基于汉语复句的语义相关度计算及类别的标识

杨进才 陈忠忠 沈显君 胡金柱
(华中师范大学计算机学院 武汉 430079)

摘 要 语义相关度计算作为中文信息处理领域中的一项关键技术,在信息检索、语义消岐、文本分类中起着重要的作用。利用汉语复句的句法理论和关系标记搭配理论,以汉语复句语料库以及搜索引擎获取的复句为语料,提出了一种基于汉语复句的语义相关度计算方法——SRCCS。本方法不仅能够计算词语的相关度,而且能够表明相关的性质与类别。与通过短文计算相关度的方法相比,本方法选取的计算对象范围更小,因而结果更准确,计算复杂度更低。在同一测试集上与搜索引擎方法的对比分析证明了基于汉语复句的语义相关度计算方法的有效性与优越性。

关键词 复句,语义相关度,关系标记,关系类别
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.051

Word Semantic Relevancy Computation and Categories Identification Based on Chinese Compound Sentences

YANG Jin-cai CHEN Zhong-zhong SHEN Xian-jun HU Jin-zhu
(School of Computer Science, Central China Normal University, Wuhan 430079, China)

Abstract As a critical technique in the field of Chinese information processing, word semantic relevancy computation plays an important role in information retrieval, ambiguity elimination, and text processing. Using syntactic theory and the collocation theory of the relation markers of Chinese compound sentences, as well as making the corpus of Chinese compound sentences and some compound sentences from search engine as the data resource, a semantic relevancy computation method was proposed based on Chinese compound sentence (SRCCS). This method can not only compute the word semantic relevancy, but also show the property and category of the word semantic relevancy. Compared with the method of short text semantic relevancy, this method chooses a smaller scope of evaluation objects, so the results are more accurate and have little computational complexity. Compared with the result by Google Distance, the new measure is more reliable and effective.

Keywords Complex sentences, Semantic relevancy, Relations marker, Relations category

1 引言

语义相关度计算是中文信息处理领域的重要研究课题之一,在信息检索、语义消岐、文本分类中起着重要的作用。信息检索系统使用相关度得分对查询进行扩展;利用词语之间的相关性协助计算机进行词义消歧;在文档自动文摘以及问答系统中常常使用相关度或相似度的得分来评估候选语句的精准程度;在拼写校正中也会用到语义相关度的计算。实际相关性很大的词语,计算得到的相关度可能很小;而实际不相关的词语,计算得到的相关度可能很大。如何利用计算机更好地计算文本或者词汇之间的语义相关度是研究的难点。

例1 ①一则破除迷信,②二则杜绝浪费。(《长江日报》1994年03月25日09版次)

例2 ①一边搞教学,②一边搞科研。(《人民日报》2001年07月07日01版次)

利用Google Distance[1]对例1中的"迷信"、"浪费"进行相关度计算,得到的相关度值相对较大。但是,"迷信"、"浪费"在例1中表现出的语义相关性不大。使用Google Distance对例2中的"教学"、"科研"进行相关度计算,得到的结果相对较小。但是,"教学"和"科研"在例2中表现为语义非常相关。

复句在汉语句法中占有重要地位,汉语句子中近2/3是复句。复句又包括无标复句和有标复句,而有标复句中的关系标记可以作为句法与语义的形式标记,便于进行句子语义计算[2]。因此,本文在研究有标复句的关系词搭配理论、层次关系的基础上,结合搜索引擎,提出了一种基于汉语复句的语义相关度计算方法,以计算不同语境下的词语间的语义相关度。

2 相关工作

2.1 语义相关度与语义相似度的概念

词语语义相关度用于表示两个词语的相关程度,它反映词语的关联程度,即看到一个词语,是不是可以想到另外一个

到稿日期:2016-04-21 返修日期:2016-07-03 本文受国家社科基金(14BYY093)资助。
杨进才(1967—),男,教授,主要研究领域为现代信息系统、中文信息处理;陈忠忠(1990—),男,硕士生,主要研究领域为中文信息处理;沈显君(1973—),男,教授,主要研究领域为人工智能、软件工程;胡金柱(1947—),男,教授,博士生导师,主要研究领域为中文信息处理、软件工程。

