

使用网络搜索引擎计算汉语词汇的语义相似度

高国强,黄吕威,陈丰钰

(武汉纺织大学 传媒学院,湖北 武汉 430073)

摘要:汉字词语的语义相似度计算是中文信息处理中的一个关键问题。文中利用网络搜索引擎提供的信息来计算汉语词对的语义相似性。首先通过程序访问搜索引擎,获取汉字词汇的搜索结果数,并依此实现了相似度计算模型 WebPMI;然后描述了根据查询返回的文本片段进行语义相关性分析的模型 CODC;最后,结合这两个模型,给出了文中算法的伪代码。实验结果显示,文中的算法较好地利用了互联网信息,实现了一种较新的汉语词汇语义相似度计算方法,接近于利用词典提供的信息计算相似度的传统算法。

关键词:相似度;搜索引擎;词典

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2014)07-0084-04

doi:10.3969/j.issn.1673-629X.2014.07.021

Calculation of Chinese Words Semantic Similarity Using Network Search Engines

GAO Guo-qiang, HUANG Lü-wei, CHEN Feng-yu

(School of Media and Communication, Wuhan Textile University, Wuhan 430073, China)

Abstract: Similarity computation of Chinese words is a key problem in Chinese information processing. It measures semantic similarity between Chinese words using the information returned by web search engines. First, implement a model named WebPMI which computes similarity using page counts, and then, describe another model named CODC which analyzes semantic similarity using text snippets. Finally, present the algorithm based on the two models. Experimental results show that this algorithm outperforms all the existing web-based semantic similarity measures for Chinese, and is close to the traditional semantic similarity measures using lexicon.

Key words: similarity; search engines; lexicon

0 引言

词汇间语义相似度的研究一直是信息检索和自然语言处理的核心部分,对于汉语来说尤其如此。词汇之间的语义相似度在时间和领域范围内是经常变动的。比如说,在互联网上“苹果”经常是“苹果电脑”的意思,然而在大部分词典中苹果是没有这种意思的。一个用户在互联网上搜索“苹果”可能就是找苹果电脑,而不是找一种水果。目前的汉语词汇语义相似度算法^[1-3]大部分使用词典来计算相似度,然而新的词汇是不断增加的,而且领域不同词汇间的相似度也有很大区别。因此手工维护词典来保证完整性是非常困难的,即使可行,代价也是高昂的。

文中结合已有的两种相似度计算机制^[4-5],提出

了使用网络搜索引擎来计算汉语词汇之间的相似度。虽然精度不如基于词典的算法,但是可以避免复杂的词法分析,也避免了维护不断出现的新词汇的代价。对于精度要求不是很高的应用该算法具有一定价值,而且随着互联网的不断扩展,这种算法的精度将不断提高。网络搜索引擎对互联网上的海量信息提供了一个有效的访问接口,比如说其提供的搜索结果数和文本片段就是很有价值的信息。一个查询的搜索结果数是包含这个查询的页面个数。查询“ P AND Q ”的搜索结果数可以被用来作为考虑词 P 和 Q 存在相关性的全局度量标准,比如查询“老师 AND 教授”在百度中的搜索结果数是 88 000 000,然而“工人 AND 教授”的仅有 7 350 000。查询“老师 AND 教授”的搜索结果数是

收稿日期:2013-09-28

修回日期:2013-12-30

网络出版时间:2014-04-24

基金项目:湖北省自然科学基金(2013CFB310);湖北教育科研项目(B2013205);湖北省高等学校 2013 年省级大学生创新创业训练计划项目(2013CXZD027);2013 年武汉纺织大学大学生创新创业训练计划项目(2013CXXL008,2013CXXL009)

作者简介:高国强(1971-),男,湖南常德人,讲师,CCF 高级会员,研究方向为数据挖掘、信息检索。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20140424.0830.074.html>

