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 [10] 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 [7] 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 dis-