Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 139–145

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Support Vector Machines for a new Hybrid Information Retrieval System

Hamid KHALIFI[a,*], Abderrahim ELQADI[b], Youssef GHANOU[a]

[a]*TIM team, EST Meknes - Moulay Ismail University, Meknes*
h.khalifi@gmail.com
youssefghanou@yahoo.fr
[b]*LASTIMI, EST Sale - Mohammed V University, Rabat*
elqadi_a@yahoo.com

Abstract

Information Retrieval systems are used to extract, from a large database, relevant information for users. When the type of data is text, the complex nature of the database makes the process of retrieving information more difficult. Generally, such processes reformulate queries according to associations among information items before the query session. In this latter, semantic relationships or other approaches such as machine learning techniques can be applied to select the appropriate results to return. This paper presents a formal model and a new search algorithm. The proposed algorithm is applied to find associations between information items, and then use them to structure search results. It incorporates a natural language preprocessing stage, a statistical representation of short documents and queries and a machine learning model to select relevant results. On a series of experiments through Yahoo dataset, the proposed hybrid information retrieval system returned significantly satisfying results.

Keywords: Information retrieval; natural language processing; unsupervised classification; supervised classification; support vector machines.

1. Introduction

Nowadays, the word is witnessing a steady increase in the volume of available data[1]. Data are stored in diverse sources and every day, users consult these sources searching various kinds of information. This had led to the emergency of robust information retrieval systems to deal with this massive amount of unstructured data.

* Corresponding author. Tel.: "+212 6 68 61 55 50" ;
E-mail address: h.khalifi@gmail.com

10.1016/j.procs.2018.01.108

To explore this profusion, some modern search engines resort to semantics analysis and lexical matching. This is due to the fact that a same concept can be expressed by using different vocabularies and language styles [2].

In parallel, text classification is a subdomain of Natural Language Processing which correlates with information retrieval. Recently, models based on machine learning have become increasingly popular [3]. While these models achieve very good performance in practice, they tend to be relatively slow, limiting their use on very large datasets [4]. Moreover, they achieve weakest performance when the work introduces semantics.

Facing this dilemma between categorization performance on the first hand and semantic analysis on the second hand, lexical matching and machine learning can be gathered into a same hybrid model in order to meet the growing information retrieval needs.

Information Retrieval techniques are not limited to unstructured text data. They also have been applied to databases of images, stored speech, and other forms of data. They are more difficult to apply when the retrieved objects introduce semantics.

Generally, an information retrieval process consists of four main steps [5]: indexing, query formulation, comparison and finally the feedback.

Language modeling approaches to information retrieval connect the problem of retrieval with that of language model estimation. The main idea of such model is to estimate a language model for each document, and then rank documents by the likelihood of the query according to the estimated language model. The core problem in language modeling is the inaccuracy due to data sparseness [6].

Unfortunately, many typical challenges are noticed in language modeling approaches especially the process of stemming for unstructured text data. Therefore, this work will primarily focus on resolving the vocabulary mismatch problem. Variant word forms will be mapped to their base form (stemming).

In this sense, stemming is a crucial pre-processing step in language modeling as well as a very common requirement of Natural Language processing applications[7].

This paper is organized as follows. Section 2 summarizes main works in the area of information retrieval, and highlights the techniques of machine learning. Section 3 describes different steps of the proposed information retrieval system. Section 4 discusses experimental results. These results indicate the outperformance of the sophisticated proposed algorithm. This section includes also an empirical evaluation, as well as a discussion of its different steps. Finally, section 5 concludes the paper with a brief presentation of our further research topics.

2. Background

Information Retrieval systems have been widely investigated during last decades. Some approaches focused on lexical and vocabulary analysis while other ones based on Machine Learning techniques.

2.1. Information Retrieval techniques

Probabilistic indexing models have been considered by a number of studies as standard models for document retrieval. Such models assume that a subset of terms occurring in the document would be significant for indexing and the documents should be approximately equal length [8]. These models attempt to explicitly compute the probability of relevance between a document and the query [9]. The proportion of relevant documents is not enough to make the approximation for the initial retrieval [5].

Most information retrieval models in the past have based their statistical independences on strong or rarely satisfied assumptions: this explains the limited performance they achieved.

The critical language issue for previous works is the term mismatch problem [10]: users do often not use the same words to refer to objects or ideas. This is compounded by synonymy and polysemy [11].

The second class of approaches uses machine learning; different techniques have also been widely used in information retrieval systems which rank documents basing on statistical computations [5]. The user query is then considered as an ideal relevant document, and a similarity measure is computed between this user query and the rest of documents.

Le and Mikolov [12] proposed an unsupervised learning method to learn a paragraph vector as a distributed