Information Retrieval with the Extra-Topical Information Extracted from Document Titles*

Manuel Montes-y-Gómez [1]
Aurelio López-López [2]
Alexander Gelbukh [1]

[1] CIC, IPN,
Laboratorio de Lenguaje Natural.
Av. Juan de Dios Bátiz, México DF.
Tel. +52 (5) 729-60-00, ext. 56544.
mmontesg@susu.inaoep.mx
gelbukh@pollux.cic.ipn.mx

[2] INAOE, Electrónica.
Luis Enrique Erro No. 1
Tonantzintla, Puebla, 72840 México.
Tel. (52 22) 472-011 Fax (52 22) 470-517
allopez@gisc1.inaoep.mx

Abstract

The document titles and their abstracts are frequently used to express the content of the documents. Many information retrieval systems obtain document keywords from them. We propose to also extract from the titles some extra-topical details about the content of the documents, for instance, the document intentions. We use special information extraction techniques for the identification of document intentions and for the construction of the extra-topical representations of the documents. A possible use for these extra-topical representations in the information retrieval is described.

1 Introduction

Unlike the structured information or formal representations, raw texts have a very complex form. This allows them to describe more completely all entities and facts, but at the same time provokes many of the difficulties in the analysis.

Nowadays, almost every raw text operation, for example, text classification, information retrieval, text indexing and text mining, is done on the basis of keywords (Salton, 1983; Feldman et. al., 1998) or, in the best case, of topics obtained from entire texts of some their parts (Guzmán, 1998). All other text characteristics, those that go beyond topicality, such as intentions, proposes, plans, etc., are usually ignored (López and Myaeng, 1996).

In this paper, we reveal the link between the document title and its author intentions. We also describe a method for the automatic extraction of the document intentions from titles. We also briefly describe how documents intentions are extracted from abstracts (López and Myaeng, 1996) and finally we propose a possible use of this information in the information retrieval process.

* This is a revised version of the paper "Document Title Patterns in Information Retrieval", Proc. of the Workshop on Text, Speech and Dialogue TDS'99, Plzen, Czech Republic, September 1999. Lecture Notes in Artifitial Intelligence 1692, Springer, 1999.

2 Intention Structure

By intention, we mean determination to do something. In this sense, intentions are related with some acts fixed in the document text. They are grammatically associated with some verbs having the main topic of the document as their subjects, such as introduce, describe or propose.

On the basis of these features, the task of determining document intentions consists of finding verbs which actions are performed by the document. For instance, the intention of some document is to *describe* something if there is some evidence in the document body that relate the document with the action *to describe*.

Under our approach, the extraction of the document intention is a little more complex that the simple identification of a verb. For us, intentions are more than mere actions stated, they additionally include an object of the action and sometimes more pieces of related information. For instance, it is not sufficient to say that the intention of some document is to describe. It is also necessary to indicate what is to be described, that is, the object of the verb, as well as how, when, or why this action is done.

3 Titles and Intentions

A title is the part of the document most heavily used for such tasks as indexing and classification. Just this prompts us to use titles for extraction of the intentions. We can note the following facts about the relation between titles and intentions (López and Montes, 1998; Montes-y-Gómez et. al., 1999):

- Intentions are associated with a noun pattern:
  - A noun is followed by a preposition *of* or *to* in the beginning of the title, for instance: An *Introduction to* a Machine-Independent Data Division.
  - A substantive coordinated group is followed by a preposition *of* or *to*, for instance: *Implementation, evaluation, and refinement of* manual SDI service.
  - The case is similar to the previous, but with a *dash* instead of conjunction. Computer *simulation - discussion of* the technique …
- Intentions are related to some gerund patterns:
  - A *gerund* is at the beginning of the title, for instance: *Proving* theorems by recognition.
  - The sequence *adjective - gerund* starts the title, for instance: *Automatic indexing* and generation of classification systems.
  - *Prepositional group with gerund* is anywhere except at the end, for instance: A language *for modeling* and simulating dynamic systems.

4 The Extraction Process

The system we developed follows a common scheme of the information extraction systems (Cowie and Lehnert**,** 1996). It contains a tagger, a filtering component, a parser,

