Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 551 – 559

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Supporting contextualized information finding with automatic excerpt categorization

Ricardo Kawase[a,*], Patrick Siehndel[a], Bernardo Pereira Nunes[b]

[a]*L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany*
[b]*Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro/RJ – Brazil, CEP 22451-900*

Abstract

The volume of information on the Web is constantly growing. Consequently, finding specific pieces of information becomes a harder task. Wikipedia, the largest online reference Website is beginning to witness this phenomenon. Learners often turn to Wikipedia in order to learn facts regarding different subjects. However, as time passes, Wikipedia articles get larger and specific information gets more difficult to be located. In this work, we propose an automatic annotation method that is able to precisely assign categories to any textual resource. Our approach relies on semantic enhanced annotations and the categorization schema of Wikipedia. The results of a user study show that our proposed method provides solid results for classifying text and provides a useful support for locating information. As implication, our research will help future learners to easily identify desired learning topics of interest in large textual resources.

Peer-review under responsibility of KES International.

Keywords: Annotation; Categorization; Wikipedia; Learning Support;

1. Introduction

Since the rise of the Web 2.0, the volume of information available has significantly grown. Users have become the core contributors to the Web information space, producing a wide range of content and transforming it into the main source of information to the most variety of topics.

In fact, the advent of the Web 2.0 has also created a cultural change in how people interact, communicate and acquire knowledge.

* Corresponding author. Tel.: +49-(0)511-762-19715 ; fax: +49-(0)511-762-19712.
E-mail address: kawase@L3S.de

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.136

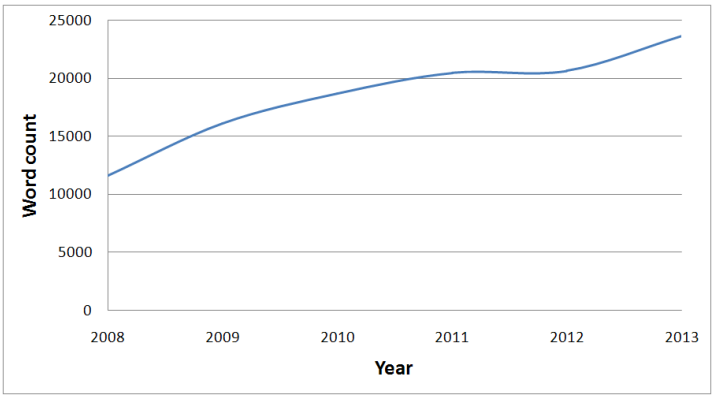

Fig. 1. Word count of Wikipedia artilcle 'Barack Obama' in the last five years.

A recent report from Pew Research Center[1] shows evidences of such user behaviors, where 92% of the adult users utilise the Web to perform online search and exchange e-mails. Although the increasing amount of the information arguably creates a richer Web, it also brings drawbacks and challenges. As more information is available, the more difficult it becomes to find, select and consume relevant contents.

This is particularly a problem in students' learning process, where a flood of information might hinder their understanding. For instance, students with attention deficit disorder may suffer even more, since they have difficulties in sustaining attention, fail to give attention to details and are easily distracted. In this manner, if the provided content is focused solely on the students' interests, or if students can focus only on excerpts of texts that they are interested in, then, the chances to get distracted is decreased.

We can illustrate the increase of information through a look at Wikipedia[2], a free encyclopedia created collaboratively by people who use it. Currently, Wikipedia has almost 30,000 active contributors, 4,4 million articles and registers over 3 million edits per month[3]. If we take into consideration the Wikipedia article of Barack Obama[4], we will find that, in terms of length (word count), his Wikipedia article page has duplicated in the last 5 years (See Figure 1) from 11,609 words (September 12, 2008) to 23,653 (September 12, 2013).

Over time, as in any other Web page, Barack Obama's article will definitely change and new content will be added. The constant growth of information may hinder the consume of information by its users and, therefore, new forms to access it must be provided. Thus, if a user is interested solely on Obama's association to *Sport* or *Education*, instead of pointing to his Wikipedia page, we must point the user to the excerpt of text in his Wikipedia page related to those topics of interest. Since an article in Wikipedia serves as a starting point for learning, delimiting its topics would facilitate and improve learning experience.

In this light, our main motivation in this paper is to extract topic-relevant information from Web pages and provide to end users an overview of the contents based on the topics they address. Our research is closely related to text segmentation, summarization and classification, however, differently from previous works in the field, our method relies on entity extraction and semantic classification.

Concisely, given a textual resource and a topic of interest, our method describes the input by selecting only the topic-relevant information. To achieve this goal, we identify the main topic subject for each paragraph.

Our high-level topic classification relies on the Wikipedia top categories which contain a broad coverage of topics that are maintained by the overall agreement of millions of contributors. This topic classification provides readers a sense making categorization that is digestible and manageable. While other approaches like clustering and Latent

[1] `http://www.pewinternet.org/Static-Pages/Trend-Data-(Adults)/Online-Activites-Total.aspx` accessed on Sept. 12
[2] `http://www.wikipedia.org`
[3] `http://stats.wikimedia.org/EN/SummaryEN.htm`
[4] `http://en.wikipedia.org/wiki/Barack_Obama`