文章编号：1003-0077(2016)04-0065-06

基于 WordNet 的中泰文跨语言文本相似度计算

石 杰[1,2]，周兰江[1,2]，线岩团[1,2]，余正涛[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 智能信息处理重点实验室，云南 昆明 650500)

摘 要：文本相似度在信息检索、文本挖掘、抄袭检测等领域有着广泛的应用。目前，大多数研究都只是针对同一种语言的文本相似度计算，关于跨语言文本相似度计算的研究则很少，不同语言之间的差异使得跨语言文本相似度计算很困难，针对这种情况，该文提出一种基于 WordNet 的中泰文跨语言文本相似度的计算方法。首先对中泰文本进行预处理和特征选择，然后利用语义词典 WordNet 将中泰文本转换成中间层语言，最后在中间层上计算中泰文本的相似度。实验结果表明，该方法准确率达到 82%。

关键词：WordNet；中间层语言；跨语言文本相似度

中图分类号：TP391 **文献标识码**：A

Chinese-Thai Cross-language Text Similarity Computing Based on WordNet

SHI Jie[1,2], ZHOU Lanjiang[1,2], XIAN Yantuan[1,2], YU Zhengtao[1,2]

(1. School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Key Laboratory of Intelligent Information Processing,
Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Text similarity calculation is widely used by information retrieval, question answering system, plagiarism detection and so on. At present, most research just aim at text similarity of the same language, and research on cross-language text similarity calculation remains an open issue. This paper propose a WordNet-based method of Chinese-Thai cross-language text similarity calculation. We apply the semantic dictionary WordNet to convert the Chinese text and Thai text into a middle layer language, and compute the text similarity between Chinese and Thai in the middle layer. Experimental results show that, this paper's method of computing the similarity between Chinese text and Thai text has 82%'s accuracy.

Key words: WordNet; middle layer language; cross-language text similarity

1 引言

文本相似度在语言学、心理学和信息理论等领域被广泛的讨论，文本相似度计算旨在比较两个文本之间的相关程度。近年来，基于同一种语言的文本相似度计算方法[1-3]日趋成熟，代表算法模型有布尔模型、向量空间模型、概率模型等。但是，对于跨语言文本相似度的研究则很少，跨语言文本相似度是指量化两个不同语言文本之间的相似性，并使量化的结果尽可能符合人工判断的结果。由于汉语和泰语在语法上存在差异，我们无法用现有的计算同一语言文本相似度的方法来计算汉泰双语文本的相似度。目前，关于跨语言文本相似度计算主要有以下几种方法：1)基于机器翻译的方法[4]。该方法将源语言文本翻译成目标语言文本，在目标语言空间计算相似度，该方法依赖机器翻译的质量，并很难扩展到多种语言；2)基于统计翻译模型的方法[5]。该方法需要两种语言之间的翻译概念词典，但是翻译概念词典需要建立大规模对齐语料库，代价很大，并

收稿日期：2014-01-04 定稿日期：2015-05-04
基金项目：国家自然科学基金(61363044)

很难扩展到多种语言；3)基于平行语料的方法[6]，该方法以两种语言的平行语料库为基础来计算相似度，该方法的准确性依赖于平行语料库的规模和质量。虽然上述方法取得了不错的效果，但是存在扩展性不足、工作量大等缺点。

Steinberger R[7]等提出一种中间层语言思想，用独立于语言的方式来表示不同语言的文本内容，在多语种词库 EUROVOC 上计算英文文本和西班牙文文本之间的相似度，该种方法不依赖于机器翻译，且有较高的扩展性和准确性，但 Steinberger 并没有把某一种具体的自然语言作为中间层语言，由此受到启发：将中间层语言具体化，将不同语言空间转换成这一具体语言空间来计算文本相似度，WordNet 的多语言版本特性使得语言空间的转换成为可能。WordNet[8]是一个使用同义词集表示概念的英文语义词典，有多语言版本，包括中文版、泰文版，中文 WordNet 的构建原则基本遵守英文 WordNet 的结构特点，将 WordNet 中的概念(同义词集合)映射为本国语言同义词集合，保留概念间的关系[9-10]，本文使用的中文 WordNet 是由东南大学开发的中文版 WordNet，泰文版 WordNet 由 AsianWordNet 提供。不同语言版本之间的 WordNet 的同义词集合的 synset_id 是对应的，通过 synset_id 将中泰文 WordNet 与英文 WordNet 对应起来。因此，本文利用多语言版本 WordNet 的 synset_id 相对应这一特性，提出了一种基于 WordNet 的中泰文跨语言文本相似度计算的方法，利用 WordNet 将中文文本和泰文文本转换成统一的中间层语言，并在中间层上计算相似度。

