Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 52 – 59

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

Ranking Sentences for Keyphrase Extraction: A Relational Data Mining Approach

Michelangelo Ceci[a,*], Corrado Loglisci[a], Lucrezia Macchia[a]

[a]*Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

Abstract

Document summarization involves reducing a text document into a short set of phrases or sentences that convey the main meaning of the text. In digital libraries, summaries can be used as concise descriptions which the user can read for a rapid comprehension of the retrieved documents. Most of the existing approaches rely on the classification algorithms which tend to generate "crisp" summaries, where the phrases are considered equally relevant and no information on their degree of importance or factor of significance is provided. Motivated by this, we present a probabilistic relational data mining method to model preference relations on sentences of document images. Preference relations are then used to rank the sentences which will form the final summary. We empirically evaluate the method on real document images.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014

Keywords: Document summarization; Ranking; Relational data mining

1. Introduction

The growing amount of documents available in digital libraries makes difficult and arduous obtaining the desired information, and therefore demands for the development of technologies to effectively support the user in a rapid comprehension once interesting documents have been retrieved. Numerous studies have been carried out in Natural Language Processing and, in particular, in the subfield of Automatic Text Summarization in order to generate a summarizing text which conveys the most salient and important information of the original document(s)[1]. A *summary* can be defined as a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). Summaries can be categorized in *extracts*, when they are created by selecting the keyphrases of the original text, and *abstracts*, when they are created by inferring the meaning of the source document or by re-generating the content of it[2]. The techniques oriented to the generation of abstracts require linguistic knowledge and sophisticated resources, and perform a deep analysis of the textual content by taking into account typical language constructs, such as discourse structure. The techniques based on the extracts rather perform a shallow analysis of the text and do not require linguistic knowledge. Although the extract-based techniques can produce summaries with evident problems of interpretation and cohesion among the selected portions

[*] Corresponding author. Tel.: +39-0805442285; fax: +39-0805442285.
E-mail address: michelangelo.ceci@uniba.it (Michelangelo Ceci).

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.011

of text, they have been proven to yield summaries whose informative level is satisfactory. This is particularly true when the extract is used as a component of another system and is not directly used by humans.

A strategy widely investigated for extractive approaches, is that of selecting the more salient sentences through Machine Learning or Data Mining algorithms which aim at either recognizing and then classifying the sentences to be included in the summary or ranking the source sentences and then selecting those with highest rank[3]. Typically, sentences are described in terms of lexical and structural features (e.g., keywords frequency, title keywords, sentence location, indicator phrases, etc.[4]) and represented as vectors of quantitative and categorical measures of those features (attribute-value representation). However, in some practical applications, it is also possible to exploit additional information conveyed by the structure of the original document. For example, in the case of document images obtained by scanning paper documents, sentences can be related to layout components or paragraphs. Another example is that of semistructured documents such as XML/HTML documents where sentences can be related to sections. In such situations, the classical attribute-value representation (based on the single-table assumption[5]), according to which sentences would be represented in a single table of a relational database (each row represents a sentence and columns correspond to properties of the sentence) appears to be too restrictive for at least three reasons. First, sentences cannot be realistically considered independent observations, because their arrangement is mutually constrained. Second, relationships among sentences in the same paragraph cannot be properly represented by a fixed number of attributes in a table. Third, the representation of properties of objects related to sentences (such as layout components or sections) would lead to redundancy problems that cause changes in the underlying probability distribution of examples. Since the single-table assumption limits the representation of relationships between examples, it also prevents the discovery of this kind of patterns, which can be very useful in the context of document summarization.

In this paper we propose an extractive approach aiming at learning to rank the source sentences extracted from document images. The proposed approach overcomes limitations posed by the single-table assumption by resorting to the relational data mining setting[5] according to which data are represented in several tables of a relational database possibly related according to foreign key constraints. This allows us to distinguish between the *reference objects* of analysis (sentences) and other *task-relevant spatial objects* (e.g. layout components), and to represent their interactions. This also allows us to represent different entities in different ways: sentences can be represented exploiting lexical and structural properties while layout components, for example, can be represented according to geometrical properties.

2. Background and Related Works

Background of this work is the Document image analysis system WISDOM++ [1] that enables the transformation of document images into XML format by means of several complex steps[6]. Initial processing steps include binarization, skew detection, noise filtering, and segmentation. The document image is then decomposed into several constituent items which represent coherent components of the documents (e.g., text lines or halftone images), without any knowledge of the specific format. This layout analysis step precedes the interpretation or understanding of document images, whose aim is that of recognizing “logic components”, that is, logically relevant layout components (e.g., title and section title of a scientific paper)[7] as well as extracting abstract relationships between layout components (e.g., reading order)[8]. By moving towards a higher level of abstraction, it is also possible to identify “semantic components” (e.g., motivations and experiments of a scientific paper) composed by several logic components (possibly belonging to different document pages) by exploiting their OCRed textual content. In this paper we add to WISDOM++ the keyphrase extraction step that is based on sentence ranking.

Sentence ranking is an approach investigated mostly with extract-based techniques which implement supervised Data Mining algorithms. This assumes the availability of a set of textual documents where ranking of the sentences of each document is given. Data Mining algorithms are then used to learn a ranking model to be applied on new documents. As in our case, in the literature, learning to rank from previously ranked sentences has been also interpreted as the problem of learning a preference function.

[1] http:/www.di.uniba.it/%7Emalerba/wisdom++/