Chapter 37
Information Retrieval (IR) and Extracting Associative Rules

Asmae Dami
Sultan Moulay Slimane University, Morocco

Mohamed Fakir
Sultan Moulay Slimane University, Morocco

Belaid Bouikhalene
Sultan Moulay Slimane University, Morocco

ABSTRACT

This chapter is located in the intersection of two research themes, namely: Information Retrieval and Knowledge Discovery from texts (Text mining). The purpose of this paper is two-fold: first, it focuses on Information Retrieval (IR) whose purpose is to implement a set of models and systems for selecting a set of documents satisfying user needs in terms of information expressed as a query. An information retrieval system is composed mainly of two processes the representation and retrieval process. The process of representation is called indexing, which allows representation of documents and queries by descriptors, or indexes. These descriptors reflect the contents of documents. The retrieval process consists on the comparison between documents representations and query representation. The second aim of this paper is to discover the relationships between terms (keywords) descriptors of documents in a document database. The correlations (relationships) between terms are extracted by using a technique of the Text mining, mainly association rules.

1. INTRODUCTION

Information plays a vital role in today's information society and we are witnessing an unprecedented explosion of its volume and its potential users. This rapid increase in the volume of information has created the problem of how to find information that interests us in this great mass of information. To address this problem a whole discipline was born. This discipline is called Information Retrieval (IR). Indeed, the main objective in the field of IR is to provide models, techniques and systems for storing and organizing masses of information and select those that respond to a user query. In general, a process of Information retrieval based on two basic steps, namely:

DOI: 10.4018/978-1-4666-9562-7.ch037

1. Indexing is a very important step in the process of IR. It is to identify and extract representative terms from the content of a document or query, which cover the most of their semantic content.
2. The step of selecting the relevant information consists of matching the descriptors extracted by the step of indexing with the descriptors of user query, in order to identify the information that respond to the needs of user query.

The information obtained by structuring a textual corpus (extraction of representative terms of document content) is only one facet of the implicit knowledge contained in a corpus. For this reason, one of the goals of text mining is to propose techniques for the extraction of implicit information in the document database.

One of the branches of text mining is concerned with the implications that describe the different correlations between the terms in the documents. These implications are called association rules.

The main problem of this paper is to extract from a textual corpus a set of useful knowledge for information retrieval system.

We defined two major objectives for extracting knowledge from textual corpus.

- The first objective is to study the information retrieval system (IRS), its functioning, its models and techniques used for evaluating information retrieval system.
- The second objective is to extract relationships between the representative terms of informational content of the corpus (index terms) using association rules.

This paper is organized as follows. The first section provides an introduction to the field of information retrieval. First, we introduce the research process, indexing process, models of IR and evaluation of information retrieval system. The second section is devoted to the presentation of the association rules. In the third section, we will introduce the concept of extracting knowledge from texts. Then, we describe a method for extracting associations between terms from a textual indexed collection using the technique of association rules. The last section will be devoted to the experimental part which we will describe the main features of our application illustrated with screenshots.

2. INFORMATION RETRIEVAL, BASIC CONCEPTS AND MODELS

The Information Retrieval (IR) (Ricardo & Berthier, 2011; Baziz, 2005) is traditionally defined as a set of techniques to select from a collection of documents, those who are likely to respond to the needs of the user. Manage texts involves storing, retrieving and exploring relevant documents.

The operation of IR (Mooers, 1948) is performed by software tools called information retrieval systems (IRS), whose goal is to find documents that satisfy user needs.

In an Information Retrieval System (IRS), the user expresses his information need as a query. The IRS tries to find all relevant documents and reject the documents that are not relevant. In practice, the set of documents returned by a query for a SRI is composed of a subset of relevant documents and a subset of irrelevant documents. These subsets determine the performance of an SRI (Karbasi, 2007; Harrathi, 2009).

2.1. Information Retrieval, Basic Concepts

The main goal of an information retrieval system (IRS) is to find the information assumed to be relevant to a user query generally expressed by a set of keywords.

According to Salton & Williamson (1968) information retrieval is a set of techniques used to select from a collection of documents, those