词语[3]。目前，作为衡量两个词语相关程度的语义相关度并没有明确的定义[4-5]。

语义相似度与语义相关度是两个不同的概念，语义相似度是两个词在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度[6]。因此语义相似度指的是词语之间的相似性，相似性反映的是词语之间的可替代性，两个相似的事物必然有其相似点。例如："中国"和"熊猫"具有很大的语义相关性，但是"中国"与"熊猫"的语义相似度很小。

语义相似度与语义相关度之间有着密切联系，语义相关度包含了语义相似度的概念[4]。所以，如果两个词语的语义相似，那么这两个词语必定语义相关；反之，如果两个词语语义相关，那么它们不一定语义相似。

2.2　现有的语义相关度计算方法

语义相关度的计算方法主要有3类：基于词典的方法、基于统计的方法、基于维基百科的方法。基于词典的方法主要是利用WordNet和HowNet两种语义词典进行计算[7-9]，如许云、樊孝忠等提出了一种基于知网的语义相关度计算[4]，文献[10]在本体技术[11]的基础上提出了词语语义相关度计算方法。基于统计的方法主要通过统计词语在语料库中出现的频率来计算语义相关度，如王治和、王凌云等提出的一种基于混合概率的潜在语义分析模型[12]。基于维基百科的方法主要是用Wiki Related算法进行语义相关度计算[13]，如万富强、吴云芳提出的基于中文维基百科的词语语义相关度计算[14]。

基于语义词典的计算方法的核心是语义词典，而语义词典在很大程度上受规模和构建者的影响，所以基于语义词典的语义相关度计算方法的正确率不高。基于统计的语义相关度计算方法一般是以计算机为载体，以承载的语言知识资源为基础，对语料信息进行挖掘，通过统计的方法计算相关度，其准确率比与基于词典的方法有了很大提高。基于维基百科的方法把需要计算语义相关度的两个词语放到维基百科中进行查找，然后通过查找得到的页面和类别等信息计算出词语的相关度。维基百科能够获取词典或语料中未收录的词，但其搜索的词大多处于不同的语境下，因而影响了准确率。

句子是由词构成的，进行语义相关度计算的两个词必然与句子的语法结构、语境等存在着联系。如果在计算两个词的语义相关度时能够考虑句子的语法结构特点，则准确率可以提高。复句由多个语义密切相关的句子组成，从复句中计算词语的相关度能将复句的语义关联性反映到词语的相关度中。

2.3　复句关系的分类

复句的关系可分为3类：因果类复句、并列类复句、转折类复句。根据有标复句中关系标记的类别划分，可以将复句划分为12个小类：因果复句、推断复句、假设复句、条件复句、目的复句、并列复句、连贯复句、递进复句、选择复句、转折复句、让步复句、假转复句。比如：

例3　①虽然身体虚弱，②但是所有的队员都打得勇猛坚定。(《长江日报》1993年10月04日03版次)

例4　①一来可以保护环境，②二来可以帮帮企业。(《长江日报》1998年11月11日09版次)

例5　①今天下雨了，②地上很潮湿。(摘自互联网)

例3中①的关系标记"虽然"与②中的关系标记"但是"形成转折关系标记搭配，即例3为转折复句。例4中①、②两个分句形成并列关系，所以例4为并列复句。例5中①、②两分句中虽然没有关系标记，但是对两个分句进行人工分析发现其为因果复句。根据对CCCS语料库(CCCS中包含658447条有标复句，约有19375158个词。本文在进行语义相关度计算时主要采取该语料库中的语句进行统计分析)中65万多条复句进行分析，与例3、例4中的关系标记搭配类似的大约有400对，并且这些关系标记之间存在着搭配距离(搭配距离是指关系标记在搭配格式中的间隔跨度，即关系标记间隔的单词数目)，在CCCS语料库中统计的关系标记搭配距离的部分数据如表1所列。而类似于例5的无标复句主要分为两类：因果类复句和并列类复句。所以，不管实际语言运用中复句如何千变万化，都离不开这几种关系[15]。

表1　关系标记的搭配距离

搭配标记	关系类别	搭配距离(平均值)
假设…就	假设关系	20.1538462
由于…从而	因果关系	21.6949153
因为…从而	因果关系	23.1652174

