Chapter 41
Finding the Semantic Relationship Between Wikipedia Articles Based on a Useful Entry Relationship

Lin-Chih Chen
National Dong Hwa University, Tawian

ABSTRACT

Wikipedia is the largest online Internet encyclopedia, and everyone can create and edit different articles. On the one hand, because it contains huge amounts of articles and there are many different language versions, it often faces synonymous and polysemy problems. On the other hand, since some of the similar Wikipedia articles may have the same topic of discussion, it needs a suitable way to identify effectively the semantic relationships between articles. This paper first uses three well-known semantic analysis models LSA, PLSA, and LDA as evaluation benchmarks. Then, it uses the entry relationship between Wikipedia articles to design its model. According to the experimental results and analysis, its model has high performance and low cost characteristics compared with other models. The advantages of its model are as follows: (1) it is a good model for finding the semantic relationships between Wikipedia articles; (2) it is suitable for dealing with huge amounts of documentation.

1. INTRODUCTION

World Wide Web (hereinafter referred to as the Web) is a distributed information-sharing platform that allows a wide range of users to share their information via the Internet. Starting with Web 1.0, users only need to use the relevant HTML syntax to create a static web page and publish the page to the web server host. Web 1.0 only provides limited one-way interaction between hosts and users (Singh, Bebi, & Gulati, 2011). That is, users cannot share their ideas directly and immediately with other users by using Web 1.0.

DOI: 10.4018/978-1-7998-0951-7.ch041

From traditional Web 1.0 to modern Web 2.0, users can use various information and communication platforms to achieve the two-way interaction between users. Until today, there are many well-known Web 2.0 platforms, such as Facebook, Twitter, YouTube, and Wikipedia, which provide users with an environment to comment, review and share ideas with other users. According to Best (2006), Web 2.0's main features compared to Web 1.0 include a rich user experience, user participation, dynamic content, metadata, Web standards, and scalability.

A widely-used Web 2.0 platform is Wikipedia, which is a free access to the collaborative Internet encyclopedia, where everyone can create and edit their articles or entries. The advantages of Wikipedia compared to other traditional encyclopedias include the following: (1) it contains almost all the possible topics in different subjects (Garcia & Ng, 2006), (2) it can respond quickly to any new event (Jokinen & Wilcock, 2012), (3) it offers various different language versions of the encyclopedia for different users (Hale, 2014).

When users search in Wikipedia, they often need to find semantic relationships that may occur between search terms. In general, users often need to consider the synonymy and polysemous relationships between search terms to help them select the most suitable Wikipedia entry. This is important because most users are difficult to distinguish between synonyms and polysemy entries in Wikipedia's 37 million entries (CBS, 2015).

Semantic analysis models are widely used to identify semantic relationships between terms (Egozi, Markovitch, & Gabrilovich, 2011; Ji, Jing, Wang, & Su, 2012; Liu, Zhang, Chang, & Sun, 2011; Lu, Mei, & Zhai, 2011; Lu, Peng, & Ip, 2010). In recent years, some well-known semantic analysis models have emerged, such as Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1998), Probabilistic LSA (PLSA) (Hofmann, 2001), and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to find possible semantic relationships between terms. However, these models lack the mechanism to find efficiently the semantic relationships between Wikipedia entries. This is also important because similar entries in Wikipedia may have the same discussion topic. Thus, in this paper, we use a new entry relationship to identify hidden semantic relationships that exist between Wikipedia articles.

The rest of this paper is organized as follows. First, in section 2, we briefly review some of the literature related to this paper. Next, in section 3, we introduce all the semantic analysis models used in this paper. Then, in section 4, we analyze and discuss the relevant experimental results. Finally, in section 5, we conclude this paper and discuss future research directions.

2. LITERATURE REVIEW

In this section, we briefly review some of the research literature related to this paper, including Wikipedia applications and semantic analysis models. In this section, we provide two tables for readers to read and compare related literature.

2.1. Wikipedia Applications

Many researchers have tried to use Wikipedia as the main source of research to solve many different information retrieval problems. Table 1 shows some recent studies on Wikipedia applications. Milne and Witten (2013) created a Wikipedia miner toolkit that allows users to combine the semantics of Wikipedia into their applications. The benefit of the toolkit is that it can efficiently classify different elements of