Contextual Exploration of Text Collections

M. Montes-y-Gómez, M. Pérez-Coutiño,
L. Villaseñor-Pineda, A. López-López.

Laboratorio de Tecnologías del Lenguaje, INAOE, Mexico.
{mmontesg,mapco,villasen,allopez}@inaoep.mx

Abstract. Nowadays there is a large amount of digital texts available for every purpose. New flexible and robust approaches are necessary for their access and analysis. This paper proposes a text exploration scheme based on hypertext, which incorporates some elements from information retrieval and text mining in order to transform the blind navigation of the hypertext into a step-by-step informed exploration. The proposed scheme is of relevance since it integrates three basic exploration functionalities, i.e. access, navigation and analysis. The paper also presents some preliminary results on the generation of hypertext from two text collections in an implementation of the scheme.

Keywords: automatic text processing, information retrieval, hypertext, text mining, metadata, and information visualization.

1 Introduction

Nowadays there is a large amount of digital texts accessible from private collections as well as from the web. However, without the proper methods for its access and analysis, all this textual data is practically useless. In order to solve this dilemma several text-exploration approaches have emerged. Three popular examples are: information retrieval, hypertext and text mining.

Information retrieval [1] addresses the problems associated with retrieval of documents from a collection in response to a user query. The goal of an information retrieval system is to search a text collection and return as result a subset of documents ordered by decreasing likelihood of being relevant to the given query.

Hypertext [8] is a general manual medium for textual exploration. Its navigational interface, browsing facility, and its graph structure allow users to handle information easily. In a hypertext system, a user explores a text collection following the links among the documents, reading their content and extracting the desired information.

Text mining [5] is concerned with the automatic discovery of interesting patterns, such as clusters, associations and deviations, from text collections. Text mining is intended for analysis tasks rather than to facilitate access. However, some of its techniques can be used as a complement for accessing large text collections.

These three text-exploration approaches are different but complementary. On one hand, information retrieval is a robust and fast approach for *information access*. However, its results are non-explicitly inter-connected and thus they can only be explored sequentially. On the other hand, hypertexts are specifically designed for non-sequential *navegation of texts collections*, but this navigation is blind (there is no precise information about the link nature or information about the document relevance

to the user information need) and the user frequently gets lost in the hyperspace. Finally, text mining techniques, in particular document clustering and association discovery [2,3], support a pattern-based browsing of the text collections. Although these techniques allow the *content analysis* of the text collections, they are difficult to incorporate on exploration situations where the processing of information is done on the fly.

This paper proposes a new approach for text exploration. This approach, named *contextual exploration*, is primarily based on hypertext, but incorporates some elements from information retrieval and text mining in order to transform the blind navigation of the hypertext in a step-by-step informed exploration. In this way, contextual exploration is a *powerful and complete* text-exploration approach that integrates the three basic functionalities for this purpose: access, navigation and analysis.

The rest of the paper is organized as follows. Section 2 introduces the concept of contextual exploration. Section 3 presents the exploration scheme, and describes their main components. Section 4 discusses some experimental results on the hypertext generation and information visualization. Finally, section 5 exposes our conclusions and future work.

2 Contextual exploration

Hypertext is one of the most popular approaches for exploring text collections. Its graph structure (where the nodes represent documents and the edges indicate some relationships among them) allows users to handle information easily. This approach models the exploration of a text collection as a *graph traversing procedure*. Therefore, in order to explore a text collection, a user must take a document as starting point, follow the links among the documents, assess their content, and hopefully extract the desired information.

Hypertext, as defined above, seems to be a general, flexible and easy-way to explore text collections. However, navigating through hypertext frequently leads to the problem of *getting lost in the hyperspace*, i.e. knowing where you are in the hypergraph and knowing how to get to the place you are actually looking for [6]. In order to ameliorate this problem we introduce the idea of contextual exploration.