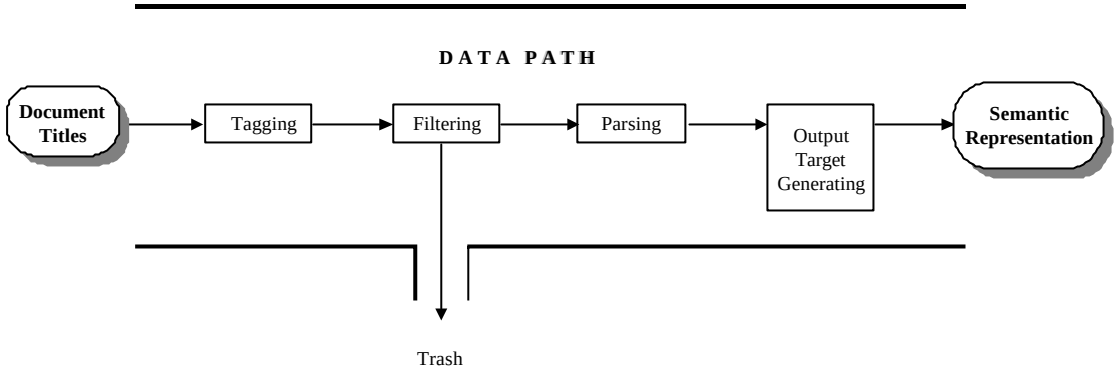

Figure 1. Intention Extraction System

and a module of generation of the output data. Figure 1 shows the general architecture of the intention extraction system.

As an example, let us process the title Algebraic Formulation of Flow Diagrams.
In the tagging module, each word is supplied with a syntactic-role tag. The Tags we are using are based on the Penn Treebank Tagset.

Algebraic|JJ formulation|NN of|IN flow|NN diagrams|NNS|$

The next component selects only the titles containing some information about intentions, that is, it selects the titles containing one of the patterns previously described. Then the chosen titles are parsed[1] and their structured representation is formed (Strzalkowski, 1992).

[[np,[n,[formulation,sg]],[adj,[algebraic]],[of,[np,[n,[diagram,pl]],[n_pos,[np,[n,[flow ,sg]]]]]]],'.']

This representation is entered to the last component, i.e., the output generator, where the structured representation of the document is transformed into a conceptual graph (Sowa, 1983). In this representation, i.e. conceptual graphs, the concepts mentioned in the title and some of their relations are described.

[flow-diagram,{}] ¬ (obj) ¬ [formulate] ® (manr) ® [algebraically]*

5 Abstracts and Intentions

The abstracts of technical and scientific texts often show a level of discourse that complements the expression of the central topic. This level is called metadiscourse, and describes the content of the text, the way the content is elaborate in the document or the participants of the communication, i.e. author and readers (Crismore, 1984).

One particular type of metadiscourse common in technical papers is the *preplan informational metadiscurse* (PIM), which are those global preliminary statements about content structure. Typically, these statements use the text as subject of the central speech

[1] The parser we are using was created in the New York University by Tomek Strzalkowski. It is based on "The Linguist String Proyect (LSP) Grammar" designed by Naomi Sager.

	CACM	CISI
Useful documents	550	252
Automatic extracted titles	512	237
Useful titles extracted by the automatic process	505	229
Useful titles ignored by the automatic process	45	23
Useless titles extracted by the automatic process	7	8
Recall	92%	90%
Precision	98%	96%

Table 1. Statistics of the Extraction System

act. For example, in the statement *this paper presents a method for detecting edges and contours in noisy pictures*, the PIM is *the paper presents.*

In (López and Myaeng, 1996) a method for the extraction of document intentions from PIM sentences is presented. This method takes an abstract, separates the useful sentences (those expressing a PIM statement), and finally transforms each PIM sentence in a conceptual graph. For example, the abstract sentence:

The automatic procedure is describe

Is transformed in the next conceptual graph:

[describe] ® (obj) ® [procedure:#] ® (attr) ® [automatic]

6 Experimental Results

The intention extraction process was tested on a collection of 4663 documents. Manual evaluations gave 802 useful documents (17.2\% of their total number), while our system 738 (15.7\%). Table 1 shows some statistics.

The low percentage of the documents that can be processed by our method does not mean its low usefulness. The described method of intention extraction from titles is to be used together with our method of intention extraction from abstracts (López and Myaeng, 1996). As it is shown in figure 2, the two methods work on nearly complementary

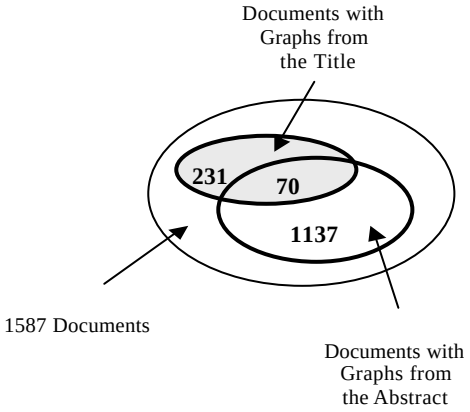

Figure 2. Intentions from titles and abstracts

distributed sets of documents and together cover up to 90\% of the collection.[2]

7 Uses in Information Retrieval

Nowadays, with the electronic information explosion caused by the Internet, increasingly diverse information is available. To handle and use such great amount of information, better search engines are necessary. The more information about documents is preserved in their formal representation used for information retrieval, the better the documents can be selected or evaluated.

Based on these ideas, we are developing a new information retrieval system. This system performs the document selection taking into account two different levels of document representation.

