Semantic Analysis Based on Human Thought Pattern

Libo Zhang, Tiejian Luo, Yihan Sun and Lin Yang
University of Chinese Academy of Sciences
Beijing, China

***Abstract*—Semantic analysis is an important component of recommendation systems and information retrieval in computer aideddetection. Previous researches have made certain breakthroughs in disease diagnosis and drugs recommended by semantic analysis. We propose a bilateral shortest paths method for computing semantic relatedness based on the human thought patterns for making sufficient use of the hyperlink structure. The proposed novel method exploits bilateral shortest paths method to calculate word similarity, and employs the method of matrix partition to calculate text similarity. Finally, an evaluation based on WS353-Ex and Lee datasets is carried out and the result shows that we obtain effective performance.**

I. Introduction

The main challenge in the field of computer aided diagnosis is how to make computer be able to think like a doctor. Human thought patterns are too intricate to imitate, but it is a basic fact that human thought patterns are firstly based on the association of experiences. Thus, endowing a computer with an associative mechanism that resembles human's is a difficult problem which remains to be solved. Association is based on the correlation between two words, paragraphs or texts. The evaluation of relatedness has important applications in both medicine and biology field like expert system, treatment recommendation, drug guidance and so on.

Traditional methods of evaluating semantic relatedness are mainly based on lexical overlap [1], which is widely used in systems to check for duplication in papers. However, if someone uses a synonym to represent the same meaning, then this traditional method may not work since the object of evaluation by lexical overlap is words, which does not take into account the relationship between words. Probabilistic Latent Semantic Analysis and Linear Discriminant Analysis utilize the relationship between words and the subject of the article, but they do not consider the relationship between words from the perspective of knowledge structure. Some new algorithms consider taking advantage of knowledge structure which is already been established by humans, such as Wikipedia, to evaluate the semantic relatedness [2], [3]. Specifically, these algorithms use links and texts contained in two article pages of Wikipedia. But there are some related works claiming that although hyperlinks can capture the knowledge structure of the real world, they cannot evaluate lexical similarity only by relying on the relationship of links [4].

As the repositories of human knowledge, Encyclopaedia Britannica and Wikipedia Encyclopedia are based on artificial arrangement and artificial editing, and organized based on vocabulary. Wikipedia has more articles with similar accuracy compared to Encyclopaedia Britannica [5]. This paper proposes that the shortest path between nodes (we call the smallest unit nodes) in Wikipedia can reflect the human associative mechanism. To demonstrate it, we present an algorithm of evaluating semantic relatedness on the basis of the shortest path, which is different from all previous algorithms.

II. Related Work

The traditional analyses on semantic relatedness need to reason and associate from human knowledge and experience, which is always an obstacle difficult to overcome for a computer, because in order to explore the relationship between two articles, the computer must acquire significant background knowledge in the related field, but previous works have ignored the background knowledge while they relied more on statistics. Some semantic relatedness research based on background knowledge cannot obtain a good result due to a limited overall base of knowledge. Researchers who analyzed semantics on the basis of web links have to some extent solved that problem of a limited overall base of knowledge. They indeed have gained some breakthroughs but those researches cannot reflect the whole picture of a certain point of knowledge due to the fact that the knowledge granularity of web links is coarse-grained and the knowledge content is sparse. In 2005, Giles [5] indicated that the breadth and depth of Wikipedia are both trustworthy and from then on, a boom of semantic analysis based on Wikipedia became apparent.

