International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:11, 2011

Graph-Based Text Similarity Measurement by Exploiting Wikipedia as Background Knowledge

Lu Zhang, Chunping Li, Jun Liu, Hui Wang

***Abstract*—**Text similarity measurement is a fundamental issue in many textual applications such as document clustering, classification, summarization and question answering. However, prevailing approaches based on Vector Space Model (VSM) more or less suffer from the limitation of Bag of Words (BOW), which ignores the semantic relationship among words. Enriching document representation with background knowledge from Wikipedia is proven to be an effective way to solve this problem, but most existing methods still cannot avoid similar flaws of BOW in a new vector space. In this paper, we propose a novel text similarity measurement which goes beyond VSM and can find semantic affinity between documents. Specifically, it is a unified graph model that exploits Wikipedia as background knowledge and synthesizes both document representation and similarity computation. The experimental results on two different datasets show that our approach significantly improves VSM-based methods in both text clustering and classification.

***Keywords*—**Text classification, Text clustering, Text similarity, Wikipedia

I. Introduction

THE notion of text similarity measurement is essential in many textual applications such as document clustering, classification, summarization and question answering. Generally, Vector Space Model (VSM)-based approaches are widely adopted in information retrieval and text mining, in which documents are represented as word vectors and cosine similarity is computed to measure the affinity between them. Although these methods are simple to implement, they often fail to capture cases in which the vector space is sparse or when there is a known relationship between words. This results in a notorious limitation of VSM – the semantic sensitivity, that is, if two documents employ distinct collections of core words to express the same topic, they will be considered to be irrelevant even though the core words they use might be synonyms or semantically associated.

One way to resolve this problem is enriching the document representation with background knowledge. Ontologies like WordNet [15] and Mesh [16] have been integrated in previous researches, but the limited coverage prevents them from being utilized in more domains. Recently, as the largest electronic knowledge repositorywith millions of articles contributed collaboratively by volunteers on the web, Wikipedia shows its advantage over other standard ontologies. In Wikipedia, each article describes a single topic (which we call it as a concept) and equivalent concepts are grouped together through concept redirect. Besides, each concept belongs to at least one category and the categories form a well-organized hierarchical structure. Compared with WordNet and Mesh, Wikipedia is much more comprehensive and up to date. All these features make it a potential external knowledge repository in text mining.

In recent years, there is a growing amount of research on how to make use of the abundant concepts, categories and links in Wikipedia to enhance text classification [1], [2], [3] and clustering [4], [5], [6], text summarization [7], information retrieval [8] and community question answering [9]. In their work, text representation is augmented or replaced with Wikipedia concepts or categories. But with regards to similarity calculation, cosine is mostly employed, which may still suffer from similar problems of Bag of Words (BOW) under the concept or category space. To the best of our knowledge, by leveraging Wikipedia as background knowledge, few researches have been done on how to measure the text similarity outside the framework of VSM.

In this paper, we aim to propose a more comprehensive text similarity measurement which goes beyond VSM and can find semantic relationship between documents. A novel graph-based model that combines both text representation and similarity computation under a unified framework is presented. Specifically, by exploiting Wikipedia concepts as background knowledge, a document-concept bipartite graph is constructed and we develop a new approach based on the graph to compute document similarity. This model can overcome the semantic sensitivity problem by utilizing background knowledge and calculating on the bipartite graph iteratively at the same time. Therefore, two documents need not share common words or concepts to attain a similarity score, as long as their connected concepts are correlated.The experimental results on two datasets, 20-Newsgroups and Yahoo! Answers, show significant improvement over other VSM-based methods.

The main contributions of our work are: (1) We propose a unified framework of graph-based text similarity measurement by leveraging Wikipedia as background knowledge, which can overcome the semantic sensitivity problem of VSM-based approaches and avoid the potential flaws of previous research focused on enriching document representation with background knowledge. (2) The experimental results of our method on two datasets show significant improvements in text clustering and classification over traditional VSM-based ones, which indicate a promising application of this model in other text mining tasks.