查询“工人 AND 教授”的12倍,这说明前者比后者语义上更加相关。

尽管使用搜索结果数可以指示词汇间存在语义相关性,但是仅仅使用结果数来测量语义相似度存在一些缺点。比如说可能存在大量不相关词对同时出现的网页,这将导致该方法变得不可信。搜索引擎返回的片段也是一种有效度量信息,比如查询“教授”的前100个最高排序结果的片段中,“老师”这个词出现的次数比“工人”多,这说明了“教授”和“老师”更加相关。某个词对同时出现的网页中可能有很多页面是它们不相关的网页,使用这种方式可以抵消不相关网页对统计结果的干扰。比如含片段“我的计算机坏了,而且又忘了买苹果”的网页,这个网页就不能表明“计算机”和“苹果”之间有相关性,这两个词汇只是偶然出现在一个网页中的,如果把这种网页作为度量参数的话就会对统计结果产生偏差。搜索引擎返回的前100个最高排序结果的网页与查询是最相关的,比如查询“计算机”返回的前100个页面应该都是和“计算机”相关的信息,前面例子的那个页面一般不会出现在前100个中。由此可见前100个页面的文本片段是非常有价值的信息,比如查询“计算机”的前100个页面片段中含有大量“苹果”词汇,说明两者之间的确有相关性。当然片段度量标准和搜索引擎效率以及网络信息容量有很大关系,这是文中提出算法的基础。就目前而言,网络的容量已经达到了一个可以用来作为度量标准的级别,而且网络是在不断发展的;搜索引擎技术也已经比较成熟,对于中文搜索,“百度”是可以信赖的工具。

1 相关工作

语义相似度测量在许多与网络相关的应用中扮演着非常重要的角色,比如在查询扩展^[6]中,一个用户的查询通过同义词被修正以提高搜索的准确性。在查询中发现合适替代词的一个方法是通过语义相似度计算来比较用户以前的查询,如果以前的查询与目前的查询语义相关,就建议用户或搜索引擎修正查询语句。文献[7]利用相似度对网络连接数据的属性特征进行选择,抽取其关键特征,并降低属性的冗余度,以优化朴素贝叶斯的分类性能。实验结果表明,他们的方法能降低分类数据的维数,提高分类的准确率。现有的相似度算法^[8-9]大部分集中在利用词典提供的信息进行度量,给定一个词典,一个计算两个词之间相似度的直接方法是发现连接词典中两个词的最短路径长度。对于汉语词汇来说,尤其如此,不过也更为复杂。

文献[10]利用HowNet作为词典库,在考虑义原距离和义原深度的条件下,进行词语相似度计算,可以

获得与人们的主观判断更接近的结果。刘群等提出的基于《知网》的词汇语义相似度计算^[11]也是利用词典信息进行计算的例子,为了计算用知识描述语言表达的两个概念的语义表达式之间的相似度,他们采用了“整体的相似度等于部分相似度加权平均”的做法。首先将一个整体分解成部分,再将两个整体的各个部分进行组合配对,通过计算每个组合对的相似度的加权平均得到整体的相似度。夏天^[12]在汉语词语义相似度计算研究中提出的算法也是利用了《知网》提供的信息。他提出了一种基于知网、面向语义、可扩展的相似度计算新方法,该方法从信息论的角度出发,定义了知网义原间的相似度计算公式,通过对未登录词进行概念切分和语义自动生成,解决了未登录词无法参与语义计算的难题,实现了任意词语在语义层面上的相似度计算。针对同义词词林的实验结果表明,该方法的准确率比现有方法高出近15个百分点。虽然文献[13]分析多种相似度计算方法,但也都是基于词典的算法。