We collect and systemize important works based on Wikipedia in the field of semantic analysis and perform the analysis from six dimensions, which is demonstrated in Table 1. Gabrilovich and Markovitch [6] put forward a method named Explicit Semantic Analysis (ESA), which gathered concepts via Wikipedia and then set up the relationship matrix between text and gathered concepts. Zesch et al. [7] proved that using data from Wikipedia could improve the result of the previous method. Hassan and Mihalcea [2] measured the cross-lingual relatedness between two concepts, and established the concept vector representations via explicit semantic analysis. Syed et al. [3] regarded Wikipedia as an ontology and combined the keywords and subjects with the text by three methods. Coursey et al. [8] built a relational graph between concepts and categories, then came up with a method to classify topic automatically. Gabrilovich and Markovitch [9] established weight vectors among concepts by means of setting up a semantic

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

interpreter and assessed the relativity by comparing vectors. The above methods only considered the relationship among categories when using article texts to calculate the relationship between concepts, while not paying sufficient attention to the path of hyperlink.

Some papers made use of path information to evaluate the semantic similarity, several examples are: Ponzetto and Strube [10] mapped concepts to the category every concept belongs to, then evaluated the correlation between two concepts based on the path length between two categories. Milne and Witten [11], [12] assessed the relatedness by calculating repetitive rate of the links contained in articles. Similar to a WikiWalk approach proposed by Yeh et al. [13], Yazdani and Popescu-Belis [4] presented a method named Visiting Probability to calculate the distance between the nodes. The methods mentioned above made use of path based only on the path's relationship among the category concepts belong to, or on the superposition of hyperlinks contained in two specific articles. Singer et al. [14] took Wikigame as the dataset, which recorded the path's relationship between two articles on the basis of clicking from people. Niebler et al. [15] noticed unconstrained navigation datasets. Wikipedia Clickstream [16] kept track of how people clicking on the next link from one Wikipedia article. Dallmann et al. [17] employed an algorithm named Random Walk to simulate human behaviors about the click of hyperlinks, with the Wikipedia Clickstream as the dataset. However, Random Walk is so random that we cannot obtain the shortest path, which is meaningful for people to make a strategic decision.

Distinct from the traditional method utilizing web links, we describe the relationship between two nodes via figuring out the connection route between them. In this paper, a new method named Bilateral Shortest Path Semantic Analysis is put forward, to be specific, depicting the relationship between two articles by utilizing link relations among nodes in structured Wikipedia.

III. The Proposed Algorithm

A. Data Set

In recent years, studies found that navigation behaviors of users in Wikipedia can improve the result of semantic analysis and the main datasets used in this method are Wikigame and Wikipedia Clickstream. However, we consider that these data cannot fully reflect human thought patterns due to the imperfection of datasets.

Taking into account the limitations of the existing datasets, we think focus should be on the Web link structure of Wikipedia, since Wikipedia is the crystallization of human wisdom accumulated over a long time. Its editing structure is highly consistent with human thought patterns. The raw dataset we employed is a complete Wikipedia dump from June 1st, 2016, which forms a 53.4 Gb XML file after being unzipped. All articles are included in this XML file and each article contains hyperlinks pointing to other articles. Traditional algorithms mainly pay attention to the relationship among articles while we hope to observe it from a smaller dimension, so that in this paper the object of our study are nodes, which are the corresponding words of the hyperlink in every article and map all link relations in Wikipedia. The number of processed nodes which we decided to employ in our study is a 12 million and this number is 5 to 10 times of the number of research objects in traditional algorithms.

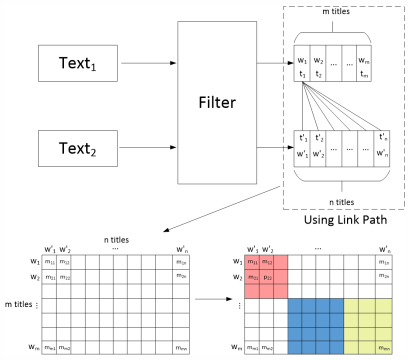

Fig. 1. The processing of our method.

B. Bilateral Shortest Path Semantic Analysis

In the research of semantic analysis based on the consideration of paths, researchers used hyperlinks to look for the path between two articles and these studies were all based on the hypothesis that such a path exists.

