TELKOMNIKA, Vol. 11, No. 4, April 2013, pp. 1909~1915
ISSN: 2302-4046

A HowNet-based Semantic Relatedness Kernel for Text Classification

Pei-ying ZHANG
College of Computer and Communication Engineering/ China University of Petroleum
QingDao, ShanDong China
e-mail: 25640521@qq.com

Abstract

The exploitation of the semantic relatedness kernel has always been an appealing subject in the context of text retrieval and information management. Typically, in text classification the documents are represented in the vector space using the bag-of-words (BOW) approach. The BOW approach does not take into account the semantic relatedness information. To further improve the text classification performance, this paper presents a new semantic-based kernel of support vector machine algorithm for text classification. This method firstly using CHI method to select document feature vectors, secondly calculates the feature vector weights using TF-IDF method, and utilizes the semantic relatedness kernel which involves the semantic similarity computation and semantic relevance computation to classify the document using support vector machines. Experimental results show that compared with the traditional support vector machine algorithm, the algorithm in the text classification achieves improved classification F1-measure.

Keywords*: semantic relatedness kernel, text classification, semantic similarity computation, semantic relevance computation, support vector machine*

Copyright © 2013 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

With the rapid growth of online information, text classification has become one of the key tools for automatically handing and organizing text information. The key steps in text classification are document representation and classifier training using a corpus of labeled documents. In the commonly used bag of words representation method, documents are represented by numeric vectors whose components are weights given to different words or terms occurring in the document. Despite its ease of use, bag of words representation can not handle word synonymy and polysemy problems and does not consider semantic relatedness between words. The lack of semantics in the bag of words representation limits the effectiveness of automatic text classification methods.

Semantic relatedness computation including semantic similarity computation and semantic relevance computation is widely used in many applications such word sense disambiguation, information retrieval and text clustering, etc. Semantic relatedness computation can be divided into two categories, statistical approaches and approaches based on semantic ontology dictionary. Statistical approaches computed semantic relatedness by using the corpus of training set. In the absence of external semantic knowledge, corpus-based statistical methods such as Latent Semantic Analysis (LSA) [1] can be applied to alleviate the synonymy problem, but the problem of polysemy still remains. But it is sensitive to the training corpus. L. Lillian [2] used joint entropy and P. Brown et al [3] used average mutual information to compute the similarity between words. J.H. Lee et al. [4] used the distance between words in WordNet to compute semantic similarity between English words. Resnik [5] used the largest amount of information of the ancestors of nodes to measure the semantic similarity of two English words. Agirre and Rigau [6] used the information such as concept distance, depth and area density of concept hierarchy tree to compute the semantic similarity between English words. Liu Qun et al. [7] exploited distance of sememes in the sememe tree to compute semantic similarity between two words for example-based machine translation. Li Sujian [8] took the semantic relevance of words as maximum sum of the sememe relevance of the word's concept. In his work, he

Received January 13, 2013; Revised February 14, 2013; Accepted February 26, 2013

computed the sememe with other sememe which is in the sememe extension set of another word.

The support vector machine (SVM) method is a new and very popular technique for text classification in the machine learning community. However, the traditional SVM algorithm is only taking into account the characteristics of words in the document such as word frequency information, without considering the semantic information which documents contained, and limited the SVM algorithm applications. In this paper, we present and evaluate a semantically-enriched BOW representation for text classification. We adopt the HowNet-based semantic relatedness measure to build a smoothing matrix and a kernel for semantically adjusting the BOW representation.

We will present in section 2 some related works. Section 3 gives our semantic mapping we called semantic smoothing that is equivalent to the definition of a semantic relatedness metric. Section 4 will devoted to the brief presentation of support vector machine with references to previous works on text classification. In section 5, experimental results concerning support vector machine will be discussed. In section 6, a conclusion and different openings to this work will be presented.

