Information Processing and Management 54 (2018) 1129–1153

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Semantic text classification: A survey of past and recent advances

Berna Altınel[*], Murat Can Ganiz

College of Engineering, Department of Computer Engineering, Marmara University, Turkey

ARTICLE INFO

Keywords:
Text classification
Semantic text classification
Knowledge-based systems
Corpus-based systems
Neural language models
Deep learning

ABSTRACT

Automatic text classification is the task of organizing documents into pre-determined classes, generally using machine learning algorithms. Generally speaking, it is one of the most important methods to organize and make use of the gigantic amounts of information that exist in unstructured textual format. Text classification is a widely studied research area of language processing and text mining. In traditional text classification, a document is represented as a bag of words where the words in other words terms are cut from their finer context i.e. their location in a sentence or in a document. Only the broader context of document is used with some type of term frequency information in the vector space. Consequently, semantics of words that can be inferred from the finer context of its location in a sentence and its relations with neighboring words are usually ignored. However, meaning of words, semantic connections between words, documents and even classes are obviously important since methods that capture semantics generally reach better classification performances. Several surveys have been published to analyze diverse approaches for the traditional text classification methods. Most of these surveys cover application of different semantic term relatedness methods in text classification up to a certain degree. However, they do not specifically target semantic text classification algorithms and their advantages over the traditional text classification. In order to fill this gap, we undertake a comprehensive discussion of semantic text classification vs. traditional text classification. This survey explores the past and recent advancements in semantic text classification and attempts to organize existing approaches under five fundamental categories; domain knowledge-based approaches, corpus-based approaches, deep learning based approaches, word/character sequence enhanced approaches and linguistic enriched approaches. Furthermore, this survey highlights the advantages of semantic text classification algorithms over the traditional text classification algorithms.

1. Introduction

1.1. Traditional text classification and its challenges

Text mining studies steadily gain importance in recent years due to the wide range of sources that produce enormous amounts of data, such as social networks, blogs/forums, web sites, e-mails, and online libraries publishing research papers. The growth of electronic textual data will no doubt continue to increase with new developments in technology such as speech to text engines and digital assistants or intelligent personal assistants. Automatically processing, organizing and handling this textual data is a fundamental problem. Text mining has several important applications like classification (i.e., supervised, unsupervised and semi-

[*] Corresponding author.
E-mail addresses: berna.altinel@marmara.edu.tr (B. Altınel), murat.ganiz@marmara.edu.tr (M.C. Ganiz).

https://doi.org/10.1016/j.ipm.2018.08.001
Received 4 August 2017; Received in revised form 28 July 2018; Accepted 6 August 2018
0306-4573/ © 2018 Elsevier Ltd. All rights reserved.

supervised classification), document filtering, summarization, and sentiment analysis/opinion classification. Natural Language Processing (NLP), Machine Learning (ML) and Data Mining (DM) methods work together to detect patterns from the different types of the documents and classify them in an automatic manner (Sebastiani, 2005).

A traditional method for representing documents is called Bag of Words (BOW). This representation technique only include information about the terms and their corresponding frequencies in a document independent of their locations in the sentence or document. It is also called the Vector Space Model (VSM) since each document is represented as a vector of term frequencies in the vocabulary. Each of these terms in the vocabulary denotes an independent (orthogonal) dimension in the vector space, which usually results in a very high dimensional document vectors with only a few of them taking a frequency value which in turn yields to high sparsity. Furthermore, this representation does not take into account semantic associations between words. For instance, two words written as a different sequence of characters constitute different orthogonal dimensions of this vector space although they may be synonymous. Additionally, order of these words in the sentences are completely lost in the BOW representation. This approach mainly emphasizes the existence of some form of frequency information of terms. The BOW methodology makes the representation of documents simpler by disregarding the following several different semantic and syntactic relations between words in natural language: Firstly, it disregards the multi-word expressions by separating them into independent terms. Secondly, it treats polysemous words (words with multiple meanings) as a single entity because the word is separated from its neighboring words that determine its sense. Thirdly, the BOW approach maps synonymous words into distinct terms (Salton & Yang, 1973).

A text classifier is expected to label textual documents with pre-determined classes with an obvious assumption that each class consist of similar documents, usually talking about a particular topic that is different from the topics of other classes. However, vector space demonstration of texts usually results in high dimensionality and consequently high sparsity. This is a big difficulty especially when there are numerous class labels but inadequate training data for each of them. Obtaining labeled quality data for training is usually very expensive in real world applications. Accordingly, an accurate text classifier should have the capability of using this semantic information.

1.2. Semantic text classification and its advantages over traditional text classification

In semantic text classification methods, semantic relations between words are considered in order to, generally, measure similarity between documents. The semantic approach focuses on meaning of the words and hidden semantic connections between words and consequently between documents. Advantages of semantic text classification over traditional text classification are listed as:

- Implicit or explicit relationship discovery between words.
- Extracting and using latent relationships between words and documents.
- Capability to generate representative keywords for the existing classes.
- Semantic understanding of text, which improves accuracy of classification.
- Ability to handle synonymy and polysemy in compare to traditional text classification algorithms since they utilize semantic relationships between words.

1.3. Overview of existing semantic text classification algorithms

In order to overcome the difficulties created by BOW feature representation as mentioned above, a number of semantic relatedness methods have been proposed to incorporate semantic relations between words in text classification. These methods can be grouped into five categories, namely; domain knowledge-based (ontology-based) methods, corpus-based methods, deep learning based methods, word/character enhanced methods and linguistic enriched methods (Fig. 1):

- *Domain knowledge-based (Language dependent) approaches*: An ontology or thesaurus is used by domain knowledge-based systems to identify concepts in documents. Examples of knowledge bases are dictionaries, thesauri and encyclopedic resources. Common knowledge bases are WordNet, Wiktionary and Wikipedia. Among them WordNet is by far the most used knowledge-base.
- *Corpus-based (Language independent) approaches*: Certain mathematical computations are performed in these systems for exposing latent similarities between words in the training corpus (Zhang, Gentile, & Ciravegna, 2012). One of the well-known corpus-based algorithms is Latent Semantics Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).
- *Deep learning based approaches*: In recent years, especially since 2006, deep learning or hierarchical learning, has gained much attention in machine learning applications. Deep learning is a hybrid research area that is in the intersection of neural networks, graphical modeling, optimization, pattern recognition, and signal processing.
- *Word/character sequence enhanced approaches*: Word/character sequence enhanced systems treat words or characters as string sequences, which are taken out from documents by traditional string-matching techniques.
- *Linguistic enriched approaches*: These approaches use lexical and syntactic rules for extracting the noun phrases, entities and terminologies from a document to develop a representation of the document. Types of semantic algorithms for text classification are shown in Fig. 1.

Many studies in the scientific literature (Aas & Eikvil, 1999; Aggarwal & Zhai, 2012; Berry, 2004; Hotho, Nürnberger, & Paaß, 2005; Sebastiani, 2005) focus on traditional methods for text mining. Furthermore, there are also surveys that focus on particular type of classification algorithms such as kernel methods (Campbell, 2002; Jäkel, Schölkopf, & Wichmann, 2007). There are also

Download English Version:

https://daneshyari.com/en/article/10997999

Download Persian Version:

https://daneshyari.com/article/10997999

Daneshyari.com