We proposed a new method to evaluate the relevance among texts on the basis of the application of Bilateral Shortest Path and the detailed algorithm is shown in Fig 1. $Text_1$ and $Text_2$ are both texts requiring relatedness estimation. We input them into a filter, which can preprocess the text, including deletion of stop word, lemmatization, and so on. Then the segmentation is achieved based on the comparison between words in text and Wikipedia articles. Next, divide $Text_1$ into the set made of m words, $T_1{=}\{w_1, w_2 \dots w_m\}$, where: w represents every word; t represents the number of times the corresponding title has occurred in $Text_1$. Similarly, divide $Text_2$ into the set made of n words, $T_2{=}\{w'_1, w'_2 \dots w'_m\}$, where: w' represents every words; t' represents the number of times the corresponding word has occurred in $Text_2$. As shown in Fig 2, we can construct a $m \times n$ matrix M (assume $m < n$) by multiple comparison between m words in $Text_1$ and n words in $Text_2$. The ith row and jth column value m_{ij} of matrix M means the correlation between title w_i and $w_j^{'}$, calculated by Bilateral Shortest Path. We put these words in two tuples $A = (w_1, \dots, w_m)$ and $B = \left(w_1^{'}, \dots, w_n^{'}\right)$.

Then we apply matrix partitioning to calculate the correlation between $Text_1$ and $Text_2$:

(1) We rearrange matrix M to separate out common words in A and B. Suppose that $|A \cap B| = p$ and then rearrange indexes of tuple A and B subject to:

$$\forall\ 1 \le k \le p,\ \ w_k = w_k^{'} \tag{1}$$

Namely, select common words in A and B, and place them in the front of respective tuple, with the same index

corresponding to same words. The submatrix of common words is shown as the red region in Fig 2.

(2) Obtain tuple $\bar{A}$ and $\bar{B}$ via deleting common words in A and B:

$$\bar{A} = (w_{p+1}, \ldots, w_m) \tag{2}$$

$$\bar{B} = \left(w_{p+1}^{'}, \ldots, w_n^{'}\right) \tag{3}$$

Any two words in tuple $\bar{A}$ and $\bar{B}$ are different from each other and then we can obtain a submatrix M_s with size of $(m-p)(n-p)$.

(3) The rearranged matrix M_s takes the form:

$$\forall\ (p+1) \leq r \leq s \leq m,\ \ t_r \geq t_s \tag{4}$$

$$\forall\ (p+1) \leq r' \leq s' \leq n,\ \ t_r^{'} \geq t_s^{'} \tag{5}$$

Where t_r denotes the weights of w_r in $Text_1$; t_s denotes the weights of w_s in $Text_1$; $t_r^{'}$ denotes the weights of $w_r^{'}$ in $Text_2$; $t_s^{'}$ denotes the weights of $w_s^{'}$ in $Text_2$. That is, rearrange $\bar{A}$ and $\bar{B}$ in order of descending weights, the rearranged submatrix is shown as the blue and yellow regions in Fig 4.

(4) Let q=min(m,n) ,and:

$$\tilde{A} = (w_{p+1}, \ldots, w_q) \tag{6}$$

$$\tilde{B} = \left(w_{p+1}^{'}, \ldots, w_q^{'}\right) \tag{7}$$

Then we obtain a q square matrix, shown as blue region in Fig 1, and in the q square matrix, we utilize the Gale-Shapley Algorithm to calculate the stable matches:

Note: For any $w \in \tilde{A}$ and $w' \in \tilde{B}$, the correlation between w and w' is independent of their order, so, $\tilde{A} - \tilde{B}$ and $\tilde{B} - \tilde{A}$ Gale-Shapley Algorithm is ought to be the same.