2. Related Works

During the last decades, a large number of text classification systems have been proposed using a variety of approaches such as support vector machine, boosting algorithms, term frequency and inverse document frequency. Most of these systems use the bag-of-words model or vector space representation by having individual words as basis representative features for the document content. While bag-of words approaches present a good performance on many machine-learning tasks due to low computational cost and inherent parallelism, their limitations are also well acknowledged. Especially, the underlying classification scheme is restricted to detecting patterns within the used terminology only, which excludes conceptual patterns as well as any semantically related words. It is thus possible to gain better results across multiple domains by utilizing an external semantic thesaurus like WordNet that defines an upper-level of relationships among most of the terms in the testing data.

The importance of embedding semantic relatedness between two text segments for text classification was initially highlighted in [9] where semantic similarity between words has been used for the semantic smoothing of the TF-IDF vectors. Semantic-aware kernels have been proposed by Mavroeidis et al. [10] who propose a generalized vector space model with WordNet senses and their hypernyms to improve text classification performance. Typically, WordNet is a database for the English language containing semantic lexicon that organizes words into groups of synsets. Every synset stands for a single word prototype that refers to a group of words that share the same meaning. In addition to making use of relations in WordNet, feature such as part-of speech tags have been considered in [11]. This motivates the intensive research carried out in this issue which had given rise to a variety of implemented systems incorporating features derived from the common semantic theosaurus WordNet and its words relations. Strictly speaking, the hierarchical organization of WordNet involves important distinction between various part-of-speech parts. Indeed, while categorization of nouns into underlying taxonomies, headed by a unique beginner such as animite or artifact is straighforward, this does not extend to verbs, which are rather partitioned into several semantic fields with many overlapping. This discrepancy between verbs and nouns obviously influences the calculus of semantic similarity, especially when dealing with sentences where the word-by-word semantic similarity has proven usually to be non-effective, which in turn influences negatively the performance of retrieval, summarization and classification tasks. This makes the debate of nouns versus verbs semantic similarity widely open. Stephan Bloehdorn and Alessandro Moschitti [12] combined syntactic and semantic kernels for text classification. Roberto Basili, Marco Cammisa and Alessandro Moschitti [13] use a semantic kernel to classify texts with very few training examples. Jamal Abdul Nasir, Asim Karim, George Tsatsaronis and Iraklis Varlamis [14] use a knowledge-based semantic kernel for text classification. Cristianini, N., Taylor, J.S. [15] use latent semantic kernel for text categorization. Shoushan Li, Rui Xia [16] propose a framework of feature selection methods for text categorization. J. Blitzer, M. Dredze and F. Pereira [17] use domain adaptation for sentiment classification. J. Brank et al [18] put forward interaction of feature selection methods and linear classification models. Literature [19-21] introduces the optimization model

and information expectation with cloud computing in china, proposes a sparse representation method. G. Forman [22] put forward a feature selection metrics for text classification. L. Marina et al [23] use conceptual extraction from ontologies for web classification.

This paper attempts to use the HowNet as the chinese semantic knowledge, caculates the similarity between words and the semantic relevance between words, combined the similarity and the relevance as the support vector machine semantic kernel for text classification.

3. Incorporation of Semantic Relatedness into Texts Metric

Term relatedness is used to design document similarities which are the critical functions of most text classification algorithms. This section firstly introduces the semantic relatedness between words based on HowNet, secondly give the description of document similarity kernels which can be incorporated to the definition of a kernel in support vector machine.

3.1. Semantic Relatedness of Words based on HowNet

Semantic relatedness of words is composed of semantic similarity and semantic relevance. Semantic similarity is the two words in different contexts can be used interchangeably without altering the text of the syntactic structure of semantic level. In HowNet, not every concept will correspond to the concept of a hierarchy tree in a node, but through a series of original meaning called sememe, uses a knowledge description language to describe a concept. These sememes through hypernym and hyponym relations organize into a hierarchy tree. We use Liu Qun [7] method to compute the semantic similarity between words.

Semantic similarity is the two words in different contexts can be used interchangeably without altering the text of the syntactic structure of semantic level. Different from traditional semantic dictionary, in HowNet, not every concept will correspond to the concept of a hierarchy tree in a node, but through a series of original meaning, the use of a knowledge description language to describe a concept. These sememes through superordinate and hyponym relations organize into a hierarchy tree. Our goal is to find a way to use this knowledge of the language that describes the similarity of two semantic expressions calculation.