The *contextual exploration* is a hypertext navigational scheme that includes complementary information for each document (node) that allows evaluating the relevance of its content against the entire collection, as well as, the content of their associated (linked) documents. Basically, it considers the following three *contexts* of information (see the figure 1):

1. *Self context*, that includes some metadata from the document currently being displayed (for instance, date, author, and main topics) in addition to the actual document.
2. *Near context*, consisting of a set of related documents with their corresponding metadata. Its purpose is as an aid to understand the nature of each relation and the relevance of each related document to the user information need.
3. *Far context*, represented by a set of pertinent topic associations (i.e., global co-occurrence relations among the topics of the current document). Its aim is to pro-

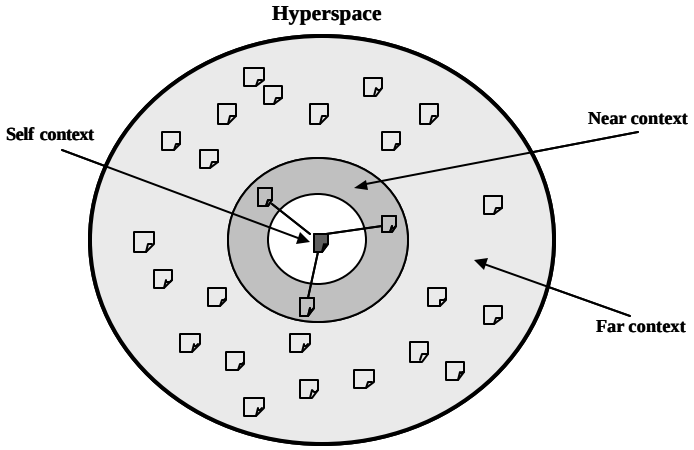

Figure 1. Three information contexts about the current document

vide a general mechanism to estimate the importance of the content of the current document in the whole collection.

Figure 3, in section 4, shows a snapshot of the interface of our system for contextual navigation of text collections. This snapshot clarifies the way in which the three kinds of contextual information are integrated.

3 Scheme description

This section describes the proposed scheme from two different perspectives. The subsection 3.1 briefly describes the *functionality* of the scheme, while the rest of the subsection discusses the applied *methods* for each one of the components.

3.1 Functional overview

The system performs two kinds of processes: off-line and on-line. The goal of the *off-line processes* is to generate a set of intermediate document representations containing information from different context levels from a given text collection (as explained in section 2). On the other hand, the *on-line processes* use these representations for two different purposes. First, to filter the information that satisfies the user query, and second, to provide the user the search results in the form of a hypertext.

The *hypertext* assembled as an answer to the user query, not only considers the content of the documents found, but also some descriptive information in the form of metadata about them (i.e., self context information) as well as a list of related documents for each one of them (near context information). Additionally, it includes information about some topic associations (far context information), which are pertinent (directly related) to the document at hand.

Figure 2 illustrates both on-line and off-line processes of the system. The components in the scheme can be extended in order to get information from more contexts. The subsequent subsections describe the goals and functionality of the four

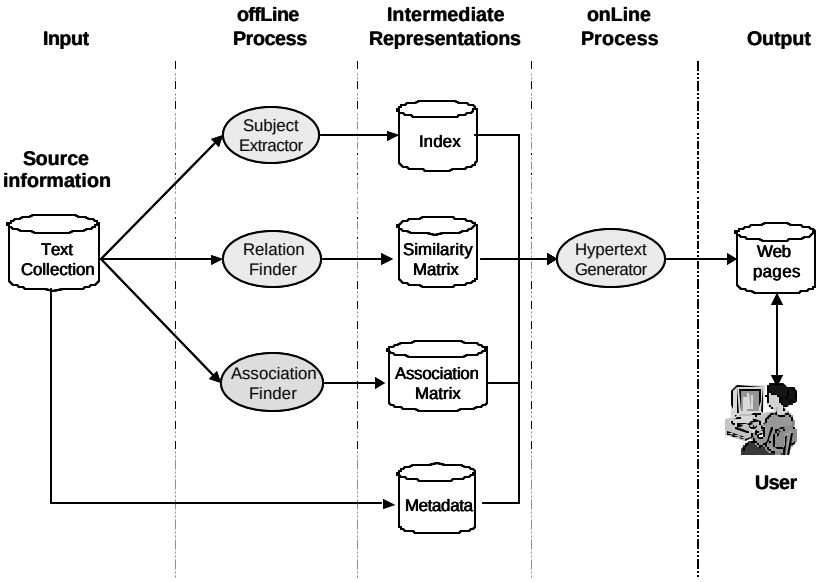

Figure 2. Scheme overview

main components: the subject extractor, the relation finder, the association finder and the hypertext generator.

3.2 Subject Extractor

This component has two main tasks: to identify the candidate topics for each document of a given collection, and to build a representation of their content.