The rest of this paper is organized as follows. Section II gives a brief introduction to related work. Section III describes our proposed approach. Section IV presents experimental results

Lu Zhang and Chunping Li are with Tsinghua National Laboratory for Information Science and Technology, School of Software, Tsinghua University, 100084, Beijing, China (e-mail: luzhang06@gmail.com,cli@tsinghua.edu.cn).
Jun Liu and Hui Wang are with School of Computing and Mathematics, University of Ulster, UK (e-mail: J.Liu@ulster.ac.uk, H.Wang@ulster.ac.uk).

and our analysis. Finally, we have the conclusion and future work in Section V.

II. Related Work

Using Wikipedia as external knowledge in text mining tasks has drawn researchers' attention recently. Gabrilovich et al. [1],[2] propose and evaluate a method to render text classification systems where traditional document representation is augmented with Wikipedia concepts. Banerjee et al. [4] employ Wikipedia concepts for short text clustering in the same manner as in [1], except that they use query strings created from document texts to retrieve relevant Wikipedia articles. Furthermore, both concept and category information are utilized in text classification and clustering in [3] and [5], [6] respectively. In [6], the authors develop two approaches, *exact-match* and *relatedness-match*, to map text documents to Wikipedia concepts, and further to Wikipedia categories.By *exact-match* scheme, each document is scanned to find Wikipedia concepts, which are mostly short phrases. The searched Wikipedia concepts are used to comprise the concept vector of the corresponding document. Compared with *relatedness-match* scheme which uses the content of each Wikipedia article, *exact-match* is more effective and efficient, and can map synonymous phrases to the same concept through the redirect links in Wikipedia. However, in these methods, documents are still treated as vectors with words replaced or expanded with concepts or categories, which induces the consequence that semantic sensitivity problem still exists under the concept or category space when applying cosine similarity to measure document affinity. Therefore, we need a better similarity measurement that can figure out this problem.

SimRank [10]is an algorithmfor measuring object similarity applicable in any domain whose central idea is that, two nodes are similar if they are related by similar nodes. Random Surfer Model [17] is the algorithm's theoretical basis. The SimRank score *s(a, b)*indicates how soon two random surfers are expected to meet at the same node if they start at nodes *a* and *b*,and randomly walk the graph backwards. Different from *co-citation* [14], SimRank is able to find the similarity between objects that are not directly referenced by the same object, but it may fail to translatesomecases in accord with our intuition. For example, it doesn't take the edge weight into consideration, meaning that we cannot tell the probabilities of a random surfer starting from the same node to its different neighbors. Therefore, SimRank++ [11] is put forward and it enhances SimRank by leveraging the edge weights and the number of two nodes' shared neighbors. The technique is applied on a click graph in query rewriting of sponsored search and attains better performance than SimRank as well as other similarity measurements like Pearson correlation, cosine and Jaccard similarity.

In our work, inspired by the ideas behind SimRank and SimRank++, a graph-based document representation is proposed on which a new algorithm can be applied to compute document similarity. Our unified model can solve the semantic sensitivity problem by the synthesis of semantic background knowledge and the link structure in document representation.

III. Graph-Based Similarity Measurement

A. Related Definitions

Wikipedia Concept: In Wikipedia, each article only describes a single topic. The title of the article, which we call Wikipedia concept, is a succinct phrase. Wikipedia articles are created in an academic manner and under strict guidelines. Therefore, compared with words or phrases extracted from plain text, Wikipedia concepts are more concise and less noisy.

Redirect Concept: In Wikipedia, not all concepts have their own corresponding articles for detailed description. Some might be redirected to another one through concept redirect. For instance, if we search concept "Tsinghua" in Wikipedia, the result page will be titled with "Tsinghua University", which is redirected from the original concept. In this case, we call "Tsinghua University" the redirect concept of "Tsinghua". According to Wikipedia guidance[1], this special mechanism groups concepts with similar meanings (e.g. alternative names, closely related words, etc.) and different forms (e.g. plurals, adjectives/adverbs pointing to noun forms, alternative spellings or punctuationetc.) together and can handle the synonymand morphologyproblems appearing in many text mining tasks.

*Document-Concept Bipartite Graph:*Based on the definitions of concept and redirect concept, we can define a document-concept bipartite graph as illustrated in Fig. 1.

