MEASURING SEMANTIC SIMILARITY BY CONTEXTUAL WORD CONNECTIONS IN CHINESE NEWS STORY SEGMENTATION

Xuecheng Nie[1,2]*, Wei Feng*[1,3,*]*, Liang Wan*[2,3]*, Lei Xie*[4]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China
[2] School of Computer Software, Tianjin University, Tianjin, China
[3] Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin, China
[4] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{xcnie,wfeng,lwan}@tju.edu.cn, lxie@nwpu.edu.cn

ABSTRACT

A lot of recent work in story segmentation focuses on developing better partitioning criteria to segment news transcripts into sequences of topically coherent stories, while simply relying on the repetition based hard word-level similarities and ignoring the semantic correlations between different words. In this paper, we propose a purely data-driven approach to measuring soft semantic word- and sentence-level similarity from a given corpus, without the guidance of linguistic knowledge, ground-truth topic labeling or story boundaries. We show that contextual word connections can help to produce semantically meaningful similarity measurement between any pair of Chinese words. Based on this, we further use a parallel all-pair SimRank algorithm to propagate such contextual similarities throughout the whole vocabulary. The resultant word semantic similarity matrix is then used to refine the classical cosine similarity measurement of sentences. Experiments on benchmark Chinese news corpora show that, story segmentation using the proposed soft semantic similarity measurement can always produce better segmentation accuracy than using the hard similarity. Specifically, we can achieve 3%–10% average F1-measure improvement to state-of-the-art NCuts based story segmentation.

Index Terms— Semantic similarity, contextual word connections, similarity propagation, story segmentation

1. INTRODUCTION

News story segmentation aims at partitioning real-world news transcripts, *e.g.* online news texts or erroneous news transcripts generated by LVCSR, into sequences of topically coherent stories. The well-segmented transcripts are an important prerequisite for a number of content-level applications, such as news topic tracking and detection [1], particular news understanding and retrieval [2, 3].

Relation to prior work. Technically, there are two important issues highly related to the performance of story segmentation: (1) how to measure semantically meaningful word- and sentence-level similarities; and (2) by what criterion to segment the input transcripts. A lot of recent efforts have focused on designing suitable partitioning criteria in story segmentation [4], *e.g.* the minimum NCuts criterion [5, 6] and the maximum lexical cohesion criterion [7]. In contrast to the widely studied segmentation criteria, many existing methods simply employ the following repetition-based hard similarity metric of any two words a and b:

$$\mathrm{sim}_{\mathrm{H}}(a,b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

It is clear that the repetition-based hard similarity, *i.e.* Eq. (1), only considers the equivalent similarity of two same words as 1, while ignoring all potential semantic correlations between different words. This may help to maintain the simplicity of the algorithm. But, clearly, it would be much desirable if we could find a feasible way to take the potential semantic correlations of different words into account.

In English, WordNet::Similarity [8] is available to measure semantic relatedness of any two English words based on the famous WordNet project [9], which can be viewed as a comprehensive lexical expert system organizing nouns, verbs, adjectives and adverbs according to their conceptual and linguistic senses. In Chinese news story segmentation, however, such general knowledge-based similarity measurement may not be the best choice, mainly due to two reasons. First, despite the existence of Chinese WordNet, the Chinese version of WordNet::Similarity is not publicly available yet. Second, for a particular news corpus, the specific corpus-related semantic similarity measurement may help to produce better segmentation accuracy than the general knowledge derived similarity metric does.

* is the corresponding author. This work is supported by the National Natural Science Foundation of China (61100121, 61100122 and 61175018), the Program for New Century Excellent Talents in University (NCET-11-0365), the research fund for The Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China (TJKLASP-2012-1 and TJKLASP-2012-2), and the Fok Ying Tung Education Foundation (131059).

