An efficient Wikipedia semantic matching approach to text document classification

Zongda Wu[a,*], Hui Zhu[b,*], Guiling Li[c], Zongmin Cui[d], Hui Huang[e], Jun Li[e], Enhong Chen[f], Guandong Xu[g]

[a]*Oujiang College, Wenzhou University, Wenzhou, Zhejiang, China*
[b]*Wenzhou Vocational College of Science and Technology, Wenzhou, Zhejiang, China*
[c]*School of Computer Science, China University of Geosciences, Wuhan, China*
[d]*School of Information Science and Technology, Jiujiang University, Jiangxi, China*
[e]*College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, Zhejiang, China*
[f]*School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China*
[g]*Faculty of Engineering and IT, University of Technology, Sydney, Australia*

Abstract

A traditional classification approach based on keyword matching represents each text document as a set of keywords, without considering the semantic information, thereby, reducing the accuracy of classification. To solve this problem, a new classification approach based on Wikipedia matching was proposed, which represents each document as a concept vector in the Wikipedia semantic space so as to understand the text semantics, and has been demonstrated to improve the accuracy of classification. However, the immense Wikipedia semantic space greatly reduces the generation efficiency of a concept vector, resulting in a negative impact on the availability of the approach in an online environment. In this paper, we propose an efficient Wikipedia semantic matching approach to document classification. First, we define several heuristic selection rules to quickly pick out related concepts for a document from the Wikipedia semantic space, making it no longer necessary to match all the concepts in the semantic space, thus greatly improving the generation efficiency of the concept vector. Second, based on the semantic representation of each text document, we compute the similarity between documents so as to accurately classify the documents. Finally, evaluation experiments demonstrate the effectiveness of our approach, i.e., which can improve the classification efficiency of the Wikipedia matching under the precondition of not compromising the classification accuracy.

Keywords: Wikipedia matching, keyword matching, document classification, semantics

1. Introduction

The rapid growth of online documents in the World Wide Web has raised an urgent demand for efficient and effective classification algorithms to help people achieve fast navigation and browsing of online documents [7, 17, 31]. In general, traditional document classification algorithms were developed based on keyword matching [18, 13], whose basic idea is to represent a document as a vector of weighted occurrence frequencies of individual keywords, and then analyze the relevance of keyword vectors to measure the text similarity of documents. However, keyword matching techniques only take into consideration the surface text information, and do not consider the semantic information contained in documents, resulting in problems such as semantic confusion caused by polysemy, and content mismatch caused by synonym, thus reducing the effectiveness of the techniques [12, 33, 5]. To solve this problem, a new technique called Wikipedia matching was proposed [10, 11, 3, 1], whose basic idea is to use the semantic concepts from Wikipedia as an intermediate reference space, upon which a document is mapped from a keyword vector to a concept vector, so as to capture the semantic information contained in the document. As pointed out in [11], compared to other knowledge repositories, Wikipedia has the following advantages: (1) it has broad knowledge coverage about different concepts; (2) its knowledge concepts are always in step with the times; and (3) it contains a lot of new terms that cannot be found in other repositories. All the advantages enable Wikipedia matching to overcome the semantic mismatch problem encountered in keyword matching [1] and as a result improve the accuracy of document similarity computation. Below, we use a simple example to show the superiority of Wikipedia matching over keyword matching.

ID	Document Content
Doc 1	Puma, an American Feline Resembling a Lion.
Doc 2	Puma, a Famous Sports Brand from German.
Doc 3	Welcome to Zoo, an Animal World.

Table 1: Three short text documents

Given three short text documents shown in Table 1, keyword matching would mistakenly think that the similarity between Doc 1 and Doc 2 is higher than that between Doc 1 and Doc 3, because there is the polysemous keyword 'Puma' contained in both Doc 1 and Doc 2. However, in the Wikipedia matching approach, the documents would be mapped into concept vectors in the Wikipedia reference space by using keyword matching.

*Corresponding author
Email address: zongda1983@163.com (Zongda Wu)

Since there are such keywords as 'Feline' and 'Lion' in Doc 1 and 'Animal' in Doc 3, for the concepts related to the topic 'Animal', their corresponding elements would have higher values in both the concept vector of Doc 1 and the concept vector of Doc 3. However, the elements corresponding to these concepts would have smaller values in the concept vector of Doc 2, because it contains no keywords related to the topic 'Animal'. As a result, the Wikipedia matching approach, which calculates the document similarity based on concept vectors, correctly concludes that the similarity between Doc 1 and Doc 3 is higher than that between Doc 1 and Doc 2. It is observed that by using the Wikipedia knowledge to analyze the semantic information behind document keywords, Wikipedia matching can overcome the semantic mismatch problem encountered in keyword matching, and thus improve the accuracy of document similarity computation and in turn the accuracy of document classification. In addition, a number of studies have also demonstrated the effectiveness of Wikipedia matching [15, 22, 4, 21].

1.1. Motivation

However, in order to generate the concept vector for a document, the Wikipedia matching approach needs to conduct full-text keyword matching over a large number of concepts in the reference space from Wikipedia, thereby reducing the running efficiency of the approach and then the availability of the approach in an online document classification environment. To improve the efficiency, a straightforward way is to select a small number of concepts from Wikipedia to construct a small-scale reference space, so as to decrease the number of keyword matching operations. For example, Pak et al. [1] proposed to choose $1,000$ feature concepts on various topics to construct the Wikipedia reference space. However, such an oversimplified way greatly limits the semantic coverage of the reference space, making it difficult to pick out related concepts for documents (i.e., the document concept vectors would be very sparse), thereby reducing the accuracy of document similarity computation. In fact, if we only choose a small part of concepts, the main advantages of Wikipedia mentioned above (e.g., broad knowledge coverage) will also cease to exist. In summary, there is a dilemma in traditional classification approaches based on Wikipedia matching. On the one hand, if only a small number of concepts are chosen from Wikipedia as the reference space so as to improve the efficiency, it would limit the coverage of semantics and thus reduce the accuracy of document similarity computation. On the other hand, if a large number of concepts are chosen so as to ensure the semantic coverage of the reference space, it would decrease the efficiency of document classification.

1.2. Contribution

To solve the contradiction between accuracy and efficiency, we propose an efficient Wikipedia matching approach to text document classification, called WMDC. The WMDC approach is developed based on several heuristic selection rules, which can quickly select a set of relevant concepts from the Wikipedia reference space for a given document, making it no longer necessary to conduct time-consuming full-text matching over all the concepts in the reference space, and thus improving the Wikipedia matching efficiency. Specifically, the main contributions of this paper are twofold. **(1) Heuristic selection rules**. In the WMDC approach, the semantic reference space is constructed on a large number of concepts from Wikipedia. To improve the efficiency, we use heuristic selection rules to quickly pick out related concepts from the reference space for a given document. Then, based on keyword matching, a concept vector can be constructed efficiently for the document, where each element is determined by the full-text keyword similarity between the document and the corresponding concept (for the concepts, which are not selected by the heuristic rules, their corresponding elements are set to zero). **(2) Wikipedia classification algorithm**. A classification algorithm is proposed, which, based on the concept vector together with the keyword vector of each document, computes the semantic similarity and textual similarity between documents, upon which all the documents can be classified accurately.

It can be observed that compared to Wikipedia matching, the major improvement of the WMDC approach is that for a given document, it only needs to match a small number of related concepts (instead of all the concepts) in the reference space, consequently improving the generation efficiency of the document concept vector. In addition, the Wikipedia reference space used by the approach is constructed on a large number of concepts, consequently ensuring the semantic coverage of the reference space and in turn the accuracy of document similarity computation. In short, the approach can better deal with the trade-off between the accuracy and efficiency of document classification.

1.3. Organization

The rest of this paper is organized as follows. Section 2 reviews related work on keyword textual matching and Wikipedia semantic matching. Section 3 describes the proposed approach, specifically including selection heuristic rules, document similarity computation and algorithm implementation. Section 4 evaluates the effectiveness of the proposed approach by experiments. Finally, we conclude our work in Section 5.

2. Related Work

In traditional text document classification methods, keyword matching techniques were widely used, where the similarity between documents is measured by analyzing the common keywords between the documents [18, 13]. Specifically, each text document is first represented as a vector of weighted occurrence frequencies of individual keywords (the vector consists of the TF-IDF [32] values of all the keywords), and then the relevance (generally, measured by the cosine metric) between keyword vectors is used as the similarity measure between documents. For example, based on keyword matching, Jung [18] proposed a log document overlap measure framework to measure the content relevance between log documents. Fan et al. [13] introduced sentiment words into keyword matching, i.e., based on keyword matching, the authors introduced a sentiment similarity measure between page documents so as to improve

the accuracy of keyword similarity computation. Ramiz [23] proposed a sentence similarity measure approach based on sentence clustering, and applied the approach into automatic document summarization. In addition, there are some methods proposed to improve the TF-IDF model [14]. However, on account of only considering surface text information and ignoring semantic information, the keyword matching techniques would lead to such problems as semantic confusion caused by polysemy and content mismatch caused by synonymy, resulting in a negative impact on the effectiveness of this kind of techniques [12, 33, 36].