(5) Rearrange tuple $\tilde{A} - \tilde{B}$ to make the matched words have the same indexes:

$$\forall\ (p+1) \leq k \leq q,\ \ w_k\ is\ matched\ with\ w_k' \tag{8}$$

(6) Record joint matrix as M', which has value m_{ij}' in the ith row and jth column. Then calculate the correlation between $Text_1$ and $Text_2$ $corr(Text_1, Text_2)$:

$$corr(A,B) = \frac{\sum_{k=1}^{q} w_k \times m_{kk}'}{\sum_{k=1}^{q} w_k} \tag{9}$$

$$w_k = \frac{1}{2}(t_k + t_k') \tag{10}$$

IV. Results

In this section, we use WS353Ex and Lee to verify the effectiveness of the Bilateral Shortest Path Semantic Analysis algorithm. The first part is the evaluation of word similarity using theWS353Ex dataset, the detailed processing is described below: Calculate the FSPL and BSPL between every word pair in WS353Ex, based on the related method proposed in Section 3, then work out their ASPL by the formula:

$$ASPL = \frac{FSPL + BSPL}{2} \tag{11}$$

TABLE I
A comparison of word similarity scores on WS353-Ex using three kinds of shortest path.

	Correlation
FSPL	0.48
BSPL	0.53
ASPL	0.84

We employ the Ridge Regression [18] to fit the score given by humans:

$$\hat{h}(w,p) = \omega_0 + \omega_1 p_1 + \ldots + \omega_m p_m \tag{12}$$

Where h represents the score given by humans; ω=$\omega_1, ..., \omega_p$ denotes coefficient; ω_0 denotes interception. We employ Ridge Regression to solve the problem of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares:

$$\min_{\omega} {||P\omega - h||_2}^2 + \alpha {||\omega||_2}^2 \tag{13}$$

Where α is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

We employ the Spearman rank correlation coefficient [4] to evaluate the accuracy of the prediction fitting model, which verifies the accuracy of the Bilateral Shortest Path Semantic Analysis algorithm. Three groups of experiments are designed in this paper, in which we respectively combine FSPL, BSPL and ASPL with the scores given by humans to estimate correlation. In experiments, we randomly select 30% of word pairs as training set and the other 70% as validation set for 500 trials and take the average of the outcomes. Ultimate results are shown in Table 3, when exclusively using FSPL the score equals 0.48, and when exclusively using BSPL the score equals 0.53. However, if we utilize ASPL, the score is up to 0.84 and remains stable. We make comparisons with some previous work, shown in Table 4. The hyperlink similarity algorithm, designed by Milne and Witten [12], improves the score to 0.79, while the algorithm proposed in this paper is obviously superior to previous algorithms from the perspective of the calculation of word similarity.

For exploring the effect in the evaluation of text similarity, we verify the ASPL algorithm using the Lee dataset [23]. This dataset contains 50 texts and their corresponding similarity scores given by humans. According to the processes in Section 3.4, we apply the Pearson correlation coefficient [4] to assess correlation between scores obtained from experiences and scores given by humans. 1000 randomly chosen pairs of texts are tested to obtain the mean of correlation, and the result is shown in Table 5. Hassan and Mihalce [2] proposed that via the training of a small document corpus, the LSA algorithm can get 0.69 points. Gabrilovich and Markovitch [9] put forward ESA algorithm and got 0.72 points. Our research analyzes Document similarity by means of the ASPL algorithm with 0.70 points obtained. The method used the

TABLE II
A COMPARISON OF WORD SIMILARITY SCORES ON WS353-EX WITH FINKELSTEIN ET AL. [19], JARMASZ [20], STRUBE AND PONZETTO [10], HUGHES AND RAMAGE [21], GABRILOVICH AND MARKOVITCH [9], AGIRRE ET AL. [22], YAZDANI AND POPESCU-BELIS [4], MILNE AND WITTEN [12].

Method	Correlation
Finkelstein et al.	0.56
Jarmasz	0.55
Strube and Ponzetto	0.48
Hughes and Ramage	0.55
Gabrilovich and Markovitch	0.75
Agirre et al.	0.78
Yazdani and Popescu-Belis	0.70
Milne and Witten	0.79
Our method	0.84

TABLE III
A COMPARISON OF TEXT SIMILARITY SCORES ON LEE WITH HASSAN AND MIHALCE [2], GABRILOVICH AND MARKOVITCH [9].