Contributions. In this paper, we propose a purely data-driven similarity measurement of Chinese words and sentences from a given corpus, correlating to their semantic relatedness, and need not the guidance of linguistic knowledge, ground-truth topic labeling, story boundaries, or any kind of expert supervision. We show that contextual word connections can help to produce a semantically reasonable similarity metric between any pair of Chinese words. Moreover, we use a parallel all-pair SimRank algorithm [10] to propagate such contextually driven similarities throughout the whole vocabulary of the corpus, which results in a corpus-dependent word semantic similarity matrix. We then use this similarity matrix to refine the widely-used cosine similarity measurement of sentences [4, 5, 6], by taking soft semantic similarities between different words into consideration. Experiments on benchmark Chinese news corpora, CCTV and TDT2, have shown that, story segmentation using the proposed soft semantic similarity metric can always produce better segmentation accuracy than using the hard similarity. Specifically, we can achieve 3%–10% average F1-measure improvement to state-of-the-art NCuts based story segmentation [5, 6].

Fig. 1. Contextual connections of word w_i in $\mathcal{T}$, with $\tau = 3$.

2. SEMANTIC SIMILARITY MEASUREMENT OF CHINESE WORDS AND SENTENCES

2.1. Contextual Word Connections

Our semantic similarity measurement is based on contextual word connections. For a given corpus $\mathcal{C} = \{\mathcal{T}_i\}_{i=1}^n$ composed of n news transcripts $\mathcal{T}_i$, let $\mathcal{V}$ be the vocabulary as the set of all words appeared in $\mathcal{C}$. For any two different words a and b in $\mathcal{V}$, we say they are *contextually connected* iff they both occur in a particular transcript $\mathcal{T} \in \mathcal{C}$ and their distance is not greater than τ (see Fig. 1).[1] Note that, there may be multiple appearances of word pair (a, b) in corpus $\mathcal{C}$ satisfying the contextual connection definition. Thus, we use $\mathrm{freq}(a, b)$ to denote the contextually concurred times of word pair (a, b). If a and b are not contextually connected, $\mathrm{freq}(a, b) = 0$. Then, we can define word *contextual similarity* as:

$$\mathrm{sim}_{\mathrm{C}}(a, b) = \frac{\mathrm{freq}(a, b)}{\mathrm{freq}_{\max} + \epsilon}, \tag{2}$$

where $\mathrm{freq}_{\max} = \max_{(i,j)}\{\mathrm{freq}(i, j)\}$, $\epsilon > 0$ is a constant ensuring $0 \le \mathrm{sim}_{\mathrm{C}}(a, b) < 1$ (if $a \neq b$). We further define $\mathrm{sim}_{\mathrm{C}}(a, a) = 1$. Accordingly, for all words in the vocabulary $\mathcal{V}$, we can construct a contextual similarity matrix

[1] Parameter τ is the cut-off threshold controlling the correlation range in the measurement. We set $\tau = 3$ in our experiments.

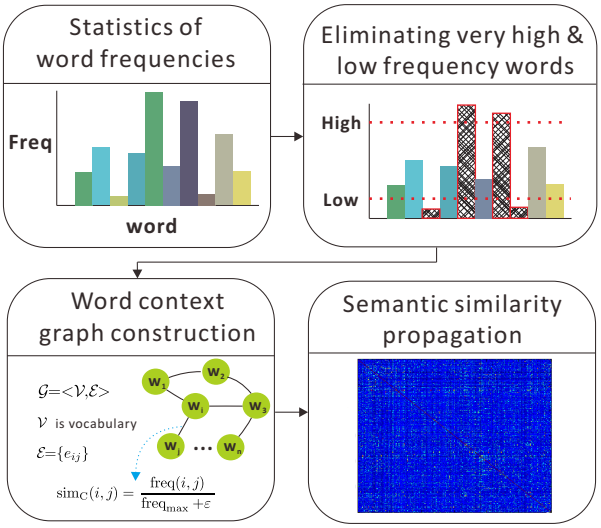

Fig. 2. The construction of word context graph and the resultant semantic similarity matrix after propagation.

$\mathbf{S}_{\mathrm{C}} = \{\mathrm{sim}_{\mathrm{C}}(i, j)\}_{(i,j)\in\mathcal{D}^2}$, which reflects the word-level semantic similarities, since, by the "bag of words" model, topically coherent words are more likely to appear in a same story, thus are more likely being contextually connected.

Note that, although it is not a brand new opinion that context correlates with semantic similarity in natural language processing [11], it has not been non-trivially used in story segmentation yet. In the next, we show how to automatically derive semantic word- and sentence-level similarities, stemming from the contextual similarity matrix $\mathbf{S}_{\mathrm{C}}$.

