Textual Document Mining Using a Graphical Interface

G. Hubert(1), J. Mothe(1)(2), A. Benammar(1), T. Dkaki(3), B. Dousset(1), S. Karouach(1)

(1) Institut de Recherche en Informatique de Toulouse, 118 Rte de Narbonne, 31062 Toulouse Cedex 4, France.
(2) Institut Universitaire de Formation des Maîtres de Toulouse, 56 Av. de l’URSS, 31400 Toulouse, France.
(3) IUT Strasbourg Sud, Université R. Schuman, 72 route du Rhin – BP315, 67411 Illkirch-Grafenstaden, France
Tel : 05 61 55 87 - Fax : 05 61 55 62 48
Email : {dkaki/dousset/mothe}@irit.fr

ABSTRACT

In this paper we present an approach in order to help the user extracting advanced information from a document set. The approach is based on the application of different mining modules that collaborate in order to provide graphical views to the user. These views correspond to a graphical summarization on document dimension correlations. To proceed, the documents are first represented according to different concept hierarchies that correspond to the document dimensions. Then, the resulting representation are mined using different modules depending on the user’s goal.

1. INTRODUCTION

The goal of an Information Retrieval System (IRS) is to retrieve the information that responds to the users’ needs from a collection of documents. Most of the IRS displays the results under the form of a list of document references and titles. Then the user browses the returned references and accesses those s/he supposes to be relevant. However, generally the list of returned references is long and it is time consuming to be processed by the users. In addition, the information access (i.e. retrieving the pages that contain the relevant raw information) is not the only goal of the users. The user can be interested in more advanced information as unknown pieces of information and relationships between information. This kind of information (named knowledge) is extracted from the raw information and the user generally discovers it manually while s/he is reading the documents. Some automatic processes have been reported in the literature. One application is the identification of a core set of authors or papers in a given field (White & McCain, 1989) via the cocitation analysis or the determination of the relative authority of Web pages based on the hyper-reference analysis (Kleinberg, 1999). Document classification is another application that can help the user when analyzing a retrieved document set. Post retrieval classification aims at grouping together documents that have been retrieved on the base of the similarity of their content (Hearst, 2000). These applications focuses on a single element of information (document content, citations or hyper-references, authors). However, documents have a wide range of dimensions that can lead to a range of relevant information extracted when analyzing a set of documents (Mothe, 2001). In this paper, we present an approach that takes advantages of different elements of information automatically extracted from the document content. Once extracted, the elements of information are analyzed using different mining components. Each component has its specific aim and the different components collaborate in order to help the user in the information-discovering task.

The paper is organized as follows. Section 2 presents the way documents are represented so that they can be automatically mined. Section 3 and 4 describe the main components of the system and how they co-operate in order to achieve an interactive document analysis.

2. DOCUMENT REPRESENTATION

Information retrieval systems generally represent documents as bags of words/terms, which can be weighted in order to represent the relative importance of those terms. In that case, the terms are automatically extracted from the document contents. Alternatively, a control vocabulary can be used. Terms from the controlled vocabulary are (often manually) assigned to each document. This approach is used in most of the specific digital libraries (Medline, Questel). The same kind of approach is used when documents are assigned to nodes of a concept hierarchy (CH) as it is the case for Web directories such as Yahoo. Whereas these systems use a single hierarchy of terms, our system is based on several ontologies (or CH), each one describing a document dimension. The multidimensionality of the documents is a key point of our approach as it corresponds to the starting point for a data mining process.
In a first phase, the system automatically assigns the documents to the different CH when possible. The nodes to which a document has been assigned correspond to the initial multidimensional document representation. Additionally, we compute a more summarized document representation, which is more adapted to mining processes, under the form of contingency tables.

2.2 Association of documents to the CH

The association of documents to the different CHs is based on Information Extraction principals. When a node from a CH is extracted from the document content, the document is automatically attached to that node. To proceed, templates are defined for each dimension (i.e. for each CH). The extraction process is based on extraction rules, which use syntactic tags from the documents and regular expression matching. Once a syntactic tag has been detected, semantic and filtering functions are used in order to complete the extraction process. Semantic functions take into account synonymy between terms, whereas filtering function eliminate the candidate terms that have been extracted but that do not belong to the corresponding CH.

2.3 Contingency tables

Whereas a bag of words is a representation adapted to information retrieval, it does not fit data mining processes. Contingency tables are a much more adapted representation (Fayyad et al, 1996). Contingency table is a mean to transform non-numerical information into numerical information. A contingency table is obtained by dividing up a population (in our case a set of documents) according to two variables or dimensions, I and J. The columns of the table correspond to the values of the variable J, whereas the lines correspond to the values of I. The intersection T_{ij} of a row i and a column j corresponds to the number of objects in the population for which the variable I has the value i and the variable J has the value j simultaneously. In our approach, a variable corresponds to a CH and the values of the variables are the nodes of the corresponding CH. A contingency table where I and J correspond to the same variable depicts co-occurrence relationships. For example, a contingency table where I and J are author names (one of the CHs) correspond to co-authoring relationships. When I and J are of different natures, some other relationships are depicted. For example, if I represents the geographic reference of the documents (e.g. the country of the author) and J represents the time (e.g. date of publication), the relative contribution of the countries on the field of the documents is depicted along time. In fact the crossing can involve any kind of information, as soon as it corresponds to a CH. Each contingency table corresponds to a 2-D document representation.

