

改进的主客观结合的词语语义相似度算法

吴旭东¹, 成卫青¹, 黄卫东²

(1 南京邮电大学 计算机学院, 江苏 南京 210003;

2. 南京邮电大学 经济与管理学院, 江苏 南京 210003)

摘要:鉴于词语表达形式与词语语义的多样性,词语语义相似度计算是自然语言处理、智能检索、文档聚类等领域的一个研究热点。文中根据词语表达方式的特点,在基于词语语义词典和基于大规模语料库这两种计算词语语义相似度方法的基础之上,提出一种改进的主观和客观相结合的词语相似度计算方法。从方法论的角度,本算法既融合了主观经验主义思想也融合了客观的理性主义思想,使得词语语义相似度的计算结果能够更加准确。实验结果表明采用文方法是有效的,能够显著提高词语语义相似度计算结果的准确性。

关键词:词语语义相似度;知网;客观相似度;主观相似度

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2012)09-0045-05

An Improved Subjective and Objective Combination Method for Measuring Word Semantic Similarity

WU Xu-dong¹, CHENG Wei-qing¹, HUANG Wei-dong²

(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. College of Economics & Management, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract: In view of the diversity of word expression form and word semantics, the word semantic calculation is a hot research topic in the fields of natural language processing, intelligent search, document clustering and so on. According to the features of word expression, based on the two methods which is based on word semantic dictionary and the other is based on large-scale corpus to calculate word semantic similarity, an improved method combining subjective and objective methods to calculate word semantic similarity is proposed. From the point of view of the methodology, the method has combined both subjective experience and objective rationality, making it possible to improve the accuracy of the word semantic similarity. Experimental results show that the proposed method is effective and can significantly improve the accuracy of the word semantic similarity.

Key words: word semantic similarity; howNet; objective similarity; subjective similarity

0 引言

词语语义相似度计算是自然语言处理、智能检索、文档聚类等领域的一个基本问题。随着互联网得到越来越广泛的应用,网络上的信息数据也呈爆炸式增长,人们开始更加重视词语语义的研究^[1]。与前些年信息匮乏相比,现在信息用户更加重视的是从大量的信息资源中发觉其所需要的信息,鉴于信息资源的异构性,使用传统的字符串匹配为基础的信息检索很难满足用

户对深层信息和知识的需求,因此语义相似度计算成为当前的一个热点研究课题。

词语语义相似度是一个主观性较强的概念,对于两个词语,所应用的语义环境不同得出的语义相似度也可能不同,在一种语境中看来非常相似的词语,在另一种语境中可能差别就很大。在基于实例的机器学习中,词语语义相似度表示两个词语在上下文中相互替换使用而不改变文本的句法和语义结构的程度,改变程度越小,词语相似度越高。

目前计算词语相似度主要有两类方法:1)根据某种已经被认可的知识结构;2)基于大规模语料库进行统计的方法。第一类方法,一般是利用由语言专家研究和总结得出的对词语进行了明确定义的词语语义词典,根据上下位关系和同位关系来计算词语相似度。

收稿日期:2012-02-02;修回日期:2012-05-10

基金项目:国家自然科学基金资助项目(61170322, 71171117);软件开发环境国家重点实验室开放课题(SKLSDE-2011KF-0X);江苏省自然科学基金资助项目(BK2010524)

作者简介:吴旭东(1986-),男,硕士生,CCF会员,研究方向为文本聚类 and 智能检索;成卫青,副教授,博士,研究方向为网络测量。

语义词典是根据词语所表达的概念和意义的关系将词语组织在一起的树状层次结构。第二类方法是一种根据词语所在的语义环境得到词语相似度的方法。该方法认为词语所在的语境可以为词语的表示提供足够的信息^[2]。常见的一种算法是构建词语向量模型,该模型首先要选择一组特征词作为基础词语,然后计算这一组特征词与每个词之间的相关性,这儿的相关性可根据基础词语和每个词语在上下文的共现率来计算,共现率高说明词语间的相关性大,共现率低说明相关性较小,于是每个词语可得出一个以基础词语为维度,以相关性程度为其维度数值的词语向量,最后利用这个基础词语向量得到词语的相似度。