In order to identify the set of topics of a document, this extractor uses a method similar to that proposed by Gay and Croft [4], where the topics are related to noun strings. Basically, this component applies a set of heuristic rules specific for Spanish, based on the proximity of words that allows identifying and extracting key phrases. These rules are driven by the occurrence of articles and the preposition *de* ('of') along with nouns or proper names. Some morphological inflection patterns (typical endings of nouns and verbs) are also taken into account. For instance, given the paragraph below, the subject extractor component selects the underlined words as candidate topics:

"*Góngora Pimentel* aseguró que estas *demandas* se resolverán en un *plazo* no mayor de 30 días y que sin duda la *demanda interpuesta* por el *PRD* ante la *Suprema Corte de Justicia* se anexará a la que presentó el *Partido Acción Nacional*".[1]

Then, based on the candidate topics, this component builds an enriched representation of the documents. This representation is expressed as a weighted vector of topics

[1] '*Góngora Pimentel* confirmed that these *demands* will be satisfied in a *period* not longer than 30 days and that without any doubt the *demand introduced* by the *PRD* to the *Justice Supreme Court* will be added to that presented by the *National Action Party*.'

in a given *n*-dimensional vector space. That is, for a given collection of documents $D=\{d_i\}$, with a corresponding set of topics $\{t_1,\ldots,t_n\}$, the new document representation is formally expressed as follows:

$$d_i \rightarrow \vec{d}_i = (w_i(t_1), w_i(t_2),\ldots, w_i(t_n)), \text{ where :}$$

$$w_i(t_j) = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}$$

In these formulas, $w_i(t_j)$ is the normalized weight of the topic j in the document i; f_{ij} is the number of occurrences of the topic j in the document i; and n is the number of topics in the whole collection.

3.3 Relation Finder

The goal of the relation finder component is to identify the most significant inter-document relations. Basically, this component finds the set of thematically related documents for each item of the given source collection.

In order to accomplish its goal, the relation finder component computes the similarity for every pair of documents in the source collection, and then determines the most important connections.

The similarity measure used is based on the Dice coefficient [7]:

$$s(d_i, d_j) = s_{ij} = \frac{1}{2} \sum_{\forall t \in d_i \cap d_j} w_i(t) + w_j(t)$$

Here, the topic $t \in d_i \cap d_j$ is a common topic of both documents d_i and d_j, and $w_k(t)$ indicates the weight of the topic t in the document d_k.

The criteria used to determine the set of related items associated to the document d_i, after computing all the similarities, is the following:

$$R_i = \{d_j \mid s_{ij} \geq s_m,\ j \neq i\}, \text{ where :}$$

$$s_m = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=i+1 \\ s_{ij}>0}}^{N} s_{ij}$$

Here, R_i is the set of thematically related documents for the document d_i, s_{ij} is the similarity measure of documents d_i and d_j, and N is the number of documents in the whole collection. Basing this criterion on the average similarity among documents allows producing an associated set of items, independently of how homogeneous is the collection. That is, even in highly heterogeneous collection (a very diverse set of topics), we can obtain existing relations.

3.4 Association Finder

This component focuses on the discovery of interesting topic associations between pairs of documents in a given text collection. We define a topic association as an expression $t_i \Rightarrow t_j$, where t_i and t_j are two different topics from the collection. This kind of associations indicates that the documents that contain the topic t_i tend to co n-

tain also the topic t_j.

Each topic association $t_i \Rightarrow t_j$ has a confidence value. This value is calculated as follows:

$$c_{ij} = \frac{|Q_{ij}|}{|Q_i|}, \text{ where :}$$
$$Q_{ij} = \{d_k | t_i, t_j \in d_k\}$$
$$Q_i = \{d_k | t_i \in d_k\}$$

Here, $c_{i,j}$ denotes the confidence value of the association $t_i \Rightarrow t_j$, and $Q_{i,j}$ and Q_i the sets of documents containing the topics t_i and t_j, and the topic t_i respectively.

The criterion used to determine the set of pertinent associations to the document d_k, is the following:

$$A_k = \{(t_i \Rightarrow t_j, c_{ij}) | c_{ij} \geq u, t_i \vee t_j \in d_k\}$$

This criterion selects as the set of pertinent associations for the document d_k, those having a confidence value greater than a predefined threshold u, and that include a topic of the document d_k.

