International Journal of Mobile Computing and Multimedia Communications
Volume 13 • Issue 2

English Article Style Recognition and Matching by Using Web Semantics

Mi Zhou, Dalian University of Science and Technology, China*
Lina Peng, Dalian Naval Academy, China

ABSTRACT

With the explosion of internet information, people feel helpless and find it difficult to choose in the face of massive information. However, the traditional method to organize a huge set of original documents is not only time-consuming and laborious, but also not ideal. The automatic text classification can liberate users from the tedious document processing work, recognize and distinguish different document contents more conveniently, make a large number of complicated documents institutionalized and systematized, and greatly improve the utilization rate of information. This paper adopts termed-based model to extract the features in web semantics to represent documents. The extracted web semantics features are used to learn a reduced support vector machine. The experimental results show that the proposed method can correctly identify most of the writing styles.

KEYWORDS

Document Type Recognition, Reduced Support Vector Machine, Vector Space Model, Web Semantics

1. INTRODUCTION

With the continuous popularization and development of information superhighway, information technology has penetrated into every corner of our social living (He 2021, Camero 2019). It has changed people's life and work style with unprecedented speed and ability. We are in an era of information explosion (Kumari 2017). On one side, the Internet contains a vast amount of information which is far beyond people's imagination. On the other hand, people often feel helpless when they face the vast ocean of information. It is called as information overload (Schmitt 2018, Swar 2017). It is a challenging task to help people effectively to manage massive information and quickly select useful information that they are interested in.

The web information is increasing, including online news, e-magazines, online technical reports, online documents, e-mail, BBS, online announcements (Yamamoto 2018). The traditional method to handle daily huge amount of information is time-consuming and laborious. The automatic text classification can directly filter and classify the document information (Kadhim 2019, Nguyen 2018). The user can only receive minor part which they are interested in. Then, users can be liberated from tedious document processing work and can easily understand and distinguish different document

DOI: 10.4018/IJMCMC.293751 *Corresponding Author

This article published as an Open Access Article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

contents. A large number of complicated documents can be regulated and systematized, and the utilization rate of information can be greatly improved.

One the other hand, while people can easily send and obtain web information, they also face with many harmful or illegal information, such as pornography, violence, and superstition. Unhealthy content such as gambling can be seen everywhere on the Internet. Even criminal text may be exist in BBS, blog, e-mail etc. to carry out reactionary propaganda, fraud, extortion, terrorist threats, drug sales and other illegal and criminal activities. Installing illegal information filtering software cannot effectively prevent the occurrence of illegal web information. Through legislative means, investigating the criminal responsibility of criminals can effectively crack down on this kind of criminal behavior. However, due to the lack of effective evidence, the criminals may be free from the evasion of legal sanction.

Text classification refers to the process of marking a free document with one or more predefined category labels according to its content information (Kowsari 2019, Mirończuk 2018). In order to correctly perform the task of text classification, it must input the useful information of the text into the computer to scientifically abstract the text and establish mathematical model to describe the text. The document expression is a key part of the text classification. The text representation refers to many representation methods and techniques of text retrieval (Wang 2020, Luo 2019). The common used text retrieval methods include: Boolean model (Lashkari 2009), vector space model (Raghavan 1986) and probabilistic model (Feng 2018). These models deal with feature weighting, category learning, and similarity calculation from different perspectives.

By closely combining with machine learning, the vector space model has been successfully used in text classification and becomes a mainstream method in the field of text classification. Vector space model (VSM) was first proposed the field of information retrieval. Then, it has been widely used in the field of text classification. In the vector space model based text classification method, the documents are converted as vector form by using term frequency and inverse document frequency. The vectors are indexed by inverted documents to calculate document similarity. Although vector space model has been solved text representation, it still needs to assume that words in the document are independent with each other to reduce the complexity of the representation. This paper adopts a termed-based model to represent the document and utilizes the extracted features to learn a reduced support vector machine to recognize the document type.

2. PROBLEM DESCRIPTION

Text classification is a supervised learning process. It learns a classification model to represent the relation between text features and text labels by a training set which consists of massive labeled documents. The features text document are input into learnt classification model to determine the document type. The text classification is mathematical mapping which maps the document to the associated type. The mapping can be represented as:

$$f : T \rightarrow L \tag{1}$$

In Equation (1), T represents the document set, while L represents the document type set.

The mapping rule of text classification is based on the data information of samples from each class to summarize the regularity of classification and establish the rules to determine the text related categories. The classification of text is based on document's content other than the data pattern in the document. It means that the concept of which type of text is related to is subjective.

From the above description, the flowchart of text classification contains two stages: training classification model and predicting future document according to learnt classification model. In the training classification model stage, the documents in the training set are represented as a unified form

11 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/english-article-style-recognition-and-matching-by-using-web-semantics/293751

Related Content

Mobile Computing for M-Commerce
Anastasis Sofokleous, Marios C. Angelidesand Christos Schizas (2009). *Mobile Computing: Concepts, Methodologies, Tools, and Applications (pp. 1584-1592).*
www.irma-international.org/chapter/mobile-computing-commerce/26608

The Benefits and Challenges of Mobile Technologies in Education: A Perspective for Sub-Saharan Africa
Julius Sonko (2015). *Promoting Active Learning through the Integration of Mobile and Ubiquitous Technologies (pp. 55-73).*
www.irma-international.org/chapter/the-benefits-and-challenges-of-mobile-technologies-in-education/115468

Towards Multimodal Mobile GIS for the Elderly
Julie Doyle, Michela Bertolottoand David Wilson (2010). *Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability (pp. 301-320).*
www.irma-international.org/chapter/towards-multimodal-mobile-gis-elderly/38546

Mobile Application Benchmarking Based on the Resource Usage Monitoring
Reza Rawassizadeh (2009). *International Journal of Mobile Computing and Multimedia Communications (pp. 64-75).*
www.irma-international.org/article/mobile-application-benchmarking-based-resource/37456

SDLC Phases of a Mobile Application
Drin Hoti, Monika Malokuand Klinton Gashi (2023). *Designing and Developing Innovative Mobile Applications (pp. 232-249).*
www.irma-international.org/chapter/sdlc-phases-of-a-mobile-application/322073