Fourth International Conference on knowledge-Based Intelligent Engineering Systems & Allied Technologies, 30th Aug-1st Sept 2000, Brighton,UK

An Intelligent Agent for Content-Based Indexing and Retrieval of Documents.

N. K. Mimouni[1], F. Marir[2] and F. Meziane[3]
1: School of Computing & Information Systems, Kingston University
2: School of Informatics & Multimedia Technology, North London University
3: Department of Computer Science, University of Salford

Abstract

The amount of information available on the Internet is currently growing at an incredible rate. However, the lack of efficient indexing is still a major barrier to effective information retrieval on the web. This paper presents the design of an intelligent agent for content-based indexing and retrieval of relevant documents from a large collection such as the Internet.
The agent aims at improving the quality of retrieval by capturing the semantics of the documents. It performs the conventional keyword based indexing and introduces a thematic relationship between parts of text using Natural Language Understanding and a linguistics theory called Rhetorical Structure Theory. The agent described in this paper will be implemented and compared against several indexing systems. It is expected to produce a satisfactory improvement over existing techniques.

Introduction and Previous Work

Day by day, the Internet is becoming more accessible, computers are becoming faster, and memory is becoming cheaper. As a consequence, even more documents are placed on the web. The Internet is currently growing at 300% per annum and if it maintains its high development speed then retrieval of relevant information will become more of a crucial issue than what it is today.

A lot of research has gone into developing retrieval systems on the web [1], [2], [3]. Despite all that, using current indexing techniques, it has been reliably estimated that on average only 30% of the returned items are relevant to the user's need, and that 70% of all relevant items in the collection are never returned [4]. These results are far from ideal considering the user is still presented with thousands of documents pertaining to a keyword query in milliseconds. Existing indexing techniques, mainly used by search engines, are keyword based. In other words, each document is represented by a set of meaningful terms (also called descriptors or keywords) that are believed to express its content. These keywords are assigned some weights depending on factors such as their frequency of occurrence (i.e. using Boolean vector based, or probabilistic methods; see [5],[6],[7]). The major drawback to keyword based methods is that they only use a small amount of the information associated with a document as the basis for relevance decisions. As a consequence, irrelevant information that uses a certain word in a different context might be retrieved or information where different words about the desired content are used might be missed. To achieve better performance, more semantic information about the documents needs to be captured. Some attempts at improving the traditional techniques using Natural Language Processing [8], logic [9] and document clustering [10] have offered some improvements.

The aim of the work presented in this paper is to design an intelligent agent for content based indexing which will enhance conventional keyword based techniques with computational and linguistic characteristics found in Rhetorical Structure Theory (RST) [11] and Natural Language Understanding (NLU). The agent will focus on capturing the content of the documents for accurate indexing and retrieval resulting in an enhanced recall. The agent is introduced in the next section followed by the conclusion and future work.

Description of the Agent

As previously pointed out, conventional indexing techniques suffer from the lack of keyword semantics. We propose to use those conventional techniques and introduce the notion of capturing the relationships between units of texts where keywords occur using RST. RST has previously been used for text generation [12], we will use it for text indexing. It represents a refinement to our indexing process because it determines the importance of a term occurrence and therefore excludes irrelevant keywords. After this initial

0-7803-6400-7/00/$10.00 ©2000 IEEE

filtration, a further refinement is the use of NLU to establish the role of the selected terms in a sentence.

With the growing number of the documents available to users and the advance in Internet technology, more robust and reliable document retrieval systems are necessary. There is a need to understand the content of a document, compare it with the meaning of the query and select it only if it is relevant. The intelligent agent presented here is concerned about the semantics, context and the structure of the documents in opposition to single term indexing. It indexes the content of the document through its three phases described below: *keyword extraction*, *capturing the document linguistic structure* and *capturing the role of the selected keyword* in the sentence where it occurs. For a better understanding of the agent's functionality, an article from The Scientific American will be used throughout the explanation of the different phases of the model.

Title: "*Lactose and Lactase*".

Abstract: " *Lactose is milk sugar, the enzyme Lactase breaks it down. For want of Lactase, most adults cannot digest milk. In populations that drink milk, the adults have more Lactase, perhaps through natural selection*".

1- Keyword Extraction

This technique is a basic keyword extraction technique, based on the term's frequency of occurrence. It operates as follows:

- Eliminate common function words form the document texts by consulting a special dictionary, or stop list, containing a list of high-frequency function words such as "*and*", "*or*", "*but*" etc.
- Compute the term frequency (*tf*) for all the remaining terms T in each document D, specifying the number of occurrences of T in D.
- Choose a threshold frequency Th, and assign to each document D all terms T for which tf > Th[2].

