基于百度百科的词语相似度计算

詹志建 梁丽娜 杨小平
(中国人民大学信息学院 北京 100872)

摘 要 词语相似度计算是自然语言处理的关键技术之一,是一个被广泛研究的基础课题。传统的词语相似度量方法大多是基于语义知识和基于语料库统计的方法,即这两类方法需要具有层次关系组织的语义词典和大规模的语料库。提出了一种新的基于百度百科的词语相似度量方法,通过分析百度百科词条信息,从表征词条的解释内容方面综合分析词条相似度,并定义了词条间的相似度计算公式,通过计算部分之间的相似度得到整体的相似度。实验结果表明,与已有的相似度计算方法对比,提出的算法更加有效合理。

关键词 词语相似度,语言网络,百度百科,向量空间模型
中图法分类号 TP311 **文献标识码** A

Word Similarity Measurement Based on BaiduBaike

ZHAN Zhi-jian LIANG Li-na YANG Xiao-ping
(School of Information, Renmin University of China, Beijing 100872, China)

Abstract Research on word similarity measurement has been popular not only in natural language processing but also in other basic research. Traditional word similarity measurements use semantic lexical or large-scale corpus. We first discussed the background of the applications of word similarity measurement, such as information retrieval, information extraction, text classification, example-based machine translation, etc. Then two strategies of word similarity measurement were summarized: one is based on ontology or a semantic taxonomy, the other is based on large collocations of words in corpus. BaiduBaike, an online open encyclopedia, could be used not only as a corpus but also a knowledge resource with rich semantic information. Based on BaiduBaike with its rich semantic information and category graph, we proposed a new method to analyze and compute Chinese word similarity from four dimensions: the baike card, the content of word, the open classification of word and the correlation words. We used language-network to choose top key terms of content of word. Based on vector space mode (VSM) theory, we calculated the similarity between parts of words. We presented a new "multi-path searching" algorithm on BaiduBaike category graph. A comprehensive similarity measuring method based on the four parts was proposed. Experiment results show that the method has a good performance.

Keywords Word similarity, Language network, BaiduBaike, VSM

1 引言

词语相似度计算研究的是词语之间的相似度量方法,是一个在语言学、心理学和信息理论等领域被广泛研究的基础课题[1]。在不同的应用中,词语相似度计算有着不同的用途:在信息检索中,词语相似度计算主要反映用户查询与文档结果集信息匹配的符合程度,被认为是改进检索效果的最好方法之一;在基于实例的机器翻译中,词语相似度计算主要用于衡量文本中词语的可替换程度;在自动问答系统中,词语相似度计算主要用于计算用户问句和领域文本内容的相似度。此外,词语相似度计算还广泛地应用于文本分类[14]、文本摘要的自动生成、词义排歧等。

百度百科是一个基于WEB2.0技术的中文百科全书,现已成为互联网上规模最大、使用最广泛的开放式中文电子百科全书,也成为由互联网用户以自由贡献、共同协作的方式构建大规模知识资源的典范。作为语料库,百度百科包含了数百万的文档资源,质量上和数量上都有着其它语料库无法比拟的优势。

本文介绍了一种新的基于百度百科的词语相似度量方法。我们的基本设想是:词语相似度要建立在表征词语的各部分相似的基础上,通过计算各部分之间的相似度得到整体的相似度。给定两个词语,通过本文的算法能够高效、自动地计算出两个词语在语义层次上的相似度,并且能够在较为广泛的应用领域内使用。

2 相关工作

国内外对词语相似度的研究方法大体上可以分为两类:基于语义知识的方法和基于大规模语料库统计的方法。基于

到稿日期:2012-08-06 返修日期:2012-11-16 本文受国家自然科学基金(70871115)资助。
詹志建(1982-),男,博士,主要研究方向为信息处理、语义计算、Web数据管理,E-mail:zhanzj@ruc.edu.cn;梁丽娜(1988-),女,硕士,主要研究方向为信息处理、语义计算、本体工程;杨小平(1956-),男,博士,教授,主要研究方向为信息系统工程、电子政务、网络安全。

语义知识的方法是利用具有层次关系组织的语义词典，依据词语之间的上下位关系和同义关系，通过计算词语在树状层次体系中的距离得到词语间的相似度。基于大规模语料库统计的方法将词语的上下文信息作为语义相似度计算的参照依据。词语向量空间模型[2]是目前基于统计的词语相似度计算策略使用比较广泛的一种。该模型事先选择一组特征词，然后计算这组特征词与每一个词的相关性，于是对于每一个词都可以得到一个相关性的特征词向量，最后将这些向量之间的相似度作为词语间的相似度。