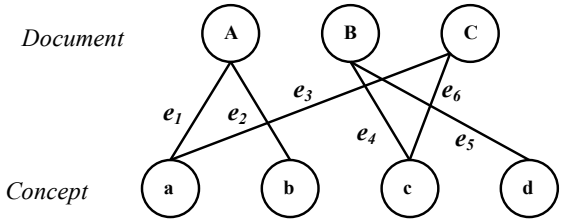

Fig. 1 A Document-Concept Bipartite Graph

In Fig. 1, the nodes on the top denote documents in the collection (e.g.in 20-Newsgroups, each node is a news article in the corpus). For each document, we extract representative keywords or phrases (filtering out the stopwords and those words or phrases with too large *idf*) first and then map them into Wikipedia concepts using the *exact-match* scheme introduced in [6]. These concepts constitute the nodes at the bottom of the bipartite graph. There is an edge between a document node and a concept node if the concept appears in the specific document. The weight of the edge is determined by the frequency of the concept's occurrence in that document, which is similar to word frequency in VSM.

More formally, the document-concept bipartite graph *G* can be defined as follows:

$$G = < V, E > \quad (1)$$

where

$$V = V_{doc} \cup V_{con} \quad (2)$$

[1]http://en.wikipedia.org/wiki/Wikipedia:Redirect

$$E \subseteq \{(v_{doc}, v_{con}) | v_{doc} \in V_{doc}, v_{con} \in V_{con}\} \quad (3)$$

$$w(i,j) = \textit{weight of edge } e_{ij}\textit{, frequency of concept } j\textit{'s occurrance in document } i\textit{, } e_{ij} \in E \quad (4)$$

$$N(x) = \{v | (x,v) \in E\} \quad (5)$$

In bipartite graph *G*, calculating the similarity between documents is equivalent to measuring the correlation between nodes $v \in V_{doc}$, namely, $sim_{doc}(v_i, v_j)$, which we will deal with in subsection B.

B. Our Approach

Our graph-based similarity measurement is based on the link structure of the document-concept bipartite graph. Motivated by the underlying idea of bipartite SimRank presented in [10] that two objects of one type are similar if they are related to similar objects of the second type, we develop a novel approach to calculate the similarities of two documents or concepts. In our scenario, the similarity of two documents is determined by the similarity of the concepts they contain; meanwhile, document similarity will influence the affinity of their associated concepts. Thus two documents need not share common concepts to attain a similarity score, as long as their connected concepts are correlated, which resolves the semantic sensitivity problem appearing often in VSM.

In general, the following formulae compute the affinity of different objects:

$$sim_{doc}(A,B) = C * \sum_i^{|N(A)|} \sum_j^{|N(B)|} \frac{w(A,i)}{\sum_{k \in N(A)} w(A,k)} \frac{w(B,j)}{\sum_{k \in N(B)} w(B,k)} sim_{con}(i,j) \quad (6)$$

$$sim_{con}(a,b) = C * \sum_I^{|N(a)|} \sum_J^{|N(b)|} \frac{w(a,I)}{\sum_{K \in N(a)} w(a,K)} \frac{w(b,J)}{\sum_{K \in N(b)} w(b,K)} sim_{doc}(I,J) \quad (7)$$

where *A*, *B* (*A*≠*B*) are document nodes and *i*, *j*, *k* are their corresponding concept nodes (analogously, *a*, *b* are concept nodes and *I*, *J*, *K* are their connected document nodes). For each node *x*, *N(x)* denotes the node set that is associated with it (node *x*'s neighbor nodes) and *w(x,y)* is the weight of the edge which connects node *x* and its neighbor *y*. Constant *C* ranging from 0 to 1 can be taken either as a confidence level or a decay factor, which reveals our assumption that the similarity of a node with itself is 1, but the affinity between different objects should be less than 1.