2.2. Word-Level Semantic Similarity

As shown in Fig. 2, based on $\mathbf{S}_{\mathrm{C}}$, we construct an undirected *word context graph* (WCG) $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, with the vocabulary $\mathcal{V}$ as its vertex set and $\mathcal{E} = \{e_{ij} | (i, j) \in \mathcal{V}^2\}$ being the edge set. For each edge e_{ij} in $\mathcal{G}$, we initialize its weight as $\mathrm{sim}_{\mathrm{C}}(i, j)$. Hence, $\mathcal{G}$ encodes the contextual similarity of all word pairs. Moreover, Fig. 2 also shows the biased word usage of news transcripts. As a result, to control the complexity, we eliminate the words with usage frequency less than 3 from $\mathcal{G}$ in our following computation, since very low frequency words (can be viewed as isolated words) have very slight correlation to other words. Their semantic similarities to any other words are close to 0. Besides, we also remove the words with very high frequencies (can be viewed as trivial words) from $\mathcal{G}$, since such words tend to be used in all kinds of news stories, thus measuring their soft semantic similarities to other words is meaningless and does not help in separating different stories. Without loss of generality, we just measure the similarities between the removed very high/low frequency words and other words by their hard similarities in Eq. (1).

For any word pair (a, b) in $\mathcal{G}$, their semantic similarity

$\text{sim}_\text{S}(a, b)$ is measured by the following three principles:

1. The semantic similarity of a word a to itself is 1, *i.e.* $\text{sim}_\text{S}(a, a) = 1$;

2. $\text{sim}_\text{S}(a, b)$ positively correlates with $\text{sim}_\text{C}(a, b)$, *i.e.* if a and b have higher contextual similarity, their semantic similarity is higher accordingly, and vice versa;

3. $\text{sim}_\text{S}(a, b)$ positively correlates with the semantic similarities of their neighbors.

Hence, we can define semantic similarity $\text{sim}_\text{S}(a, b)$ in the following iterative propagation form:

$$\text{sim}_\text{S}^{(0)}(a, b) = \text{sim}_\text{C}(a, b), \tag{3}$$

$$\text{sim}_\text{S}^{(t+1)}(a, b) = \frac{c}{Z} \sum_{\substack{u \sim a \\ v \sim b}} \text{sim}_\text{S}^{(t)}(u, v) w_{ua} w_{vb}, \tag{4}$$

$$\text{sim}_\text{S}(a, b) = \lim_{t \to \infty} \text{sim}_\text{S}^{(t)}(a, b), \tag{5}$$

where $u \sim a$ and $v \sim b$ indicate that u and v are neighboring vertices of word a and b in the graph $\mathcal{G}$ respectively, $w_{ua} = \text{sim}_\text{C}(u, a)$, $w_{vb} = \text{sim}_\text{C}(v, b)$, $Z = \sum_{u \sim a, v \sim b} w_{ua} w_{vb}$ is the normalization factor, $c = 0.5$ is a constant controlling factor, $\text{sim}_\text{S}^{(t)}(a, b)$ is the semantic similarity of (a, b) after t iterations of propagation.

Semantic Similarity Propagation. Eqs. (3)–(5) actually define a similarity propagation process throughout the graph $\mathcal{G}$, which conforms to the SimRank measurement [10]. Note, the SimRank algorithm measures similarities according to the rule of "two objects are similar if they are related to similar objects" that is equivalent to principle #3 of our similarity metric $\text{sim}_\text{S}(a, b)$. Hence, based on the SimRank algorithm, the complexity of computing semantic similarity $\text{sim}_\text{S}(\cdot, \cdot)$ for all word pairs in the vocabulary by Eq. (4) is $\text{O}(k|\mathcal{V}|^2)$, where k is the average vertex degree in $\mathcal{G}$, $|\mathcal{V}|$ is the vocabulary size. By eliminating very high/low frequency words, we downsize the vocabulary in semantic similarity propagation, thus helping to speed-up the computation. Furthermore, since the parallel *single-pair* SimRank computation has been proven to be very efficient in [12], we implement a parallel *all-pair* SimRank-based semantic similarity propagation process in GPU. Due to the independence of updating $\text{sim}_\text{S}(a, b)$ and $\text{sim}_\text{S}(a', b')$ by Eq. (4), our parallel all-pair SimRank implementation is about 10^3 times faster than CPU implementation, and averages 100 times faster than calling $|\mathcal{V}|(|\mathcal{V}| - 1)$ times of single-pair parallel SimRank [12].