3　基于汉语复句的语义相关度计算

语义相关度算法的主要思想是利用汉语复句的关系标记、搭配理论分析复句的语境信息，并结合搜索引擎提取出给定词语的统计信息和语义信息。通过利用汉语复句的搭配理论，可以对给定词语所处的语义环境进行建模分析，对给定的语料库分析统计词语在语料库中的频数、搭配距离以及在网页中出现的频数等统计信息。因为频数和搭配距离从不同方面表示了词语的语义信息，所以在基于汉语复句的语义相关度计算时，对两者进行了综合考虑。对相关度计算公式中涉及的变量进行如下定义。

定义1　在汉语复句语料库中，假设复句语料库的总词量为$W1$，给定词$c1$在语料库中出现的频数记为$f(c1)$，给定词$c2$在复句语料库中出现的频数记为$f(c2)$，给定词$c1$与$c2$同时在一条复句中出现的频数记为$f(c1,c2)$。

汉语复句中同一词语对出现在不同的语句中，由于词语间隔距离不同，导致语义相关度不同。因此，计算语义相关度时应考虑特定词语间的间隔。

例6　①她一边玩苹果手机，②一边听音乐。(摘自互联网)

例7　①她一边吃苹果，②一边玩手机。(摘自互联网)

例6中的"苹果"与"手机"之间的词语间隔为0，例7中的"苹果"与"手机"间的词语间隔为2。通过人工分析，例6中的"苹果"、"手机"与例7中的"苹果"、"手机"的语义相关性不同，前者的语义相关性大于后者的语义相关性。

定义2　假设给定词$c1$和$c2$之间的间隔的单词数是d，那么将d作为词$c1$和$c2$之间的跨度。

例8　①不仅使蔬菜市场丰富多彩，②同时也使菜农收入明显提高。(《人民日报》)

对例8进行分词后得到的语句为"不仅/c 使/v 蔬菜/n

市场/n 丰富多彩/i,/wp 同时/c 也/d 使/v 菜农/n 收入/n 明显/a 提高/v。/wp"。假如给定的两个词为"蔬菜"、"菜农","蔬菜"与"菜农"之间的间隔单词数为 5,那么"蔬菜"和"菜农"的跨度为 5。

定义 3 将给定词 $c1$ 与 $c2$ 所在的复句中关系标记间的搭配距离记为 m。

m 的取值是根据同一关系标记搭配在 CCCS 语料库中各分句中词语间的搭配距离而计算的平均值。

汉语复句语料库虽然是研究汉语复句的重要语料库,但是语料量有限,给定的词可能在语料库中未收录。本文使用搜索引擎来解决词未收录的问题。

定义 4 假设搜索引擎的总索引量为 $W2$,利用搜索引擎搜索给定词 $c1$,其在网页中出现的次数记为 $f(c1')$。利用搜索引擎搜索给定词 $c2$,其在网页中出现的次数记为 $f(c2')$。给定词 $c1$ 与 $c2$ 在网页中同时出现的次数记为 $f(c1',c2')$。

根据定义 1—定义 4,令 $W=W1+W2$,$F(c1)=f(c1)+f(c1')$,$F(c2)=f(c2)+f(c2')$,$F(c1,c2)=f(c1,c2)+f(c1',c2')$。其中,$F(c1)$与 $F(c2)$为 $c1$ 与 $c2$ 在语料库和搜索引擎中共现的次数,$F(c1,c2)$为 $c1$ 与 $c2$ 在语料库和搜索引擎中共现的次数。对词语出现的频数取对数再进行概率计算,$c1$ 的概率计算公式为:

$$P(c1)=\frac{\log F(c1)}{\log W} \tag{1}$$

同理,$c2$ 的概率计算公式为:

$$P(c2)=\frac{\log F(c2)}{\log W} \tag{2}$$

根据关系标记搭配理论,在一定的语境下,词语共现在一定程度上可以反映词 $c1$ 和 $c2$ 的语义相关度。基于此,本文将词语共现作为语义相关度计算的考虑因素之一,提出如下语义相关度计算公式:

$$Rel(c1,c2)=P(c1)*P(c2)*\alpha^2+\beta*\frac{\log F(c1,c2)}{m*m+d*d} \tag{3}$$

其中,α,β是可调节参数,$\alpha+\beta=1$;$F(c1,c2)$表示复句中词 $c1$ 和 $c2$ 共现的次数;m 为表 1 中关系标记的搭配距离;d 为 $c1$ 与 $c2$ 之间的跨度。这个公式考虑了词语共现、关系标记搭配距离以及词间距对语义相关度的影响。