From Equation (6) and (7), we can find that the similarity between two documents or concepts is the weighted average similarity of the concepts or documents they are associated with. Note that if a node has no neighbor, we set its similarity with all the other nodes to be 0. Distinguished from SimRank, we take the edge weight into account to denote the importance of different concepts. For example, if concept *a* appears more frequently in document *A* than concept *b*, *a* should contribute more than *b* when calculating *A*'s similarity with other documents. This design is in correspondence with our intuition. Moreover, factors like "spread" and "evidence" introduced in SimRank++ are not involved in our model, for they are not so important in our case and taking them into consideration increases the computational cost.

In practice, we use Equation (8) to compute the similarity among documents. Initially, $R_0(x,x) = 1$ for any node $x \in V$ and $R_0(x,y) = 0$ for all node pairs (*x*, *y*) where $x \neq y$. It is proven in [10] that after sufficient number of iterations, each *R* score will converge. In our experiment, we iterate Equation (8) until the difference of average similarity of all document pairs in the last two iterations is less than 10^{-5}.

$$R_{k+1}(x,y) = \begin{cases} 1 & x = y \\ 0 & N(x) = \emptyset \; or N(y) = \emptyset \\ C * \sum_i^{|N(x)|} \sum_j^{|N(y)|} \frac{w(x,i)}{\sum_{k \in N(x)} w(x,k)} \frac{w(y,j)}{\sum_{k \in N(y)} w(y,k)} R_k(i,j) & else \end{cases} \quad (8)$$

The key difference of our graph model and other traditional VSM-based methods is that, in our model, the semantic relationship between concepts can be inferred by iteratively calculating their similarity on the graph, which will influence the affinity between documents thereafter. For instance, as illustrated in Fig. 1, two documents *A* and *B* do not share any common concept, in which case their cosine similarity is zero. Nevertheless, their associated concepts *a* and *c* have co-occurred in document *C*, which indicates some kind of correlation between them. Because of this semantic relatedness, we will derive a similarity score between *A* and *B* in our presented model.

IV. Evaluation

We evaluate our proposed similarity measurement on two text mining tasks, i.e., clustering and categorization. Our model outputs document similarity directly, which causes popular methods for text clustering (e.g., K-means) and categorization (e.g., SVM) inapplicable. In our experiment, we choose Agglomerative Hierarchical Clustering (AHC) for clustering and K Nearest Neighbors (KNN) for categorization.

A. Datasets

*Wikipedia.*Wikipedia releases its database dumps [1] periodically. We download its English dump released on April 5th 2011, which is comprised of more than 3.6 million concepts, and use them in the *exact-match* concept mapping.

*20-Newsgroups.*The 20-Newsgroups dataset[2] is a collection ofapproximately 20,000 newsgroup documents, partitioned evenly across 20 different topics. The collection has become a benchmark for text clustering and categorization. In our experiment, for efficiency and reliability, we uniformly partition it into 10 subsets with each one containing 100 documents randomly picked from every class. Experiments are conducted on each subset separately. The overall performance on the whole collection is evaluated by the average ofall the

[1] http://dumps.wikimedia.org/enwiki/
[2] http://people.csail.mit.edu/jrennie/20Newsgroups/

subsets and significance test is done as well.

*Yahoo! Answers.*Yahoo! Answers[1] is a famous Community Question Answering site and all the questions posted to it are organized into hierarchical categories. Under the domain of *Internet*, we choose questions from its seven categories, including *Facebook*, *Google*, *MySpace*, *Wikipedia*, *Flickr*, *MSN* and *YouTube*, for experiments. This corpus is divided into 10 parts in a similar manner as 20-Newsgroups.

B. Experimental Design

There are four clustering and classification schemes in our experiments, i.e., *word-cosine*, *concept-cosine*, *word-concept-cosine* and *concept-graph*. Representation and similarity measurement of these schemes are shown in Table I.