To overcome the semantic mismatch problem encountered in keyword matching, a new technique called Wikipedia matching was proposed, whose basic idea is to use a large number of concepts from Wikipedia to construct a reference space, upon which each document is mapped from a keyword vector to a concept vector, so as to capture the semantic information contained in the document. A number of studies have demonstrated the effectiveness of Wikipedia matching to overcome the semantic mismatch problem [15, 22, 4, 2]. The Wikipedia matching approach was proposed for the first time in [10, 11], and has been applied into many fields. For example, Pak et al. [1] proposed a contextual advertising approach based on Wikipedia matching so as to embed candidate ads into related pages. Hu et al [15] proposed a document clustering approach based on Wikipedia matching, to enrich the feature vector of a text document by using the correlation information between concept articles. Aiming at the patent literature search problem, it was proposed in [4] to use the semantic annotation information in Wikipedia to extend user search terms, so as to disambiguate the search words. Qureshi [21] proposed a text mining approach based on Wikipedia matching, and has demonstrated the superiority of the approach over traditional text mining approaches. Hu et al. [16] proposed to interpret the intent of a user query by using the Wikipedia knowledge, so as to reduce the workload of training a query intent classifier. In the approach, each user query would be mapped into a set of concepts and categories of Wikipedia so as to interpret the user query intent. In addition to Wikipedia, there are some semantic matching methods based on other knowledge repositories [9, 25]. However, the immense Wikipedia reference space results in that it needs to conduct a large number of full-text keyword matching operations to generate a document concept vector, thereby reducing the Wikipedia matching efficiency, and in turn the availability of the approach in an online environment of document classification.

3. Proposed Approach

In order to overcome the contradiction between the accuracy and efficiency of document classification, we propose an efficient Wikipedia matching approach to document classification, called WMDC. The framework of the WMDC approach is presented in Figure 1, which consists of the following several steps. First, we select a sufficient number of knowledge concepts from Wikipedia to construct a semantic space with hyper high dimension, used as the intermediate reference of document similarity

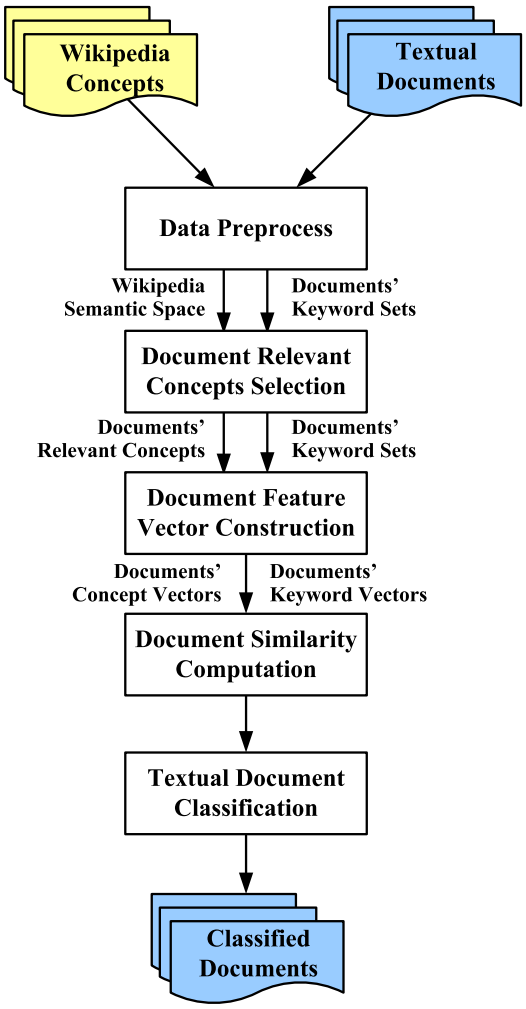


Figure 1: The Framework of the WMDC Approach

computation. Second, for any given document, we use well-designed heuristic selection rules to pick out related concepts from the Wikipedia reference space, and then based on the keyword matching technique, we compute the relevance between the document and each of the selected concepts (for each unselected concept, its relevance to the document is set to zero), so as to construct the document concept vector. Third, based on Wikipedia matching together with keyword matching, after obtaining the concept vector and keyword vector of each document, we compute the semantic similarity and textual similarity between any two documents as the similarity measure of documents. Finally, based on the similarity measure, we classify all the documents. In the WMDC approach, we particularly take two aspects of similarity measures between documents into consideration: (1) the **textual similarity** based on keyword matching, so as to capture the textual commonness of documents; and (2) the **semantic similarity** based on Wikipedia matching, so as to capture the relevance between documents from the semantic perspectives of knowledge ontology. More importantly, in the approach, we design several heuristic selection rules which can quickly pick out related concepts for a document from the reference space, consequently improving the efficiency of document

Symbol	Explanation
$\mathcal{W}$	A concept consisting of keywords, i.e., $\mathcal{W} = \{k\}$
$\mathcal{D}$	A document consisting of keywords, i.e., $\mathcal{D} = \{k\}$
$\mathcal{T}$	A concept title consisting of keywords, i.e., $\mathcal{T} = \{k\}$
$\mathbb{W}$	A Wikipedia reference space consisting of semantic concepts, i.e., $\mathbb{W} = \{\mathcal{W}\}$
$\mathbb{D}$	An unclassified document set consisting of text documents, i.e., $\mathbb{D} = \{\mathcal{D}\}$
$\mathbb{T}(\mathcal{W})$	A title set consisting of all the titles of a concept $\mathcal{W}$, i.e., $\mathbb{T}(\mathcal{W}) = \{\mathcal{T}\}$
$\mathbb{W}(\mathcal{D})$	A concept set consisting of the concepts related to a document $\mathcal{D}$, i.e., $\mathbb{W}(\mathcal{D}) = \{\mathcal{W}\}$ ($\mathbb{W}(\mathcal{D}) \subset \mathbb{W}$)
$\mathbf{K}(\mathcal{D})$	A keyword vector consisting of the TF-IDF values of all the keywords contained in a document $\mathcal{D}$, i.e., $\mathbf{K}(\mathcal{D}) = \langle \mathbf{tfidf}(k, \mathcal{D})\rangle_{k\in\mathcal{D}}$
$\mathbf{W}(\mathcal{D})$	A concept vector consisting of the keyword matching values between all the reference concepts and a document $\mathcal{D}$, i.e., $\mathbf{W}(\mathcal{D}) = \langle \mathbf{sim}(\mathcal{W}, \mathcal{D})\rangle_{\mathcal{W}\in\mathbb{W}}$

Table 2: Symbols and their meanings

similarity computation under the precondition of not compromising the accuracy.

In this section, we first describe data preprocessing, i.e., how to construct a semantic reference space in advance according to Wikipedia. Second, we formulate heuristic selection rules, which can quickly pick out related reference concepts for a given document so as to improve the Wikipedia semantic matching efficiency. Third, we introduce how to construct a document concept vector based on the Wikipedia semantic space, and then establish the similarity measure between two documents. Finally, we describe the algorithm implementation of the proposed approach. In Table 2, we describe some key symbols used in this paper.

3.1. Data Preprocessing

Wikipedia is one of the world's largest human repositories, which consists of a large number of concepts (whose number is close to ten millions and still increasing quickly), thus it has very broad coverage of diverse concepts. In Wikipedia, each concept is described by one article, but summarized by one or more article titles. Wikipedia can be edited by volunteers around the world, and thus its knowledge concepts can be updated timely. In advance, we extract all the concepts from Wikipedia (in the experiments, we select a total of one million concepts). Then, according to the following three steps, we preprocess the selected concepts so as to construct a semantic reference space.

- **Word segmentation**. This step aims to map each concept $\mathcal{W}$ into a set of individual keywords. Because words in English are generally separated by such special characters as space and tabs, the word segmentation process is relatively simple. In this paper, we use the famous NLTK word segmentation[1]. In addition, in the word segmentation process, we turn each word to lowercase to facilitate subsequent processing.

- **Remove stopwords**. Stopwords are the words, which contain no concrete meanings (e.g., prepositions, pronouns, articles). Thus, stopwords need to be removed from each concept document so as to avoid a negative impact on the approach. Here, we use the stop list supplied by NLTK to remove stopwords for each concept outputted by the step of word segmentation.

- **Stemming**. Each word has itself word stem, so stemming means to change words in different tenses (e.g., past tense, continuous tense) and different parts of speech (e.g., noun, verb) to their stems. A stemming operation can centralize the language information to reduce the calculation scales of the subsequent steps. Here, we use the famous Snowball frame[2] for stemming.

Now, each concept $\mathcal{W}$ extracted from Wikipedia is represented as a set of individual keywords, denoted by $\mathcal{W} = \{k\}$, i.e., we obtain a Wikipedia reference space, denoted by $\mathbb{W} = \{\mathcal{W}\}$. It should be pointed out that the above preprocess for concepts was completed in advance, so it does not decrease the efficiency of online document classification. Moreover, to improve the efficiency of the subsequent Wikipedia matching operations to map each document to a concept vector in the Wikipedia reference space, we have in advance mapped each concept $\mathcal{W} \in \mathbb{W}$ to a keyword vector, denoted by $\mathbf{K}(\mathcal{W}) = \langle \mathbf{tfidf}(k, \mathcal{W})\rangle_{k\in\mathcal{W}}$ (the keyword vector mapping process will be introduced in Section 3.3).