3.5 Hypertext Generator

The output of the system is a hypertext document that unifies the information from the three context levels (self, near and far) in a single interface. For the case of the far-context information, i.e., the topic associations, the interface only displays those associations with a confidence value greater than a user-specified threshold u and related to the content of the current document (refer to section 3.4).

The proposed interface is based on a template that fulfills the standard XHTML 1.0 proposed by the World Wide Web Consortium (W3C), and includes the following set of metadata: title, creator, publisher, date, subject and relation. It also contains the source document and a pointer to the document metadata representation that could be later accessed by software agents.

The output corresponding to the example text is showed in the figure 3.

4 Experimental Results

4.1 The test collections

In order to prove the functionality of the proposed system, we analyzed two document collections: *News94 and ExcelNews*. These collections are in raw text format (i.e. ASCII). They differ in their topics and in the document average size. Following, we describe the main characteristics of these collections. More details are in table 1.

Collection News94

News94 is a set of 94 news documents. The average size per document is 3.44 Kb, and the biggest document size is 18 Kb. This collection is a subset of the ExcelNews

data set.

Collection ExcelNews

This collection consists of 1,357 documents. These documents contain national and international news from 1998 to 2000 as well as cultural notes about literature, science and technology. The document average size is 3.52 Kb, and the biggest document size is 28 Kb.

An important characteristic of the ExcelNews collection is the variety of writing styles and lexical forms of its documents, causing a large distribution of terms in the vocabulary.

Table 1. Main data of test collections

Collection	Size (Mb)	Number of documents	Average document size	Number of pages	Number of lexical forms	Number of terms
News94	372 Kb	94	3.44 Kb	124	11,562	29,611
ExcelNews	4.81	1357	3.52 Kb	1,642	41,717	391,003

4.2 Results

Table 2 summarizes the results obtained from the preprocessing of the test collections (offline processes). These results consider three main aspects: (1) the topic distribution of the test collections, (2) the required time for their analysis, and (3) the connectivity level of the resulting hypertext document sets. In addition, the table 3 shows some topic associations and their confidence values.

Table 2. Main results from the collection analysis

Collection	Topics	Instances of topics	Indexing time	Searching time	Connected documents	Relations	Average of related documents
News94	2,571	4,874	0''.26	0''.55	90	459	5
ExcelNews	24,298	72,983	3''.56	3'50''.59	1350	47,486	35

Table 3. Sample of topic associations

Recesión (*recession*)→ Estados Unidos (*United States*)	1
Banco Mundial (*World Bank*)→ Fobaproa	1
Neoliberal (*neoliberal*)→ Ernesto Zedillo	0.75
Amor (*love*)→ Novela (*novel*)	0.75
Oriente Medio (*middle east*)→ Estado de Israel (*State of Israel*)	0.66
PNR (*acronim of National Revolutionary Party*)→ Plebiscito (*plebiscite*)	0.66
Tercer Mundo (*Third World*)→ Aliado (*ally*)	0.66
Narcotráfico (*narcotraffic*)→ Estados Unidos (*United States*)	0.66
Naciones Unidas (*United Nations*)→ Guerra (*war*)	0.66