TABLE I
FOUR CLUSTERING AND CLASSIFICATION EXPERIMENT SCHEMES

Scheme	Representation	Similarity Mea-surement
word-cosine	word vector	cosine
concept-cosine	concept vector	cosine
word-concept-cosine	(1 - α) * word vector + α * concept vector (α = 0.1, 0.2, ..., 0.9)	cosine
concept-graph	document-concept bipartite graph	graph-based similarity mea-surement

Among these schemes, *word-cosine* is the baseline. *Concept-cosine* and *word-concept-cosine* are two schemes based on cosine similarity competitive with our proposed *concept-graph* scheme. The three VSM-based schemes use *tf-idf* as the weight of a word or concept and*concept-graph* employs concept frequency as the edge weight. *Word-concept-cosine* models document as a linear combination of word vector and concept vector, and constant $\alpha = 0.1, 0.2, ..., 0.9$ indicates the concept's significance in the combined vector. The best outcome over all values of α is viewed as the final result of this scheme. In *concept-graph*, we try different values of C ranging from 0.1 to 1.0 on the training set and choose the best oneto apply on the testing set, which is viewed as the scheme's final result.

C. Evaluation Metrics

*Clustering.*The Agglomerative Hierarchical Clustering result is evaluated by five metrics, i.e., *Entropy*, *Purity*, *Cophenetic Correlation*(*Cophenet*), *F-measure* and *Normalized Mutual Information* (*NMI*). Among these metrics, *Entropy* and *Purity* measure how the classes of objects are distributed within each cluster; *Cophenetic Correlation* is a special metric in Agglomerative Hierarchical Clustering, which measures how faithfully the hierarchical tree represents the dissimilarities among observations; *F-measure*combines the information of *Precision* and *Recall* which is extensively applied in information retrieval; *NMI* measures both homogeneity (the extent to which clusters contain only objects from a single class) and completeness (the extent to which all objects from a single class are assigned to a single cluster), and is not affected by the number of clusters. Let $L=\{w_1,w_2,...\}$ be the set of clusters, $C=\{c_1,c_2,...\}$ be the set of classes and Nbe the number of objects, some of the metrics are defined as follows:

$$Entropy = \sum_{j=1}^{|L|} \frac{|w_j|}{N}\left(-\frac{1}{log|C|}\sum_{i=1}^{|C|}\frac{|w_j \cap c_i|}{|w_j|} log \frac{|w_j \cap c_i|}{|w_j|}\right) \quad (9)$$

$$Purity = \frac{1}{N}\sum_{j=1}^{L}(|w_j| * P_j) \quad (10)$$

where

$$P_j = \frac{1}{|w_j|} max_i(w_j, c_i) \quad (11)$$

For the equation of *NMI* [13], let n_h be the number of objects in class h, n_l the number of objects in the l-th cluster, and $n_{h,l}$ the number of objects in cluster l but with class label h, the definition of *NMI* is:

$$NMI = \sum_{h,l} \frac{n_{h,l} * log\left(\frac{n * n_{h,l}}{n_h * n_l}\right)}{\sqrt{\left(\sum_h n_h * log\left(\frac{n_h}{n}\right)\right) * \left(\sum_l n_l * log\left(\frac{n_l}{n}\right)\right)}} \quad (12)$$

*Classification.*As for KNN classification, commonly used metrics such as *Precision*, *Recall* and *F-measure* are adopted to measure its result. P_i, R_i and $F_i$represent the *Precision*, *Recall* and *F-measure* value of class i:

$$P_i = \frac{number\ of\ correctly\ classified\ objects\ in\ class\ i}{number\ of\ objects\ classified\ as\ i} \quad (13)$$

$$R_i = \frac{number\ of\ correctly\ classified\ objects\ in\ class\ i}{number\ of\ objects\ whose\ real\ class\ label\ is\ i} \quad (14)$$

$$F_i = \frac{2 * P_i * R_i}{P_i + R_i} \quad (15)$$

The overall value of each metric is the weighted average of the individual class's P_i, R_i and F_i over all classes.

D. Text Clustering Results

Table II shows the results of Agglomerative Hierarchical Clustering (*average* linkage is used)on both datasets. The values in bold are improved results compared to the baseline, and "*" indicates the improvement is statistically significant with a confidence level of 95% (p value is 0.05). The meanings of formats and symbols are the samein Table III.

The clustering results are measured by five metrics, i.e., *Entropy*, *Purity*, *Cophenet*, *F-measure* and *NMI*, which evaluate the clustering quality in terms of homogeneity or completeness as mentioned in subsection C. Among these metrics, *NMI*is an increasingly popular indicator because it is more comprehensive and does not necessarily become greater when the number of clusters increases. All these metrics range from 0 to 1, and except *Entropy*, the higher their values, the better the quality is.