In a practical application, the Wikipedia reference space $\mathbb{W}$ may be dynamically adjusted (because the articles in Wikipedia will be updated or added). In this case, we only need to recreate the keyword vectors for the updated concepts in $\mathbb{W}$, or create the keyword vectors for the added concepts and then insert them into $\mathbb{W}$. Since the above space update operations are completed offline, we here no longer present the specific algorithm (refer to [27, 29] for efficiently dealing with online changing semantic space).

3.2. Heuristic Selection

In Wikipedia semantic matching, each text document $\mathcal{D}$ has to be mapped into a concept vector in the Wikipedia reference space $\mathbb{W}$, where each element corresponds to a concept (i.e., a dimension) $\mathcal{W}$ ($\mathcal{W} \in \mathbb{W}$), and the element value denotes the content relevance (measured using keyword matching) between the document $\mathcal{D}$ and the concept $\mathcal{W}$. The Wikipedia reference space contains a large number of concepts (millions) and each concept contains a number of keywords (thousands). Hence, in order to map a document $\mathcal{D}$ into a concept vector in the reference space $\mathbb{W}$, we need to conduct a large number of full-text keyword matching operations (the concept vector mapping process will be introduced in Section 3.3), thereby reducing the Wikipedia matching efficiency and then the

[1] http://www.nltk.org

[2] http://snowball.tartarus.org/texts/introduction.html

availability of the approach in an online environment of text document classification. However, we observe that the concept vector generated by using Wikipedia matching for a document is very sparse, i.e., there is only a small number of concepts in the reference space really related to the document. As a result, if we can quickly pick out the related concepts for a document, we only need to conduct time-consuming full-text keyword matching over the related concepts (instead of all the concepts in the reference space), which certainly will improve the efficiency of document concept vector generation, and in turn the efficiency of document classification. Moreover, we also observe that in Wikipedia, each concept is described by one article of some length, but summarized by one or more article titles of short length [26]. As a result, by using the succinct well-formed concept titles, we can quickly in advance judge whether a concept is related to a document. The heuristic selection rules used in our approach are designed according to the above observation.

Observation 1. Given a document and a Wikipedia concept, if there is one title (or some part of one title) of the concept appearing in the document, then it is likely that the concept is related to the document; otherwise, it has a small probability that the concept is related to the document.

For example, for a text document about the topic 'Zhejiang Travel', we suppose that a phrase 'West Lake' is contained in the document and another keyword 'Database' is not. Then, it has a high probability that the concept named by 'West Lake' is related to the document, and it has a small probability for the concept 'Database'. Below, based on Observation 1, we formulate several well-designed heuristic selection rules, which can efficiently determine a set of related reference concepts for a given document. In addition, due to the reasons of spelling, abbreviation, synonym and so on, a concept in Wikipedia may have more than one title (i.e., a concept may be summarized by several phrases simultaneously). Below, we use $\mathbb{T}(\mathcal{W}) = \{\mathcal{T}\}$ to denote the title set consisting of all the titles of a concept $\mathcal{W}$.

Definition 1 (Full-Title-Relevance). Given a text document $\mathcal{D}$ and a Wikipedia concept $\mathcal{W}$, the full-title-relevance between them can be measured by the occurrence frequency of each title of the concept $\mathcal{W}$ in the document $\mathcal{D}$, i.e.,

$$Re^{(1)}(\mathcal{D}, \mathcal{W}) = \sum_{\mathcal{T} \in \mathbb{T}(\mathcal{W})} \frac{|\mathcal{T}|}{|\mathcal{D}|} \sqrt{\mathbf{num}(\mathcal{T}, \mathcal{D})}$$

wherein, $\mathbf{num}(\mathcal{T}, \mathcal{D})$ denotes the number of occurrences of each concept title $\mathcal{T}$ in the document $\mathcal{D}$, $|\mathcal{D}|$ denotes the size of the document $\mathcal{D}$ (i.e., the number of keywords contained in $\mathcal{D}$) and $|\mathcal{T}|$ denotes the size of each concept title $\mathcal{T}$.

Rule 1 (Full-Title-Selection). Given a text document $\mathcal{D}$, it is deemed by the rule that a concept $\mathcal{W}$ is related to the document $\mathcal{D}$, if and only if the full-title-relevance between them is greater than a given threshold θ_1 ($\theta_1 \geq 0$). Thus, a set of related reference concepts determined by the rule for the document $\mathcal{D}$ can be formulated as

$$\mathbb{W}^{(1)}(\mathcal{D}) = \left\{ \mathcal{W} \,|\, \mathcal{W} \in \mathbb{W} \wedge Re^{(1)}(\mathcal{D}, \mathcal{W}) > \theta_1 \right\}$$

Suppose that the threshold θ_1 is set to 0, and the document $\mathcal{D}$ only contains one sentence 'Hangzhou is Famous for the West Lake'. A concept $\mathcal{W}$ is deemed by Rule 1 to be related to the document $\mathcal{D}$, if and only if there is at least one title $\mathcal{T}$ of $\mathcal{W}$ appearing in $\mathcal{D}$. For example, for two concepts named by 'West Lake' and 'Hangzhou', respectively, they are deemed to be related to the document $\mathcal{D}$, i.e., they belong to the set $\mathbb{W}^{(1)}(\mathcal{D})$.

Definition 2 (All-Keyword-Relevance). Given a document $\mathcal{D}$ and a concept $\mathcal{W}$, the all-keyword-relevance between them can be measured by the occurrence frequencies of all the keywords of each title of the concept $\mathcal{W}$ in the document $\mathcal{D}$, i.e.,

$$Re^{(2)}(\mathcal{D}, \mathcal{W}) = \sum_{\mathcal{T} \in \mathbb{T}(\mathcal{W})} \frac{|\mathcal{T}|}{|\mathcal{D}|} \min_{k \in \mathcal{T}} \sqrt{\mathbf{num}(k, \mathcal{D})}$$

wherein, $\mathbf{num}(k, \mathcal{D})$ denotes the number of occurrences of each keyword k of a concept title $\mathcal{T}$ in the document $\mathcal{D}$.

Rule 2 (All-Keyword-Selection). Given a document $\mathcal{D}$, any concept $\mathcal{W} \in \mathbb{W}$ is deemed to be related to the document $\mathcal{D}$, if and only if the all-keyword-relevance between them is greater than a given threshold θ_2 ($\theta_2 \geq 0$). Thus, a set of related reference concepts determined by the rule for the document $\mathcal{D}$ can be formulated as

$$\mathbb{W}^{(2)}(\mathcal{D}) = \left\{ \mathcal{W} \,|\, \mathcal{W} \in \mathbb{W} \wedge Re^{(2)}(\mathcal{D}, \mathcal{W}) > \theta_2 \right\}$$

Suppose that the threshold θ_2 is set to 0, and the document $\mathcal{D}$ contains one sentence as 'Puma is a Famous Sports Company from German'. For two concepts $\mathcal{W}_1$ and $\mathcal{W}_2$ named by 'Ford Puma' and 'Public Company', respectively, Rule 2 deems that the concept $\mathcal{W}_2$ is related to the document $\mathcal{D}$, because there is at least one title in $\mathcal{W}_2$, whose keywords are all contained in $\mathcal{D}$ (note that 'Public' is not a keyword). However, the concept $\mathcal{W}_1$ is deemed to be unrelated to $\mathcal{D}$, because the keyword 'Ford' is not contained in $\mathcal{D}$. Formally, we have that $\mathcal{W}_2 \in \mathbb{W}^{(2)}(\mathcal{D})$ and $\mathcal{W}_1 \notin \mathbb{W}^{(2)}(\mathcal{D})$.

Definition 3 (Any-Keyword-Relevance). Given a document $\mathcal{D}$ and a concept $\mathcal{W}$, the any-keyword-relevance between them can be measured by the occurrence frequencies of any title keyword of the concept $\mathcal{W}$ in the document $\mathcal{D}$, i.e.,

$$Re^{(3)}(\mathcal{D}, \mathcal{W}) = \sum_{\mathcal{T} \in \mathbb{T}(\mathcal{W})} \frac{|\mathcal{T}|}{|\mathcal{D}|} \sum_{k \in \mathcal{T}} \sqrt{\mathbf{num}(k, \mathcal{D})}$$

Rule 3 (Any-Keyword-Selection). Given a document $\mathcal{D}$, a concept $\mathcal{W}$ is deemed to be related to the document $\mathcal{D}$, if and only if the any-keyword-relevance between them is greater than a given threshold θ_3 ($\theta_3 \geq 0$). Thus, a set of related concepts selected by the rule for the document $\mathcal{D}$ can be formulated as

$$\mathbb{W}^{(3)}(\mathcal{D}) = \left\{ \mathcal{W} \,|\, \mathcal{W} \in \mathbb{W} \wedge Re^{(3)}(\mathcal{D}, \mathcal{W}) > \theta_3 \right\}$$

Suppose that the threshold θ_3 is set to 0, and the document $\mathcal{D}$ only contains a sentence 'Zhejiang Province located in the Yangtze River port, is a strong economical province in China'. Any concept $\mathcal{W}$ is deemed by Rule 3 to be related to the document $\mathcal{D}$, if and only if there is at least one title keyword of the concept $\mathcal{W}$ appearing in the document $\mathcal{D}$. For example, for two concepts named by 'Mainland China' and 'Yangtze River', respectively, they are both deemed to be related to the document, i.e., they belong to the set $\mathbb{W}^{(3)}(\mathcal{D})$.