最近,一些研究正关注于利用网络信息来测量语义相似度。Danushka等^[4]提出了利用网络搜索引擎测量词对间的相似度。对于两个词 P 和 Q ,根据查询 P 、 Q 、 P AND Q 的搜索结果数按四种不同计算公式计算归一相似值,并作为一个四维向量保存。然后获取查询 P AND Q 的排序前100个搜索结果的片段,对这些片段进行语义分析,提取词 P 和 Q 之间的语义模式,比如 P is Q 、 P and Q are等等,并对这些模式按出现频率排序。最后对片段中出现的模式选择前200个形成一个向量并与前面的搜索结果数向量合并为一个向量 F ,再利用SVM(支持向量机)对向量 F 进行处理得出词对间的归一相似度。这种方法有较高的相关性,接近于利用词典的算法,但是模式提取算法非常复杂,而且要利用大量同义和反义词对进行参考模式库生成。同时,生成参考模式库时,参考词对数规模和选取词对的标准都是启发式的,这是该算法实用性的一个问题。对于汉语词汇尤其如此,首先是提取模式复杂,因为汉语间没有空格,要提取就要使用分词处理,这将导致准确率下降。再者就是如何选择汉语词对及规模生成参考模式库,这都是非常复杂的问题,也导致了如果利用此种方式处理汉语词汇将使算法不实用。陈信希等^[5]提出了一种利用网络搜索计算词汇相关性的相关性双重检测算法(CODC)。对词 P 和 Q ,首先使用搜索引擎分别获得查询 P 、 Q 的搜索结果数和前100个搜索结果的返回片段,然后对 P 的片段统计出现词 Q 的次数,对 Q 的片段统计出现词 P 的次数,然后将这些数据输入一个设计好的模型得出归一的相似度。CODC算法在利用网络信息进行相似度计算的算法中有较高

的相关性,但它的缺点是,很多词对不能统计相似度。尤其对语义相差较大或者网络信息覆盖较弱的词对,对这些情况 CODC 算法会出错而给出一个为零的相似度,这是因为 P 的片段中可能根本就不会出现词 Q 。对于 CODC 算法的简洁性,文中的算法进行了采用,但对于其不足的地方文中进行了改进。

2 相似度计算方法

文中提出的算法,对于给定的一组汉语词对,利用了网络搜索引擎提供的搜索结果数和前 100 个搜索结果片段这两种信息,并分别处理后集成在一起形成归一的相似度。在 2.1 节,描述了一种使用搜索结果数计算相似度分值的模型。在 2.2 节,文中首先分别获取查询结果片段中另外一个词出现的次数,然后利用了 CODC 算法^[5]提供的模型计算出归一的相似度。在 2.3 节,给出了文中的相似度算法伪代码,该算法首先利用 2.1 节的方法,然后利用 2.2 节的方法,并在综合两种方法的基础上给出一个最终的词对相似度。

2.1 基于搜索结果数的相似度模型

查询“ P AND Q ”的搜索结果数能够作为两个词 P 和 Q 的近似相关度,但是仅仅考虑查询“ P AND Q ”的搜索结果数来计算语义相似度是不准确的。例如,在百度中搜索“大学 AND 讲师”返回搜索结果数是 1 170 000 条,而搜索“大学 AND 苹果”却有 5 720 000 条。虽然相比于“苹果”,“讲师”这个词在语义上与“大学”更相近,但是查询“大学 AND 苹果”在百度中的搜索结果数却远远大于查询“大学 AND 讲师”的结果数。所以在进行相似度计算时,不仅要考虑查询“ P AND Q ”的搜索结果数,而且还要考虑查询 P 、 Q 的搜索结果数。

文中扩展了现有的比较流行的相关性计算方法:逐点共有信息 (PMI) 算法,通过使用搜索结果数修改 PMI 为 WebPMI 算法来计算语义相似度。这里定义 $H(P)$ 表示使用搜索引擎查询 P 的搜索结果数, $P \cap Q$ 表示联合查询“ P AND Q ”。WebPMI 算法关于词 P 和 Q 的相似度为 WebPMI(P, Q),定义在式(1)中。

$$\text{WebPMI}(P, Q) = \begin{cases} 0 & H(P \cap Q) \leq c \\ \log_2 \left(\frac{\frac{H(P \cap Q)}{N}}{\frac{H(P)}{N} \frac{H(Q)}{N}} \right) & \text{otherwise} \end{cases} \quad (1)$$