[1] http://answers.yahoo.com/

TABLE II
AGGLOMERATIVE HIERARCHICAL CLUSTERING RESULTS ON TWO DATASETS

20-Newsgroups					
Scheme	**Entropy**	**Purity**	**F-Measure**	**NMI**	**Co-phenet**
word-cosine	0.7414	0.3031	0.3548	0.2776	0.7119
concept-cosine	**0.7377**	0.2849	0.3472	**0.2916***	0.6922
word-concept-cosine	**0.7208***	**0.3203**	**0.3762***	**0.3039***	0.7119
concept-graph	**0.7019***	**0.3475***	**0.3878***	**0.3255***	**0.7447***
Yahoo! Answers					
Scheme	**Entropy**	**Purity**	**F-Measure**	**NMI**	**Co-phenet**
word-cosine	0.9465	0.1959	0.2850	0.1108	0.6415
concept-cosine	**0.9196**	**0.2208**	**0.3080**	**0.1537**	0.6357
word-concept-cosine	**0.8718***	**0.2732***	**0.3557***	**0.2045***	**0.6471***
concept-graph	**0.7623***	**0.3611***	**0.4366***	**0.3392***	**0.6959***

It can be observed from the results that our method is able to improve the baseline significantly compared with *concept-cosine* and *word-concept-cosine*. Regarding *NMI*, the relative improvements of our model are 17.26% and 206.14% on 20-Newsgroups and Yahoo! Answers respectively, whereas*word-concept-cosine*, which achieves the best performance among VSM-based schemes, improves the baseline only by 9.47% and 84.57%. Besides, for *Cophenet*, which is a special metric in Agglomerative Hierarchical Clustering, our approach significantly improves the baseline by 4.6% whereas the best result of *word-concept-cosine* scheme is just the same as that of baseline.

In comparison with the best agglomerative clustering result (with *NMI* = 0.166 and 15.28% improvement to baseline under *Word_Concept_Category* match scheme) of *exact-match* illustrated in [6], our approach can achieve higher clustering quality with *NMI* = 0.326 and better relative improvement to the baseline.

E. Text Categorization Results

The results of KNN are shown in Table . Evaluation metrics are *Precision*, *Recall* and *F-measure*. As is known to all, choosing different values of K will lead to distinct outcomes and different computational complexities; generally, larger values of K reduce the effect of noise on the classification, but make boundaries between classes less distinct. We try K = 5, 10, 15, 20 on both datasets and for different values of K, the performance of our method is always better than the baseline. Due to the limitation of space, we select the outcomes of K = 15 for 20-Newsgroups and K = 10 for Yahoo! Answers to show, which can achieve a good balance between effectiveness and efficiency. In each subset, we apply *leave-one-out* cross-validation on all the samples and use the average result for error estimate.

Similar to clustering, our proposed approach also performs better than the other three VSM-based ones on the task of classification. On 20-Newsgroups dataset, all methods leveraging Wikipedia concepts as background knowledge outperform the baseline, but the relative *F-measure* improvements of *concept-cosine* and *word-concept-cosine* are 2.95% ($p = 10^{-4}$) and 3.63% ($p = 1.5*10^{-3}$), whereas the amelioration of our model can reach to 5.52% with $p< 10^{-4}$ in significance test. On Yahoo! Answers corpus, results of *word-cosine* and *concept-cosine* are at a comparable level and *word-concept-cosine* can improve the baseline by 0.94% ($p = 6.9*10^{-3}$)slightly in *F-Measure*. Our proposed *concept-graph* is superior to all of them by a 3.81% increase ($p = 3*10^{-4}$) in *F-Measure* to the baseline.