According to the above, we can know that under the same threshold value (i.e., $\theta_1 = \theta_2 = \theta_3$), for the three related concept sets $\mathbb{W}^{(1)}(\mathcal{D})$, $\mathbb{W}^{(2)}(\mathcal{D})$ and $\mathbb{W}^{(3)}(\mathcal{D})$, which are determined by the heuristic selection rules for a document $\mathcal{D}$, respectively, they should satisfy that $\mathbb{W}^{(1)}(\mathcal{D}) \subseteq \mathbb{W}^{(2)}(\mathcal{D}) \subseteq \mathbb{W}^{(3)}(\mathcal{D})$. In addition, we can observe that compared to a concept consisting of thousands of keywords, a concept title consists of much fewer keywords (generally less than 5), resulting in that the heuristic selection rules developed based on the concept title information can be carried out efficiently (see the time complexity analysis in Section 3.4). Now, according to one of the heuristic selection rules, we can quickly pick out a set of related concepts for a given document $\mathcal{D}$, denoted by $\mathbb{W}(\mathcal{D})$ uniformly, which is an important intermediate reference of document similarity computation.

3.3. Document Similarity Computation

In the fields of information retrieval and text mining, TF-IDF is an important weight often used together with the cosine metric to compute the textual similarity between documents [32]. Given two documents $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathbb{W}(\mathcal{D}_1)$ and $\mathbb{W}(\mathcal{D}_2)$ denote their concept sets (which are determined by a heuristic selection rule). In this subsection, we mainly describe how to construct the concept vectors $\mathbf{W}(\mathcal{D}_1)$ and $\mathbf{W}(\mathcal{D}_2)$, according to the concept sets $\mathbb{W}(\mathcal{D}_1)$ and $\mathbb{W}(\mathcal{D}_2)$ together with the TF-IDF weight, so as to measure the similarity between the documents $\mathcal{D}_1$ and $\mathcal{D}_2$. Here, we particularly take two aspects of document similarity measures into consideration, i.e., the textual similarity based on keyword matching, and the semantic similarity based on Wikipedia matching.

Given a document $\mathcal{D}$, based on the data preprocess technique introduced in Section 3.1, we first map it into a keyword set, denoted by $\mathcal{D} = \{k\}$. Then, we compute the TF-IDF value for each keyword k contained in the document $\mathcal{D}$, i.e.,

$$\mathbf{tfidf}(k, \mathcal{D}) = \mathbf{tf}(k, \mathcal{D}) \cdot \mathbf{idf}(k)$$

wherein, $\mathbf{tf}(k, \mathcal{D})$ denotes the number of occurrences of the keyword k in the document $\mathcal{D}$, and $\mathbf{idf}(k)$ can be calculated as follows

$$\mathbf{idf}(k) = \log \frac{|\mathbb{D}|}{|\{\mathcal{D}_i \,|\, k \in \mathcal{D}_i \wedge \mathcal{D}_i \in \mathbb{D}\}| + 1}$$

According to the above, we can further map the document $\mathcal{D}$ into a keyword vector, denoted by $\mathbf{K}(\mathcal{D}) = \langle \mathbf{tfidf}(k, \mathcal{D}) \rangle_{k \in \mathcal{D}}$, which consists of the TF-IDF values of all the keywords contained in the document $\mathcal{D}$. However, similarly, each concept $\mathcal{W}$ in the Wikipedia reference space $\mathbb{W}$ has also been in advance mapped into a keyword vector $\mathbf{K}(\mathcal{W}) = \langle \mathbf{tfidf}(k, \mathcal{W}) \rangle_{k \in \mathcal{W}}$. As a result, together with the related concept set $\mathbb{W}(\mathcal{D})$ determined by a heuristic selection rule for $\mathcal{D}$, we can define the content relevance between the document $\mathcal{D}$ and each concept $\mathcal{W} \in \mathbb{W}$, and thereby construct the document concept vector $\mathbf{W}(\mathcal{D})$.

Definition 4 (Concept Relevance). Given a document $\mathcal{D}$ and a Wikipedia concept $\mathcal{W} \in \mathbb{W}$, let $\mathbb{W}(\mathcal{D})$ denote the related concept set determined by a selection heuristic rule for $\mathcal{D}$. Then, the content relevance between the document $\mathcal{D}$ and the concept $\mathcal{W}$ is defined as

$$\mathbf{sim}(\mathcal{D}, \mathcal{W}) = \begin{cases} \mathbf{cos}\angle\mathbf{K}(\mathcal{D}), \mathbf{K}(\mathcal{W}), & \text{if } \mathcal{W} \in \mathbb{W}(\mathcal{D}) \\ 0, & \text{otherwise} \end{cases}$$

wherein, $\mathbf{cos}\angle\mathbf{K}(\mathcal{D}), \mathbf{K}(\mathcal{W})$ denotes the cosine similarity between the vectors $\mathbf{K}(\mathcal{D})$ and $\mathbf{K}(\mathcal{W})$, i.e.,

$$\mathbf{cos}\angle\mathbf{K}(\mathcal{D}), \mathbf{K}(\mathcal{W}) = \frac{\sum\limits_{k \in \mathcal{D} \cap \mathcal{W}} \mathbf{tfidf}(k, \mathcal{D}) \cdot \mathbf{tfidf}(k, \mathcal{W})}{\sqrt{\sum\limits_{k \in \mathcal{D}} \mathbf{tfidf}(k, \mathcal{D})^2} \sqrt{\sum\limits_{k \in \mathcal{W}} \mathbf{tfidf}(k, \mathcal{W})^2}}$$

Note that besides the cosine metric, there are other similarity metrics that may have better performance. However, in keeping with the traditional Wikipedia matching approach [3] (which is developed based on the cosine metric), we here also use the cosine metric.

Definition 5 (Concept Vector). Given a text document $\mathcal{D}$, the concept vector of $\mathcal{D}$ in the Wikipedia reference space $\mathbb{W}$ is defined as $\mathbf{W}(\mathcal{D}) = \langle \mathbf{sim}(\mathcal{D}, \mathcal{W}) \rangle_{\mathcal{W} \in \mathbb{W}}$.

It is observed that the relevance computation (a.k.a., the full-text keyword matching operation) between a document and a concept is somewhat time-consuming. However, more importantly, to generate a document concept vector, it needs to conduct such full-text keyword matching operations on all the concepts in the reference space of hyper high dimension, which certainly will greatly reduce the efficiency of document concept vector generation. In order to improve the efficiency, in Definition 4, for each concept $\mathcal{W}$ in the reference space $\mathbb{W}$, if it is not contained in the related concept set $\mathbb{W}(\mathcal{D})$ (i.e., $\mathcal{W} \notin \mathbb{W}(\mathcal{D})$), then it is deemed to be unrelated to the document $\mathcal{D}$, and its relevance to $\mathcal{D}$ is uniformly set to zero. As a result, this makes that we only need to conduct full-text keyword matching over the related concepts in $\mathbb{W}(\mathcal{D})$, thereby greatly improving the efficiency of document concept vector generation (since $|\mathbb{W}(\mathcal{D})| \ll |\mathbb{W}|$). Now, the two documents $\mathcal{D}_1$ and $\mathcal{D}_2$ are represented as two concept vectors $\mathbf{W}(\mathcal{D}_1)$ and $\mathbf{W}(\mathcal{D}_2)$, upon which we can compute the document semantic similarity, and then together with the document textual similarity, we can further compute the overall document similarity.

Definition 6 (Semantic Similarity). Given two text documents $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathbf{W}(\mathcal{D}_1)$ and $\mathbf{W}(\mathcal{D}_2)$ denote two concept vectors of the documents in the Wikipedia reference space. Then,

Algorithm 1 A Wikipedia semantic matching approach to text document classification (WMDC)

Input: a set of unclassified text documents, $\mathbb{D} = \{\mathcal{D}\}$.
Output: a set of classified text documents.

