ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 7, No. 1, 2012, *pp*. 025-036

Document Classification in Summarization

Georgios Mamakis[1], Athanasios G. Malamos[2], J. Andrew Ware[3], Ioanna Karelli[4]

[1]Faculty of Advanced Technology, University of Glamorgan, Trefforest, Wales and Department of Applied Informatics and Multimedia. Technological Educational Institute of Crete, Heraklion Crete, Greece Email: gmamakis@epp.teicrete.gr

[2]Department of Applied Informatics and Multimedia. Technological Educational Institute of Crete, Heraklion Crete, Greece, Email: amalamos@epp.teicrete.gr

[3]Faculty of Advanced Technology, University of Glamorgan, Trefforest, Wales, Email: jaware@glam.ac.uk

[4]Faculty of Philology, University of Crete, Rethymnon, Crete, Email: iwannakarelli@gmail.com

(*Received August 13, 2011, accepted September 15, 2011*)

Abstract. Document classification and document summarization have a fairly indirect relation as document classification fall into classification problems as opposed to document summarization, where it is treated as a problem of semantics. A major part of the summarization process is the identification of the topic or topics that are discussed in a random document. With that in mind, we try to discover whether document classification can assist in supervised document summarization. Our approach considers a set of classes, in which a document may be classified in, and a novel summarization scheme adapted to extract summaries according the results of the classification. The system is evaluated against a number of supervises and unsupervised approaches and yields significant results.

Keywords: Document classification, supervised document summarization, statistics

1. Introduction

One of the major areas of data engineering both nowadays and in the past is text management. Important work has been undertaken in the area since early in the history of Information Technology. Text management includes subjects as document classification and document summarization. Document classification refers to the automatic assignment of a random document to one (single-label) or more (multi-label) classes. Applications of document classification include spam mail recognition and decision support systems. Document summarization, on the other hand, refers to the extraction or generation of text from one or multiple sources, in a shortened form compared to the original source(s). In this paper, we examine whether the use of document classification can result in better summaries, or if it can yield significant results. Our motivation came from remarks regarding document summarization. The main motivation came from a generic summarization procedure template that was first proposed by Lin and Hovy in [1]. The authors proposed that one of the important factors in document summarization is the identification of the topics that are present in a document. Identifying the topics discussed in a document enables to some extend the identification of the important words that will assist in the final extraction or generation of the summary. In addition to that, Moens et al. [2] undertook research utilizing a classification scheme to decide whether a random word in a document is a topic word (term) or not. Moreover, research by Barzilay et al. [3] tried to investigate news articles in conjunction to the topic they describe. However, their scope was to exploit characteristics that were domain-dependent, e.g. the pattern of authoring behind earthquake articles (location, size, victims). These research approaches led us to consider whether classification is an appropriate assisting tool in summarization tasks, not only in deciding if a random word is a topic word or a random sentence is a potential summary sentence as implied by [4] and [5], but rather in applying an adaptive approach on word importance, based on the class a random document may belong to. Therefore, instead of searching for the extraction of terminology that would result in identifying the potential topics of a document, we consider a set of class thesauri consisted of what we have automatically identified as terminology in the classifier training phase, classify the random document according to the lexicon that it adapts best to, and use this reference lexicon in extracting the most important sentences of the document as a summary.

The rest of the paper is organized as follows: In section 2, we provide background information on text classification, and insight on previous work we have undertaken in the area. In section 3, we present several document summarization approaches, and categorize them according to the scope and approach used, while in section 4, we validate theoretically our approach in supervised document summarization using classification, and analyze the main concepts behind our algorithms. In section 5, we provide an extended description of the limitations and algorithms that apply to our approach, while in section 6 we proved experimental results comparing several supervised and unsupervised algorithms. The final section of this paper concludes with future work in the area, underlining the feasibility of our approach.

2. Document Classification

One of the major problems in Machine Learning (ML) is deciding on the labeling of random input text into categories. Text classification or categorization has been an intriguing task, given that such decisions may not be always obvious. In order to tackle such problems, a number of approaches have been proposed such as statistical approaches, vector space models, artificial intelligence, decision trees and rules-based methods. Statistical classifiers are the most widely used generation of classifiers, since they are very efficient, very easy to construct and perform extremely quickly. Statistical classifiers include, among others, classifiers such as Naïve Bayes Classifier (NBC), Language Models and regression algorithms. Each algorithm provides a different approach in extracting the class of random input data, according to the number of labels it can assign. Thus, a second distinction, apart from the technology utilized, in document classification is referring to single-label classification, where the random case is assigned exactly one label,, and multi-label classification where the classifier can assign random input to a set of potential classes.

A typical classifier consists of two discrete modules:

- A training phase, where the classifier is provided with a number of features and the class they correspond to, constructing a classification decision space
- An application phase, where the classifier decides on the class a random feature set approximates best, using the classification decision

Document classification is a special case of classification algorithms, where the input features are the document words and the output is a class or set of classes where a random document may belong to. However, generic classification approaches apply as well. The most commonly used statistical algorithm is NBC. NBC is a supervised statistical classification algorithm based on the Bayes theorem of statistical independence, assuming that each input feature value is statistical independent to any other input feature in the same feature set. Despite the naivety in such an approach, NBC has been proven to operate very efficiently [6], outperforming more complex algorithms and approaches. Researchers that have modified and enhanced NBC in the past are [7,8,9]. However, it has been fairly recently suggested in [10], that NBC has a major drawback in its operation, that occurs when the set of training classes distribution is uneven. In such cases NBC behaves in a biased manner towards larger datasets. This has also been experimentally proven in our case as shown in [11].

Another commonly used statistical classification algorithm is Language Models (LMs). LMs are statistical models that instead of assuming statistical independence among features, use n-grams of features in both training and evaluation phase. The efficiency of LMs lies in the fact that they consider not only the existence of one word, but the co-existence of a sequence of words as e.g. San Francisco or Mona Lisa. Extensive work on LMs has been undertaken by [12] and [13]. It has been stated that uni-gram LMs approximate efficiently NBC results [14].

A common characteristic of both algorithms is that they are single-label classifiers. Multi-label classification is largely considered as an extension to single-label classification. Multi-label classification is generally achieved through a series of binary classifiers over multi-label training datasets to identify the classes a random document may belong to. Examples of research in the area includes modified kNN (k-Nearest Neighbors) approaches as the ones proposed by Cheng and Hullermeier in [15] and Zhang and Zhou in [16], or adaptations on algorithms such as SVM, proposed by Godbole and Sarawagi in [17].

3. Document Summarization

Document Summarization refers to the process of extracting or generating shortened content from one or various sources. This content generally either answers to specific user questions or offers a more generic covering as many topics as possible. The size of the summary can be either proportional to the original