information Retrieval Systems by using statistics and probability theory, logic, computational linguistics and various aspects of artificial intelligence. However no references has been found for RST based indexing with dynamic weight assignment technique.

THE PROPOSED INDEXING TECHNIQUE

It has already been mentioned in Introduction Section that the existing indexing technique suffer from a lot of problems like retrieval of irrelevant information and missing of capturing the semantic information. The proposed indexing technique first time in the history is presenting concept of indexing by using Rhetorical Structure Theory (RST) in which we can query the data by using keyword and also the rhetorical relations. The processing of indexing by using RST is complex and requires certain other steps.

It involves text segmentation, rhetorical relation finding, and rhetorical parsing tree. All these steps with the proposed algorithm have already been presented in the previous papers [14-16].

The text was broken into small segments [14] on the basis of Cue phrases and Punctuations. These obtained segments were passed to the relation finder [15]. The relation finder algorithm uses the technique of Natural Language processing, cue phrases, and punctuations for finding the relations present in discourse.

The obtained relations were then used for the construction of Rhetorical Parsing Tree [16]. The nodes of the tree contains the relations and the text spans. The concept of the strong node and week node [15] was introduced to Asses the initial weights at this stage. This initial weights assessment basically enables us to the make the assignment of weights dynamic and its implementation is given in the proposed algorithm.

This initial weight is between 0 and 1.We assigned the root node 1 and assign 0.9 to the nucleus and 0.5 to the satellite of parent. These values of weights can be changeable. The weight assigned to the child nodes is calculated on the basis of the following formula.

Initial weight of child node =
*Weight of the child node * weight of the parent node*

If one node has two relationships then parent value is assigned to them. Otherwise the weight of all the children is calculated. This weight is attached with the index terms obtained from the text spans. The actual weight is assigned to the index terms on the basis of initial weight assessment and term frequency.

Its formula is as follows:

Actual weight of the index term =
*Initial weight assessment * term frequency.*

Indexer takes document id, vocabulary id and weight and maintains the knowledge base. It takes the document id as an input for determining which word exist in which document and takes vocabulary id because knowledge base contains collections of words. And id is assigned to words so that redundancy doesn't

occur and less space is consumed. Weight is semantic based and shows the occurrence of the important index terms on the basis of semantics in the document. Knowledge base contains document vocabulary, and dynamic weight assessment in normalized form.

PROPOSED SOLUTION FOR KEY WORD BASED INDEXING

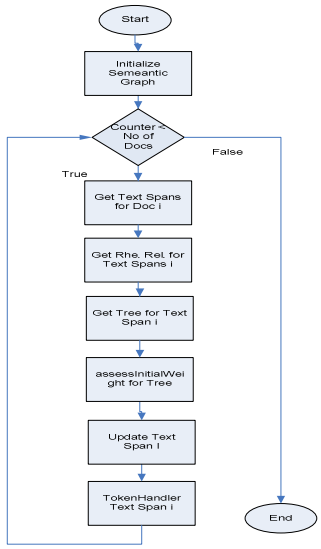
The proposed algorithm works as under:
Procedure Indexer takes input collection of Documents and uses a different utility procedures getTextSpans, which extract the text spans from the collection, getRelations which hypothesis the rhetorical relations while makeTree is utilized to make Rhetorical Text Tree for the respective document which in tern calls assessInitialWeight to get initial weight assessment to get initial weight assessed Text Spans. These Text Spans are handled by tokenHalder which manipulates the knowledgebase.

The proposed solution is using the basic database structure for Knowledgebase representation.

Table 2:

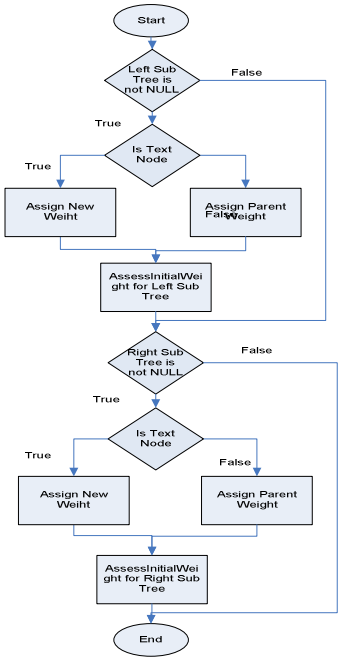
Document Collection	Vocabulary List	Knowledge base
ID	ID	Document ID
Text	Text	Vocabulary ID
Title		Weight

Step 1:
Procedure Indexer takes input collection of Documents and uses a different utility procedures getTextSpans, getRelations, makeTree and assessInitialWeight to get initial weight assessment to get initial weight assessed Text Spans. These Text Spans are handled by tokenHalder which manipulates the knowledgebase.

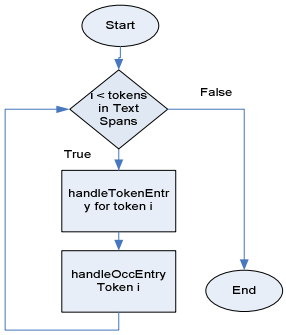


Step2: Procedure assessInitialWeight takes Rhetorical Tree as input along with Nucleous Ration and Stalite Ratio and assigns the initial weight assessment. It works in recursive manners and uses the in-order traversing mechanism.