本文第二节主要介绍中泰文本相似度计算的过程，第三节对本文的算法进行测试与评估。

2 中泰文本相似度计算过程

2.1 文本预处理

尽管原始文本包含所有的文本信息，但是目前的自然语言处理技术无法完全处理这些文本信息，因此，需要对文本进行预处理。传统的文本预处理主要是去掉停用词，如“的”“地”等。由于本文的方法需要对词的语义进行分析，因此需要对一些地名、人名等特殊词进行处理，将这些特殊词统一转换成特定的字符串，在进行特征选择时，将这些特殊词项忽略，避免噪声干扰。

2.2 文本特征选择

经过文本预处理后，需要进行文本特征选择。特征选择的目的是选择对相似度计算真正有贡献的特征项，被选中的特征项应能表征原始文本的主题。本文提取词作为文本的特征，将每个文档看成一个词袋，对于中文文档和泰文文档，通过分词，去掉停用词后，都可以形成一个特征词集。然后通过文本频度的选择方法去掉干扰原始文本主题的无用词。文档频度(Document Frequency，DF)是指整个文本集合中包含特征词 t 的文本个数，DF 大于某一阈值则去掉，DF 越高，说明 t 在越多的文本出现；DF 小于某一阈值也去掉，要么是稀有词或噪声。

2.3 中、泰语言空间的转换

考虑到不同语言之间存在很大的差异性，无法在不同语言层完成相似度计算，本文提出一种中间层语言的思想，即将不同语言转换成统一的中间层语言，在中间层上实现中泰文跨语言文本相似度计算。转换模型如图 1 所示。

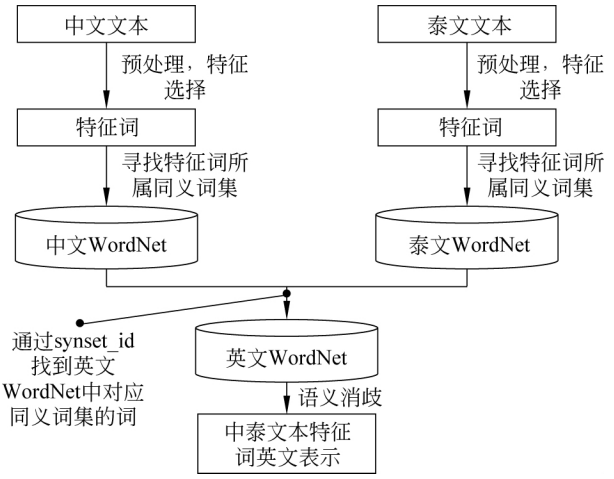

图 1 中、泰语言空间转换

通过图 1 的方式将中文和泰文转换成统一的中间层英语语言空间，我们只需在英语空间上计算中泰文文本的相似度即可。

2.4 语义消歧

在 WordNet 中，任意词可能同时属于若干个同义词集，即若干个语义，要想计算相似度，首先要确定特征词在上下文语境中的语义。因此，对经过 2.3 转换成英文形式的中、泰文本特征词进行语义消歧是计算中泰文本相似度的关键步骤。本文提出

了一种基于 WordNet 语义密度的消歧算法，语义相关性是本文消歧算法的重要依据，这样定义语义相关性：一个词的某个义项与上下文其他词的义项更相关，那么该义项更可能是符合上下文语义的正确义项。该消歧算法首先定义一个语义哈希函数 $\varphi(x)=h$，h 为一整数，将 WordNet 的同义词集 x 映射到一个数值空间，$\varphi(x)$ 满足：

1）任意 $x_1 \neq x_2$，有 $\varphi(x_1) \neq \varphi(x_2)$；

2）任意 x_1, x_2, x_3，若 $d(x_1, x_2) < d(x_1, x_3)$，则有 $P(|\varphi(x_1)-\varphi(x_2)| < P|\varphi(x_1)-\varphi(x_3)|) > 1-\theta, \theta \to 0$。