TABLE III
K NEAREST NEIGHBORS RESULTS ON TWO DATASETS

20-Newsgroup (K=15)			
Scheme	**Precision**	**Recall**	**F-measure**
word-cosine	0.5798	0.6321	0.6000
concept-cosine	**0.5968***	**0.6506***	**0.6177***
word-concept-cosine	**0.6012***	**0.6546***	**0.6218***
concept-graph	**0.6123***	**0.6717***	**0.6331***
Yahoo! Answers (K=10)			
Scheme	**Precision**	**Recall**	**F-measure**
word-cosine	0.7378	0.8592	0.7902
concept-cosine	0.7321	0.8536	0.7839
word-concept-cosine	**0.7451***	**0.8672***	**0.7976***
concept-graph	**0.7794***	**0.8850***	**0.8203***

F. Discussion

The experimental results confirm that using Wikipedia concepts as replacement or additional features in document representation does help in both text clustering and categorization. Moreover, by exploiting background knowledge from Wikipedia, our graph-based similarity measurement indeed generates better document similarity and thereby enhance text mining tasks.Concerning clustering, our model attains a much more prominent improvement on Yahoo! Answers. This may be owing to the fact that most articles in Yahoo! Answers are shorter than those in 20-Newsgroups, and we derive less representative words or concepts from the former collection. Therefore, in VSM-based methods, documents only share a few words or concepts which results in a low cosine score even though they belong to the same category. This increases the probability of two documents being mistakenly assigned to different clusters, which will impact the overall result since Agglomerative Hierarchical Clustering is a greedy algorithm. However, our approach can take full advantage of the relationship between concepts, with a relative low possibility of wrong assignment and a much higher *NMI* score than the baseline. On the other hand, the outcome of KNN on Yahoo! Answers is not strongly influenced by the dataset's characteristic because KNN selects K candidates to vote for a document's class label, and in our experiment ($K \geq 5$), a wrong candidate will slightly affect the final result.In our model, the decay factor *C* indicates the dissimilarity between different objects, so choosing distinct values of *C* will induce divergent outcomes (see Fig. 2 and Fig. 3). This is especially true in Agglomerative Hierarchical Clustering, where *concept-graph* can be weak effective in some cases (points in shadow) if $C \geq 0.6$. Meanwhile, the results of KNN are less sensitive to the values of *C*. So according to Fig. 2 and Fig. 3, we set $C = 0.4$ on both datasets in Agglomerative Hierarchical Clustering and $C = 0.6$ and $C = 0.9$ on 20-Newsgroups and Yahoo! Answers in KNN.

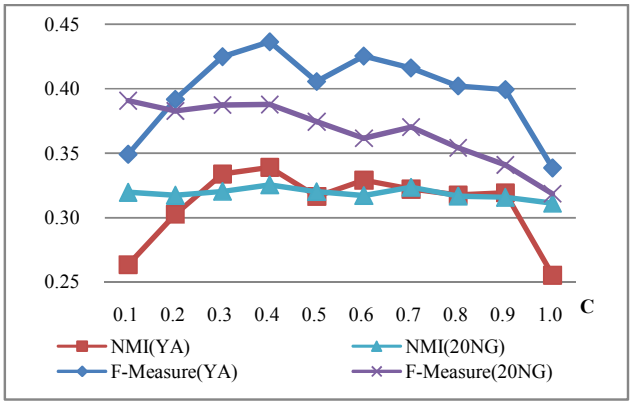

Fig. 2Clustering Results for Different Values of *C*

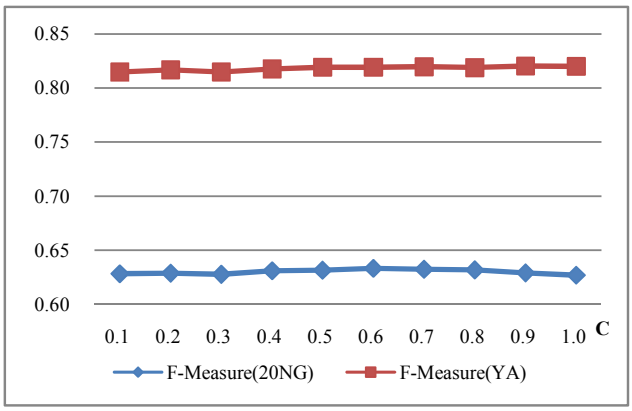

Fig. 3CategorizationResults for Different Values of *C*

As for the performance of our approach, we can divide it into two steps. First, constructing the document-concept bipartite graph costs as equal time as generating word or concept vectors. Next, in similarity computation, although our approach consumes longer time than VSM-based methods due to the iterative calculation, the time cost is still acceptable. Furthermore, our proposed algorithm can be easily implemented in parallel (e.g. under the framework of Map-Reduce), which will show a promising performance enhancement when dealing with large scale data.

