Expert Systems With Applications 94 (2018) 126–136

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Bees swarm optimization guided by data mining techniques for document information retrieval

Youcef Djenouri[a,*], Asma Belhadi[b], Riadh Belkebir[c]

[a] Ulsan National Institute of Science and Technology, Ulsan, South Korea
[b] RIMA Lab, USTHB, Algiers, Algeria
[c] LRIA Lab, USTHB, Algiers, Algeria

ARTICLE INFO

Article history:
Received 27 July 2017
Revised 15 October 2017
Accepted 16 October 2017

Keywords:
Information retrieval
Data mining
Big data
BSO algorithm
Bio-inspired methods

ABSTRACT

This paper explores advances in the data mining field to solve the fundamental Document Information Retrieval problem. In the proposed approach, useful knowledge is first discovered by using data mining techniques, then swarms use this knowledge to explore the whole space of documents intelligently. We have investigated two data mining techniques in the preprocessing step. The first one aims to split the collection of documents into similar clusters by using the K-means algorithm, while the second one extracts the most closed frequent terms on each cluster already created using the DCI_Closed algorithm. For the solving step, BSO (Bees Swarm Optimization) is used to explore the cluster of documents deeply. The proposed approach has been evaluated on well-known collections such as CACM (Collection of ACM), TREC (Text REtrieval Conference), Webdocs, and Wikilinks, and it has been compared to state-of-the-art data mining, bio-inspired and other documents information retrieval based approaches. The results show that the proposed approach improves the quality of returned documents considerably, with a competitive computational time compared to state-of-the-art approaches.

1. Introduction

Let $D = \{d_1, d_2 \dots d_m\}$ be the set of m documents and let $T = \{t_1, t_2 \dots t_n\}$ be the set of n terms. Each document is composed of the subset of terms included in T. The user's request Req is represented by the set of terms. The Document Information Retrieval (DIR) is the process of finding relevant documents, according to the user's request Req, from a collection of documents D (Salton & McGill, 1986). The classical approach for DIR problem scans all documents and computes the score between each one and the given user's request. The ranking function is then used to keep only the relevant documents (Salton & McGill, 1986). This process has a polynomial complexity, however, when dealing with a large number of documents, the runtime will be extremely high.

In the last decade, many data mining based approaches have been used as preprocessing step to solve the IR problem by discovering knowledge from collections of documents. The extracted knowledge is then used for the solving step. Indeed, these approaches split the collection of documents into many groups. After that, the search process is performed on each group using high-performance computing tools. Two main approaches have been proposed. The works reported in Steinbach, Karypis, Kumar et al. (2000), Cai, He, and Han (2005), Hammouda and Kamel (2004), Hatamlou, Abdullah, and Nezamabadi-Pour (2012), Mahdavi and Abolhassani (2009) use the K-means algorithm (MacQueen, 1965) to group the clusters into k disjoints clusters, where each group contains similar documents. Whereas, the works reported in Beil, Ester, and Xu (2002), Fung, Wang, and Ester (2003), Yu, Searsmith, Li, and Han (2004), Menezes et al. (2010) use Frequent Patterns Mining (FPM) (Zaki & Hsiao, 2002) to discover the frequent terms in the collection. Afterward, the top k frequent patterns are used to create the groups of documents. These approaches decompose the initial problem into many subproblems, where each of which could be solved independently. However, the runtime of the DIR problem is still prohibitive, especially when dealing with massive number of documents existing in the World Wide Web (WWW).

More recently, some bio-inspired approaches have been developed to handle the DIR problem. The evolutionary-based approaches (Benetos, Dixon, Giannoulis, Kirchhoff, & Klapuri, 2013; Blei, 2012; Fan, Gordon, & Pathak, 2004; Lin, Chen, & Wu, 2014) and the swarm intelligence approaches (Buscher, Dengel, Biedert, & Elst, 2012; Collett, Graham, Harris, & Hempel-de Ibarra, 2006; Picard, Revel, & Cord, 2012; Zhang, Mei, Liu, Tao, & Zhou, 2011) proposed in the literature transform the DIR problem into an op-

* Corresponding author.
E-mail address: ydjenouri@unist.ac.kr (A. Belhadi).

https://doi.org/10.1016/j.eswa.2017.10.042
0957-4174/

timization problem and consider the collection of documents as a space of solutions. The intensification and diversification strategies of these approaches allow finding an approximate subset of documents in a reasonable computational time. Moreover, the communication between the swarms permits to find high-quality solutions compared to the evolutionary ones. Nevertheless, the process is still stochastic, and when the space of solutions is large, the swarms are disoriented, which degrades the quality of the final subset of documents returned by the swarms. Motivated by the success of BSO in dealing with many optimization problems, this paper aims to develop and investigate a new BSO algorithm for solving the DIR problem. The proposed algorithm exploits the information extracted in the pre-processing step by using both clustering and closed frequent itemset mining techniques to guide the swarms in the exploration of the space of documents.