4 实验结果与分析

4.1 实验测试集

与传统的基于语料库的 α 语义相关度分析方法不同,基于汉语复句的语义相关度计算方法不仅依赖于语料库,而且依赖于搜索引擎。为了保证提取语义信息的有效性,本文选取的语料库是汉语复句语料库 CCCS,搜索引擎为百度搜索引擎。本文研究的是从汉语复句中计算两个词语的语义相关度,传统的标准测试集是人为挑选的一些具有代表性的词语对,但这些词语对大多数未出现在同一复句中,无法满足实验的需要。本文根据传统标准测集(30 对词语)的构造方法新建一个基于汉语复句的测试集。本文从汉语复句中选取了 30 对词语构成实验测试集进行测试,实验测试集同时包括相关、一般相关、不相关 3 种特性。由于搜索引擎中的信息不断变化,因此语义相关度的值是动态的,具有时效性。

4.2 实验结果

为了更好地说明基于汉语复句的语义相关度计算方法的可行性,本文采用 Google Distance 进行对比分析。Google Distance 使用 Google 搜索引擎,为了保证实验环境的一致性,本文将 Google Distance 的语义相关度计算公式运用到百度搜索引擎上,下文将该方法称为 Baidu Distance。在计算出复句中两个特定词之间的语义相关度的同时,根据复句关系词的类别标出相关度的类别。相关度用向量 $V(v_1,v_2,\cdots,v_{13})$表示。其中,v_1-v_{12}分别表示 12 种关系类别(因果、推断、假设、条件、目的、并列、连贯、递进、选择、转折、让步、假转)的值;v_{13}为其他不明确类型的相关度,例如无标复句的关联类别。

例 9 ①既有硬件,②又有软件。(并列复句)(《长江日报》1991 年 01 月 11 日 01 版次)

例 10 ①如果把计算机硬件比作人的身躯,②那么软件就如人的思想和智慧。(假设复句)(《人民日报》1992 年 11 月 20 日)

例 11 ①无论"硬件"还是"软件",②都离不开法制。(让步复句)(《人民日报》2002 年 09 月 23 日 02 版次)

例 12 ①这就不仅是硬件的取舍、人员的改组,②而且是软件的更新、机制的重构。(递进复句)(《长江日报》1993 年 10 月 15 日 01 版次)

例 13 ①做硬件的人少,②做软件的人多。(无标复句)(《人民日报》)

使用式(3)对例 9 中的"软件"、"硬件"两词进行计算,得到的结果为(0,0,0,0,0,0.17101624,0,0,0,0,0,0,0),其中的 0.17101624 代表"软件"与"硬件"的并列复句的语义相关度。使用式(3)对例 10—例 13 中的"硬件"和"软件"进行计算,得到的结果分别为(0,0,0.067419951,0,0,0,0,0,0,0,0,0),(0,0,0,0,0,0,0,0,0,0,0.067419951,0,0),(0,0,0,0,0,0,0,0,0.067419951,0,0,0,0,0),(0,0,0,0,0,0,0,0,0,0,0,0,0.067419951)。根据计算结果,分别标识例 9—例 13 的关系类别为并列、假设、让步、递进以及无标复句。

实验过程中,将式(3)中的 α 设为 0.3,实验结果的部分数据如表 2 所列。

实验结果中的 Baidu Distance 是将谷歌距离公式运用到百度搜索引擎上,SRCCS 是本文所提方法——基于汉语复句的语义相关度计算方法。为了证明 SRCCS 的可行性,本文对表 2 中的部分数据进行分析。

例 14 ①夫妻俩一边听收音机,②一边打开箱子。(《人民日报》2000 年 01 月 13 日 06 版次)

例 15 ①因为落后,②所以挨打。(《人民日报》1990 年 10 月 17 日)

