A Laplacian Eigenmaps Based Semantic Similarity Measure between Words

Yuming Wu[1 2],Cungen Cao[1],Shi Wang[1] and Dongsheng Wang[1 2]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road Zhongguancun,Beijing 100-190,China

[2] Graduate University of Chinese Academy of Sciences No. 19 Yu Quan Road, Shi Jing Shan Distinct, Beijing 100-049,China

Abstract. The measurement of semantic similarity between words is very important in many applicaitons. In this paper, we propose a method based on Laplacian eigenmaps to measure semantic similarity between words. First, we attach semantic features to each word. Second, a similarity matrix ,which semantic features are encoded into, is calculated in the original high-dimensional space. Finally, with the aid of Laplacian eigenmaps, we recalculate the similarities in the target low-dimensional space. The experiment on the Miller-Charles benchmark shows that the similarity measurement in the low-dimensional space achieves a correlation coefficient of 0.812, in contrast with the correlation coefficient of 0.683 calculated in the high-dimensional space, implying a significant improvement of 18.9%.

1 Introduction

Similarity measurement plays an important role in many areas, especially in semantic related applications [7]. So, the objective similarity measurement has to take more features of semantic level into consideration. For the purpose of attaching semantic features to words, we should have a knowledge source, from which we get semantic features and provide a flexible way to represent them, which can be extended to other knowledge sources without much modification.

In this paper, we propose a new method based on Laplacian eigenmaps [2] to define the semantic similarity between words. First, we use an online dictionary as a knowledge source, which is in semi-structured text format. Several types of interpretations can be extracted directly from the webpages. Then, these interpretations are transformed into a set of attribute-value pairs. These attribute-value pairs are used as semantic features, which are represented in a high dimensional space. After that, the Laplacian eigenmaps based method is adopted to find the intrinsic coordinates in low dimensional space. Finally the similarities are recalculated under the intrinsic coordinates in the low-dimensional space.

The remainder of the paper is organized as follows. In section 2 we describe the background. Section 3 makes further analysis to knowledge sources and features representation. In section 4 we describe the materials and methods. Section

5 gives the experimental result and compares against other methods on Miller-Charles benchmark dataset. Finally, we discuss the shortcomings of the proposed method and conclude this paper.

2 Background and Related Work

There has been a great deal of work on semantic similarity measurement. Some of them incorporate semantic features in the definition of similarity measure [6] [9] [3]. In Lins paper [6], an information-theoretic definition of similarity was proposed, the similarity of two objects, i.e. A and B only depends on the distributions of $common(A; B)$ and $description(A; B)$. This method is generic, but for some specific similarity, such as semantic similarity, Lins method may be unsuitable. As there is no an unified approach to encode the various semantic information in knowledge sources into the distribution, and when the data is sparse, the real distribution is hard to approximate. Resnik [9] presented a measure of semantic similarity in an IS-A taxonomy, based on the notion of information content. This measure depends on the structure taxonomy. In Chens [3] work, a context vector model to measure word similarity was adopted. The syntactic and semantic similarity is balanced by using related syntactic contexts only. Two major differences in these methods are the approach of using knowledge and representing semantic features. A more detail analysis on them is given in next section.

3 Knowledge Sources and Features Representation

Knowledge sources vary in their format, either structured or unstructured. WordNet [4]and HowNet[7] are structured knowledge sources, in which concepts are related in some manners. For example, the concepts in WordNet are related through several types of semantic relations, such as hypernym, antonym, holonym and meronym. The context around a word also provides strong and consistent clues to the sense of it [11]. Today, much text format materials are available on the Web. The context of a word can be relatively reliably extracted from those materials.

In this paper, we use an online dictionary[3] as knowledge source . In the online dictionary, there are plenty of items to interpret the sense of a given word. These items interpret the word from different aspects, such as interpretations in another language, sample sentences to describe the usage of the word, and further interpretations in same language and synonyms. After extractions, for each word, a collection of attribute-value pairs are obtained. For instance, the word "car" has an attribute-value pair $< hasSynonym, "automobile" >$, which stands for that "car" and "automobile" is synonymous to each other. These attribute-value pairs are taken as the semantic features. The concrete examples are show in section 4.