1: According to the concepts of Wikipedia, we construct a semantic reference space in advance, denoted by $\mathbb{W} = \{\mathcal{W}\}$;
2: **for all** $\mathcal{D} \in \mathbb{D}$ **do**
3: Preprocess the document $\mathcal{D}$, i.e., we map it into a keyword set, and then map it into a keyword vector $\mathbf{K}(\mathcal{D})$;
4: Construct a keyword hash index for the document $\mathcal{D}$;
5: Initialize the related concept set $\mathbb{W}(\mathcal{D})$ as an empty set;
6: **for all** $\mathcal{W} \in \mathbb{W}$ **do**
7: According to a heuristic selection rule (Rule 1, 2 or 3), together with the keyword hash index of $\mathcal{D}$, we judge whether the concept $\mathcal{W}$ is related to $\mathcal{D}$; if yes, then we add $\mathcal{W}$ into the related concept set $\mathbb{W}(\mathcal{D})$;
8: **end for**
9: According to the keyword vector $\mathbf{K}(\mathcal{D})$ and the keyword vector $\mathbf{K}(\mathcal{W})$ of each concept $\mathcal{W} \in \mathbb{W}$, we use Definition 4 to map the document $\mathcal{D}$ to a concept vector $\mathbf{W}(\mathcal{D})$;
10: **end for**
11: According to the keyword vector $\mathbf{K}(\mathcal{D})$ and concept vector $\mathbf{W}(\mathcal{D})$ of each document $\mathcal{D} \in \mathbb{D}$, together with Definition 8, we compute the overall similarity between any two documents, upon which we classify all the documents in $\mathbb{D}$.

the semantic similarity between the two documents can be computed as follows

$$\mathbf{sim^w}(\mathcal{D}_1, \mathcal{D}_2) = \mathbf{cos}\angle\mathbf{W}(\mathcal{D}_1), \mathbf{W}(\mathcal{D}_2)$$

Definition 7 (Textual Similarity). Given text documents $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathbf{K}(\mathcal{D}_1)$ and $\mathbf{K}(\mathcal{D}_2)$ denote two keyword vectors of the documents. Then, the textual similarity between the two documents can be computed as follows

$$\mathbf{sim^k}(\mathcal{D}_1, \mathcal{D}_2) = \mathbf{cos}\angle\mathbf{K}(\mathcal{D}_1), \mathbf{K}(\mathcal{D}_2)$$

Definition 8 (Document Similarity). Given two text documents $\mathcal{D}_1$ and $\mathcal{D}_2$, based on the semantic similarity and textual similarity between them, the overall similarity between the two documents can be computed as follows

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = \alpha \cdot \mathbf{sim^w}(\mathcal{D}_1, \mathcal{D}_2) + (1 - \alpha) \cdot \mathbf{sim^k}(\mathcal{D}_1, \mathcal{D}_2)$$

where α ($0 \leq \alpha \leq 1$) denotes a balance parameter, and the greater the parameter value, the higher the weight of the semantic similarity, and otherwise the higher the weight of the textual similarity. In the experiments, α is set to 0.5.

3.4. Algorithm Implementation

According to data preprocessing, heuristic selection and document similarity computation, Algorithm 1 presents an efficient Wikipedia semantic matching approach to document classification, called WMDC. From the algorithm description, it can be seen that at Line 4, we construct a keyword hash index for each document $\mathcal{D} \in \mathbb{D}$, upon which given any keyword k, we can quickly judge whether the keyword k is contained in the document $\mathcal{D}$, thereby enabling the subsequent heuristic selection rule (Line 7) to quickly pick out a set $\mathbb{W}(\mathcal{D})$ of concepts related to the document $\mathcal{D}$ from the Wikipedia reference space $\mathbb{W}$. In addition, we also sort all the keywords of each concept in the Wikipedia reference space, and sort all the keywords of each input document, so as to improve the efficiency of similarity computation between a concept $\mathcal{W}$ and a document $\mathcal{D}$ at Line 9 (i.e., Line 9 only needs to scan all the keywords of the concept $\mathcal{W}$ and the document $\mathcal{D}$ once). It should be pointed out that the construction of the Wikipedia semantic reference space $\mathbb{W}$ (Line 1) is completed in advance, so it does not affect the efficiency of online document classification.

Below, we analyze the online execution efficiency of the WMDC approach. Let n_1 denote the size of the reference space $\mathbb{W}$, i.e., $n_1 = |\mathbb{W}|$, n_2 denote the average number of keywords contained in each concept $\mathcal{W}$, i.e., $n_2 = \frac{1}{|\mathbb{W}|}\sum_{\mathcal{W}\in\mathbb{W}}|\mathcal{W}|$, m_1 denote the size of the input document set $\mathbb{D}$, i.e., $m_1 = |\mathbb{D}|$, and m_2 denote the average number of keywords contained in each document $\mathcal{D}$, i.e., $m_2 = \frac{1}{|\mathbb{D}|}\sum_{\mathcal{D}\in\mathbb{D}}|\mathcal{D}|$. If the selectivity of a heuristic selection rule to the reference space is equal to σ, i.e., $\sigma = \frac{|\mathbb{W}(\mathcal{D})|}{|\mathbb{W}|}$, then from Algorithm 1, we conclude that the time complexity of constructing a document concept vector is equal to $O(\sigma n_1(n_2+m_2))$. Thus, the time complexity of document classification is equal to $O(\sigma n_1 m_1(n_2+m_2))$. However, if without the heuristic selection rules, the time complexity of document classification is equal to $O(n_1 m_1(n_2+m_2))$. Because only a small number of related concepts will be picked out by a heuristic selection rule from the reference space, we have that $\sigma n_1 \ll n_1$ (thousands versus millions), consequently improving the efficiency of online document classification. In addition, we also see that given a document set $\mathbb{D}$, the time complexity of its classification is proportional to the size of the Wikipedia reference space $\mathbb{W}$, i.e., the classification efficiency for a given document set can scale up when there is a large Wikipedia reference space.

4. Evaluation Experiment

It has been demonstrated that by mapping documents into concept vectors in the Wikipedia reference space, Wikipedia semantic matching can capture the semantic information behind keywords, and thus improve the accuracy of document similarity computation and then the accuracy of document classification. Hence, the experimental evaluations in this section are divided into two parts: (1) the first part focuses on the effectiveness of the heuristic selection rules, i.e., whether they can effectively pick out the related concepts for a document so as to ensure the efficiency and quality of document concept vector construction; and (2) the second part further evaluates the effectiveness of our approach by the comparison with Wikipedia matching, i.e., whether our approach can improve the classification efficiency, without compromising the classification accuracy.

4.1. Experimental Setup

Before the experimental evaluations, we briefly describe the experimental setup, i.e., the Wikipedia dataset, the text document set, classification algorithm and system configuration.

(1) Wikipedia dataset. As an open project, all the Wikipedia datasets can be directly downloaded from the Internet[3]. The dataset we used was published in 2011-10-07, which consists of about $7,512,630$ title words and $3,304,175$ concepts. In order to obtain them, we first executed the SQL script[4] to initialize an empty database. Second, we imported data into the database by executing the scripts[5] on relevant tables (e.g., Page). Finally, with the help of the database, we obtained a large number of concepts and their titles (actually, to improve the experiment efficiency, we chose about one million concepts to construct the Wikipedia reference space). In addition, the description of how to construct the Wikipedia reference space based on these concepts has been presented in Section 3.1.

(2) Document dataset. The document set used in the experiments was Reuters-21578[6], which consists of $11,367$ manually labeled text documents that are divided into 82 clusters. In order to enhance the experimental effect, we removed those clusters with less than 15 documents or more than 200 documents, leaving 30 clusters comprising of about $1,600$ documents.

(3) Classification algorithm. In our approach, the classification algorithm we used is K-Means [24, 20], which is a famous algorithm developed based on a textual similarity function, and has been demonstrated to be accurate and efficient. It should be pointed out that our approach is open, i.e., other classification algorithms can also be used, e.g., the fine-designed algorithms presented in [28, 30, 19]. In the experiments, the reason of using K-Means is to simplify the approach implementation.

(4) System configuration. In the experiments, all the algorithms were implemented by using the Java programming language. The experiments were performed on a Java Virtual Machine (version 1.7.0_07) with an Intel i7-5500U CPU and 2 GB of maximum working memory.

4.2. Rule Effectiveness Evaluation

In order to generate a document concept vector efficiently, according to the title information of each concept, our approach uses a heuristic selection rule to quickly pick out the related concepts from the reference space. It can be seen that on the one hand, the more relevant the concepts picked out, the better the quality of concept vector construction, and in turn the better the accuracy of document similarity computation; but on the other hand, the fewer the related concepts picked out, the better the efficiency of concept vector construction, and in turn the better the efficiency of document similarity computation. Therefore, we use two factors to measure the effectiveness of a heuristic selection rule, i.e., selectivity and relevance.

[3]http://dumps.wikimedia.org/enwiki/20111007/

[4]http://en.wikipedia.org/wiki/wikipedia:database_download

[5]http://dumps.wikimedia.org/enwiki/20111007/

[6]http://www.daviddlewis.com/resources/testcollections/reuters21578/

Metric 1 (Selectivity). Given a heuristic selection rule R, a document set $\mathbb{D}$ and the Wikipedia reference space $\mathbb{W}$, $\mathbb{W}(\mathcal{D})$ denotes a set of related concepts selected by the rule R for any document $\mathcal{D} \in \mathbb{D}$. Then, the selectivity (maximum, average and minimum) of R over $\mathbb{W}$ and $\mathbb{D}$ can be defined as follows.

$$\textbf{MAX\_SE}(R, \mathbb{W}, \mathbb{D}) = \max_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D})|}{|\mathbb{W}|}$$

$$\textbf{AVE\_SE}(R, \mathbb{W}, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D})|}{|\mathbb{W}|}$$

$$\textbf{MIN\_SE}(R, \mathbb{W}, \mathbb{D}) = \min_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D})|}{|\mathbb{W}|}$$

Metric 2 (Relevance). Given a heuristic selection rule R, a document set $\mathbb{D}$ and the Wikipedia reference space $\mathbb{W}$, $\mathbb{W}(\mathcal{D})$ denotes a set of related concepts selected by the rule R for any document $\mathcal{D} \in \mathbb{D}$, and $\mathbb{W}^*(\mathcal{D})$ denotes a set of all the concepts really related to $\mathcal{D}$ in the reference space $\mathbb{W}$. Then, the relevance (maximum, average and minimum) of R over $\mathbb{W}$ and $\mathbb{D}$ can be defined as follows.