例 16 ①生也罢,②死也罢。(《长江日报》1997 年 09 月 19 日 20 版次)

例 1 中"迷信"和"浪费"两个词使用 Baidu Distance 计算的结果大于 SRCCS 的计算结果,而"迷信"和"浪费"两个词的相关联程度很小,SRCCS 的准确性更好。例 14 中"收音机"和"箱子"的相关性很小,所以两者的相关度计算值较小,而

Baidu Distance 与 SRCCS 的计算结果相近。例 2、例 15、例 16 中的“教学”和“科研”、“落后”和“挨打”、“生”和“死”都具有很强的相关性，使用 SRCCS 计算的结果值远大于 Baidu Distance 计算的结果值。

表 2　实验结果示例

Sentence	Word1	Word2	Baidu Distance	SRCCS	Relation type
一则破除迷信，二则杜绝浪费。	迷信	浪费	0.668880221	0.210963693	并列
夫妻俩一边听收音机，一边打开箱子。	收音机	箱子	0.104441007	0.198964865	并列
因为落后，所以挨打。	落后	挨打	0.059510998	0.165406900	因果
生也罢，死也罢。	生	死	0.478531511	0.946296045	并列
一边搞教学，一边搞科研。	教学	科研	0.064944762	0.372080425	并列
苦也罢，乐也罢。	苦	乐	0.733374885	0.779757151	并列
佛寺也罢，道观也罢。	佛寺	道观	0.359492080	0.831867921	并列
挖土机也罢，航海也罢。	挖土机	航海	0.346382955	0.703743418	并列
不仅有水果，还有其他食物。	水果	食物	0.0000013204	0.192715845	递进
学校也好，学生也好。	学校	学生	0.00000538686	0.824212977	并列
因为安装了电话，所以大家沟通起来很方便。	电话	沟通	0.258285839	0.115529521	因果
杨杰既当医生，又当护士。	医生	护士	0.552933873	0.176452828	并列
既有学生，也有教授。	学生	教授	0.299208982	0.173606028	并列
书店不仅有售书的职能，而且还起到图书馆的作用。	书	图书馆	0.139803291	0.131380186	递进
如果你钱多，就把它存入银行或者多买些有用的东西。	钱	银行	0.284490197	0.100770536	假设
既聪明能干，学习成绩又好的学生。	聪明	学生	0.220181461	0.136862220	并列
一边看，一边骂。	看	骂	0.568842090	0.386127377	并列
一边上高山，一边上电梯。	高山	电梯	0.187161991	0.295395139	并列
一边是诗，一边是作者。	诗	作者	0.289833402	0.327798811	并列
我一边翻看，一边问。	翻看	问	0.459549528	0.383595638	并列
一边治理，一边破坏。	治理	破坏	0.317125768	0.393792894	并列
他一边玩手机，一边吃苹果。	手机	苹果	0.263367022	0.346646830	并列
要既抓生产，又抓节约。	生产	节约	0.221231336	0.196047861	并列
既要热烈，又要安全。	热烈	安全	0.562015400	0.194022978	并列
既有软件，又有硬件。	软件	硬件	0.105700089	0.214804425	并列
这既不卫生，又妨碍交通。	卫生	交通	0.381406957	0.187041822	并列
既费了钱，又费了时间。	钱	时间	0.263407083	0.211046184	并列
既治理污染，又保护耕地。	治理	保护	0.192521527	0.339831449	并列
人既是目的，又是手段。	目的	手段	0.093369448	0.211937044	并列
因为无愧，所以无惧	无愧	无惧	0.095005547	0.391099392	因果

4.3　实验分析

为了将基于汉语复句的语义相关度计算方法得到的结果和利用 Baidu Distance 计算的结果进行对比评估，本文采用皮尔逊相关系数作为度量标准。皮尔逊相关系数是一种线性相关系数，是一个用来反映两个变量线性相关程度的统计量，计算公式为：

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S_X}\right)\left(\frac{Y_i-\overline{Y}}{S_Y}\right) \tag{4}$$

其中，r 表示相关系数，n 表示样本量，X_i 和 Y_i 表示两个变量的观测值，$\overline{X}$ 和 $\overline{Y}$ 表示两个变量的均值，S_X 和 S_Y 是两个变量的方差。r 描述了两个变量间线性相关的强弱程度[19]。r 的绝对值越大，表明相关性越强；反之则越小。