式中, N 是搜索引擎已索引文档数。对于 N 值的给定直接影响结果的准确性,尽管估计一个搜索引擎索引文档数量是一个有趣的任务,但这超过了文中的范围。在式(1)中文中根据百度的报告给定 $N = 2 \times$

10^{10} 。 c 是一个门限值,如果查询 $P \cap Q$ 的搜索结果数小于门限值,就不使用计算模型,直接给一个结果为 0 的相似度。这是因为两个词可能纯粹偶然地出现在一个页面上,为了避免这种干扰,文中实验给定 $c = 10$ 。

2.2 使用片段信息计算相似度

文本片段是和搜索结果数一起由搜索引擎返回的,片段提供了一个词和另外的词相关的信息。比如查询“老师”的返回片段为“【讲师简介】林老师 著名实战派 全面预算管理与成本控制实务操作专家”,这个片段中有很多词汇,比如“讲师”、“著名”、“专家”等等。可以说“讲师”可能和“老师”有一定关联,如果考虑查询“老师”的大量片段,而且“讲师”在这些片段中不断出现,就可以把这种统计结果作为相关的一种度量。为了利用片段提供的这种信息,文中根据 CODC 算法定义词 P 和 Q 的语义相似度为 CODC(P, Q),定义在式(2)中。

$$\text{CODC}(P, Q) = \begin{cases} 0 & f(P@Q) = 0 \\ \log \left[\frac{f(P@Q)}{H(P)} \times \frac{f(Q@P)}{H(Q)} \right]^a & \text{otherwise} \end{cases} \quad (2)$$

式中, $f(P@Q)$ 表示百度中搜索查询 Q 的前 100 个返回片段中出现词 P 的次数。 a 是式(2)的一个常量,根据经验在实验中设置为 0.15。如果 $f(P@Q)$ 或者 $f(Q@P)$ 等于 0,那么 CODC(P, Q) 将等于 0,这是一种极端情况,也表明词 P 和 Q 之间没有关联。如果 $f(P@Q)$ 等于 $H(Q)$ 且 $f(Q@P)$ 等于 $H(P)$, CODC(P, Q) 将等于 1,表明词 P 和 Q 之间有非常强的关联,最极端的情况就是同一个词,此时 $f(P@P)$ 肯定等于 $H(P)$,从而就有 CODC(P, P) 等于 1。虽然从理论上分析,式(2)能计算所有词对的相似度,但是实验的结果却显示,利用网络信息作为数据源时,式(2)对于语义关联不强的词对结果几乎都为零。这可能与网络信息的容量以及搜索引擎的效率有关,比如词对“工人”和“技术”,因为查询“工人”的前 100 个结果的片段中没有出现“技术”,所以 CODC(工人,技术)等于零。直观上看这样的度量是不准确的,为了修正这种偏差,文中的算法结合了 2.1 节和 2.2 节两种方法,将在 2.3 中给出描述。

2.3 计算相似度算法

对于语义相关性比较强的词对, CODC 模型的相关度要比 WebPMI 模型的高,但是对于语义相关性较弱的词对, CODC 模型不能度量。文中的算法采用两种模型,对于给定的词对 (P, Q),首先判断 $f(P@Q)$ 或 $f(Q@P)$ 是否等于零,如果等于零算法使用 WebPMI 模型计算相似度,否则使用 CODC 模型计算相似度。相似度算法伪代码描述如下。


```
算法: GetSim(  $P, Q$  )。  
 $D \leftarrow \text{GetSnippets}( P )$   
if  $Q$  not in  $D$  then  
  Sim = WebPMI(  $P, Q$  )  
else  
  Sim = CODC(  $P, Q$  )  
end if  
return( Sim )
```

其中函数 $\text{GetSnippets}(P)$ 是从搜索引擎获得查询 P 的前 100 个搜索结果的片段,并存放在数据集 D 中。因为 $f(P@Q)$ 或 $f(Q@P)$ 等于零,都有 $\text{CODC}(P, Q) = 0$,所以算法中可以随便取一个查询的片段进行分析,这里取的是查询 P 的片段。Sim 是相似度变量,根据片段分析结果采用不同的计算模型,最后返回相似度 Sim。