对于基于词语语义词典的方法,分为基于距离的语义相似度测量和基于信息内容的语义相似度测量。前者利用词语在词典中的语义距离(词语之间的跳数)来计算相似度。后者首先计算语义词典中用于表示词语的信息内容各个部分概念的语义距离,然后对各部分概念的语义距离进行加权整合即得到词语语义相似度。这种方法要利用语言专家已经建立好的词典,其优点是词语表示较为直观,语义相似度计算方便,但是由于词典是由人为建立的,词语的定义由语言专家进行定义,所以主观色彩较浓,受到人主观影响较大,带有一种经验主义色彩。对于基于大规模语料库进行统计的方法,要利用大规模的语料库,它的优点是能够比较客观地反应词语的形态,有理性主义色彩,具有较强的客观性。但是词语语义相似度计算的结果依赖语料库的优劣。

文中在以上两种方法的基础之上提出一种改进的基于主观与客观相结合的词语语义相似度计算方法,利用这种方法既在一定程度上消除主观主义色彩,使得计算结果增加客观性,也使词语语义相似度计算结果对于语料库的依赖降低,即使语料库质量不高也可以在一定程度上提升词语语义相似度的计算结果的准确程度。

1 改进的主客观结合的词语语义相似度

词语语义相似度值可以用 $[0,1]$ 之间的数字表示,当计算得出的是0时说明两个词语语义没有相似性,如果计算得出的结果为1说明两个词语在语义方面是完全相等的,在文本中可以互相代替而对于文本的结构和表达意思不会产生改变。结果越接近1说明相似性越高,越接近0说明相似性越低。

鉴于运用词典计算词语相似度主观性太强,带有过于强的经验主义色彩,而运用大规模语料库基于统计的方法对于语料库的依赖程度过大,词语表达时的噪声过大,文中提出一种改进的基于主客观相结合的

词语语义相似度计算方法:即把基于词语词典的词语语义主观相似度和基于大规模语料库的词语语义客观相似度相结合,得到词语的语义相似度。文中在主观相似度计算方面利用已知的词典,利用基于信息内容的语义相似度测量方法来进行相似度计算。在客观相似度计算方面进行一些适当的修改,在利用语料库的基础之上不利用词语—词语向量进行相似度计算,而是运用搜索平台,对词语进行检索,对于搜索出的文本建立词语—文本向量,文中认同这样一种观点:通过搜索得出的文本可以表示词语语义。最后通过计算文本的相似度得到词语的相似度。将主观相似度和客观相似度加权求和,即得到所求词语语义相似度。

文中将词语语义相似度分为两个部分,分别称之为词语语义主观相似度和词语语义客观相似度。对于词语 w 和 v ,其相似度记为 $\text{Sim}(w,v)$,其语义主观相似度记为 $\text{Sim}_{\text{sub}}(w,v)$,语义客观相似度记为 $\text{Sim}_{\text{obj}}(w,v)$ 。最后把 $\text{Sim}_{\text{sub}}(w,v)$ 和 $\text{Sim}_{\text{obj}}(w,v)$ 加权求和就得到 $\text{Sim}(w,v)$ 的值:

$$\text{Sim}(w,v) = \alpha \text{Sim}_{\text{obj}}(w,v) + \beta \text{Sim}_{\text{sub}}(w,v) \quad (1)$$

其中 $0 \leq \text{Sim}_{\text{obj}}(w,v), \text{Sim}_{\text{sub}}(w,v) \leq 1$,其中 $\alpha + \beta = 1$ 。

权值 α 和 β 是可调节参数,可以根据实际应用情况设置其值。如果 α 等于1就说明只是计算词语语义的主观相似度,如果 β 等于1就说明只是计算词语语义客观相似度。文中为了达到主观与客观相似度平衡的目的将 α 和 β 均设置为0.5。下文将对词语语义的主客观相似度的计算分别进行详细的介绍。

1.1 词语语义主观相似度计算

文中在词语语义主观相似度计算方面运用的方法是基于信息内容的语义相似度测量方法,运用的语义词典是《知网》。基于信息内容的词语语义相似度测量是根据整体相似度可由部分相似度合成的思想而来的。《知网》由于其自身的特色,很适合作为基于信息内容的语义相似度测量的词典,因此文中计算词语语义主观相似度运用《知网》作为语义词典。

1.1.1 知网简介

《知网》是一个英汉双语的语义词典,它以词语所表示的概念以及这些概念所具有的特征为基础,并且能够表达概念之间的关系^[3]。在《知网》中对于词语进行解释的基本单位是概念,根据词语的语义,每个词语将由几个概念来表示,而利用“义原”作为表达概念的单位^[3],譬如词语“学生”表示为:

学生:human|人,*study|学,education|教育

其中逗号分隔的内容就是用于表示词语的最小意义单位——义原,符号‘|’和‘*’用于表示义原与词语的关系。在《知网》中大约有1500个义原,分为以

下 8 类:

- 1)Event|事件;
- 2)Entity|实体;
- 3)Attribute|属性;
- 4)Attribut Value|属性值;
- 5)Secondary Feature|次要特征;
- 6)ProperNoun|专有名词;
- 7)Event Role & Features|事件角色及特征;
- 8)Sytax|语法。

义原还可以分为三组:用来描述语义特征的“基本义原”,包括前 6 类义原;用于描述概念之间关系的“关系义原”,只包括第 7 类义原;第三组为“语法学义原”,用于描述词语的语法特征,包括第 8 类义原。此外,《知网》还用一些符号来描述语义,如表 1 所示。

表 1 《知网》知识描述语言中的符号及其意义

,	多个属性之间表示“和”关系
#	表示与其相关
\$	表示“被该动词”处置(即是该动词的受事者)
*	表示“该动词”的代理人,经历者,或者工具(即是该动词的施事者)
+	对动词类,表示它所标记的角色是一种隐形的,实际上几乎在语言中不出现
&	表示“属于”,主要用于概念的属性
-	表示很可能的,多半是,多半有
@	表示事件的时间和地点
?	表示实体的材料
{ }	对于事件,一个事件类的必要角色应放置在 { } 中;将一个单词或一个表达式放在 { } 中,作为一个定义
()	置于其中的应该是一个特有的单词和表达式,如(China 中国)
^	表示否定,如不存在或没有
!	表示一个概念的特有属性,如“味道”对于“食物”,“高度”对于“山脉”,“湿度”对于“天象”等
%	表示是其部分
[]	表示概念的共有属性
<>	作为一个角色使用时,表示这个角色是必不可少的,但它的“施事者”没有指明,如:“走狗”含义中就不会指明“施事者”

譬如学生:human|人,*study|学,education|教育,“学”前面的符号“*”就表示“学生”是“学”这个动作的施事者。再比如风衣:clothing|衣物,#body|身体,*obstruct|阻止,#wind|风,可以解释为“衣物”是第一基本义原,“身体”和“风”表示是与其相关的,“阻止”表示“风衣”是“阻止”这个动作的施事者。

1.1.2 基于《知网》词语语义相似度计算

在《知网》中把所有义原组织成一个网状结构,义原之间的相似度根据义原在网状结构中的距离进行计算,定义义原 p_1 和义原 p_2 ,其相似度值为 $\text{Sim}(p_1, p_2)$,其语义距离为 $\text{Dis}(p_1, p_2)$,则得到公式:

$$\text{Sim}(p_1, p_2) = \frac{\partial}{\partial + \text{Dis}(p_1 + p_2)}$$

(2)

其中 ∂ 是可调节参数。

文献[4 ~ 11],都是基于《知网》利用公式(2)进行词语相似度计算,文献[12]所用方法也可运用在《知网》平台上,文献[13]利用《WordNet》平台进行词语语义相似度计算,对于运用《知网》进行计算也有一定的借鉴意义。

文中使用文献[3]中的方法计算词语语义相似度。对于词语 w 和 v ,定义其语义相似度值为 $\text{Sim}(w, v)$ 。根据义原表示词语的作用和描述符号的不同,将义原分为第一基本义原、其它独立义原、关系义原和符号义原四类^[3]。两个词语的相似度表示为:

$$\text{Sim}(w, v) = \sum_{i=1}^4 \beta_i \text{Sim}_i(w, v)$$

(3)

