Semantic Relatedness Estimation using the Layout Information of Wikipedia Articles

Patrick Chan, Osaka University, Suita, Japan
Yoshinori Hijikata, Osaka University, Suita, Japan
Toshiya Kuramochi, Osaka University, Suita, Japan
Shogo Nishida, Osaka University, Suita, Japan

ABSTRACT

Computing the semantic relatedness between two words or phrases is an important problem in fields such as information retrieval and natural language processing. Explicit Semantic Analysis (ESA), a state-of-the-art approach to solve the problem uses word frequency to estimate relevance. Therefore, the relevance of words with low frequency cannot always be well estimated. To improve the relevance estimate of low-frequency words and concepts, the authors apply regression to word frequency, its location in an article, and its text style to calculate the relevance. The relevance value is subsequently used to compute semantic relatedness. Empirical evaluation shows that, for low-frequency words, the authors’ method achieves better estimate of semantic relatedness over ESA. Furthermore, when all words of the dataset are considered, the combination of the authors’ proposed method and the conventional approach outperforms the conventional approach alone.

Keywords: Explicit Semantic Analysis (ESA), Layout Information, Semantic Relatedness Estimation, Wikipedia Article, Word Frequency

INTRODUCTION

Semantic relatedness has a wide range of applications such as search, text summarization, and word sense disambiguation. It generally represents how much a word or phrase has a logical or causal connection to another word or phrase. To compute semantic relatedness, previous works made use of various linguistic resources such as WordNet and Wikipedia. They used the information about the graph built from a data source or the word frequency in a text corpus. This paper describes the result obtained using a new type of information, page layout information of Wikipedia, to improve the estimation of semantic relatedness.

Semantic relatedness applications take words or phrases as input, extract the highly semantically related words, and use the related words for their own needs. For example, a search engine generates a limited selection of results with the search terms alone, but if it uses the
related words of the search terms as well, it can produce a diverse set of results.

Many approaches have been used to estimate semantic relatedness. Among these methods, Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2007) is a Wikipedia-mining-based method that has recently become popular. It models a word as a vector of concepts, each of which is represented by a Wikipedia article. Each vector element shows the relevance between the word and the concept, which is the word’s normalized TFIDF (Karen, 1972) value in the corresponding Wikipedia article. Finally, it calculates the semantic relatedness from the cosine similarity between two concept vectors. Not only word frequency but also layout information, such as the word text style and its location in an article, are probably related to the relevance between a word and a concept. For example, the topmost section of a Wikipedia article, regarded as the summary, usually contains carefully chosen, descriptive words explaining the concept. Bold words, normally used for emphasis, might be related more to the concept than other words. Therefore, we aim at obtaining a better relevance estimate using TFIDF and an article’s layout information.

The following contributions are made by this paper.

• For words with low frequency, our proposed method achieves a higher correlation than that of ESA. Moreover, for all word pairs on the benchmark, the use of both our proposed method and ESA together results in a higher correlation than that of ESA.
• This report is the first of research work analyzing the page layout information of Wikipedia and using it to solve a research problem. The research problem we solve is semantic relatedness.
• We apply a more suitable statistical significance test to our result than our closely related work (Gabrilovich & Markovitch, 2007). Whereas Gabrilovich and Markovitch (2007) applied the test of statistically significant difference between two Pearson correlation coefficients on two Spearman’s rank correlation coefficients and claimed statistically significant difference between the Pearson correlations as the point of superiority of their method, we apply the statistical significance test designed for Spearman’s rank correlation coefficients.

The rest of the paper is organized as described below. Firstly, we present a description of the related work. Then, we give an overview of Wikipedia layout information and explain the preprocessing method of Wikipedia articles and our extraction method of layout information. The next section includes an overall description and details of our proposed method. We elaborate the experimental dataset, procedure, and results. Finally, we present some conclusions and future works.

RELATED WORK

This section presents a review of previously established approaches to semantic relatedness problems. Firstly, we specifically examine recent approaches that use Wikipedia to compute semantic relatedness. Then, we review the approaches that use search queries as a source to compute semantic relatedness. We also introduce approaches that use other knowledge bases to compute semantic relatedness. Lastly, we explain our position in these research fields.

Wikipedia Mining Approaches

Previous approaches to computing semantic relatedness by Wikipedia mining have measured relatedness from two perspectives. One perspective uses a Wikipedia article as an independent concept. Another perspective constructs a graph with nodes connected when a Wikipedia link exists from one article to another or when the articles share a category. The respective approaches pursued by Gabrilovich and Markovitch (2007) and Radinsky et al. (2011) treat a Wikipedia article as a concept, whereas those by Ito et al. (2008), Strube and Ponzetto (2006) and Hecht and Witten (2008) build a graph from Wikipedia. The former
Semantic Manipulations and Formal Ontology for Machine Learning based on Concept Algebra
www.igi-global.com/chapter/semantic-manipulations-formal-ontology-machine/72291?camid=4v1a