3 实 验

词语相似度计算的结果评价,最好是放到实际的系统中(比如基于实例的机器翻译系统),观察不同的相似度计算方法对实际系统的性能的影响。但这需要一个完整的应用程序,在条件不具备的情况下,文中采用了模拟几个不同算法,然后进行比较的方法。为了获取搜索引擎的结果,文中利用 perl 脚本调用文本浏览器 w3m 来访问百度,然后通过程序分析返回结果,从中获取搜索结果数以及片段。对于片段内容,开发了一组程序进行内容挖掘。

文中另外选取了三个相似度算法,一个是刘群等提出的算法,被称之为《知网》算法,这是基于词典模拟的算法;另一个是 Sahami^[14]等提出的利用搜索引擎计算相似度的算法,被称之为 Sahami 算法;还有一个也是基于搜索引擎的算法,叫做 Web Overlap,它仅仅使用了搜索结果数。在对比实验中文中的算法被称为 Proposed Algorithm。根据词义相关性的强弱分别选择了一些词对进行测试,实验结果如表 1 所示。

由表 1 可以看出《知网》算法效果最好,相关度达到了 0.834,接近于人工识别的程度。在非词典的算法中,文中提出的算法效果最好,达到了一个 0.713 的相关度。虽然与人工识别的 0.9 相关度还有一定距离,但是和利用词典的算法差距已经不是太大。文中的算法简单实用,不需要维护词典,而且随着互联网的发展,其精度将越来越高。

4 结束语

文中提出了一种利用互联网信息计算汉语词汇语义相似度的算法,该算法的特点是不需要维护词典或者其他的数据库,算法简单实用。使用网络搜索引擎

表 1 语义相似度实验结果

词 对	Web	Sahami	《知网》	Proposed
	Overlap	Algorithm	算法	Algorithm
束缚 - 微笑	0.036	0.090	0.000	0.207
公鸡 - 航海	0.021	0.197	0.017	0.228
中午 - 绳索	0.060	0.082	0.018	0.101
玻璃 - 魔术	0.408	0.143	0.180	0.598
和尚 - 奴隶	0.067	0.095	0.375	0.000
海岸 - 森林	0.320	0.248	0.405	0.417
和尚 - 圣贤	0.023	0.045	0.328	0.610
少年 - 奇才	0.070	0.149	0.220	0.426
森林 - 墓地	0.246	0.000	0.547	0.494
食物 - 公鸡	0.425	0.075	0.060	0.207
海岸 - 悬崖	0.279	0.293	0.874	0.350
轿车 - 旅行	0.378	0.189	0.286	0.290
起重机 - 工具	0.119	0.152	0.133	0.193
弟弟 - 少年	0.369	0.236	0.344	0.379
鸟 - 起重机	0.226	0.223	0.879	0.515
鸟 - 公 鸡	0.162	0.058	0.593	0.502
食物 - 水果	1.000	0.181	0.998	0.338
兄弟 - 和尚	0.340	0.267	0.377	0.547
庇护 - 医院	0.102	0.212	0.773	0.813
家具 - 炉子	0.118	0.310	0.889	0.928
魔术师 - 神汉	0.383	0.233	1.000	0.671
旅行 - 航行	0.182	0.524	0.996	0.417
海岸 - 海滨	0.521	0.381	0.945	0.518
工具 - 设备	0.517	0.419	0.684	0.419
男孩 - 小伙子	0.601	0.471	0.974	0.631
汽车 - 轿车	0.834	1.000	0.980	0.686
午夜 - 半夜	0.135	0.289	0.819	0.856
宝石 - 珠宝	0.094	0.211	0.686	1.000
相关度	0.382	0.579	0.834	0.713