其中 $\sum_{i=1}^4 \beta_i = 1, \beta_1 > \beta_2 > \beta_3 > \beta_4$ 。

根据以上方法即可以求得词语语义主观相似度 $\text{Sim}_{\text{sub}}(w, v)$ 。

1.2 改进的词语语义客观相似度计算

1.2.1 词语语义客观相似度介绍

词语语义客观相似度计算一般采用先根据词语所在的语义环境得到词语表达式,再计算相似度的方法^[1]。词语—词语向量模型是目前基于统计的词语相似度计算方法中使用较多的一种模型。利用该模型,胡俊峰等对唐诗宋词中词汇的语义相似度进行了研究^[14]。文中在词语—词语向量模型的基础上进行一定的修改,即根据一定的关键词提取算法对文章提取关键词,以关键词来表示文章,对文章用关键词建立索引,然后利用搜索平台对于需要计算相似度的词语进行搜索,以搜索出来的文本表示词语,就构建出词语—文本向量。譬如在搜索平台输入“经理”这个词语,搜索到《禅师杰克逊下赛季回 NBA 湖人经理:密友说他想回来》和《男人应当总经理》这两篇文章,则词语—文本向量就是“经理={ ‘禅师杰克逊下赛季回 NBA 湖人经理:密友说他想回来’, ‘男人应当总经理’ }”。即在文中认同这样一个观点:以某个词语为关键词的文章所构建的向量可以表达该词语的语义。利用词语—文本向量的一个优势是在进行主客观结合的语义相似度计算之前已经对词语在一定程度上进行了去噪处理。譬如对于词语 w ,设它的词语向量是 (w_1, w_2, \cdots, w_n) ,对于某向量值,比如 w_2 , w 和 w_2 关联度并不大,但是由于文章结构原因 w_2 却是表达 w 的一个维度,这就造成了语义噪声,如果这样的语义噪声较多就会对最后的语义相似度计算影响较大。而利用词语—文本向量时,在消除一些出现频数较大但是不具备实际意义的虚词之后,定义词语 v ,其文本向量是 (t_1, t_2, \cdots, t_n) ,只有当 v 为某文本的关键词时,该文本才可以作为 v 的一个维数,而只有文本和词语关联性较大时词语

才可以作为文本的关键词,这样就在一定程度上消除了语义噪声。

1.2.2 改进的词语语义客观相似度计算方法

定义 1 记以词语 w 为关键词的文本向量是 $\{t_1, t_2, \dots, t_n\}$, 称向量 $\{t_1, t_2, \dots, t_n\}$ 是词语 w 的信息内容, 简记为 $\text{Info}(w)$ 。

定义 2 将词语 w_1 和 w_2 的信息内容的并集称作词语覆盖信息。称词语覆盖信息的维数为词语覆盖信息量, 记为 $\text{Info}_{\text{cov}}(w_1, w_2)$, 即:

$$\text{Info}_{\text{cov}}(w_1, w_2) = |\text{Info}(w_1) \cup \text{Info}(w_2)| \quad (4)$$

定义 3 将词语 w_1 和 w_2 的信息内容的交集称作词语共有信息。称词语共有信息的维数为词语共有信息量, 简记为 $\text{Info}_{\text{com}}(w_1, w_2)$, 即:

$$\text{Info}_{\text{com}}(w_1, w_2) = |\text{Info}(w_1) \cap \text{Info}(w_2)| \quad (5)$$

定义 4 将词语 w 和 v 的词语共有信息量与词语覆盖信息量的比值称作词语语义客观相似度, 记作 $\text{Sim}_{\text{obj}}(w, v)$, 即:

$$\text{Sim}_{\text{obj}}(w, v) = \frac{\text{Info}_{\text{com}}(w, v)}{\text{Info}_{\text{cov}}(w, v)} \quad (6)$$

由公式 (4)、(5) 和 (6) 易知 $0 \leq \text{Sim}_{\text{obj}}(w, v) \leq 1$ 。当两个词语语义完全相同时, 那么这两个词语的信息内容应该是相同的, 因此词语的覆盖信息就和词语的共有信息相同, 此时计算得出的词语语义相似度为 1; 当词语语义完全不相同, 词语的共有信息为 0, 所计算出的词语相似度值也为 0; 词语语义相似度越大, 词语共有信息与词语覆盖信息差别则越小, 计算所得相似度则越接近 1; 词语相似度越小, 词语共有信息与词语覆盖信息差别则越大, 计算所得相似度则越接近于 0。

2 实验测试

2.1 实验设置

文中算法实验验证共分三个步骤:

第一步: 计算词语语义主观相似度;

第二步: 计算词语语义客观相似度;

第三步: 将词语语义主观相似度和词语语义客观相似度加权求和得到词语语义相似度, 即利用公式 (1) 得到所求结果。

词语语义主观相似度的计算首先利用《知网》获得词语的定义信息, 然后利用上文 1.1.2 节中介绍的义原分类方法将义原分类, 再利用义原网状结构和公式 (2) 获得义原间相似度, 最后利用公式 (3) 得到词语

语义主观相似度。

词语语义客观相似度算法实验由以下几部分构成: 构建语料库、使用关键词提取器对文本进行关键词提取、建立文本索引、进行搜索、通过搜索结果建立词语—文本向量、最后利用公式 (4)、(5) 和 (6) 即可得到词语语义客观相似度。