3. MINING COMPONENTS

The mining components all use the representation of the documents resulting from the document analysis (i.e. a 2-D contingency tables, where each dimension corresponds to a CH). The components co-operate in order to mine the corresponding document set. Each component is specialized in a specific mining task. In this section we present the main mining components of the system.

3.1 GeoECD

GeoECD visualizes data that are logically geo-referenced under the form of colored maps. Thus GeoECD can only be activated when the document representation includes a reference to a geographic part of the world. The intensity of the color in the obtained maps reflects the relative contribution of the continents, country or groups of countries to the data. This contribution can be weighted using features related to the countries (e.g. population, GNP, surface). Indeed, the visualized data can be either "absolute" contingencies (e.g. number of foot and mouse disease cases reported) or "relative" contingencies (e.g. number of foot and mouse disease cases reported divided by the total number of animals; the later element being a feature of the countries).

3.2 Clustering

The clustering module groups the data according to the agglomerative hierarchical clustering method. The individuals (rows of a 2-D document representation) are clustered according to the variables (columns of the 2-D document representation). Several modules are available, depending on the method used (single, average or
complete linkage). The results are visualized under the form of a dendrogram. This dendrogram can be cut at different levels, depending on the size/consistency of the classes one prefers.

3.3 Factorial analysis

This module mines the data using the correspondence factorial analysis (CFA) defined by (Benzécri, 1973); this method belongs to the family of analysis methods that benefits from on the mathematical properties of the Singular Value Decomposition (SVD) of matrices (Eckart & Young, 1936). The general goal of this class of methods is to represent vectors (the individuals) initially represented in an N dimensional space (variables) into a space with a smaller dimension. The CFA uses the SVD, but instead of being based on the absolute object representation (the contingencies), the SVD is based on the object profiles (similar to probabilities or % of contribution). Another specificity of CFA is the distance measure used to compare profiles. The measure that is used ($\chi^2$) aims at favoring the specificities instead of the too recurrent phenomena. In addition, $\chi^2$ makes it possible to represent the objects and the characteristics in the same reduced space.

To CFA is associated a x-dimensional graphical representation. The first axes resulting from the decomposition (the axes that correspond to the highest values) should be chosen for the representation as they maximize the initial distances between the individuals. In our system, the factorial analysis module uses a 4-D graphical representation.

3.4 LinkMap

LinkMap uses metrics in order to detect links between elements of information and visualizes them in a 4D view. The links are computed using different metrics applied to contingency tables. The gray level used to color the links in the views represents the link’s weight: a white link is weak whereas a black one is very strong.

The classification, factorial methods and map representation are not new; classification have largely been used in IRS; factorial methods are used in many other fields that implies the object analysis and maps are similar to the ones used in Geographic Information Systems. The main original point is that these methods are components of an interactive interface and that they can be combined in order to explore a document set.

4. INTERACTIVE MINING

4.1 The document source

The documents used are scientific publications taken from INRA organisms all over the world. The documents are extracted from publications on biomaterials of the Current Contents database. We treat 2454 documents.

4.2 The document representation

2-D document representations are extracted as described in section 2.3. The dimensions as well as the number of different values for each dimension are given in the following Table 1.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Number of different values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author's country : PA</td>
<td>40</td>
</tr>
<tr>
<td>Publication source : SO</td>
<td>164</td>
</tr>
<tr>
<td>Keywords : KP</td>
<td>923</td>
</tr>
</tbody>
</table>

Table 1. Document dimensions

4.3. Example of analysis

In this example, the document set is mined according to the two following dimensions: the country the authors belongs to (PA) and the publication source (SO). Firstly, the classification module is applied in order to group together countries that have the same behaviour with regard to the publication source. Figure 1 displays the resulting dendrogram. The dendrogram can interactively being cut at any level according to the number/size of classes the user prefers.
The result of this classification can be sent to the GeoECD module as shown Figure 2. In that case, to each class resulting from the clustering module is associated a colour.

Additionally, it is possible to detect what are the countries that have a specific behavior and to extract the reason of these specificities. This is done applying the factorial analysis method on the same document representation. The 4-D representation that results from the factorial analysis is shown Figure 3.
In Figure 3, the journal ACTA-ALIMENTARIA has a specific behavior (far from the center and close to Hungary) which can be explained by the fact that it is a Hungarian journal. The real links that exists between the elements of information can be graphically displayed using the linkMap module (see Figure 4).

In this example, we focused on the United Kingdom’s case. The gray level shows that there are more common journals between authors of the United Kingdom authors and authors of the USA than between authors of the United Kingdom and authors of Japan.

5. CONCLUSION
In this paper, we presented an approach to mine a document set according to the different document dimensions. The document dimensions are automatically extracted with regard to different concept hierarchies that correspond to views the user can have on the document set. This initial document representation is then transformed into 2-D document representations (contingency tables) that behave as input of different mining modules. We describe the main mining modules that are provided to the user in order to help him/her mining the document set. We gave an example on how the different modules can co-operate. The two dimensions used in the example correspond respectively to the country of the author and to the publication source. However, they could have been any of the document dimensions. The same kind of analysis would then have been extracted.

Acknowledgement
Research outlined in this paper is part of the IRAIA project supported by the European Commission under the Fifth Framework Programme (IST-1999-10602). However, in this paper, we described our approach which does not necessarily correspond to the IRAIA consortium view.

REFERENCES