其中 $d(x_i, x_j)$ 是义项 x_i，x_j 在 WordNet 中的语义距离，表示 x_i，x_j 所在节点最短路径的带权长度，两个义项在 WordNet 中的意义差别越大，语义距离就越大。通过 $\varphi(x)$，就能直接的将 $d(x_i, x_j)$ 转化为 $\varphi(x)$ 的差。通过研究发现，用 128 位整数对 WordNet 的树状结构编码，就能获得满足本文需求的语义哈希 $\varphi(x)$。举个例子，entity 根下的树的语义哈希编码如图 2 所示。

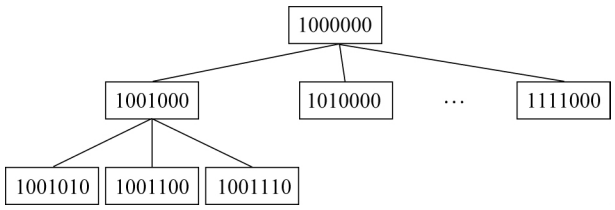

图 2　WordNet 语义哈希编码示意图

有了语义距离和语义哈希，我们就可以定义语义密度来量化一组词之间的语义相关性。对于一组同义词集 $w_1, w_2, \cdots, w_n$，它们的语义密度 density $(w_1, w_2, \cdots, w_n)$ 可以由 n^3 与包含所有 $w_1, w_2, \cdots, w_n$ 的最小子树的“体积”$V_{min}(w_1, w_2, \cdots, w_n)$ 的商，如式(1)所示。

$$density(w_1, w_2, \cdots, w_n) = \frac{n^3}{V_{min}} = \frac{n^3}{1.2max(b-72.0)} \tag{1}$$

其中，b 是 $\varphi(w_1), \varphi(w_2), \cdots, \varphi(w_n)$ 的最低相同位位置。可以这样理解语义密度，语义密度越大，这组词更相关。

有了语义密度，接下来进行语义消歧。从消歧上下文窗口中的所有词可能的义项中选取任意个，计算它们的语义密度，语义密度最大的子树所包含的义项即是消歧结果。例如，取窗口大小为 31，窗口为：$W=\{w_1, w_2, \cdots, w_{15}, w_{16}, w_{17}, \cdots, w_{31}\}$，$w_{16}$ 即为被消歧的词。完成一次消歧后，窗口向右移动一位，直到所有的词都完成消歧。具体消歧步骤如下：首先将 W 中每个词的同义词集合并成一个大的候选集 $C=\{s_{i1}, s_{i2}, \cdots\}$，得到 C 后，将 C 中所有同义词集按语义哈希值 $\varphi(x)$ 排序，得到 $C=\{s_1, s_2, \cdots, s_n\}$，其中 $\varphi(s_1) < \varphi(s_2) < \cdots < \varphi(s_n)$，然后计算 C 中任意两个同义词集的语义密度，选取语义密度最大的子树下的同义词集的义项作为消歧结果。

2.5　中、泰文本相似度计算

计算两个文档相似度一般用它们对应向量的夹角余弦值来表示，如式(2)所示。

$$Sim(D_i, D_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_{ik} * w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}} \tag{2}$$

其中 W_{ik} 和 W_{jk} 分别表示文本 D_i 和 D_j 第 K 个特征词的权值，权值计算采用 IDF-TF 算法。这种计算相似度的方法的假设前提是：词与词之间是没有语义关系的。但是现实文本中的词往往都是有关联的，比如同义关系、上下位关系等。因此，本文使用语义词典 WordNet 来计算中、泰文本特征词之间的相似度。

基于 WordNet 的词语语义相似度计算，目前有两大类算法：基于路径、基于信息内容(Information Content，IC)。本文采用基于 IC 的相似度算法。