After similarity propagation, we can finally construct the *semantic similarity matrix* $\mathbf{S}_\text{S} = \{\mathbf{S}_\text{S}(i, j)\}_{i,j \in |\mathcal{V}|}$ as either $\text{sim}_\text{S}(i, j)$ (if both i and j are not very high/low frequency words) or $\text{sim}_\text{H}(i, j)$ otherwise. Similarly, we can also define hard similarity matrix as $\mathbf{S}_\text{H} = \{\text{sim}_\text{H}(i, j)\}_{i,j \in |\mathcal{V}|} = \mathbf{I}$.

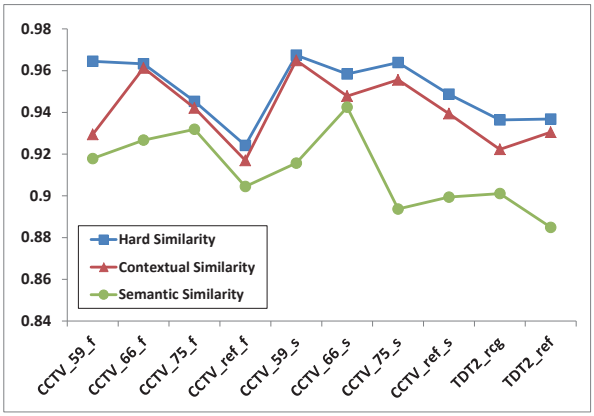

Fig. 3. The average inter- and intra-story similarity ratio on benchmark datasets CCTV and TDT2.

2.3. Sentence-Level Semantic Similarity

In story segmentation, besides word-level similarity, we need also measuring semantic similarities at sentence level [5, 6]. For a given vocabulary $\mathcal{V}$, a sentence $\mathbf{s}_i = \{w_1, \cdots, w_L\}$, *i.e.* a word sequence of length L, can be represented by the word frequency vector $\mathbf{f}_i$ that records the appearance times of each word of $\mathcal{V}$ in the sentence. With a given word-level similarity matrix $\mathbf{S}$, we can measure similarity between sentences s_i and s_j as

$$\text{Sim}(\mathbf{s}_i, \mathbf{s}_j | \mathbf{S}) = \frac{\mathbf{f}_i^T \mathbf{S} \mathbf{f}_j}{\|\mathbf{f}_i\| \|\mathbf{f}_j\|}, \tag{6}$$

where $\|\mathbf{f}_i\|$ and $\|\mathbf{f}_j\|$ denote the L_2 norm of $\mathbf{f}_i$ and $\mathbf{f}_j$, respectively. Note that, when using the hard similarity matrix $\mathbf{S}_\text{H}$, Eq. (6) reduces to the classical widely-used cosine similarity [4, 5, 6], and $\text{Sim}(\mathbf{s}_i, \mathbf{s}_j | \mathbf{S}_\text{S})$ corresponds to our proposed sentence-level soft semantic similarity measurement.

In practice, to segment a particular transcript $\mathcal{T}$, we need only describing sentences as word frequency vectors over the local vocabulary $\mathcal{V}_\mathcal{T}$ of the input transcript rather than over the whole vocabulary $\mathcal{V}$.

2.4. Validation and Discussion

In story segmentation, the goodness of a similarity metric, *e.g.* the proposed semantic similarity $\mathbf{S}_\text{S}$ and hard similarity $\mathbf{S}_\text{H}$, can be empirically evaluated by the ratio of average sentence-level inter- and intra-story similarities, based on particular corpus $\mathcal{C}$ with ground-truth story labelings available:

$$\text{R}(\mathcal{C} | \mathbf{S}) = \frac{\exp\left(\text{mean}_{\text{lab}(\mathbf{s}_i) \neq \text{lab}(\mathbf{s}_j)} \text{Sim}(\mathbf{s}_i, \mathbf{s}_j | \mathbf{S})\right)}{\exp\left(\text{mean}_{\text{lab}(\mathbf{s}_i) = \text{lab}(\mathbf{s}_j)} \text{Sim}(\mathbf{s}_i, \mathbf{s}_j | \mathbf{S})\right)}, \tag{7}$$

where $\text{lab}(\mathbf{s}_i)$ stands for the story id of sentence $\mathbf{s}_i$. Clearly, lower R ratio corresponds to better discriminative ability in story segmentation.