Let's take the above example consisting of the text entitled "*Lactose and Lactase*". After performing the initial keyword extraction and the appropriate weight calculations, a table resembling the one in figure 1 would result. According to the table, the document entitled "*Lactose and Lactase*" will be retrieved as an answer to a query about "*milk*" because the word "*milk*" appears as one of its most frequent keywords. This is not the case, the document does not discuss the subject of milk, it discusses topics that are related to milk. Also, this document could be matched against a query about "*adults*" for the same reasons, while it does not discuss a topic specifically related to adult life. This is because single term indexing is not sufficient for representing a document's theme. Hence the need to capture the semantic relationship between parts of text.

Keyword	Weight
milk	0.3
lactase	0.3
adults	0.2
lactose	0.1

Figure 1: Term weighting approaches.

2- Capturing the document linguistic structure

Efficient document structuring goes back as far as Aristotle [13], who recognised that in coherent documents, parts of text can be related in a number of ways. A number of authors have pursued this idea and developed theories to relate sentences. One of them is Rhetorical Structure Theory [11].

RST is a descriptive theory of a major aspect of organisation of natural text. It is a linguistically useful method for describing texts and characterising their structure. RST explains a range of possibilities of structure through comparing various sorts of "building blocks" which can be observed to occur in documents.

Using RST, two spans of text (virtually always adjacent, but exceptions can be found) are related such that one of them has a specific relation relative to the other. Some RST relations are presented in figure 2. A paradigm case is a claim followed by an evidence for the claim. This essentiality is represented by calling the claim span a *nucleus* and the evidence span a *satellite*. The order of spans is not constrained, but there are more likely and less likely orders for all of the relations. Using the relations in figure 2, the RST analysis of a text can be illustrated. The abstract has been broken into numbered units for analysis [14].

The analysis process, presented in figure 3, is intended to give a structured, definite way to state what some of the text content is. The analyst (generally called the observer in RST papers) is saying that the two units that explain the terms lactose and lactase ("*Lactose is milk sugar*" and "*the enzyme Lactase breaks it down*") are intended to facilitate the understanding of the rest of the text.

Relation Name	Nucleus	Satellite
Contrast	One alternative	The other alternative
Elaboration	basic information	Additional information
Background	text whose understanding 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.

Figure 2: Some common relations between spans.

Also, that units (4) (*"For want of Lactase most adults cannot digest milk"*) and (5) (*"In Populations that drink milk the adults have more lactase; Perhaps through natural selection"*) are in a contrast relation.

Now that a brief explanation of RST has been presented, its use in document retrieval is explained in the next section.

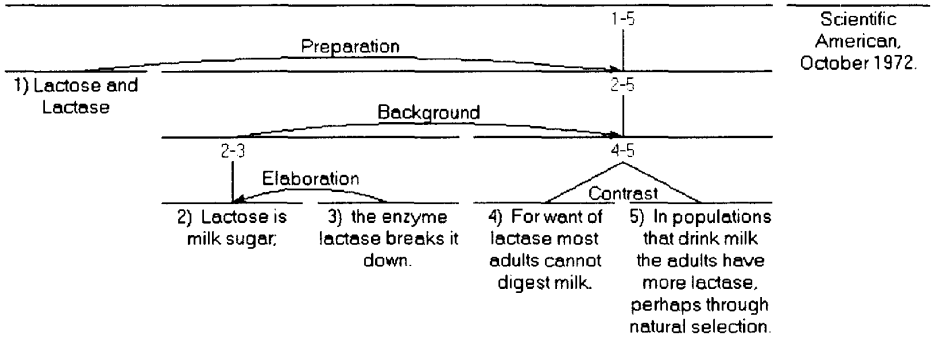

Figure 3: An RST analysis of the *"lactose and Lactase"* text.

RST in document indexing

Using conventional indexing techniques, documents are represented using keywords. Queries are also represented using keywords. For retrieval, a similarity computation is performed between the two sets of keywords and if they are sufficiently similar then the document is retrieved. These methods are term based, so documents that are not relevant but use a certain term are often retrieved. This is partly due to the lack of *semantic relationships* between different parts of texts.

Because RST provides an analysis for any coherent carefully written text, and because such an analysis provides a motivated account of why each element of the text has been included by the author. RST gives an account of textual coherence that is independent of the lexical and grammatical forms of the text. Therefore, it could be used for identifying relationships between the different units where certain keywords appear, stressing the importance of some and disregarding some. This could mean a major refinement to the number of documents retrieved resulting in an enhanced retrieval precision.

