ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 4, 2019, *pp*.279-290

An approach based on Combination of Features for automatic news retrieval

Mohammad Moradi [1], Elham Ghanbari [1 +], Mehrdad Maeen [1] and Sasan Harifi [2]

[1] *Department of Computer Engineering, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran*

2 Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran

(Received October 15 2019, accepted December 26 2019)

Abstract. Nowadays, according to the increasingly increasing information, the importance of its presentation is also increasing. The internet has become one of the main sources of information for users and their favorite topics. It also provides access to more information. Understanding this information is very important for providing the best set of information resources for users. Content providers now need a precise and efficient way to retrieve news with the least human help. Data mining has led to the emergence of new methods for detecting related and unrelated documents. Although the conceptual relationship between documents may be negligible, it is important to provide useful information and relevant content to users. In this paper, a new approach based on the Combination of Features (CoF) for information retrieval operations is introduced. Along with introducing this new approach, we proposed a dataset by identifying the most commonly used keywords in documents and using the most appropriate documents to help them with the abundance of vocabulary. Then, using the proposed approach, techniques of text categorization, evaluation criteria and ranking algorithms, the data were analyzed and examined. The evaluation results show that using the combination of features approach improves the quality and effects on efficient ranking.

Keywords: Information retrieval, news retrieval, combination of features, ranking news, dataset, benchmark dataset.

1. Introduction

At the moment, content is produced more than ever. This increase in the amount of media that is continuously generated creates a requirement for a new type of filtering service. Today, many companies perform data sorting for their customers by analyzing big data stream. But they handle them manually on a large scale, which is very difficult and inefficient. To avoid this, one solution is to use an automated data retrieval system that constantly collects and processes the information. The term "text mining" is often used to describe tasks in retrieving information about extracting useful information from large amounts of text. A subcategory in text mining is the classification of the text [1]. Text categorization is a process of assigning texts to one or more categories in a set of possible categories.

In the last 30 years, however, we are steadily moving towards solutions using machine learning, including training classifiers. Many improvements have been made in the field of text classification using machine learning algorithms. Many of the research in text categorization learning systems focuses on the classification of the subject or context of the text. An unknown problem is classification based on the format for the created text; In other words, identifying whether a text is a news article, or a piece of comment, or another template [2, 3].

Information Extraction (IE) is a sub-area of natural language processing. IE is assigned to the problem of identifying the entities mentioned in the natural language texts, the relationships between them, and the events in which they participate. IE systems are able to integrate scattered information across different documents. Structural extraction techniques have evolved considerably over the past decades [2]. There are two important challenges in IE. One of them is different ways of expressing reality. Another challenge that has been shared almost entirely with the tasks of natural language processing is the natural forms of natural languages that can have a vague structure and concept [4]. Several different approaches have been proposed to solve the IE challenges. They are classified according to different dimensions. Some of the classifications are related to the type of inputs [5]. The others are related to the type of technology used [2, 6], and another also related to the

+ Corresponding author. *E-mail address*: el.ghanbari@iausr.ac.ir

degree of system automation [3, 7]. Successful extraction of information has expanded the scope of IE. This scope is including unstructured sources and noise sources, which resulted in the provision of statistical learning algorithms.

IE and Information Retrieval (IR) are two distinct disciplines. The former deals with automatically extracting parts of unstructured or semi-structured information and storing them in a structured database and the later aims at retrieving the right content from a pool of contents in an efficient way. Both IE and IR are usually done one step before any data mining task. However they finds different aspects in the news context. Also most of the improvements are proposed for the news. They provide the necessary foundation for the news mining and retrieval [8-10].

News retrieval process starts with the request of users. Then the retrieval system must present best fitted news contents. In general, news retrieval is divided into four parts includes reducing the domain of retrieval, ranking news, results filtering, and web extraction.

- *Reducing the domain of retrieval*: IR systems can be highlighted by their scale of operation in three categories: "personal", "organizational, institutional, specific domain" and "web". On the web scale, reducing the recovery domain is a big sign. Searching the whole web has a negative effect on the performance of systems due to the large size of the web. For this reason, reduction of extraction is very important. This should be done based on specific news features. Web site news is divided into two groups, special news websites and public news websites. Web site news with special domain focus only on categories like sports, entertainment, politics, etc. Public news websites publish their news content in distinct categories that are called news services. Such news is distinguished by the tag. A news retrieval system should use these tags and classifications to reduce the domain of retrieval. [11].
- *Ranking news*: Calculation of similarity with the target of ranking news is one step of the IR system. Some studies have special attention to the content of the news [12-15].
- *Results filtering*: [16, 17]: Text level analysis, according to user requests, means filtering out the results. Filtering news articles based on their quality, novelty and relevance to the topic of the search is usually done.
- *Web extraction*: Web extraction is the automatic extraction of selected parts from a set of documents as unstructured or semi-structured information from the web. Its purpose is structured storage for ease of access in the future [18].

Generally, the data preparation phase for retrieval is divided into three categories, including supervised learning, unsupervised learning and semi-supervised learning. In supervised learning, their inputs and outputs are used to construct a model to find a generalized approximation function that is appropriate to the behavior of the data. In unsupervised learning, only inputs are known, so a model tries to clustering the data based on basic patterns. Semi-supervised learning uses combination of both labeled and unlabeled data to learning a model. Some machine learning methods include Support Vector Machine [19-21], Naïve Bayes [22-25], Nearest Neighbors [26, 27], Decision tree [25], and Ensemble Learning (Bagging and Boosting) [25].

The text categorization process is includes preprocessing, feature selection, and learning [28]. Preprocessing is a process in which stop words are deleted from the text. These words, which exist throughout the document, do not help distinguish a text from another text. Other work done in the preprocessing is mark and delete digits and return to the stem of the words. Feature selection is a process in which the weight of a word is specified in a document. We must point out that in most research, authors have used only one or two features to perform processing operations on data.

In this paper, using the different collection of features that used in other research, is proposed. Along with the proposed method which is called Combination of Features (CoF), a dataset is presented to test the news ranking techniques and proposed CoF method. This dataset called IRNA News, is our attempt to help researchers. The main purposes of the CoF method are to help the analyst and create better view for the documents. This method also helps machine learning algorithms for performing ranking operations. The main purposes of the dataset are provide a reference dataset for evaluating research, and help new researchers get started in the information retrieval field. Also, this dataset eases the development of ranking algorithms. Researchers can focus on algorithm development, and do not need to worry about experimental setup because the process of creating dataset and extracting features is done. However, this dataset can be used in many research areas such as data clustering and classification algorithms for English documents, algorithms for stemming English language, analyzing English language, and so on.