Fig. 3 shows the comparative R ratios on several benchmark datasets of the hard similarity $\mathbf{S}_{\mathrm{H}}$, contextual similarity $\mathbf{S}_{\mathrm{C}}$, and the proposed semantic similarity $\mathbf{S}_{\mathrm{S}}$. We can observe that R ratio of the proposed similarity measurement is apparently lower than the other two, and the contextual similarity is slightly better than hard similarity, *i.e.* $\mathrm{R}(\cdot|\mathbf{S}_{\mathrm{H}}) > \mathrm{R}(\cdot|\mathbf{S}_{\mathrm{C}}) > \mathrm{R}(\cdot|\mathbf{S}_{\mathrm{S}})$, which partially demonstrates the effectiveness of the proposed semantic similarity measurement.

3. EXPERIMENTAL RESULTS

We have tested the performance of the proposed semantic similarity metric $\mathbf{S}_{\mathrm{S}}$ in Chinese news story segmentation. We also used the hard and contextual similarities $\mathbf{S}_{\mathrm{H}}$ and $\mathbf{S}_{\mathrm{C}}$ as two baseline measurements. For the fairness of comparison, we applied all three similarity measurements in the state-of-the-art NCuts based story segmentation method [5, 6], using the sentence-level similarity measurement of Eq. (6). Note that, using hard similarity $\mathbf{S}_{\mathrm{H}}$ exactly corresponds to the original algorithm of [5, 6]. Our evaluation was based on two benchmark Chinese news corpora CCTV and TDT2. The CCTV corpus covers 71 news episodes with 27 hours of Mandarin broadcast news (including both long and shorter news datasets) and contains three different ASR rates 59%, 66% and 75% together with the ground-truth transcripts without ASR error (denoted as CCTV-59-f/s, CCTV-66-f/s, CCTV-75-f/s and CCTV-ref-f/s, respectively); while the TDT2 corpus contains 177 audio recordings of VOA Mandarin broadcast news accompanied by story boundaries with manual word transcripts and LVCSR transcripts (denoted as TDT2-ref and TDT2-rcg, respectively).

DataSet	$\mathbf{S}_{\mathrm{H}}$	$\mathbf{S}_{\mathrm{C}}$	$\mathbf{S}_{\mathrm{S}}$	Imp
CCTV-59-f	0.6775	0.6833	**0.7144**	0.0369 (5.4%)
CCTV-66-f	0.6589	0.6628	**0.6938**	0.0349 (5.2%)
CCTV-75-f	0.6599	0.6898	**0.7122**	0.0523 (7.9%)
CCTV-ref-f	0.6770	0.6971	**0.7439**	0.0669 (9.9%)
CCTV-59-s	0.6275	0.6345	**0.6552**	0.0277 (4.4%)
CCTV-66-s	0.6371	0.6542	**0.6657**	0.0286 (4.5%)
CCTV-75-s	0.6390	0.6497	**0.6738**	0.0348 (5.4%)
CCTV-ref-s	0.6987	0.7060	**0.7180**	0.0193 (2.8%)
TDT2-rcg	0.6532	0.6863	**0.6886**	0.0354 (5.4%)
TDT2-ref	0.6736	0.7089	**0.7137**	0.0401 (6.0%)

Table 1. Average F1-measure of story segmentation using hard similarity $\mathbf{S}_{\mathrm{H}}$, contextual similarity $\mathbf{S}_{\mathrm{C}}$, and the proposed semantic similarity $\mathbf{S}_{\mathrm{S}}$, respectively, on benchmark corpora CCTV and TDT2.