本文使用标准测试集作为实验对象，分别使用 SRCCS 和 Baidu Distance 两种方法进行相关度计算，将两种方法计算的结果作为变量并使用皮尔逊相关系数进行评估。下面将式(3)中的 α 分别取为 0.1～0.9，将 SRCCS 计算的结果和 Baidu Distance 计算的结果进行比较，得到的结果如图 1 和图 2 所示。

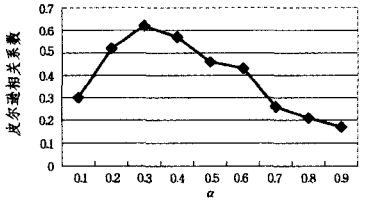

图 1　不同 α 取值下的皮尔逊相关系数

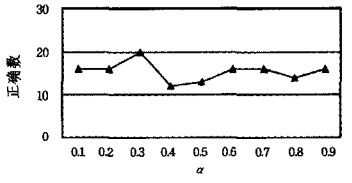

图 2　SRCCS 与 Baidu Distance 实验结果分析图

由图 1 和图 2 可知，当 r=0.62 时，即基于汉语复句的语义相关度计算方法中 α 取 0.3 时，SRCC 的实验结果与 Baidu Distance 的计算结果相比效果更好。

结束语　本文利用汉语复句关系标记及搭配理论的研究成果，并结合搜索引擎，提出了一种基于汉语复句的语义相关度计算方法。该方法不仅计算出了两个词语的相关度，同时标明了相关度的“性质与类别”，有利于更加准确地理解词语的语义。

由于复句语料库句子的有限性，而任意两个词语的组合数巨大，因此需要借助搜索引擎来搜索含有两个词语的复句。对词语在搜索引擎中搜索到的句子中共现的频率，与已有的语料库中共现的频率，需要设置适当的系数进行调和。词语共现在不同类别复句中，复句的类别对整体相关度的影响也不同，需要设计相应的系数。本文仅仅考虑了两个分句之间同种语法成分的词语间的相关度，如作为谓语的动词和作为宾语的名词。如何计算不同词性词语的相关度则需要进一步研究。

参 考 文 献

[1] KJOS-HANSSEN B, Evangelista A J. Google distance between words[OL]. http://math. hawaii. edu/~ bjoern/Publications/Evangelista_Kjos-Hanssen. pdf.

[2] 姚双云. 复句关系标记的搭配研究[M]. 武汉:华中师范大学出版社,2008.

[3] YOU B. Measuring Semantic Relatedness between Words[D]. Wuhan:Central China Normal University Press,2013. (in Chinese)
游博. 词语语义相关度计算研究[D]. 武汉:华中师范大学,2013.

[4] XU Y, FAN X Z, ZHANG F. Semantic Relevancy Computing Based on Hownet[J]. Transactions of Beijing Institute of Technology,2005,25(5):411-414. (in Chinese)
许云,樊孝忠,张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报,2005,25(5):411-414.

[5] WANG H L, LV Q, XU R. Computation model of Chinese semantic relevancy based on HowNet[C]// The National Academic Conference on Information Retrieval and Information Content Security. 2007. (in Chinese)
王红玲,吕强,徐瑞. 一种基于知网的中文语义相关度计算模型[C]//全国信息检索与内容安全学术会议. 2007.

[6] WANG J H, ZUO W L, YAN Z. Word Semantic Similarity Measurement Based on Naive Bayes Model[J]. Journal of Computer Research and Development, 2015, 52(7): 1499-1509. (in Chinese)
王俊华,左万利,闫昭. 基于朴素贝叶斯模型的单词语义相似度度量[J]. 计算机研究与发展,2015,52(7):1499-1509.

[7] AOUICHA M B, TAIEB M A H, HAMADOU A B. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness[J]. Applied Intelligence, 2016, 45(2):1-37.

[8] LI W, YANG C, FU X. Combining How Net and Extension Strategy Generation Method to Improve Customer Values[J]. Procedia Computer Science, 2015, 55:451-460.