The main contributions of this paper are the following:

- The proposed approach improves the preprocessing step of existing information retrieval approaches by applying both clustering and closed frequent itemset mining to extract knowledge from a collection of documents. Indeed, the K-means algorithm is first applied to generate k clusters; then FPM is performed on each cluster to extract the frequent patterns between highly correlated documents. To reduce the dimensionality of the frequent patterns, the closed algorithm (Lucchese, Orlando, & Perego, 2006) is used to extract only the most frequent patterns.
- Two bees swarm optimization algorithms are proposed by developing different heuristics which allow the swarms to well explore the search space. During the search process, the bees are guided by the knowledge extracted in the preprocessing step.
- To validate the runtime performance and the quality of returned documents, extensive experiments have been done on medium, large and big collections. The results show that our approach outperforms the data mining based approach when using large collections. Moreover, it outperforms the bio-inspired and other document information retrieval based approaches in terms of solution's quality using big collections and it is very competitive to them regarding the runtime process.

The remainder of the paper is organized as follows. Section 2 discusses the most used data mining and bio-inspired-based approaches for the document information retrieval problem. Section 3 presents our main proposition regarding the information retrieval problem followed by a detailed description of each components of our framework in Sections 3.1.1 and 3.2. In Section 4, the performance evaluation of the proposed approach is presented. Finally, Section 5, concludes the work by giving some remarks and future perspectives.

2. Related work

The DIR problem has been tackled in the literature using several approaches (Blei, Ng, & Jordan, 2003; Croft & Harper, 1979; Lafferty & Zhai, 2017; Ponte & Croft, 1998; Wei & Croft, 2006). In this work, we try to investigate the efficiency of using only data-mining and bio-inspired approaches for the DIR problem since the aim of these approaches is to provide approximate solutions in a reasonable time complexity. In the following, we present some data mining and bio-inspired approaches for the DIR problem.

2.1. Data mining-based approaches for DIR

In recent literature, some data mining-based approaches have been proposed to improve the information retrieval process. In the following, we discuss the most used approaches for DIR problem.

Beil et al. (2002) developed the first ARM algorithm for information retrieval called HFTC (Hierarchical Frequent Term-based Clustering). The algorithm starts by extracting frequent itemsets using Apriori algorithm. The itemsets are modeled by the terms of the collection. Then, the most frequent itemsets are considered as clusters where each frequent itemset is one cluster containing the documents that verify it.

Fung et al. (2003) proposed a new approach called FIHC (Frequent Itemset-based Hierarchical Clustering). The frequent itemsets are used to construct the hierarchical tree representing the collection. Using the frequents itemsets in the classification of the documents, the experiments reveal that the execution time of the user's requests has been reduced.

Yu et al. (2004) presented a new mining rules algorithm called TDC (Topic Document Clustering) to improve the quality of the classification of documents. It generates dynamically the different topics of the document's base using only the closet frequent itemsets. This can reduce the execution time comparing with the FIHC algorithm. TDC uses an intelligent structure that allows constructing the different links between each itemset of size k with the itemsets of size k-1 hierarchically. This approach gives high precision, but it could cause an overlap between clusters when the terms of the documents are highly linked.

In Babashzadeh, Daoud, and Huang (2013), the authors proposed a new algorithm for text processing called ARMIR (Association Rule Mining for Information Retrieval). In this approach, a given request is modeled by a set of concepts where the relations between concepts of the same request are determined by an association rules mining process. In Veloso, Almeida, Gonçalves, and Meira Jr (2008), a ranking function is proposed to sort the documents. A rules mining process is applied for training documents. The consequent part of each rule represents the scores of documents containing the terms of the antecedent part of this rule.

Another algorithm called LATRE (Lazy Associative Tag REcommender) has been proposed in Menezes et al. (2010). It extracts the association rules from the training set of documents. The obtained rules provide the keywords of a given object. This algorithm handles the pretreatment phase of information retrieval process. Furthermore, the set of relevant tags is associated with each document. This operation reduces the response time of different requests efficiently.

In Zhong, Li, and Wu (2012), the authors proposed PTM (Pattern Taxonomy Mining) algorithm to improve the comprehension of the user's request using a patterns mining algorithm. The pattern taxonomy of terms is discovered by applying the closed algorithm in the training set of documents. This technique reduces the noise between the user's request and the set of terms in the collection of documents.

In Joachims (2002) a classifier-based approach called SVMIR (Support Vector Machine for Information Retrieval) has been proposed. A Support Vector Machine learning algorithm is developed using clickthrough existing data. The results reveal that this algorithm creates groups of users according to their preferences which enhances the performance of the given information retrieval process automatically.

In Lan, Tan, Su, and Lu (2009), a new supervised term weighting approach called KNNIR (K-Nearest Neighbors for Information Retrieval) is proposed. It combines the support vector model representation and KNN (K-Nearest Neighbors) algorithm to compute the weight of each term in the given documents. The weights of the training terms are first computed, then the KNN classifier is launched to compute the score between each test term and the training terms.