Table 1 shows the average F1-measure scores on 10 particular datasets in CCTV and TDT2 corpora for the three kinds of similarity measurement. We can clearly see that our semantic similarity metric always helps to produce better segmentation accuracy than the other two measurements using exactly the same segmentation algorithm. On average, we obtained 3%–10% improvement in F1-measure by simply using the proposed similarity measurement $\mathbf{S}_{\mathrm{S}}$ in our experiments. We also find that the contextual similarity $\mathbf{S}_{\mathrm{C}}$ is better than the hard similarity $\mathbf{S}_{\mathrm{H}}$. This demonstrates the benefit of considering contextually inferred similarities in story segmentation.

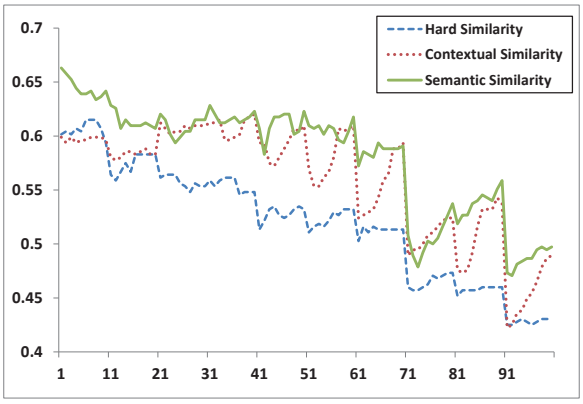

Fig. 4. Segmentation accuracy of three types similarity measurements using 100 groups of random parameters for the N-Cuts algorithm on CCTV-75-s dataset.

In Fig. 4, we conducted a more strict experiment by comparing the segmentation accuracy using 100 groups of randomly generated parameters. That is, all three similarity measurements were compared using the same randomly generated segmentation parameters. As shown in Fig. 4, the proposed semantic similarity $\mathbf{S}_{\mathrm{S}}$ always outperforms the widely-used hard similarity $\mathbf{S}_{\mathrm{H}}$, and the contextual similarity $\mathbf{S}_{\mathrm{C}}$ is better than $\mathbf{S}_{\mathrm{H}}$ on average.

4. CONCLUSION

In this paper, we have proposed a purely data-driven similarity measurement of Chinese words and sentences from a given corpus. Our approach does not rely on any kind of expert supervision, or the guidance of ground-truth topic labeling and story boundaries. We show that contextual word connections positively correlate to the semantic relatedness of words and sentences. Based on the contextual similarity, we further use a parallel all-pair SimRank method to effectively propagate the sparse semantic relatedness to the whole vocabulary. We then extend the classical widely-used cosine similarity to measure soft semantic sentence-level similarities. Extensive experiments on benchmark corpora have shown that, in story segmentation, the proposed soft semantic similarity metric can always produce better segmentation accuracy than using the hard similarity by the state-of-the-art NCuts algorithm [5, 6]. Our experiments also validated the superior effectiveness of reasonably measured soft similarities in story segmentation. In the future, we plan to explore the application of our approach to story segmentation in other languages.

5. REFERENCES

[1] J. Allan, Ed., *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers, 2002.

[2] W. Feng, X. Nie, L. Wan, L. Xie, and J. Jiang, "Lexical story co-segmentation of Chinese broadcast news," in *INTERSPEECH*, 2012.

[3] S. Banerjee and I.A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings," in *INTERSPEECH*, 2006.

[4] L. Xie, J. Zeng, and W. Feng, "Multi-scale TextTiling for automatic story segmentation in Chinese broadcast news," in *AIRS*, 2008.

[5] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *ACL*, 2006.

[6] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A subword normalized cut approach to automatic story segmentation of chinese broadcast news," in *AIRS*, 2009.

[7] Z. Liu, L. Xie, and W. Feng, "Maximum lexical cohesion for fine-grained news story segmentation," in *INTERSPEECH*, 2010.

[8] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity - measuring the relatedness of concepts," in *AAAI (Intelligent Systems Demonstration)*, 2004.

[9] Christiane Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[10] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *ACM SIGKDD*, 2002.

[11] G.A. Miller and W.G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[12] G. He, H. Feng, C. Li, and H. Chen, "Parallel SimRank computation on large graphs with iterative aggregation," in *ACM SIGKDD*, 2010.