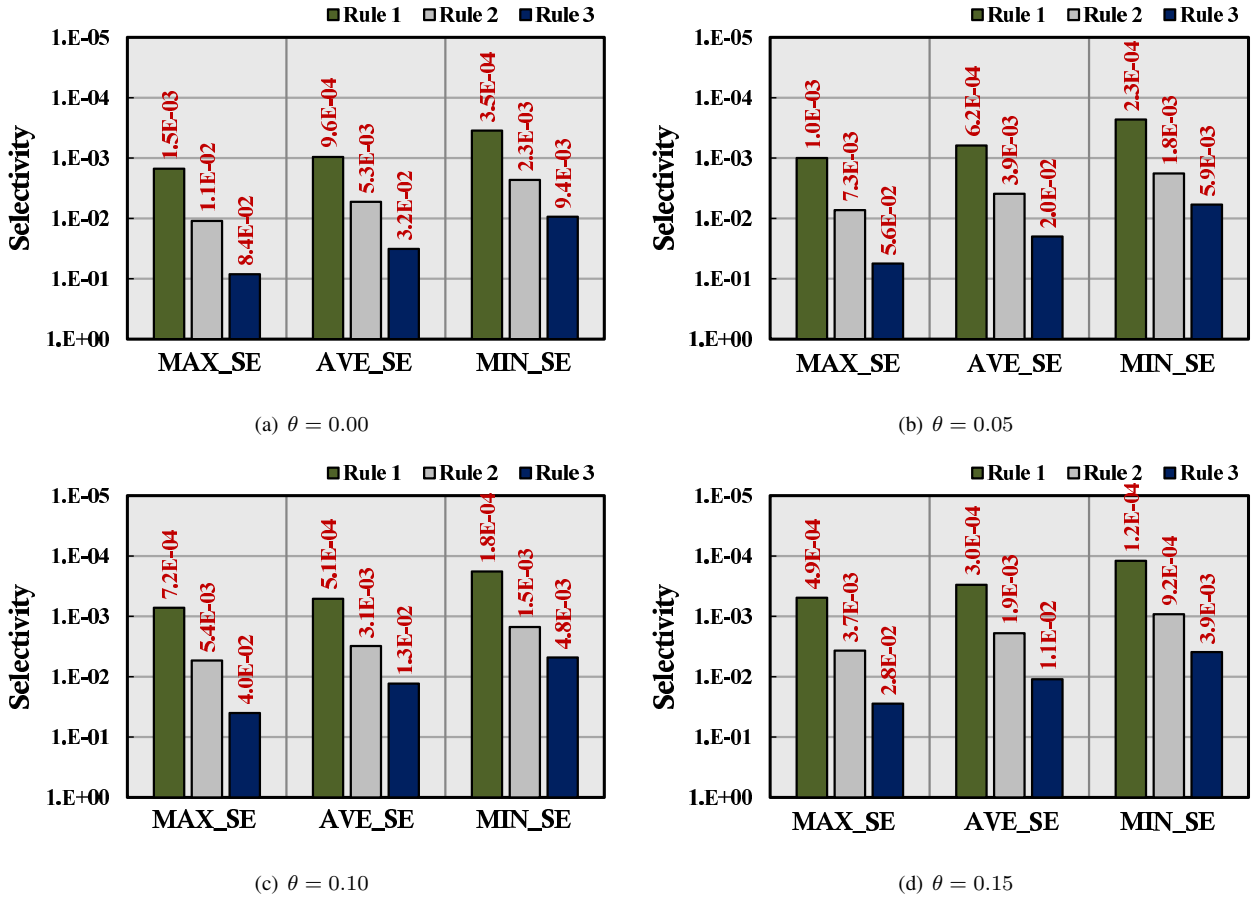
$$\textbf{MAX\_RE}(R, \mathbb{W}, \mathbb{D}) = \max_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D}) \cap \mathbb{W}^*(\mathcal{D})|}{|\mathbb{W}^*(\mathcal{D})|}$$

$$\textbf{AVE\_RE}(R, \mathbb{W}, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D}) \cap \mathbb{W}^*(\mathcal{D})|}{|\mathbb{W}^*(\mathcal{D})|}$$

$$\textbf{MIN\_RE}(R, \mathbb{W}, \mathbb{D}) = \min_{\mathcal{D} \in \mathbb{D}} \frac{|\mathbb{W}(\mathcal{D}) \cap \mathbb{W}^*(\mathcal{D})|}{|\mathbb{W}^*(\mathcal{D})|}$$

From Metrics 1 and 2, we see that (1) the smaller the rule selectivity, the better the efficiency, and (2) the higher the rule relevance, the better the quality. However, the two metrics conflict with each other. In this group of experiments, we aim to evaluate the effectiveness of each heuristic selection rule in terms of selectivity and relevance. In the experiments, we set four different threshold values for each heuristic selection rule, which are $\theta = 0 \cdot \theta^*$, $\theta = 0.05 \cdot \theta^*$, $\theta = 0.1 \cdot \theta^*$ and $\theta = 0.15 \cdot \theta^*$, where the values are set based on the maximum threshold value θ^*. Here, the threshold θ^* means the maximum value that the threshold θ can be set under the precondition of ensuring the concept set $\mathbb{W}(\mathcal{D})$ non-empty. The selectivity experimental results are shown in Figure 2, and the relevance experimental results are shown in Figure 3.

From the four subfigures of Figure 2 (the caption of each subfigure denotes the threshold value used in the experiments, and the Y axis is in a logarithmic descending order), we have the following observations. First, all the three heuristic selection rules perform well in terms of selectivity, where the maximum selectivity, minimum selectivity and average selectivity of each rule are all less than 0.09. Second, for the same threshold value, the selectivity of Rule 1 (Full-Title) is less than that of Rule 2 (All-Keyword), the selectivity of Rule 2 is less than that of Rule 3 (Any-Keyword), and the selectivity of each rule decreases with the increasing of the threshold value. Finally, from the minimum selectivity metric, we see that each rule can effectively pick out the related concepts for each document. This is mainly due to the broad coverage of semantic concepts of



(a) $\theta = 0.00$

(b) $\theta = 0.05$

(c) $\theta = 0.10$

(d) $\theta = 0.15$

Figure 2: Selectivity Evaluation Results

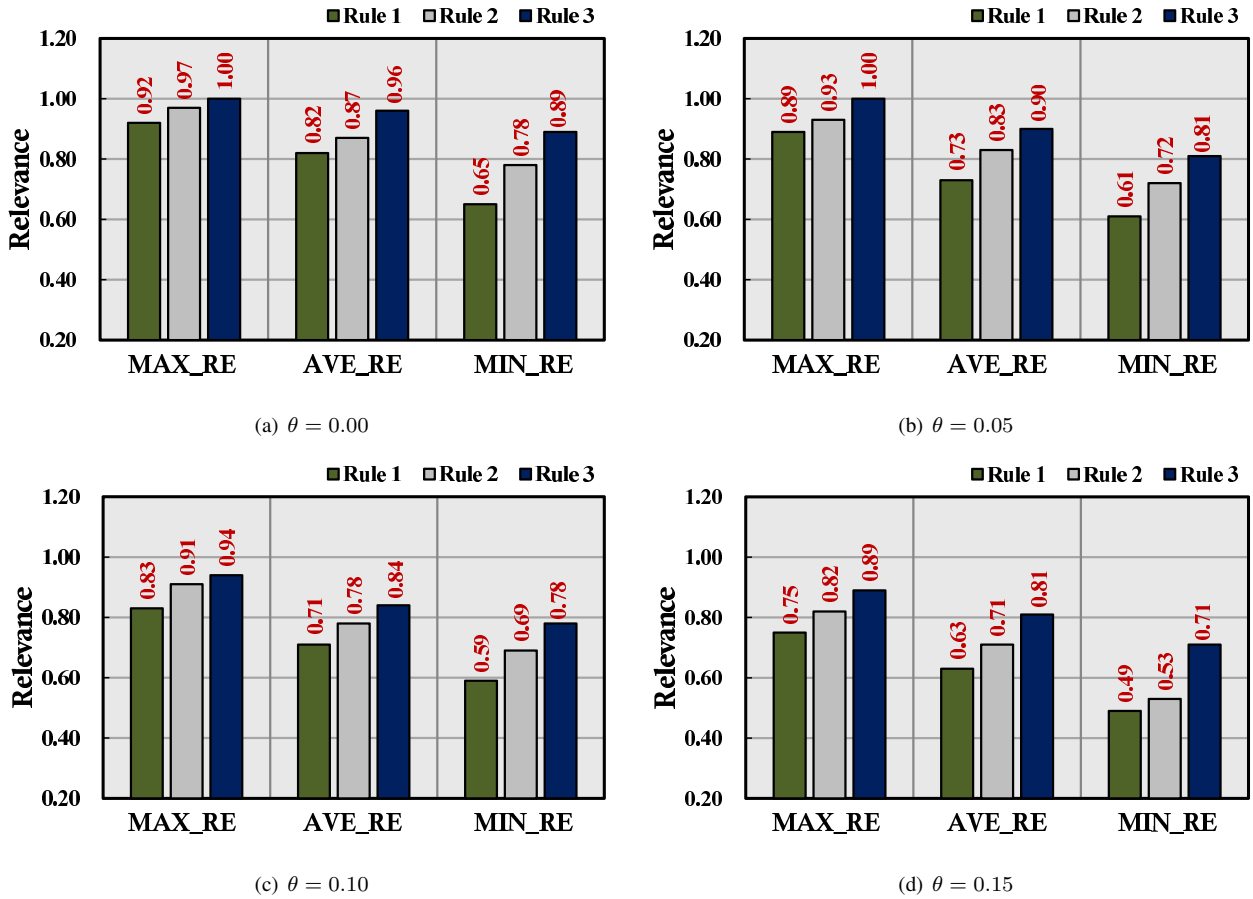
the Wikipedia reference space. From the above experiments, we conclude that our approach can greatly reduce the number of reference concepts involved in full-text keyword matching, which certainly will improve the generation efficiency of a document concept vector, where Rule 1 (Full-Title) has the best generation efficiency.