基于信息内容的相似度算法是以 WordNet 中每个概念的 IC 值作为参数，由 Resnik[11] 首次提出。IC 表示为 $-\lg p(c)$（在信息论中，称为自信息）。Resnik 认为，两概念的相似度由包含两概念的最深层的公共父节点来决定，只需求出该公共父节点的特征值，就可以得到两概念的相似度值。Resnik 的算法模型如式(3)所示。

$$sim_{RES}(c_1, c_2) = -\lg p(lso(c_1, c_2)) = IC(lso(c_1, c_2)) \tag{3}$$

$lso(c_1, c_2)$ 表示概念 c_1, c_2 在 is_a 树中最深层的公共父节点，$p(c)$ 表示遇到概念 c 的实例的概率。该类代表算法为 Lin 算法[12]。Lin 的语义相似度算法考虑定义一个通用的计算相似度的方法，算法模型如式(4)所示。

$$sim_{LIN} = \frac{2 * \lg p(lso(c_1, c_2))}{\lg p(c_1) + \lg p(c_2)} = \frac{2 * IC(lso(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{4}$$

基于 IC 的相似度算法的性能优越性主要是由概念 IC 值的精确性和将 IC 参数引入算法的合理性

来决定。因此,对IC参数模型进行改进,可以提高算法的性能。Nuno[13]对IC模型的改进算法如式(5)所示。

$$IC(c) = -\lg p(c) = 1 - \frac{\lg(hypo(c)+1)}{\lg(max_{wn})} \quad (5)$$

hypo(c)表示概念c的所有子节点,$\max_{WN}$表示分类树中所有概念的数目。Nuno的模型只是考虑概念的子节点数是有局限性的,本文给出一种改进过的IC求解模型,将概念在分类树中的深度考虑在内,算法模型如式(6)所示。

$$IC_{NEW}(c) = k\left(1 - \frac{\lg(hypo(c)+1)}{\lg(max_{wn})}\right) + (1-k)\frac{\log(deep(c))}{\log(deep_{max})} \quad (6)$$

k介于0到1之间,本文取$k=0.5$。

考虑到Lin算法的通用性,将式(6)带入式(4),得出新的求解相似度模型如式(7)所示。

$$sim_{NEW}(c_1,c_2) = \frac{2 * IC_{NEW}(lso(c_1,c_2))}{IC_{NEW}(c_1)+IC_{NEW}(c_2)} \quad (7)$$

式(7)是对WordNet中两个概念求相似度,求解词相似度算法如式(8)所示。

$$sim_{NEW}(w_1,w_2) = \max[sim_{new}(c_{1i},c_{2j})] \quad (8)$$

其中,c_{1i},c_{2j}为w_1,w_2的若干概念。

假设中文文本CH的特征词$\{CW_1, CW_2, \cdots, CW_n\}$,转换成中间层语言,进行语义消歧后得到对应的英语义项$\{CE_W_1, CE_W_2, \cdots, CE_W_n\}$;泰文文本T的特征词为$\{TW_1, TW_2, \cdots, TW_k\}$,用同样的方式得到英语义项$\{TE_W_1, TE_W_2, \cdots, TE_W_k\}$,结合式(8),则求解$CH$和$T$的相似度的公式如式(9)所示。

$$sim(CH,T) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} sim_{NEW}(CE\_W_i, TE\_W_j)}{n * k} \quad (9)$$

计算结果介于0到1之间,0表示不相似,1表示完全相似,数值越大表示两个文本越相似。

例如,一篇中文文本CH的特征词为(篮球,季后赛,比分,胜利),对应的英文特征词为(basketball, playoff, score, win);一篇泰文文本T1的特征词(บาสเกตบอล, รอบชิงชนะเลิศ, แชมป์),对应的英文特征词(basketball, final, champion)。这两组英文特征词在WordNet的is_a树的位置如图3所示。