作为访问互联网信息的接口是一种新颖的尝试,通过文中的算法可以看出这种尝试是成功的。文中利用搜索引擎返回的搜索结果数和文本片段,给出了两个相似度计算模型。实验结果表明文中的算法接近于利用词典的算法,具有一定实用价值。未来进一步要做的是如果提高效率,毕竟对于有的应用文中的算法精度是不够的,以后可能的重点在分析文本片段上面。

参考文献:

[1] 孙昌年,郑 诚,夏青松. 基于 LDA 的中文文本相似度计算[J]. 计算机技术与发展,2013,23(1):217-220.
[2] 杨方颖,蒋正翔,张姗姗. 基于本体结构的语义相似度计算[J]. 计算机技术与发展,2013,23(7):52-56.
[3] 王 桐,王 磊,吴吉义,等. WordNet 中的综合概念语义相似度计算方法[J]. 北京邮电大学学报,2013,36(2):98-101.
[4] Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using web search engines[C]//Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada: [s. n.],2007:757-766.

and regard it as the characteristic vectors of support vector machine. It overcomes the lack of clustering algorithm on a variety of characteristics through the method of integration of the two kinds of image features. It is an attempt of the image segmentation of integration of a variety of characteristics and methods. This article discusses the segmentation algorithm only for grayscale ima-

ges. The characteristic referenced finite (an image contains color, grayscale, texture and other features). Extracted from the image to the appropriate segmentation features, as well as a combination of important features as support vector machine characteristic components for further research.

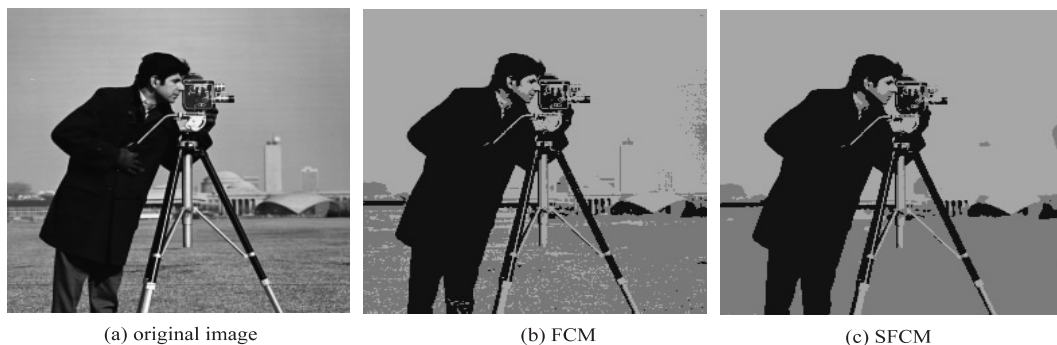



Figure 1 Segmentation results on standard image (cameraman)

参考文献:

- [1] 章毓晋. 图像分割[M]. 北京:科学出版社,2001.
- [2] 李弼程,柳葆芳. 基于二维直方图的模糊门限分割方法[J]. 数据采集与处理,2000,15(3):324-329.
- [3] Duda R O, Hart P E. Pattern classification and scene analysis [M]. New York: John Wiley & Sons, 1973.
- [4] Gomez-Moreno H, Gll-Jmenez P, Lafuente-Arroyo S, et al. Color images segmentation using the support vector machines [C]//Proc of recent advances in intelligent systems and signal processing. USA: WSES Preas, 2003:151-155.
- [5] 潘 晨, 闫相国, 郑崇勋, 等. 利用单类支持向量机分割血细胞图像[J]. 西安交通大学学报, 2005, 39(2): 150-153.
- [6] 陈 强, 周则明, 屈颖歌, 等. 左心室核磁共振图像的自动分割[J]. 计算机学报, 2005, 28(6): 991-999.
- [7] 张国宣, 孔 锐, 施泽生, 等. 一种新的结合纹理特征的 SVM 图像分割方法[J]. 中国图象图形学报, 2003, 8(z1): 441-444.
- [8] Rajpoot K M, Rajpoot N M. Wavelets and support vector machine for texture classification [C]//Proceedings of 8th IEEE international multitopic conference. [s. l.]: IEEE, 2004: 328-333.
- [9] 张道强. 基于核的联想记忆及聚类算法的研究与应用 [D]. 南京: 南京航空航天大学, 2004.
- [10] Krinidi M, Pitas I. Color texture segmentation based on the modal energy of deformable surfaces[J]. IEEE Trans on Image Processing, 2009, 18(7): 1613-1622.
- [11] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.
- [9] 冉 婕, 孙 瑜. 语义检索中的词语相似度计算研究[J]. 计算机技术与发展, 2011, 21(4): 94-97.
- [10] 廖志芳, 邱丽霞, 谢岳山, 等. 一种频率增强的语句语义相似度计算[J]. 湖南大学学报(自然科学版), 2013, 40(2): 82-88.
- [11] 刘 群, 李素建. 基于《知网》的词汇语义相似度计算 [C]//第3届中文词汇语义学研讨会. 出版地不详: 出版者不详, 2002.
- [12] 夏 天. 汉语词语语义相似度计算研究[J]. 计算机工程, 2007, 33(6): 191-194.
- [13] 魏凯斌, 冉延平, 余 牛. 语义相似度的计算方法研究与分析[J]. 计算机技术与发展, 2010, 20(7): 102-105.
- [14] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets [C]//Proceedings of the 15th international conference on World Wide Web. New York, NY, USA: ACM, 2006: 377-386.

(上接第87页)

- [5] Chen H H, Lin M S, Wei Y C. Novel association measures using web search with double checking [C]//Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. Sydney, Australia: [s. n.], 2006: 1009-1016.
- [6] Mitra M, Singhal A, Buckley C. Improving automatic query expansion [C]//Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. Melbourne, Australia: [s. n.], 1998: 206-214.
- [7] 王春东, 陈英辉, 常 青, 等. 基于特征相似度的贝叶斯网络入侵检测方法[J]. 计算机工程, 2011, 37(21): 102-104.
- [8] 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(1): 60-64.

作者: 高国强, 黄吕威, 陈丰钰, GAO Guo-qiang, HUANG Lü-wei, CHEN Feng-yu
作者单位: 武汉纺织大学 传媒学院, 湖北 武汉, 430073
刊名: 计算机技术与发展 
英文刊名: Computer Technology and Development
年, 卷(期): 2014(7)

参考文献(14条)

1. 孙昌年;郑诚;夏青松 基于LDA的中文文本相似度计算 2013(01)
2. 杨方颖;蒋正翔;张姗姗 基于本体结构的语义相似度计算 2013(07)
3. 王桐;王磊;吴吉义 WordNet中的综合概念语义相似度计算方法 2013(02)
4. Bollegala D;Matsuo Y;Ishizuka M Measuring semantic similarity between words using web search engines 2007
5. Chen H H;Lin M S;Wei Y C Novel association measures using web search with double checking 2006
6. Mitra M;Singhal A;Buckley C Improving automatic query expansion 1998
7. 王春东;陈英辉;常青 基于特征相似度的贝叶斯网络入侵检测方法 2011(21)
8. 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 2003(01)
9. 冉婕;孙瑜 语义检索中的词语相似度计算研究 2011(04)
10. 廖志芳;邱丽霞;谢岳山 一种频率增强的语句语义相似度计算 2013(02)
11. 刘群;李素建 基于《知网》的词汇语义相似度计算 2002
12. 夏天 汉语词语语义相似度计算研究 2007(06)
13. 魏凯斌;冉延平;余牛 语义相似度的计算方法研究与分析 2010(07)
14. Sahami M;Heilman T D A web-based kernel function for measuring the similarity of short text snippets 2006

引用本文格式: 高国强, 黄吕威, 陈丰钰, GAO Guo-qiang, HUANG Lü-wei, CHEN Feng-yu 使用网络搜索引擎计算汉语词汇的语义相似度[期刊论文]-计算机技术与发展 2014(7)