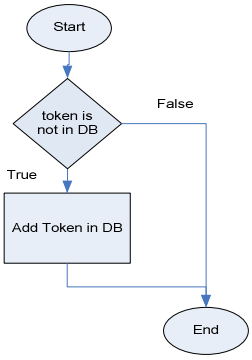
Process Name assessInitialWeight
Output CTreeNode Tree with Initial Weight assessment
Input CTreeNode tr
Float nRatio (Nucleus Ration)
Float sRatio (Satellite Ration)



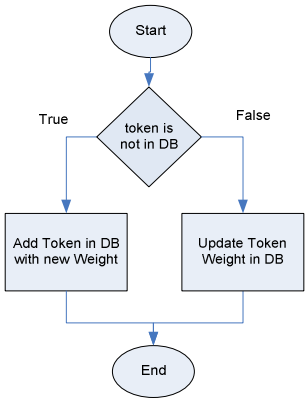
Step3:Procedure tokenHandler takes document ID, Text Span and initial Weight as input and uses handleVocEntry and handleOccEntry procedures to handle Vocabulary and Knowledgebase respectively.
Procedure Name tokenHandler
Output Updated Knowledge-Base with Occurrence and optionally new keyword
Input int did
string st
float weight



Step4: Procedure handleVocEntry adds the new word in Vocabulary List if required and returns the respective Vocabulary ID.
Procdure Name handleVocEntry
OutPut int tokenID
Input string token



Step5:Procedure handleOccEntry adds the new entry in Knowledgebase and also take care of existence of Vocabulary List.
Procedure Name handleOccEntry
Output Updated Knowledgebase.
Input int vid
int did
float weight
float TermFrequency



PROPOSED CONCEPT FOR RELATION BASED INDEXING

To understand the semantic of the document and retrieving only the relevant information, retrieval system that is concerned with the semantic and discourse structure works on the basis of relations the proposed algorithm can follow the following steps

Segmentation: The techniques presented in the paper [14] will be used to segment the text to identify the elementary units in the text.

Relation finder: From these small segments the relations those exist between different parts of text will be identified. The technique has been elaborated in the paper [15].

Parser: A parsing tree consisting of text spans and relations can be built by using the technique presented in the paper [16]. The text spans has been obtained from Segmentation and relation from Relation Finder.

Database for the rhetorical relations and text spans: The obtained text spans and relations will be put into a database. The relational model can be used for this purpose. The tables will be as following

Table 2: Document table

Document ID	Document URL
S1	www.scipub.org

Table 3: Text spans table

Document ID	Span
S1g1	Text Span1
S1g2	Text Span2
S1g3	Text Span3
….	…..
S2g1	Text Span1.
S2g2	.
S2g3	
……	

The obtained relations will be put into the following Table

Table 3: Relations table

Relation ID	ID1 of Text Span1	ID2 of Text Span1	…….
Relation 1	ID Text Span1	ID Text Span1	……
Relation2	ID Text Span2	ID Text Span2	……
Relation3	ID Text Span3	ID Text Span3 ID	……
Relation4	ID Text Span4	Text Span4	……

The relation tables can be manipulated by using SQL. The query can be made to find out the rhetorical relevant documents to the query and search on the relation's table will result a high precision.

CONCLUSION AND FUTURE WORK

A new indexing technique by using RST based on dynamic weight assignment has been presented in this paper, which has been successfully implemented. A concept of the indexing technique using relations has been presented. It is concluded that the system has high degree of precision than the system that use traditional indexing techniques. The algorithm can be enhanced to accommodate other kinds of documents like multimedia, images etc. as well.

The concept of noise word and stemming can improve the efficiency of the proposed algorithm. The proposed concept can be implemented, which will have high precision. We have only considered Boolean Extended model. Research can be carried for certain flexibilities to accommodate other models as well in the proposed algorithm.

REFERENCES

1. Korfhage, R., 1997. Information Storage and Retrieval.

2. Jacoby, J. and V. Slamecka,. Indexer consistency under minimal conditions. Report no. RADC TR 62-426. Documentation, Inc., Bethesda, Maryland, AD-288 087.
3. Cooper, W.S., 1969. Is inter indexer consistency a hobgoblin? American Documentation 20. 3: 268-278.
4. Salton, G., 1969. A comparison between manual and automatic indexing methods. American Documentation 1: 61-71.
5. Preschel, B.M., 1997. Indexer consistency in perception of concepts and choice of terminology. Final report. School of Library Science, Columbia University.
6. Borko, H., 1979. Inter-indexer consistency. Cranfield Conference.
7. Weinberg, B.H., 1987. Why indexing fails the researcher. In Proc. 50th ASIS Annu. Meeting, Boston, pp: 241-244.
8. Luhn, H.P., 1957. A statistical approach to mechanized encoding and searching of literary information. IBM J. Res. and Develop., 1: 309-17.
9. Mirna, A. and W.B. Croft, 1997. Retrieval Effectiveness Of Various Indexing Techniques On Indonesian News Articles.
10. Shah, A. and M. Shoaib, 2005. Sources of irrelevancy in Information Retreival Systems. The 2005 Intl. Multi Conf. in Computer Sci. Computer Engg., USA.
11. Shoaib, M. and A. Shah, 2005. A dynamic weight assignment approach for IR systems. Ist Intl. Conf. Computer and Communication Technol., IEEE, Pakistan.
12. Mann, W. and S. Thompson, 1987. Rhetorical Structure Theory: A Theory of Text Organization. Marina del Rey, CA: Information Sciences Institute.
13. Salton, G., 1971. The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall, Inc..
14. Shoaib, M. and A. Shah, 2005. A Methodology to Segment the Text for Index Terms. Unpublished research paper.
15. Shoaib, M. and A. Shah, 2005. Recognition of Rhetorical Relations in Text using Rhetorical Structure Theory. Unpublished research paper.
16. Shoaib, M. and A. Shah, 2005. A Text Parsing Algorithm for discourse tree in the framework of Rhetorical Structure Theory. Unpublished research paper.