From Figure 3 (note that the Y axis is in a linear ascending order), we have the following observations. First, the concept sets determined by the heuristic selection rules have high relevance to the documents (the maximum relevance, minimum relevance and average relevance are all greater than 0.5), i.e., most of the concepts related to the documents can be picked out by the rules from the Wikipedia reference space. Second, for the same threshold value, the relevance of Rule 1 is less than that of Rule 2, the relevance of Rule 2 is less than that of Rule 3, and the relevance of each rule decreases with the increasing of the threshold value. Finally, even if the threshold parameter of each rule is set to the minimum value, we cannot guarantee to obtain all the related concepts for a document, which may lead to a negative impact on the effectiveness of the Wikipedia semantic matching to a certain degree. However, from the experimental results presented in Section 4.3, we conclude that such a negative impact is not remarkable.

From the above experiment results, we conclude that if we only consider the efficiency factor, then Rule 1 is better than Rule 2 and Rule 2 is better than Rule 3; otherwise, if we only consider the quality factor, then Rule 1 is worse than Rule 2 and Rule 2 is worse than Rule 3. As a result, what heuristic selection rule we should choose depends on our actual demand (i.e., give priority to efficiency or quality). Overall, our proposed approach can accurately find out the related concepts for a given document, which certainly will ensure the quality of document concept vector construction and in turn the accuracy of document classification.

4.3. Classification Effectiveness Evaluation

A number of semantic matching approaches have been proposed for text similarity computation [8, 34, 35, 6]. However, a number of studies [10, 11, 3, 1] have demonstrated that the Wikipedia semantic matching approach can better capture the semantic information contained in text documents, resulting in better accuracy on document classification than other algorithms. Thus, in the experiments, we do not compare our approach with other algorithms again, and only compare against the Wikipedia semantic matching approach and the keyword matching approach (it is used as the baseline). In this group of experiments, we aim to further evaluate the effectiveness of our approach, i.e., whether the approach can improve the running efficiency, under the precondition of not compromising the accuracy of document classification. Here,



(a) $\theta = 0.00$

(b) $\theta = 0.05$

(c) $\theta = 0.10$

(d) $\theta = 0.15$

Figure 3: Relevance Evaluation Results

we use the metric 'purity' similar to that mentioned in [3] to measure the accuracy of document classification. In addition, for evaluating the efficiency of our approach, we use the metric 'vector construction time' to measure the efficiency of concept vector construction, which only focuses on Lines 3 to 9 in Algorithm 1, without considering the efficiency of document classification at Line 11.

Metric 3 (Purity). For a document set $\mathbb{D}$, $\mathbf{D}^m = \{\mathbb{D}_i^m\}_{i=1}^n$ ($\mathbb{D}_i^m \subset \mathbb{D}$; $\mathbb{D}_i^m \cap \mathbb{D}_j^m = \oslash$; $\cup_{i=1}^n \mathbb{D}_i^m = \mathbb{D}$) denotes a manual classification result, and $\mathbf{D}^a = \{\mathbb{D}_i^a\}_{i=1}^n$ denotes the classification result of an approach R ($\mathbb{D}_i^a \subset \mathbb{D}$; $\mathbb{D}_i^a \cap \mathbb{D}_j^a = \oslash$; $\cup_{i=1}^n \mathbb{D}_i^a = \mathbb{D}$). Then, the accuracy (maximum, average and minimum) of document classification of the approach R can be measured as follows.

$$\textbf{MAX\_PU}(R, \mathbf{D}^a, \mathbf{D}^m) = \max_{\mathbb{D}_i^a \in \mathbf{D}^a} \max_{\mathbb{D}_j^m \in \mathbf{D}^m} \frac{|\mathbb{D}_i^a \cap \mathbb{D}_j^m|}{|\mathbb{D}_i^a|}$$

$$\textbf{AVE\_PU}(R, \mathbf{D}^a, \mathbf{D}^m) = \frac{1}{n} \sum_{\mathbb{D}_i^a \in \mathbf{D}^a} \max_{\mathbb{D}_j^m \in \mathbf{D}^m} \frac{|\mathbb{D}_i^a \cap \mathbb{D}_j^m|}{|\mathbb{D}_i^a|}$$

$$\textbf{MIN\_PU}(R, \mathbf{D}^a, \mathbf{D}^m) = \min_{\mathbb{D}_i^a \in \mathbf{D}^a} \max_{\mathbb{D}_j^m \in \mathbf{D}^m} \frac{|\mathbb{D}_i^a \cap \mathbb{D}_j^m|}{|\mathbb{D}_i^a|}$$

Metric 4 (Vector Generation Time). For a document set $\mathbb{D}$, $\textbf{time}(\mathcal{D})$ denotes the time of an approach R spent on mapping any document $\mathcal{D} \in \mathbb{D}$ into a concept vector in the reference space (or a keyword vector for the keyword matching approach). Then, the feature vector generation time of the approach R over $\mathbb{D}$ can be measured as follows.

$$\textbf{MAX\_TM}(R, \mathbb{D}) = \max_{\mathcal{D} \in \mathbb{D}} \textbf{time}(\mathcal{D})$$

$$\textbf{AVE\_TM}(R, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{\mathcal{D} \in \mathbb{D}} \textbf{time}(\mathcal{D})$$

$$\textbf{MIN\_TM}(R, \mathbb{D}) = \min_{\mathcal{D} \in \mathbb{D}} \textbf{time}(\mathcal{D})$$

From Metric 3, we see that the greater the purity of document classification, the better the accuracy of document classification. In the experiments, we set the same threshold value for each heuristic selection rule (i.e., it is set to zero). The experimental results are shown in Figure 4, where 'Wiki' denotes the Wikipedia matching approach without using any heuristic selection rule, and 'Keyw' denotes the keyword matching approach. From Figure 4, we see that compared to the document classification approach only based on Wikipedia semantic matching, our approach exhibits similar performance on the accuracy of document classification; and compared to the baseline keyword matching approach, our approach is better. Specifically, the compromise on the accuracy of document classification

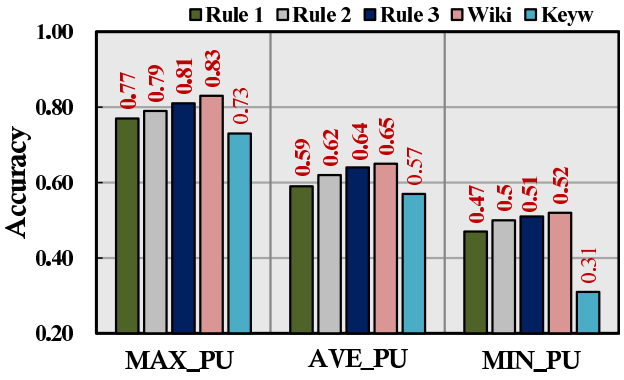


Figure 4: Accuracy Evaluation Results

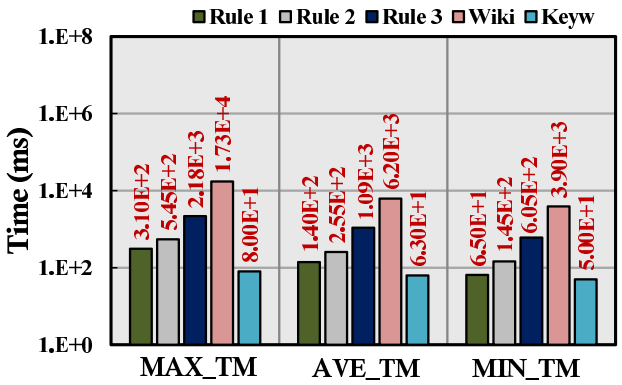


Figure 5: Efficiency Evaluation Results

caused by each heuristic selection rule is less than 5%. In addition, we also see that for the three heuristic selection rules, their performances on classification accuracy are similar to each other, where Rule 1 (Full-Title) has the best accuracy and Rule 3 (Any-Keyword) has the worst. Here, the performance differences of the heuristic selection rules are mainly caused by their relevance differences (see Section 4.2), i.e., some reference concepts actually related to the documents are unselected by the rules. However, from the above similar performances, we also see that the relevance of each unselected concept to the document is smaller (i.e., the full-text matching value is approximately equal to zero).