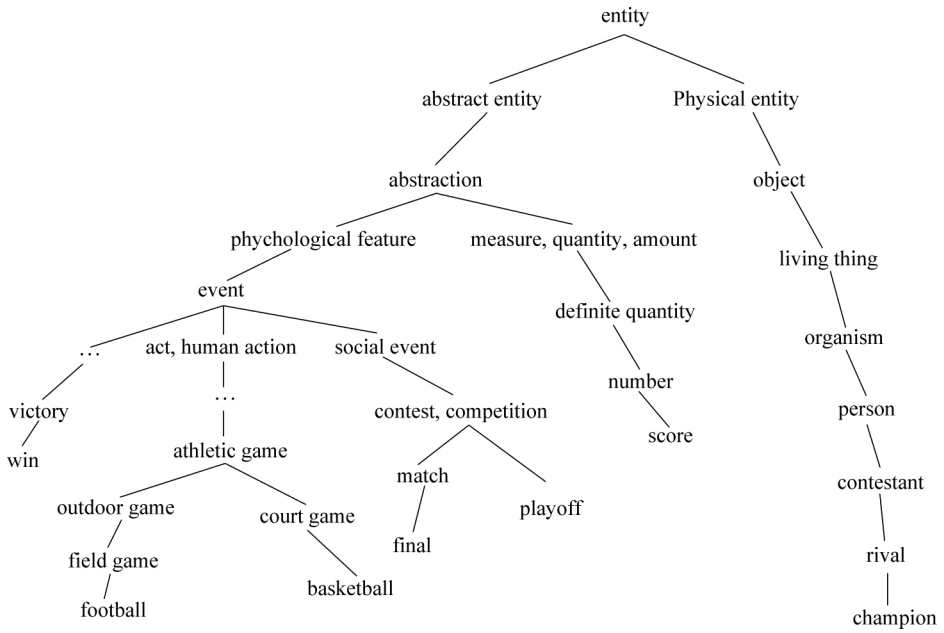

图3 WordNet is_a树

从图3中可以看出,basketball和champion的最深层公共父节点是entity,由Nuno式(5)求得IC(entity)=0,即概念越抽象,所包含的信息内容就越小,代入式(4),求得sim_{LIN}(basketball, champion)=0,这显然不符合人工对这两个词的语义相似度的判别,因此,使用改进过的IC求解模型能够提高精确性。用改进过的式(9)求得CH和T1的语义相似度为0.783。另选取一篇泰文文本T2,特征词(คู่มือท่องเที่ยว,นักท่องเที่ยว,ท่าอากาศยาน),对应英文特征词(guide, tourist, airport),计算CH和T2的相似度为0.091。因此,中文文本CH与泰文文本T1的相似度要高于CH与T2的相似度,这与我

们人工的判断一致。

3　实验及分析

首先对本文提出的语义消歧算法进行实验测试，为后文计算中泰文本相似度实验提供更准确的特征词义项。实验选用一个公开的语义标注语料库 SemCor，SemCor 的单词语义是基于 WordNet 标注的，用词性标注工具 TreeTagger 进行 POS 标注，将标注结果作为消歧算法的输入，将算法的消歧结果与 SemCor 中人工标注的结果进行对比，得到本文消歧算法的准确率。表 1 列出了 SemCor 中前 10 篇文档的消歧准确率。作为对照，表中基准列表示随机猜测时的消歧准确率。例如，一个词有五个同义词集(即义项)，那么随机猜测的准确率为 20%，即基准为 20%。

表 1　消歧实验结果

文档	准确率/%	基准/%
1	63.45	20.45
2	58.33	21.36
3	42.23	22.77
4	47.32	19.81
5	55.65	23.67
6	50.77	20.84
7	57.35	23.61
8	45.19	22.88
9	47.85	18.01
10	49.99	20.92

消歧算法在 SemCor 上的平均准确率达到 51.8%。

接下来对本文计算中泰文本相似度进行试验。本文实验的文本数为 1 000 篇文本，中文文本 900 篇，泰文文本 100 篇。其中，600 篇中文文本为噪音文本，构成噪声集；另外，300 篇中文文本和 100 篇泰文文本构成标准集，并按中文文本和泰文文本两两间的相似度可分为 20 类，每个类中有 13 到 17 篇中文文本不等，也可以这样理解，在标准集中，每篇泰文文本都有 13 到 17 篇人为觉得相似的中文文本。将噪声集和标准集混合构成测试集进行试验，如下：

从标准集 100 篇泰文文本中顺序抽出一篇文本，然后计算这篇泰文文本与测试集中文文本之间的相似度，按照相似度大小排序，输出相似度最大的前 17 个，然后人为观察输出结果，如果与该篇泰文文本属于同一类的中文文本都被输出，则认为本次计算相似度成功。本文使用空间余弦的相似度算法与本文的算法作比较。

