Chapter 24
Using Association Rules for Query Reformulation

Ismaïl Biskri
University of Quebec at Trois-Rivieres, Canada

Louis Rompré
University of Quebec at Montreal, Canada

ABSTRACT

In this paper the authors will present research on the combination of two methods of data mining: text classification and maximal association rules. Text classification has been the focus of interest of many researchers for a long time. However, the results take the form of lists of words (classes) that people often do not know what to do with. The use of maximal association rules induced a number of advantages: (1) the detection of dependencies and correlations between the relevant units of information (words) of different classes, (2) the extraction of hidden knowledge, often relevant, from a large volume of data. The authors will show how this combination can improve the process of information retrieval.

INTRODUCTION

The ever increasing importance of internet penetration and the growing size of electronic documents has made information retrieval a major scientific discipline in computer science, all while access to relevant information has become difficult, having become an informational tide that is occasionally reduced to nothing more than noise.

Information retrieval consists of selecting the documents or segments of text likely to respond to the needs of a user from a document database. This operation is carried out by way of digital tools that are sometimes associated with linguistic tools in order to refine the granularity of the results given certain points of view (Desclés & Djioua, 2009) or logical tools in question-answer format, or even tools proper to Semantic Web. However, we knowingly omit a presentation of the contributions of these linguistic methods, logical methods, and Semantic Web, due to a concern for not weighing down the writing in this chapter, since we are primarily interested in the numerical side.

Formally, there are three main elements that stand out with regards to information retrieval:

DOI: 10.4018/978-1-4666-2455-9.ch024

1. The group of documents.
2. The information needs of the users.
3. The relevance of the documents or segments of text that an information retrieval system returns given the needs expressed by the user.

The last two aspects necessarily rely on the user. Not only does the user define their needs, but they also validate the relevance of the documents returned. To express their needs, a user formulates a query that often (but not always) takes the form of key words submitted to an information retrieval system based either on a Boolean model, a vector model, or a probabilistic model (Boughanem & Savoy, 2008). However, it is often difficult for a user to find key words that allow them to express their exact needs. In many cases, the user is confronted by a lack of knowledge on the subject of interest in their information search on the one hand, and on the other hand, by results that may be biased, as is the case with search engines on the Web. Thus, retrieving relevant documents from the first search is almost impossible. Therefore, there is a need to carry out a reformulation of the query either by using completely different key words, or by expanding the initial query with the addition of new key words (El Amrani et al., 2004).

In the case of expanding the query, two variants are possible:

1. The first is manual. The user chooses terms that are judged relevant in the documents that are also judged relevant in order to strengthen the query. This strategy is simple and computationally costs the least. However, it does not allow for a general view of the group of documents returned by the retrieval system considering their large numbers, and given that it is not humanly possible. Quite often, the user only consults the first few documents, and only judges these few.
2. The second is semi-automatic. The terms added to the initial query are chosen by the user from a thesaurus (which may be constructed manually) or from similarity classes of documents and co-occurrences of terms obtained following a classification applied to a group of documents, obtained following the initial request as in clustering engines. A process of classifying textual data from web sites can help the user of a search engine to better identify the target site or to better formulate a query. Indeed, the lexical units which co-occur with the keywords submitted to the search engine can provide more details concerning the documents to which access is desired. However, the interpretation of similarity classes is a nontrivial exercise. The classes of similarity are usually presented as lists of words that occur together. These lists are often very large and their vocabulary is very noisy.

In this chapter we will show how maximal association rules can improve the semi-automatic reformulation of a query in order to access target documents more quickly.

MAXIMAL ASSOCIATION RULES

A brief survey of the literature on data mining (Amir & Aumann, 2005) teaches us that association rules allow for a representation of regularities in the co-occurrence of data (in the general sense of the term) in transactions, regardless of their nature. Thus, data that regularly appear together are structured in so-called association rules. An association rule is expressed as $X \Longrightarrow Y$. This is read as follows: each time that X is encountered in a transaction, so is Y. There are also ways to measure the quality of these association rules: the measure of Support and the measure of Confidence.

The concept of association rule emerges mainly from the late 60 (Hajek et al., 1966) with the introduction of the concept of the support and the confidence. Interest in this concept was