Chapter 38
Novel Text Summarization Techniques for Contextual Advertising

Giuliano Armano
University of Cagliari, Italy

Alessandro Giuliani
University of Cagliari, Italy

ABSTRACT

Recently, there has been a renewed interest on automatic text summarization techniques. The Internet has caused a continuous growth of information overload, focusing the attention on retrieval and filtering needs. Since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. This chapter concentrates on studying and developing techniques for summarizing Webpages. In particular, the focus is the field of contextual advertising, the task of automatically suggesting ads within the content of a generic Webpage. Several novel text summarization techniques are proposed, comparing them with state of the art techniques and assessing whether the proposed techniques can be successfully applied to contextual advertising. Comparative experimental results are also reported and discussed. Results highlight the improvements of the proposals with respect to well-known text summarization techniques.

INTRODUCTION

The creation of a shortened but meaningful version of a text by a computer program, called Text Summarization (TS), is an old challenge in text mining. Given a text, its summary (i.e., a non-redundant extract from the original text) is returned.

This chapter is focused on studying and developing techniques for summarizing webpages. In particular, the interest is concentrated on the field of Contextual Advertising (CA), a form of online advertising. CA is the task of automatically suggesting ads within the content of a generic webpage. A commercial intermediary, the ad network, is usually in charge of optimizing the selection of ads with

DOI: 10.4018/978-1-5225-5191-1.ch038

the twofold goal of increasing revenue (shared between publisher and ad network) and improving user experience. Ads are selected and served by automated systems based on the content displayed to the user (Anagnostopoulos et al., 2007), (Broder et al., 2007), (Deepayan et al., 2008), (Lacerda, et al., 2006) (Ribeiro-Neto et al, 2005).

The motivation of the proposed research activity is that nowadays, ad networks need to deal in real time with a large amount of data, involving billions of pages and ads. Hence, efficiency and computational costs are crucial factors in the choice of methods and algorithms. A common methodology for Web advertising in real time is focused on the contributions of the different fragments of a webpage. The methodology relies on the adoption of a text summarization task for summarizing the webpage. Extraction-based techniques are usually adopted with the goal of fulfilling the given real time constraint. This methodology allows to identify short but informative excerpts of the webpage by selecting meaningful blocks of text. Several novel TS techniques that take into account further fragments, such as the title of the webpage, are proposed. Experiments confirm the effectiveness of the proposed techniques with respect to the state-of-the-art techniques. A further proposal is motivated by the evolution of the Web. In fact, classical techniques are often not easily applicable to dynamic webpages that typically rely on Microsoft Silverlight, Adobe Flash, Adobe Shock-wave, or contain applets written in Java. Conventional parsing methods are often not applicable to a webpage created on-the-fly. Therefore, snippets –i.e., page excerpts provided together with user query results by search engines– should be adopted to perform text summarization on webpages. Each study is conducted along two directions: comparing the proposed approach with classic text summarization technique and assessing whether the proposals can be successfully applied to CA. After a brief survey of relevant related work on text summarization and Contextual Advertising, the proposed techniques are compared with classic methods. Then, a description about several Contextual Advertising systems, implemented according to the proposed techniques, is provided. Experimental results, obtained by running the systems on relevant data, are also reported and discussed. A discussion on further issues and relevant solutions ends the chapter.

BACKGROUND

This section is aimed at giving a summary about TS and CA, recalling the main contributions in these fields. Then, the section provides the reasons why TS should be performed by a CA system. Finally, a generic architecture for CA is illustrated, describing also the baseline system to be used as starting point for experiments.

Text Summarization

Text summarization is an old challenge in text mining. During the 60’s, a large amount of scientific papers and books have been digitally stored and made searchable. Due to the limitation of storage capacity, documents were stored, indexed, and made searchable only through their summaries. For this reason, the automatic creation of text summaries became a primary task and several techniques were defined and developed.