Using the RST analysis in figure 3, a preliminary selection excludes the word *"adult"* for the reason that a contrast (and only a contrast) occurs between the sentences where the word appears (units 4 and 5). A preliminary refinement is produced eliminating one keyword[1]. (See figure 4 for details).

However, The word *"milk"* is still identified as a key term because of its frequency of occurrence. The RST analysis reports two relationships between the different occurrences of the word *"milk"*, elaboration and background (see figure 3). Also, *"Lactase"* is still not identified as a major part of the text's topic, it has a relatively low weight so we need to process further. Using RST on its own is indeed an improvement but it is not sufficient to prove the document is not about milk. NLU is needed to clarify the role of terms in sentences.

Keyword	Weight
milk	0.3
lactase	0.3
adult	0
lactose	0.1

[1] This will involve some mathematical calculations.

Figure 4: results of applying RST.

3- Capturing the documents theme

Further investigation landed on Natural Language Understanding techniques [15]. These techniques aim at resolving ambiguity and determining the theme of the text through exploring the roles of certain terms in a text. By applying them to selected pieces of text in opposition to the whole document, considerable processing time is gained. In our example, NLU could be used for the first sentence ("*Lactose is milk sugar*") to confirm that the subject is indeed Lactase and not milk. Hence, the weight given to the word "*milk*" is decreased. Furthermore, NLU techniques are performed on the title identifying the term "*lactase*" as a subject term and its weight is increased. The final table would resemble the one below. The table truly represents the content of the text and is indeed a considerable refinement.

Keyword	Weight
lactase	0.5
milk	0.2
lactose	0.3
adult	0

Figure 5: results of applying NLU.

Conclusion and Future work

In this paper, we have presented an intelligent agent for content-based indexing of documents in a large collection. It introduces an indexing method which is a combination of traditional keyword based techniques, a text description theory called Rhetorical Structure Theory and Natural Language Understanding. RST was created in the late eighties for text generation primarily, we propose to employ it for text indexing. Using RST, keywords and the weights assigned to them could be refined to produce better retrieval precision. Once RST is applied, NLU techniques are used on selected pieces of text (rather than the whole document) for additional refinement. This could involve some pre-processing time during indexing. This time might be saved during retrieval.

At the moment, the project is still at its conceptual level. Future work includes the development of some experiments based on different document topics and sizes. A comparative study on bigger collections will follow. The study will compare the results of the agent to other retrieval systems including the ones used by search engines. As the research progresses, more results will be obtained.

References

1] Pollitt, A. S, *Information Storage and Retrieval Systems*. Ellis Horwood Ltd., Chichester, The UK, 1989.

[2] Salton Gerard, *Automatic Text Processing*, Addison-Wesley, USA, 1989.

[3] Frants, Valery.I, Shapiro, Jacob and Voiskunskii, Vladimir G, *Automated Information Retrieval: theory and Methods*, Academic Press, California, 1997.

[4] Sparck-Jones, Karen & Willet, Peter, *Readings in Information Retrieval*. Morgan Kauffman, California, USA, 1997.

[5] Liu, G.Z. Semantic vector space model: implementation and evaluation. *Journal of the American Society for Information Science*, 48(5), 395-417, 1997.

[6] Korfhage, R. *Information Storage and Retrieval*. John Wiley and Sons, London, 1997.

[7] Losee, R.M. Comparing boolean and probabilistic information retrieval systems across queries and disciplines, *Journal of the American Society for Information Science*, 48(2), 143-156, 1997.

[8] Smeaton, A.F. Progress in the application of natural language processing to information retrieval, *The Computer Journal*, 35, 268-278, 1992.

[9] Lalmas, Mounia and Bruza, Peter, D. The use of logic in information retrieval modelling. *The Knowledge Engineering Review*, 13(3), 263-295, 1998.

[10] Hagen, Eric, *An Information Retrieval System for performing Hierarchical document Clustering*, Thesis, Dartmouth college, 1997

[11] Mann, W.C., & Thompson, S.A. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, **8** (3). 243-28, 1988.

[12] Rosener,D and Stede, M. Customizing RST for the Automatic Production of Technical Manuals. *Lecture Notes in AI*, 587, 199-214, 1992.

[13]Aristotle. *The Rhetoric*, in W. Rhys Roberts (translator), The Rhetoric and Poetics of Aristotle, Random House, New York, 1954.

[14] www.sil.org/linguistics/rst/index.htm

[15]Vadera, Sand Meziane, F. From English to Formal Specifications. *The Computer Journal*, 37(9), 1994