Method	Correlation
Hassan and Mihalce	0.69
Gabrilovich and Markovitch	0.72
ASPL	0.70

ASPL to obtain excellent effect in the evaluation of Word similarity while the effect is limited on text similarity. We think this is mainly because of the sparse training data, which is only based on dataset WS353-Ex with 344 pairs of words.

V. CONCLUSION

In this paper, we proposed an effective method for semantic relatedness analysis based on the Wikipedia hyperlink network. This is the first paper considering bilateral shortest paths length between two words as different, then verifies the universal connectivity and shows that the average shortest path length between two nodes is 3.84. In consideration of the bilateral thought patterns of humans, we apply bilateral shortest paths to design a state of art approach, which can make full use of potential human characteristics in the Wikipedia structure. This paper employs bilateral shortest paths length to analyze word similarity and then achieves outstanding performance. Based on this foundation, we put forward a method to analyze text similarity using Matrix partition and the effect is close to the advanced standards. The greatest contributions of this work are to obtain the excellent result using only Wikipedia hyperlinks with a small number of training data. In future work, we hope to improve the quality of the evaluation in two ways: 1) Combine other textual characteristics and add more structured knowledge Storage; 2) Add more training data and improve the evaluation model.

REFERENCES

[1] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[2] S. Hassan and R. Mihalcea, "Cross-lingual semantic relatedness using encyclopedic knowledge," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1192–1201.

[3] Z. S. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents." in *ICWSM*, 2008.

[4] M. Yazdani and A. Popescu-Belis, "Computing text semantic relatedness using the contents and links of a hypertext encyclopedia," *Artificial Intelligence*, vol. 194, pp. 176–202, 2013.

[5] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, 2005.

[6] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJcAI*, vol. 7, 2007, pp. 1606–1611.

[7] T. Zesch, C. Müller, and I. Gurevych, "Using wiktionary for computing semantic relatedness." in *AAAI*, vol. 8, 2008, pp. 861–866.

[8] K. Coursey, R. Mihalcea, and W. Moen, "Using encyclopedic knowledge for automatic topic identification," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009, pp. 210–218.

[9] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research*, vol. 34, pp. 443–498, 2009.

[10] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *AAAI*, vol. 6, 2006, pp. 1419–1424.

[11] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, 2008, pp. 25–30.

[12] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, vol. 194, pp. 222–239, 2013.

[13] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, "Wikiwalk: random walks on wikipedia for semantic relatedness," in *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 2009, pp. 41–49.

[14] P. Singer, T. Niebler, M. Strohmaier, and A. Hotho, "Computing semantic relatedness from human navigational paths: A case study on wikipedia," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 9, no. 4, pp. 41–70, 2013.

[15] T. Niebler, D. Schlör, M. Becker, and A. Hotho, "Extracting semantics from unconstrained navigation on wikipedia," *KI-Künstliche Intelligenz*, pp. 1–6, 2015.

[16] E. Wulczyn and D. Taraborelli, "Wikipedia clickstream," *Retrieved February*, vol. 17, 2015.

[17] A. Dallmann, T. Niebler, F. Lemmerich, and A. Hotho, "Extracting semantics from random walks on wikipedia: Comparing learning and counting methods," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[18] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[19] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.

[20] M. Jarmasz, "Roget's thesaurus as a lexical resource for natural language processing," *arXiv preprint arXiv:1204.0140*, 2012.

[21] T. Hughes and D. Ramage, "Lexical semantic relatedness with random graph walks." in *EMNLP-CoNLL*, 2007, pp. 581–589.

[22] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 19–27.

[23] M. Lee, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text document similarity," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2005, pp. 1254–1259.