Definition 1: Let two Chinese words to be W_1 and W_2, if W_1 has n meaning items: S_{11}, S_{12}, … , S_{1n}, W_2 has m meaning items: S_{21}, S_{22}, … , S_{2m}, W_1 and W_2 is the similarity of the senses maximum similarity, can be calculated as formula (1):

$$Sim(W_1, W_2) = \max_{i=1..n, j=1..m} Sim(S_{1i}, S_{2j}) \quad (1)$$

Thus, the similarity between words can be the issue boils down to the similarity between two concepts. Since all concepts are ultimately attributed to the orginal with the meaning (in some places with a specific word) to represent, so the orginal meaning of the concept of similarity calculation is the basis for calculating similarity. Since all of sememe composite a tree of hierarchy based on a relationship between the upper and lower, where a simple calculation by semantic distance similarity approach.

Definition 2: Let two sememes at this level in the orginal path of distance d, according to the formula (1), which was the original meaning of these two semantic distances between p1 and p2, the similarity between p1 and p2 is calculated as formula (2):

$$Similarity(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (2)$$

Where p1 and p2 represent two of which the sememe, d is the p1 and p2 in the context of the original path hierarchy in length, is a positive integer, αis an adjustable parameter.

For Chinese text classification, content words are the main basis for classification, so the notional calculation of similarity is the key.

Definition 3: Let the semantics of the expression notional concept consists of four parts. Independent meaning that the original description of the first type: the two concepts of similarity in this part of the record for the Sim1(s1, s2); other independent descriptive meaning

of the original: the first independent semantic meaning of the expression, in addition to the original than all the other independent meaning of the original (or specific words), this part of the two concepts of similarity is denoted Sim2(s1, s2); relational original descriptive meaning: the semantic relationship between the expression of all meaning with the original description of the type, the two concepts, this part of the similarity denoted Sim3(s1, s2); symbolic meaning of the original descriptive meaning of the symbol, the two concepts of similarity in this part of the record for the Sim4(s1, s2). The similarity between semantic expressions of two concepts is defined as formula (3):

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(S_1, S_2) \quad (3)$$

Where $\beta_i (1 \leq i \leq 4)$ is and adjustable parameters and satisfy the equation: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$.

Definition 4: The semantic relevance computation between words is by means of document [8] method, the formula is as follows:

$$\begin{aligned} &\text{Rele}(c_1, c_2) = \text{Rele}(def(c_1), def(c_2)) \\ &\text{Rele}(def(c_1), def(c_2)) \approx \sum_{s_i \in def(c_1), s_j \in def(c_2)} \max \text{Rele}(s_i, s_j) \\ &\text{def(c)} = \{s_i \mid \text{REL(c,s}_i)\} \\ &\text{Rele}(s_i, s_j) = \omega_s sim(s_i, s_j) + \omega_a asso(s_i, s_j) \end{aligned} \quad (4)$$

Where Rele(c1,c2) denotes the semantic relevance of two words, def(c) denotes the interpretation sememe set of word c. Readers can consult document [8] for more details.

Definition 3: The semantic relatedness between words can be defined using the semantic similarity and semantic relevancy, the computation is defined as formula (5):

$$SR(w_1, w_2) = (1 - \gamma) \times \text{Sim}(w_1, w_2) + \gamma \times \text{Rele}(w_1, w_2) \quad (5)$$

In formula (4), $SR(w_1,w_2)$ is the semantic relatedness between two words w_1 and w_2, $Sim(w_1,w_2)$ is the semantic similarity which defined by formula (1), $Rele(w_1,w_2)$ is the semantic relevancy between words which defined by formula (3). γ is a parameter to scale the weights of the two parts.