[9] XIANG C C, SUI Z F, ZHAN W D. On Mapping between HowNet and CCD[J]. Journal of Chinese Information Processing, 2015, 29(3):44-51. (in Chinese)
向春丞,穗志方,詹卫东. HowNet 与 CCD 映射方法研究[J]. 中文信息学报,2015,29(3):44-51.

[10] KIMTANI D K, CHOUDHURY J, CHAKRABARTY A. Improvement in Word Sense Disambiguation by introducing enhancements in English WordNetStructure [J]. International Journal on Computer Science & Engineering, 2012, 4(7): 1366-1370.

[11] XIAO S, HU J Z, YAO S Y, et al. Objectorient ontology modeling for tag complex sentence[J]. Application Research of Computer, 2010, 27(2):552-554. (in Chinese)
肖升,胡金柱,姚双云,等. 面向对象有标复句本体建模[J]. 计算机应用研究,2010,27(2):552-554.

[12] WANG Z H, WANG L Y, DANG H, et al. Web Clustering Based on Hybrid Probabilistic Latent Semantic Analysis Model [J]. Journal of Computer Applications, 2012, 32(11): 3018-3022. (in Chinese)
王治和,王凌云,党辉,等. 基于混合概率潜在语义分析模型的 Web 聚类[J]. 计算机应用,2012,32(11):3018-3022.

[13] STRUBE B M, PONZETTO S P. WikiRelate! Computing semantic relatedness using Wikipedia[C]// Proc. of AAAI-06. 2015:1419-1424.

[14] WAN F Q, WU Y F. Computing Lexical Semantic relevancy with Chinese Wikipedia[J]. Journal of Chinese Information Processing, 2013, 27(6):31-37, 109. (in Chinese)
万富强,吴云芳. 基于中文维基百科的词语语义相关度计算[J]. 中文信息学报,2013,27(6):31-37,109.

[15] 邢福义. 汉语复句研究[M]. 北京:商务印书馆,2001.

[16] CRISTIANINI N, SHAWE-TAYLOR J, LODHI H. Latent semantic kernels[J]. Journal of Intelligent Information Systems, 2002, 18(2/3):127-152.

(上接第 279 页)

[7] WU J, CUI Z M, SHI Y J, et al. Local Density-based Similarity Matrix Construction for Spectral Clustering [J]. Journal on Communication, 2013(3):14-22. (in Chinese)
吴健,崔志明,时玉杰,等. 基于局部密度构造相似矩阵的谱聚类算法[J]. 通信学报,2013(3):14-22.

[8] HUANG L, LI R, CHEN H, et al. Detecting network com-munities using regularized spectral clustering algorithm[J]. Artificial Intelligence Review, 2014, 41(4):579-594.

[9] SHI J, MAILIK J. Normalized cuts and image segmentation[J]. IEEE transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[10] LIANG H, XUEPING G U. Black-Start Network Partitioning Based on Spectral Clustering [J]. Power System Technology, 2013, 37(2):372-377.

[11] PAN XY, LIU F, JIAO L C. Density Sensitive Based Multi-Agent Evolutionary Clustering algorithm [J]. Journal of Software, 2010, 21(10):2420-2431. (in Chinese)
潘晓英,刘芳,焦李成. 密度敏感的多智能体进化聚类算法[J]. 软件学报,2010,21(10):2420-2431.

[12] LIU Y, CAI D, LI C. Density Sensitive Hashing [J]. IEEE Transactions on Cybernetics, 2012, 44(8):1362-1371.

[13] YANG P, ZHU Q, HUANG B. Spectral clustering with density sensitive similarity function [J]. Knowledge-Based Systems, 2011, 24(5):621- 628.

[14] CHUNG F R K. Spectral graph Theory [J]. Regional Conference, 1997, 7(1):158.

[15] MOHAR B. The Laplacian spectrum of graphs[J]. Graph Theory, Combinatorics, and Applications, 1991, 2(7):871- 898.

[16] FAK K. On the theorem of weyl concerning eigenvalues of linear transformations[J]. Proceedings of National Academy of Sciences, 1950, 35(11):652-655.

[17] BOYD, STEPHEN, VANDENBERGHE, et al. Convex Optimization[M]//Cambridge University Press. 2004:1859.