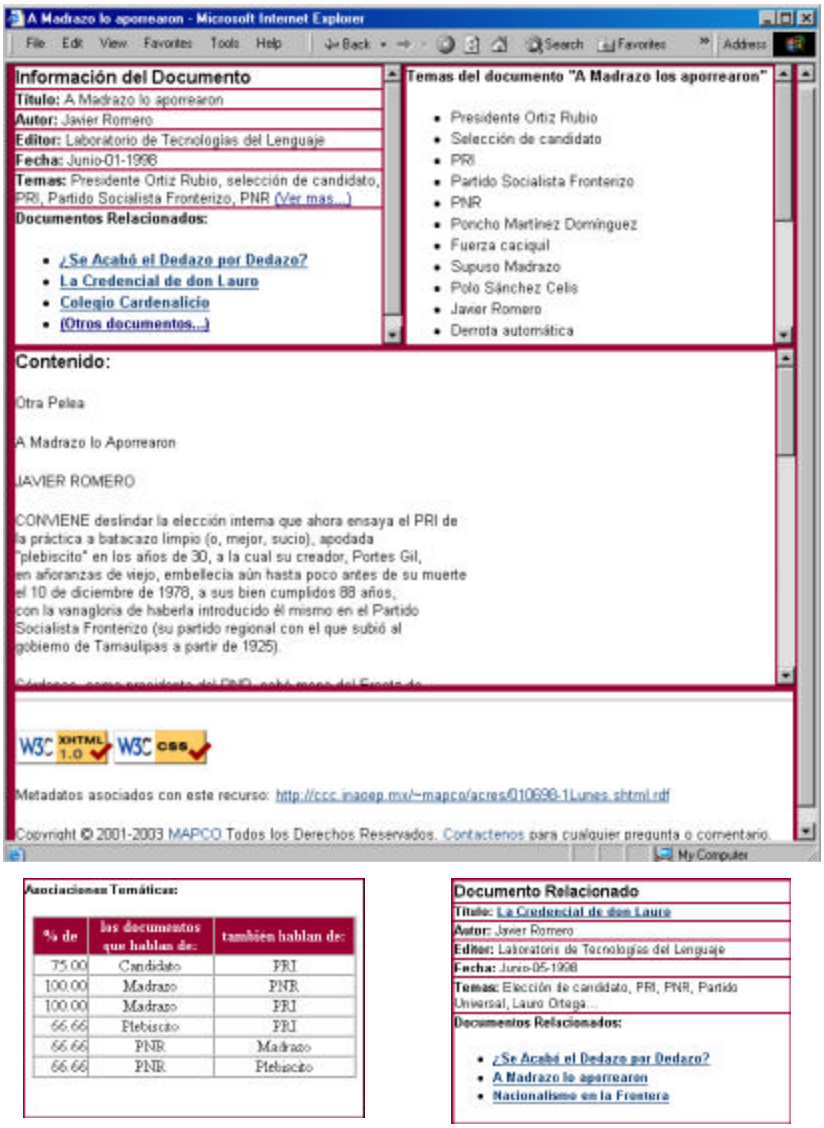

Figure 3. A sample page of hypertext generated, and two different uses of the detail region

Figure 3 shows a sample page of hypertext gathered from the given input collection (in this case, from News94). This interface has three regions. The *content region* shows the complete document content (self-context information). The *metadata region* considers descriptive data from the current document as well as the links to its related documents (self- and near-context information). Finally, the *detail region*

provides additional details about the content of the current document, or the metadata of the related documents, or even the set of pertinent topic associations (near- and far-context information). The second row of the figure 3 illustrates two of these different uses.

Conclusions and future work

Hypertext is a medium for textual exploration typically built by hand. Its navigational interface, browsing facility, and its graph structure allow users to handle information easily. However, navigating through a hypertext frequently leads to the problem of getting lost in the hyperspace.

This paper proposed a new scheme for textual exploration that extents the traditional approach of hypertext. This scheme, called *contextual exploration*, incorporates some elements from information retrieval and text mining in order to transform the blind navigation of the hypertext in a step-by-step *informed navigation*.

The contextual exploration scheme includes complementary context information for each document that allows evaluating the relevance of its content against the entire collection, as well as, the content of their linked documents.

The proposed scheme is of relevance since it integrates three basic exploration functionalities: access, navigation and analysis; in a single interface.

As future work, we plan to evaluate the quality of the generated hypertext and the usability of the system. Currently, we are designing some experiments with several users and different contexts (in particular, we are interested in exploring news about natural disasters).

Acknowledgements. This work has been partly supported by the CONACYT through project grant U39957-Y and scholarship for the second author, and by the Language Technologies Laboratory of INAOE.

References

1. Baeza-Yates and Ribeiro-Neto. *Modern Information Retrieval*, Addison-Wesley, 1999.
2. Cutting, Karger, Pedersen, and Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 1992.
3. Feldman, Klösgen, Yehuda, Kedar and Reznikov. Pattern Based Browsing in Document Collections, *Proc. of the 1st Conference on Principles of Knowledge Discovery and Data Mining (PKDD'97)*, 1997.
4. Gay and Croft. Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management* 26(1): 21-38, 1990.
5. Hearst. Untangling Text Data Mining, *Proc. of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
6. Levene and Loizou. Navigation in hypertext is easy only sometimes. *SIAM Journal on Computing*, 29(3):728-760, 1999.
7. Lin. An Information-Theoretic Definition of Similarity, *Proc. of the International Conference on Machine Learning*, 1998.
8. Shneiderman and Kearsley. *Hipertext Hands On!*, Addison Wesley, 1989.