From Metric 4, we see that the smaller the vector generation time, the higher the efficiency of document classification. The experimental results are shown in Figure 5 (the unit of the Y axis is millisecond). From Figure 5, we can see that compared to the Wikipedia semantic matching, our approach can effectively improve the generation efficiency of document concept vector. Especially, when using Rule 1, the efficiency of our approach is close to that of the keyword matching approach. This is due to that the heuristic selection rules can quickly capture the related concepts for each document, and thus reduce the number of reference concepts involved into the full-text keyword matching operations. In addition, from Figure 5, we can also see that due to the different selectivity values, the heuristic selection rules exhibit different vector generation times, where Rule 1 has the best efficiency and Rule 3 has the worst efficiency.

Finally, from the above experiments, we conclude that the proposed approach can improve the efficiency of Wikipedia semantic matching under the precondition of not compromising the accuracy of document classification.

5. Conclusion

In this paper, by using Wikipedia semantic matching together with keyword matching, we proposed an efficient Wikipedia semantic matching approach to text document classification, called WMDC. In the WMDC approach, we formulate several heuristic selection rules to quickly pick out related concepts for a document from the large-scale Wikipedia reference space, making it no longer necessary to conduct full-text keyword matching over all the concepts in the reference space, and consequently improving the efficiency of document classification. Finally, the evaluation experiments demonstrated the accuracy and efficiency of the WMDC approach: (1) the heuristic selection rules used in the approach can greatly reduce the number of reference concepts involved in full-text keyword matching, and thus improve the generation efficiency of a document concept vector; (2) the heuristic selection rules can accurately select related concepts for a document, and thus ensure the generation quality of a document concept vector; and (3) the document classification approach can well improve the efficiency of document classification, without the need to compromise the accuracy of document classification. Therefore, the proposed approach can satisfy the requirements of online document classification in terms of efficiency and accuracy.

Acknowledgment

We thank anonymous reviewers for their valuable comments. The work is supported by the Zhejiang Provincial Natural Science Foundation of China (LY15F020020, LQ13F020011 and LQ16G010006), the Jiangxi Provincial Natural Science Foundation of China (20161BAB202036), the Wenzhou Science and Technology Program (G20160006 and Y20160070), the National Natural Science Foundation of China (61402337 and 61572367) and the Visiting Scholar Program of Zhejiang Provincial Education Department (FX2014211).

References

[1] N. P. Alexander, C. Chin-Wan, A wikipedia matching approach to contextual advertising, World Wide Web 13 (3) (2010) 251–274.

[2] J. Amir H., M. Fariborz, M. R. Meybodi, Conceptual feature generation for textual information using a conceptual network constructed from wikipedia, Expert Systems 33 (1) (2016) 92–106.

[3] H. Anna, M. David, F. Eibe, H. W. Ian, Clustering documents using a wikipedia-based concept representation, Advances in Knowledge Discovery and Data Mining (2009) 628–636.

[4] A.-S. Bashar, M. Sung-Hyon, Wikipedia-based query phrase expansion in patent class search, Information Retrieval 17 (5-6) (2014) 430–451.

[5] L. Bing, S. Jiang, W. Lam, Adaptive concept resolution for document representation and its applications in text mining, Knowledge-Based Systems 74 (2015) 1–13.
[6] A. Broder, M. Fontoura, V. Josifovski, A semantic approach to contextual advertising, in: Proc. of SIGIR, 2007, pp. 559–566.
[7] L. Chi, B. Li, X. Zhu, Context-preserving hashing for fast text, in: Proc. of SDM, 2014, pp. 100–108.
[8] K. Dave, S. Lawrence, D. M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in: Proc. of WWW, 2003, pp. 519–528.
[9] A. Eneko, S. Aitor, Personalizing pagerank for word sense disambiguation, in: Proc. of IACL, 2009, pp. 33–41.
[10] G. Evgeniy, M. Shaul, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in: Proc. of IJCAI, 2007, pp. 1606–1611.
[11] G. Evgeniy, M. Shaul, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research 34 (4) (2009) 443–498.
[12] M. Fan, Q. Zhou, T. Fang Zheng, Mining the personal interests of microbloggers via exploiting wikipedia knowledge, Lecture Notes in Computer Science 8404.
[13] T.-K. Fan, C.-H. Chang, Sentiment-oriented contextual advertising, Knowledge and Information Systems 23 (3) (2010) 321–344.
[14] G. Forman, Bns feature scaling: An improved representation over tf-idf for svm text classification, in: Proc. of CIKM, 2009, pp. 263–270.
[15] J. Hu, L. Fang, L. Fang, Enhancing text clustering by leveraging wikipedia semantics, in: Proc. of SIGIR, 2008, pp. 179–186.
[16] J. Hu, G. Wang, L. Fred, Understanding user's query intent with wikipedia, in: Proc. of WWW, 2009, pp. 471–480.
[17] J. Huang, M. Peng, H. Wang, J. Cao, G. Wang, A probabilistic method for emerging topic tracking in microblog stream, World Wide Web 1 (2016) 1–26.
[18] J. J. Jason, Knowledge distribution via shared context between blog-based knowledge management systems: A case study of collaborative tagging, Expert Systems with Applications 36 (7) (2009) 10627–10633.
[19] M. E. Kabir, H. Wang, E. Bertino, Efficient systematic clustering method for k-anonymization, Acta Informatica 48 (1) (2011) 1–26.
[20] R. Md Anisur, I. Md Zahidul, Kernel penalized k-means: A feature selection method based on kernel k-means, Information Sciences 322 (20) (2014) 150–160.
[21] Q. Muhammad Atif, Utilizing wikipedia for text mining applications, ACM SIGIR Forum 49 (2) (2016) 151–151.
[22] R. Pum-Mo, J. Myung-Gil, K. Hyun-Ki, Open domain question answering using wikipedia-based knowledge model, Information Processing and Management 50 (5) (2014) 683–692.
[23] M. A. Ramiz, A new sentence similarity measure and sentence based extractive technique for automatic text summarization, Expert Systems with Applications 36 (4) (2009) 7764–7772.
[24] S. Rudolf, S. Kristian, Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters, Knowledge-Based Systems 57 (2014) 1–7.
[25] R. Santosh Kumar, S. Shailendra, B. P. Joshi, A semantic approach for question classification using wordnet and wikipedia, Pattern Recognition Letters 31 (13) (2010) 1935–1943.
[26] D. Vrandecic, M. Krtotzsch, Wikidata: A free collaborative knowledgebase, Communications of the ACM 57 (10) (2014) 78–85.
[27] F. Wang, Z. Wang, S. Wang, Z. Li, Exploiting description knowledge for keyphrase extraction, in: Proc. of PRICAI, 2014, pp. 130–142.
[28] M. Wang, W. Fu, S. Hao, Scalable semi-supervised learning by efficient anchor graph regularization, IEEE Transactions on Knowledge and Data Engineering 28 (7) (2016) 1864–1877.
[29] M. Wang, W. Li, D. Liu, Facilitating image search with a scalable and compact semantic mapping, IEEE Transactions on Cybernetics 45 (6) (2015) 2564–2577.
[30] M. Wang, X. Liu, X. Wu, Visual classification by ℓ_1-hypergraph modeling, IEEE Transactions on Knowledge and Data Engineering 27 (9) (2015) 2564–2577.
[31] Y. Wei, J. Wei, Z. Yang, Joint probability consistent relation analysis for document representation, in: Proc. of DASFAA, 2016, pp. 517–532.
[32] H. C. Wu, P. L. Robert Wing, W. Kam Fai, Interpreting tf-idf term weights as making relevance decisions, ACM Transactions on Information Systems 26 (3) (2008) 13.
[33] G. Xu, Z. Wu, G. Li, E. Chen, Improving contextual advertising matching by using wikipedia thesaurus knowledge, Knowledge and Information Systems 43 (3) (2015) 599–631.
[34] I. Yoo, X. Hu, I.-Y. Song, Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering, in: Proc. of KDD, 2006, pp. 791–796.
[35] X. Zhang, L. Jing, H. X., A comparative study of ontology based term similarity measures on document clustering, in: Proc. of DASFAA, 2007, pp. 115–126.
[36] H. Zheng, Z. Li, S. Wang, Y. Zhao, J. Zhou, Aggregating inter-sentence information to enhance relation extraction, in: Proc. of AAAI, 2016, pp. 3108–3115.