3.2. Document Similarity Kernel

Given two document d_1 and $d_2 \in D$ (the document set), we define their similarity as:

$$K(d_1, d_2) = \sum_{w_1 \in d_1, w_2 \in d_2} \lambda_1 \lambda_2 \times SR(w_1, w_2) \quad (6)$$

Where λ_1 and λ_2 are the weights of the word (features) w_1 and w_2 in the document d_1 and d_2, respectively, and $SR(w_1,w_2)$ is a term similarity function which defined by formula (5). The above document similarity could be used in kernel based support vector machines if it is a valid kernel function, i.e. if it satisfies the Mercer's condition [15]. Such conditions establish that the Gram matrix, G=K(di, dj) must be positive semi-definite. It has been shown in [15] that the matric G formed by the kernel function (Equation 2) with the outer matrix product $K(d_1,d_2)$ is indeed a positive semi-definite matrix.

4. Support Vector Machine

Support vector machines were introduced by Boser, Guyon and Vapnik in the seminal paper which can map the input data into a new space using kernel function centered into

support vectors and then make a linear separation in the new space. Let us define $\{(x_i,y_i), i=1 \ldots l\}$ the training sample for a binary classification problem. If the vector α and the scalar b are the parameters of the output hyperplane, f the SVM function is defined as formula (7):

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i \bullet y_i \bullet K(x, x_i) + b) \quad (7)$$

The introduction principle derived to determine weights of output hyperplane (and support data) is the maximization of the margin between the output hyperplane and the data encoded in the hidden layer of the network. To deal with non separable data, the margin concept was softenized [24] in order to accept some points that are on the wrong side of the margin frontiers.

To implement our approach, we have chosen the radial basis kernel that usually gets very good performance with few tuning and which is still a reproducing kernel when a metric is used as the argument of exponential.

$$K(x, y) = \exp(\gamma \| x - y \|^2) \quad (8)$$

After semantically smoothing the vectors, we get

$$K(x, y) = \exp(\gamma \| (x - y)^T \bullet SR \bullet (x - y) \|^2) \quad (9)$$

In formula (9), SR is the matrix whose element can be defined through the semantic relatedness between feature vectors. Other kernels based on the usual definition of similarity between two documents could be used as well, together with the semantic proximity matrix.

5. Experiments and Results

In order to the validity of the classification algorithm to measure and verify, we compared the commonly vector space model to the semantic-relatedness vector space model, the feature vector weight using TF-IDF method and feature selection algorithm using CHI method.

5.1. Performance Measures

To evaluate performance of the text classification system, we use the standard information retrieval measures that are precision, recall and F1-measure. The F1-measure is a kind of average of precision and recall.

Precision is defined as the ratio of correct classification of documents into categories to the total number of attempted classifications, is defined by formula (10):

$$precision = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (10)$$

Recall is defined as the ratio of correct classification of documents into categories to the total number of labeled data in the testing set, is defined by formula (11):

$$recall = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (11)$$

F1-measure is defined as the harmonic mean of precision and recall. Hence, a good classifier is assumed to have a high F1-measure, which indicates that the classifier performs well with respect to both precision and recall, is defined by formula (12):

$$F1\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \quad (12)$$

5.2. Experiments and Results

To test our proposed system, we used the proposed method compared with CHI*TF-IDF method and CHI*TF-CRF method using the F-1 measurement, the results are listed in Table 1.

Table 1. The Classification Results using the F-1 Measure

Class category	CHI+TF-IDF+SVM	CHI+TF-IDF+SVM_{SR}
IT	78.24	83.43
Education	86.14	90.24
Entertainment	85.86	92.41
Culture	79.45	84.12
History	81.26	91.24
Average	82.19	88.29

6. Conclusion

In this paper, we present a semantic relatedness kernel for smoothing the bag of words (BOW) representation. Firstly uses the CHI method to select document feature vectors, secondly calculate the feature vector weights using the TF-IDF method, and conducts two experiments using support vector machine to compare the F1 measure of two classification system between commonly text classification method and the semantic relatedness kernel text classification method. We find that semantic relatedness enhanced representation produces significant improvement in the F1-measure using support vector machine classifier.

As a next step, we will extend the BOW representation by incorporating discrimination information for text classification and compare our representation approaches for text classification task. In the future, we will study on the semantic text representation unit to denote the document feature vector and utilize the feature vector representation approaches for text classification.