The first level is the traditional keyword document representation. It serves to select all documents potentially related to the topic(s) mentioned in the user's query. The second level is formed with the conceptual graphs reflecting some document details, particularly the document intentions. It complements the topical information about the documents and provides a different way to evaluate the relevance of the document for the query.

Figure 3 shows the general architecture of our information retrieval system with two-level document selection. In this system, the query-processing module analyses the query and extracts from it a list of topics (keywords). The keyword search finds all relevant documents for such a keyword-only query. Then, the information extraction module constructs the conceptual graphs of the query and the retrieved documents, according to the process described in sections 4 (López and Montes, 1998; Montes-y-Gómez et al., 1999) and in section 5 (López and Myaeng, 1996). Then the document intention searched by the system-user and expressed in the query graph is compared with the intention graphs of the documents and finally, the documents are ordered by their value similarity with the query, that is, documents with the searched intention are first presented to the user.

This method of information retrieval is somewhat slower than the traditional search

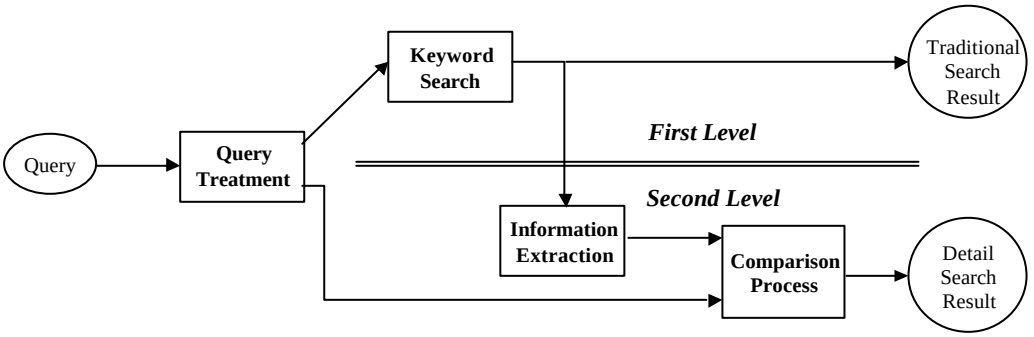

Figure 3. Information retrieval at two levels.

[2] This porcentage was obtained from the analysis of the documents of the CACM collection containing both titles and abstracts.

method, but it allows to do a different kind of search (based in extra-topical details) and also improves the precision of the results.

Conclusions

With this article and with some other previous works (López and Myaeng, 199; Montes-y-Gómez et. al., 1999; López and Montes, 1998), we try to break down the keyword representation paradigm and begin to use other document characteristics, mainly extra-topical details of texts.

In this paper, we described the relations between the document titles and the document intentions, and demonstrated how these intentions are reflected in titles.

We presented a method to extract automatically the document intentions from the document titles, and also explained how document intentions can be also extracted from the abstracts.

Our test demonstrated that our method of extraction of document intentions from titles has good precision and recall, and that joining the information coming from the titles and the abstracts a good coverage of the documents is reached.

Finally we proposed an application for the document intentions in the framework of information retrieval, and explained how this kind of extra-topical information (intentions) allows retrieving documents in a more restrictive form.

References

[Cowie & Lehnert, 1996] Jim Cowie and Wendy Lehnert, Information Extraction, Communications of the ACM, Vol.39, No.1, January 1996.

[Crismore, 1984] A. Crismore, The Rethoric of Textbooks: Metadiscourse, Journal of Curriculum Studies, Vol. 16, Num. 3, 1984.

[Feldman et. al., 1998] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, O. Zamir, Text Mining at the Term Level, Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Nantes, France, September 23-26, 1998.

[Guzmán, 1998] Adolfo Guzmán, Finding the main Themes in a Spanish Document, Expert Systems with Applications 14, pp 139-148, 1998.

[López-López and Myaeng, 1995] Aurelio López López, and Sung H. Myaeng, Extending the Capabilities of Retrieval Systems by a Two Level Representation of Content, Proceedings of the 1st Australian Document Computing Symposium, 1995.

[López and Montes, 1998] Aurelio López López and Manuel Montes y Gómez, Nominalization in titles: A Way to Extract Document Details, Memorias del Simposium Internacional de Computación CIC'98, México, D.F., 1998.

[Montes-y-Gómez et al, 1999] Manuel Montes-y-Gómez, Alexandre F. Gelbukh and Aurelio López-López, Extraction of Document Intentions from Titles, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications (IJCAI-99), Stockholm, Sweden, August 2, 1999.

[Salton, 1983] Gerald Salton, Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[Sowa, 1982] John F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, 1983.

[Strzalkowski, 1992] Strzalkowski T., TTP: A fast and Robust Parser for Natural Language, PROTEUS, Project memorandum #43-A, March, 1992.