V. Conclusion And Future Work

In this paper, we propose a novel graph-based text similarity measurement using Wikipedia as background knowledge. It is a unified model synthesizing both text representation and similarity calculation, which overcomes the semantic sensitivity problem of BOW and can find semantic affinity among documents. Experimental results on two different datasets show that our presented approach outperforms VSM-based methods in both text clustering and classification.

In future work, we plan to integrate more information such as categories and links in Wikipedia into our model. More text mining tasks like graph-clustering and text summarization will also be performed to prove the effectiveness of our model.

Acknowledgement

This work was granted by NFSC-RS Joint Project under No. 60911130419 and China National 973 Project under No.2010CB328003.

References

[1] E.Gabrilovich andS.Markovitch, “Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge,”in*Proceedings of the 21st National Conference on Artificial Intelligence*, Boston,2006, pp. 787–788.

[2] E.Gabrilovich andS.Markovitch, “Computing semantic relatedness using Wikipedia-based explicit semantic analysis,”in*Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, 2007, pp. 1606–1611.

[3] P.Wang andC.Domeniconi, “Building semantic kernels for text classification using Wikipedia,”in*Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, 2008, pp. 713–721.

[4] S.Banerjee, K.Ramanathanand A.Gupta, “Clustering short texts using Wikipedia,”in*Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, 2007, pp. 787–788.

[5] J.Hu, L.Fang,Y.Cao, et al.,“Enhancing text clustering by leveraging Wikipedia semantics,”in*Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, 2008, pp. 179–186.

[6] X.Hu, X.Zhang, C.Lu,E. K. Park and X. Zhou,“Exploiting Wikipedia as external knowledge for document clustering,”in*Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Paris, 2009, pp. 389–396.

[7] Y.Miao andC.Li, “Enhancing query-oriented summarization based on sentence wikification,”in*Workshop of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[8] Y.Li,W.P.R.Luk,K.S.E.Ho and F.L.K. Chung,“Improving weak ad-hoc queries using Wikipedia as external corpus,”in*Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, 2007, pp. 797–798.

[9] Y.Miao andC.Li, “Mining Wikipedia and Yahoo! Answers for question expansion in opinion QA,”in*Advances in Knowledge Discovery and Data Mining*, vol. 6118/2010, pp. 367–374. Springer, 2010.

[10] G.Jeh andJ.Widom, “SimRank: Ameasure of structural-context similarity,”in: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton,2002, pp. 538–543.

[11] I.Antonellis, H.Garcia-Molina andC.-C.Chang, “Simrank++: Query rewriting through link analysis of the click graph,”in *Proceedings of the Very Large Databases*,vol.1, iss.1, pp. 408–421, 2008.

[12] D.Lizorkin, P.Velikhov, M.Grinev andD.Turdakov, “Accuracy estimate and optimization techniques for Simrankcomputation,”in*Proceedings of the Very Large Databases*,vol.1, iss.1, pp.422–433, 2008.

[13] S.Zhong andJ.Ghosh, “Generative model-based document clustering: A comparative study,”in*Knowledge and Information Systems*, vol.8, no.3, pp.374–384, Springer, 2005.

[14] H.Small, “Co-citation in the scientific literature: A newmeasure of the relationship between two documents,”*Journal of American Society for Information Science*,vol.24,iss.4,pp. 265–269, 1973.

[15] A.Hotho, S.Staab andG.Stumme, “Wordnet improves text document clustering,”in*Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[16] I.Yoo, X.Hu and I.-Y.Song, “Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering,”in*Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*,Philadelphia, 2006, pp. 791–796.

[17] L´aszl´o andLov´asz,“Random walks on graphs: A survey,”*Bolyai Society Mathematical Studies*, vol.2, pp.1–46, 1993.