Acknowledgements

This work is supported by “the Fundamental Research Funds for the Central Universities” of China University of Petroleum (East China). The authors are grateful for the anonymous reviewers who made constructive comments. First Author is corresponding author.

References

[1] Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA. Indexing by Latent Semantic Analysis. *JASIS*. 1990; 41(6): 391-407.
[2] Lee LJ. Similarity-based approaches to natural language processing. *Harvard University Technical Report TR-11-97*. 1997.
[3] PB. *Word sense disambiguation using tactical methods*. Proceedings of 29th Meeting of the Association for Computational Linguistics (ACL291). 1991; 201-207.
[4] HLJ. Information Retrieval based on conceptual distance in ISA hierarchies. *Journal of Documentation*. 1993.
[5] RP. Semantic similarity in Taxonomy: an information-based measure and its application to problems of ambiguity in Natural Language. *Journal of Artificial Intelligence Research*. 1999; (11): 95-130.
[6] Agirre E, Rigau G. *A proposal for word sense disambiguation using conceptual distance.* International Conference on Recent Advances in Natural Language Processing RANLP 95; 1995.
[7] Liu Qun, Li Sujian. Word Similarity Computing Based on HowNet. *Computational Linguistics and Information Processing*. 2002; 7: 59-76.
[8] Li Sujian. Research of relevancy between sentences based on semantic computation. *Computer Engineering and Applications*. 2002; 38(7): 75-76.
[9] Siolas G, d’Alche-Buc F. *Support vector machines based on a semantic kernel for text categorization*. Proceeding of IEEE IJCNN. 2000; 205-209.
[10] Mavroeidis D, Tsatsaronis G, Vazirgiannis M, Theobald M, Weikum G. Word sense disambiguation for exploiting hierarchical thesauri in text classification. *LNCS (LNAI)*. 2005; 3721: 181-192.

[11] Padmaraju D, V Varma. Applying lexical semantics to improve text classification. Proceedings of second symposium on Indian Morphology. *Phonology and Language Engineering*. 2005: 94-98.
[12] Stephan Bloehdorn, Alessandro Moschitti. Combined Syntactic and Semantic Kernels for Text Classification*. ECIR 2007, LNCS.* 2007: 307-318.
[13] Roberto Basili, Marco Cammisa, Alessandro Moschitti. A semantic kernel to classify texts with very few training examples. *Informatica*. 2006; (30): 163-172.
[14] Jamal Abdul Nasir, Asim Karim, George Tsatsaronis, Iraklis Varlamis. A knowledge-based semantic kernel for text classification. *SPIRE 2011, LNCS 7024.* 2011: 261-266.
[15] Cristianini N, Taylor JS, Lodhi H. *Latent Semantic Kernels*. Proceeding of the Eighteenth International Conference on Machine Learning. 2001: 66-73.
[16] Shoushan Li, Rui Xia. *A framework of feature selection methods for text categorization*. Proceedings of the 47 th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. 2009: 692-700.
[17] J Blitzer, M Dredze, F Pereira. *Domain adaptation for sentiment classification*. In Proceedings of ACL-07, the 45th Meeting of the Association for computational Linguistics.
[18] J Brank, M Grobelnik, N Milic-Frayling, D Mladenic. *Interaction of feature selection methods and linear classification models*. In Workshop on Text Learning Held at ICML. 2002.
[19] Seman, Kamaruzzaman, Puspita, Fitri Maya et al. An improved optimization model of internet charging scheme in multi service networks. *Telkomnika*. 2012: 592-598.
[20] Zhexi Yang. Information expectation with cloud computing in China. *Telkomnika*. 2012: 876-882.
[21] Zhang Xinsheng. Sparse representation for detection of microcalcification clusters. *Telkomnika*. 2012: 545-550.
[22] G Forman. An extensive empirical study of feature selection metrics for text classification. The *Journal of machine learning research.* 2003; 3(1): 1289-1305.
[23] L Marina, L Mark, K Slava. Classification of web documents using concept extraction from ontologies. *Lecture Notes in Computer Science*. 2007; 4476: 287-292.
[24] V Vapnik. The nature of statistical learning. *Springer.* 1995.