语料库由笔者从新浪、百度等大型网站随机抽取的上百篇各类新闻稿件构成; 关键词提取运用已有的关键词提取器, 此软件是以词语的频数来计算该词语是否是文章的关键词; 建立文本索引和对索引进行搜索则是利用 Lucene 全文搜索组件自己进行编程实现的。Lucene 全文搜索组件是 Jakarta Apache 的开源项目, 由资深全文索引/检索专家 Doug Cutting 贡献, 主要解决各种中小型企业应用程序加入全文检索功能。譬如输入“父亲”和“母亲”两个词语, 就可得到以这两个词为关键词的文章, 实验结果如图 1 所示:

```
正在检索关键字: 父亲
检索完成, 用时12毫秒
这是第0个检索到的结果, 文件名为: C:\file\儿子欠下35万赌债逼死父亲 合肥一老人宾馆自杀致脑死亡.txt
这是第1个检索到的结果, 文件名为: C:\file\当男人成为父亲.txt
这是第2个检索到的结果, 文件名为: C:\file\父亲—男人最温情的名字.txt
这是第3个检索到的结果, 文件名为: C:\file\父亲请做宝宝最好的玩伴.txt
这是第4个检索到的结果, 文件名为: C:\file\研究生将患脑梗父亲带身边照顾.txt
这是第5个检索到的结果, 文件名为: C:\file\面对孩子, 男人如何做父亲.txt
这是第6个检索到的结果, 文件名为: C:\file\七兄妹请来剧团 六天大戏为父亲祝寿.txt
这是第7个检索到的结果, 文件名为: C:\file\布吕尼产女 萨科奇成法首位在任时荣升父亲总统.txt
这是第8个检索到的结果, 文件名为: C:\file\最新研究发现: 当爸爸死于心脏病的可能性较低.txt

-----
正在检索关键字: 母亲
检索完成, 用时1毫秒
这是第0个检索到的结果, 文件名为: C:\file\当男人成为父亲.txt
这是第1个检索到的结果, 文件名为: C:\file\父亲—男人最温情的名字.txt
```

图 1 实验结果

于是利用公式 (4)、(5) 和 (6) 可得父亲和母亲的客观相似度值为 0.222。

改进的词语语义相似度计算方法完整的实验测试步骤如图 2 所示。

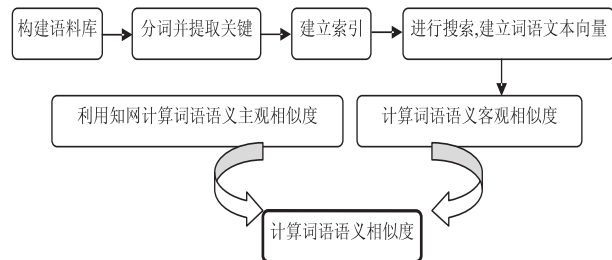


图 2 改进的词语语义相似度计算

2.2 实验结果与分析

在实验中, 公式 (1) 中的参数 α 和 β 都设为 0.5, 公式 (3) 中的参数 $\beta_1, \beta_2, \beta_3, \beta_4$ 分别设为 0.5、0.2、0.17、0.13。分别运用文中所提出的主客观相结合的词语语义相似度方法, 文献 [3] 提出的只运用主观相似度计算方法 (方法 1) 和仅使用《知网》语义表达式中第一独立义原计算词语相似度的方法 (方法 3), 以及文中

提出的改进的客观相似度计算方法(方法2)对一些词语计算词语语义相似度,实验结果如表2所示。

表2 实验结果对比

词1	词2	文中方法	方法1	方法2	方法3
男人	女人	0.6197	0.668	0.5714	1
男人	父亲	0.58	1	0.16	1
男人	苹果	0.0474	0.004	0.0909	0.285
男人	经理	0.1993	0.351	0.0476	1
男人	工作	0.0925	0.035	0.15	0.186
男人	责任	0.05025	0.005	0.1	0.016
父亲	母亲	0.445	0.668	0.222	1
中国	美国	0.4256	0.744	0.1071	1
中国	日本	0.56	1	0.12	1