实验结果计算公式如式(10)所示。

$$\text{正确率}=\frac{\text{测试结果正确的文本}}{\text{被测的文本}}\times 100\% \quad (10)$$

实验数据如表 2 所示。

表 2　实验结果对比表

计算方法	测试泰文文本/个	结果正确的泰文文本/个	正确率/%
空间向量余弦算法	100	49	49
基于 WordNet 的算法	100	82	82

实验结果表明：本文所采用计算中泰文跨语言文本相似度的方法更接近人工评断的结果。

4　结束语

本文提出了一种基于 WordNet 的中泰文跨语言文本相似度计算方法，通过将中泰文本转换成中间层语言空间，并在中间层计算中泰文本的相似度。实验结果表明本文提出的方法取得了较好的结果。在以后的工作中，考虑进一步改进 IC 模型，将 WordNet 中概念的子节点的空间结构加入模型中，这样做的目的是获得一个更加精确的 IC 值，提高本文算法的精确度。

参考文献

[1] 李红莲，何伟，袁保宗. 一种文本相似度及其在语音识别中的应用[J]. 中文信息学报，2003，17(01)：60-64.

[2] 宋玲，马军，连莉，张志军. 文档相似度综合计算研究[J]. 计算机工程与应用，2006，30：160-163.

[3] 金博，史彦军，滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报，2005，02：291-297.

[4] Maike Erdmann, Andrew Finch, et al. Calculating Wikipedia Article Similarity Using Machine Translation Evaluation Metrics[C]//Proceedings of the 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA '11). IEEE Computer Society, Washington, DC, USA, 2011: 620-625.

[5] Barrón-Cedeno A, Rosso P, Pinto D, et al. On Cross-

lingual Plagiarism Analysis using a Statistical Model [C]//Proceedings of the PAN. 2008.

[6] Potthast M, Stein B, Anderka M. A Wikipedia-based multilingual retrieval model[M]. Advances in Information Retrieval. Springer Berlin Heidelberg, 2008: 522-530.

[7] Steinberger R, Pouliquen B, Hagman J. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc[M]. Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2002: 415-424.

[8] Miller G A. WordNet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[9] 王石,曹存根. WNCT:一种WordNet概念自动翻译方法[J]. 中文信息学报,2009,23(4):63-70.

[10] 张俐,李晶皎,胡明涵,姚天顺. 中文WordNet的研究及实现[J]. 东北大学学报,2003,04:327-329.

[11] Resnik P. Using information content to evaluate semantic similarity in a taxonomy[J]. arXiv preprint cmp-lg/9511007, 1995.

[12] Lin D. An information-theoretic definition of similarity[C]//Proceedings of the ICML. 1998, 98: 296-304.

[13] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet [C]//Proceedings of the ECAI. 2004, 16: 1089.

石杰(1989—),硕士研究生,主要研究领域为自然语言处理与嵌入式系统研究。
E-mail: 254089809@qq.com

线岩团(1981—),讲师,主要研究领域为信息检索、自然语言处理。
E-mail: 195426286@qq.com

周兰江(1964—),通信作者,硕士生导师,副教授,主要研究领域为自然语言处理与嵌入式系统研究。
E-mail: 915090822@qq.com

(上接第64页)

[29] Huanshen Jia, Haixing Zhao. Spanning Trees in a Class of Four Regular Small World Network Computer Science and Application[J]. 2014,4(4): 43-49.

[30] Ke Zhang, Haixing Zhao, Faxu Li, et al. A Kind of Deterministic Small-World Networks Model and Analysis of Their Characteristics[J]. 2014,4(4): 27-31.

[31] 刘云. 汉语虚词知识库的建设[M]. 武汉: 华中师范大学出版社. 2009: 204-302.

[32] Hu Quan, Liu yanshen, et al. Research on the Automatic Identification method of the Collocation of Relative Word in Chinese Complex Sentences[C]//Proceedings of the 2013 3rd International Conference on Advanced Materials and Information Technology Processing, Los Angeles, CA, USA. 2013,10: 1-2.

胡泉(1980—),博士,讲师,主要研究领域为计算机软件工程和中文信息处理。
E-mail: 123750955@qq.com

李源(1972—),博士,副教授,主要研究领域为计算机软件工程和中文信息处理。
E-mail: yuanli@mail.ccnu.edu.cn

谢芳(1981—),博士,讲师,主要研究领域为业务流程建模、自然语言处理。
E-mail: 33460694@qq.com