[3] http://www.dict.cn

4 Materials and Methods

4.1 Interpretations in Online Dictionary

In a dictionary, a word or phrase is interpreted from multiple aspects in detail. These interpretations or descriptions are rich in semantic information. For the word ”car”, for example, we list some interpretations and descriptions, as depicted in Figure 1, to illustrate how the semantic features appear in an online dictionary.

Word	Car
Word Senses:	A wheeled vehicle adapted to the rails of railroad.
Sample Sentences:	The car in front of me stopped suddenly and I had to brake .
Synonyms:	auto, automobile, machine, motorcar, gondola.

Fig. 1. Sample interpretations for word ”car”

As shown in figure 1,there are three subtitles the webpage, i.e. ”word senses”, ”sample sentences” and synonyms. The word ”car” is interpreted with several meanings, word senses, sample sentences, synonyms and so on. The words ”wheel”, and ”vehicle”, under the subtitle of ”Word Senses”, have close links to the word ”car”. We take the subtitle as attribute name, and take each word under a subtitle as an attribute value of the corresponding attribute. Note that the word itself can be also used as an attribute value. Sample attribute-value pairs for the word ”car” are listed as Table 1 The context is a rich semantic source for a specific word, while there are also some stop words , such as ”of”,” but”, and ”in” in the context, which have little contribution to similarity between two words. Therefore, we leave out all these stop words in the process of measuring similarity.

Table 1. Sample attribute-value pairs for word ”car”

Attribute Name	Attribute Value
Word Sense	motor
Synonyms	auto

4.2 Definition of Similarity Measure

Let W be the set of words, and f is a mapping from a word to an attribute-value pairs set.

$$f(w_i) = \{< attribute_k, value_j >\} \tag{1}$$

$$W(w_i, w_j) = \frac{|f(w_i) \cap f(w_i)|}{|f(w_i) \cup f(w_i)|} \quad (2)$$

where $| \, . \, |$ refers cardinality of a set. As we take these semantic features independently, the dimension of space, in which these words are represented, is identical to the number of independent semantic features.

The number of attribute-value pairs is necessarily large. So the words are represented as points in the high dimensional space. Because the quantity of the words is limited, the words are very sparse in the high-dimensional space. Intuitively, there should exist a low dimensional embedding for the set of all the words.

A natural problem arises: how to find the optimal embedding in a low-dimensional space? Due to [2][5], the optimal embedding can be found by solve the following generalized eigenvalue problem.

$$Ly = \lambda D y \quad (3)$$

where $L = D - W$ is called Laplacian matrix and $D(D_{ii} = \sum_j W_{ji})$ is the diagonal weight matrix , Give the dimension of the target space, let M be the matrix with column as the first m eigenvectors which satisfy the formula (3). Then, the optimal map should map x_i to be y_i, which is the ith row of matrix M.

Given the representations in a low-dimensional Euclidean space, the improved semantic similarity between y_i and y_j can be calculated as:

$$Sim_{improved}(y_i, y_j) = e^{-\|y_i - y_i\|^2} \quad (4)$$

where $\| \, . \, \|$ refers to the 2-norm in Euclidean space. Now, we give the full process of how to measure the semantic similarity between two words. A more formal description of this process will be shown in the algorithm "semantic similarity based on Laplacian eigenmaps". The main ideas behind this algorithm are as follows. Firstly, encode the local relevance into a similarity matrix, in which each element is the basic similarity calculated by formula (2). Then we use Laplacian eigenmaps to find another representation in low dimensional space. Finally, we recalculate the semantic similarity in low dimensional space. Figure 2 show the algorithm of "Semantic Similarity based on Laplacian eigenmaps" .