对比4种方法的实验结果,可以发现:方法3因把第一义原相似度看作词语语义相似度,使得计算结果过于粗糙,譬如语义并不完全相同的“男人”和“女人”的词语语义相似度却为1;方法2只依赖语料库,使得词语语义相似度计算结果不是很稳定,语料库的质量和关键词提取算法以及提取关键词的维数对最后的结果有较大的影响,造成有些词语语义相似度计算结果较为准确,而有些偏差就较大;对于方法1,由于人为的主观性会产生较大影响,使得词语相似度结果脱离了语境,譬如“中国”和“日本”两个词相似度计算结果为1,而这两个词虽然都代表亚洲的在文化等各方面都较为相近的国家,但是在某些方面语义显然有较大的差别,比如说“日本发动卢沟桥事变”,如果将“日本”替换为“中国”的话就是“中国发动卢沟桥事变”,虽然句子的结构没有发生变化可是语义却发生了很大的改变。通过实验可以看到,由于文中所提方法结合了主观相似度和客观相似度,使得词语语义相似度结果更加合理,譬如“中国”与“日本”这两个词就得到了很好的区别。

3 结束语

文中在其它学者的研究基础之上,提出了一种改进的主观和客观相结合的词语语义相似度计算方法,基于词语共有信息量和词语覆盖信息量计算词语语义客观相似度,将理性主义和经验主义相结合,使词语语义相似度计算结果的准确性有了显著的改进。

对于现阶段海量的数据信息,词语相似度计算及其相关研究是一个较为热门的研究方向。陈远翔等在文献[9]中将词语相似度应用于文本分类,刘青磊和顾小丰在文献[15]中对句子相似度计算进行了研究和总结并且对于词语消歧给出了实例和计算方法,文

献[16]介绍了语义相似度测量在语义网方面的应用,文献[17]研究了基于语义词典和语料库的文本相似度计算。文中所提方法均可结合以上研究领域特点进行应用,下一步将基于词语相似度对句子相似度、文本相似度、语义网络等展开研究。

参考文献:

[1] 秦春秀,赵捧未,刘怀亮. 词语相似度计算[J]. 信息系统, 2007,30(1):105-108.

[2] 杨哲. 基于启发式规则的本体概念语义相似度匹配[J]. 计算机应用,2007,27(12):2919-2921.

[3] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理, 2002(7):59-76.

[4] 皮慧娟. 基于马尔科夫模型的词汇语义相似度计算[J]. 沈阳大学学报,2010,22(1):5-8.

[5] 张晓李,张蕾,王西锋. 基于知识图的汉语词语间语义相似度计算[J]. 计算机工程与应用,2007,43(8):160-163.

[6] 赵应秋,罗军,张君艳. 基于知网的词语语义相关度计算[J]. 信息技术,2010(3):90-93.

[7] 曹立勇,郑诚. 基于知网的语义相似度的改进算法[J]. 电子技术研发,2010,47(5):1-3.

[8] 唐歆瑜,乐文忠,李志成,等. 基于知网语义相似度计算的 特征降维方法研究[J]. 科学技术与工程,2006,21(6): 3442-3445.

[9] 陈远翔,张月国,李生红,等. 基于知网语义相似度计算的 文本特征提取[J]. 信息安全与通信保密,2009(5):89- 91.

[10] 林丽,薛方,任仲晟. 一种改进的基于《知网》的词语相 似度计算方法[J]. 计算机应用,2009,29(1):217-220.

[11] 江敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的 词语语义相似度计算[J]. 中文信息学报,2008,22(5):84- 89.

[12] 赵军,胡栓柱,樊兴华. 一种新的词语相似度计算方法 [J]. 重庆邮电大学学报,2009,21(4):528-532.

[13] Richardson R, Smeaton A, Murphy J. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words[R]. Dublin, Ireland: Dublin City University, 1994.

[14] 胡俊峰,俞士汶. 唐诗宋词中词汇语义相似度的统计分 析及应用[J]. 中文信息学报,2002(4):40-45.

[15] 刘青磊,顾小丰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报,2010,24(6):31-36.

[16] Hau J, Lee W, Darlington J. A Semantic Similarity Measure for Sematic Web Services. www2005[C]. Chiba, Japan: [s. n.], 2005.

[17] Mihalcea R, Corley C, Strapparava C. Corpus - based and Knowledge-based Measures of Text Sematic Similarity[EB/ OL]. 1995. <http://www.aaai.org>.

改进的主客观结合的词语语义相似度算法

作者:	吴旭东 , 成卫青 , 黄卫东
作者单位:	吴旭东, 成卫青(南京邮电大学 计算机学院, 江苏 南京 210003) , 黄卫东(南京邮电大学 经济与管理学院, 江苏 南京 210003)
刊名:	计算机技术与发展
英文刊名:	Computer Technology and Development
年, 卷(期):	2012(9)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201209014.aspx