5 Experiments and Discussions

We take Miller-Charles dataset [8] as a benchmark dataset. The dataset has been used in several works [6] [9]. The Miller-Charles dataset can be considered as a reliable benchmark for evaluating semantic similarity measures. As shown in Figure 3, the correlation increases rapidly when the dimension of the target space is from 1 to 10, and achieves the maximal value of 0.812 when the dimension is 11. After the dimension exceeds 11, the correlation coefficient decreases steadily. These experimental results coincide with the intuition that there is a

```
Algorithm : Semantic Similarity Based on Laplacian Eigenmaps
Input:
  (1) a set of words $S = \{word_i\}$, (2) the dimension $N$ of target space.
Output: the similarity matrix for all words in S.
Procedure:
  W are calculated by the formula [2];
  $D = diagonal(W)$; $L = D - W$;
  Calculate the first $N$ eigenvectors $v_1, ..., v_N$ which satisfy the generalized
eigenproblem Lv=Dv;
  Let $U$ be the matrix with $v_1, ..., v_N$ as columns
   Let $y_i$ be the ith row of matrix U;
  for each word-pair $< w_i, w_j > \in S \times S$
    Sim_Matrix (i,j) = Simimproved($y_i$ ,$y_j$) as defined in formula (4) ;
  end for each
Return Sim_Matrix;
```

Fig. 2. Semantic Similarity Based on Laplacian Eigenmaps

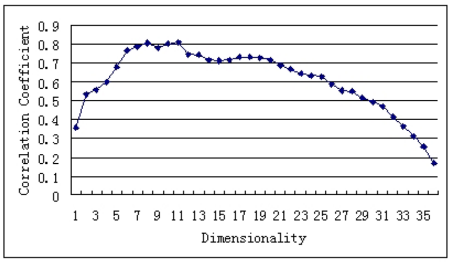

Fig. 3. Semantic Similarity Based on Laplacian Eigenmaps

low dimensional representation for the semantic features in the high-dimensional space. Table 2 presents a comparison of the proposed method with several other methods, including CODC [3], SemSim [1] , Lin [6] and Resink [9]. We get a correlation coefficient of 0.812 on Miller-Charlers dataset when the dimension of target space is 11. As shown in Table 2, the correlation coefficient of the proposed method is slightly lower than those obtained using the methods of SimSem and Lin, and higher than those obtained using the two other methods.

Table 2. Similarity measure comparison on Miller-Charles' dataset($d = 11$)

	Miller-Charlers	CO DC	Sem Sim	Lin	Resnik	Proposed Method
Correlation Coefficient	1	0.693	0.834	0.823	0.775	0.812

6 Conclusion and Future Work

In this paper, we proposed a method based on Laplacian eigenmaps to measure the semantic similarity between words. The main contributions of our work are listed as follows.

First, our method takes semantic features into consideration in a natural way when measuring similarity between words. These semantic features are organized as attribute-value pairs. Our method is very flexible and easy to extend, because there is no dependence on the structure of semantic features.

Second, the problem of data sparseness was avoided, because the final similarities were calculated in low dimensional space.

Experimental results on the Miller-Charles dataset achieve a correlation coefficient of 0.812, showing that the proposed method outperforms the traditional corpus-based and thesauri-based measures. The future work will concentrate on the following two directions. One is to transform other knowledge sources into

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant No. 60773059.

References

1. D. Bollegala, Y. Matsuo, M. Ishizuka : Measuring semantic similarity between words using web search engines, Proc. of 16th WWW, 2007, p.757-766
2. M Belkin, P Niyogi, : Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation, vol. **15**, 2003, p.1373-1396.
3. K. Chen, J. You: A study on word similarity using context vector models, Computational Linguistics and Chinese Language Processing, Vol. **7**, 2002, p.37-58

4. Fellbaum, Christiane (editor): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA. 1998.
5. Fan RK Chung: Spectral Graph Theory, Conference Board of the Mathematical Sciences, AMS, 1997.
6. D. Lin.: An information-theoretic definition of similarity, Proc of 15th ICML, Madison, WI, 296-304, 1998
7. Qun Liu, Sujian Li: Word Similarity Computing Based on How-net, Computational Linguistics and Chinese Language Processing, Taiwan, China, 2002(7), p.59-76
8. G. Miller and W. Charles: Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1-28, 1998.
9. P. Resnik: Using information content to evaluate semantic similarity, Proc 14th IJCAI, 448-453, Montreal, 1995
10. R. Richardson, A. Smeaton, and J. Murphy: Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words, Working Paper CA-1294, Dublin City University, 1994.
11. D Yarowsky: Unsupervised word sense disambiguation rivalling supervised method, Proc. of the 33rd ACL, 26-30 June 1995, p.189-196