典型的语义词典有英文的 WordNet、FrameNet、MindNet 等，中文的《知网》、《同义词词林》等。R. Rada 和 J. H. Lee 通过计算在 WordNet 中词节点之间的上下位关系所构成的最短路径来计算英文词语之间的相似度[3,4]。Agirre 和 Rigau 除了考虑 WordNet 中词语节点间的路径长度外，还考虑了层次树的深度和区域密度等[5]。Sussna M. 在考察 WordNet 词义网密度、节点深度、链接类型等因素后，提出了一种基于词义网边的词语之间的相似度量方法[6]。

在汉语词语相似度计算研究方面，刘群等人在分析《知网》结构基础上，提出了一种基于《知网》的词语语义相似度量方法[7]。该方法利用义原的上下位关系计算义原相似度，进而得到词语的相似度。王斌等人通过计算《同义词词林》树形图中节点之间的路径距离得到词语相似度[8]。李素建等人提出综合利用《知网》和《同义词词林》来计算词语间的相似度[9]。

典型的统计语料库有 COBUILD、ACL/DCI 语料库等。P. Brown 等人采用互信息来计算词语之间的相似度[10]。胡俊峰等人利用上下文的词汇向量空间模型来近似地描述词汇的语义，在此基础上定义词汇的相似关系[11]。

基于语义词典的方法比较直观且简单有效，但受人的主观影响较大，有时不能反映客观现实。此外，对于大多数语义词典来说，无法做到收录现实应用中所有的词语，必然会有部分词语不在词典中，导致无法计算相似度。基于语料库统计的方法比较依赖于语料库的优劣，存在数据稀疏的问题，有时会出现明显的错误。

基于上述对词语相似度计算方法的分析，本文提出了一种基于百度百科的词语相似度量方法。本文的贡献有以下几个方面：首先通过建立语言网络选择表征词语的文本的重要词项，有效降低了文本模型的维度，为文本的相似度计算提供了一个合适的表征模型；其次通过分析百度百科的分类图，综合多条关联路径，计算词语间的相似度。最后通过比较几种常见的词语相似度量结果，验证了本文算法的有效性和合理性。

3 基于百度百科的相似度计算

3.1 百度百科词条介绍

百度百科是一部内容开放、自由的网络百科全书，其所含内容的基础分割单位是词条。一个词条一般由以下若干部分组成：百科名片、词条正文、正文图片与图册、地图、词条内链、参考资料、开放分类、相关词条、扩展阅读等。部分词条可能只包含词条名称、词条正文、开放分类和相关词条等。百度百科收录的内容包括具体事物、知名人物、抽象概念、文学著作、热点事件、汉语字词或特定主题的组合等。每个词条对应一个单一的主题。词条样例如表 1 所列。

表 1 百度百科词条样例

词条名称	航程
百科名片	无
词条正文	[voyage；passage；range；distance by air or sea] 指飞机的续航距离，船舶中途不补充燃料可以运行的最大距离 1. 船舶或飞机由起点到终点的距离。田野《火烧岛》：“火烧岛，距离台湾只有几小时的航程。” 2. 前进的路程。《诗刊》1977 年第 9 期：“高高举起铁拳头，永为革命指航程。”
参考资料	百科词典 http://dict.baidu.com/s？wd＝%BA%BD%B3%CC
扩展阅读	无
开放分类	词语，中文，词典，汉语词典，汉英词典
相关词条	航班 里程 航迹

对百科词条具有解释意义的是 4 部分内容：百科名片、词条正文、开放分类和相关词条。本文正是在两个词条的这 4 部分之间建立起一一对应的关系，在对应的部分之间进行计算，通过计算部分之间的相似度加权得到整体的相似度。

3.2 百科名片相似度计算

百科名片是一篇对词条的概括性描述的短文本，由 250 字以内的文字叙述和一幅插图组成。我们通过建立向量空间模型计算百科名片之间的相似度。首先对待计算相似度的词条 A 和 B 的百科名片分词，然后删除对应于停用词列表中的停用词，再统计词频，建立分词向量 $T_a(t_1,t_2,\cdots,t_m)$ 和 $T_b(t_1,t_2,\cdots,t_n)$，给每个分词赋以词频权重表征文本，即 $(w_1,w_2,\cdots,w_k)$，得到基于向量空间的百科名片相似度公式：

$$\mathrm{Sim}_1=\frac{\sum_{i=1}^{k}W_{Ai}\times W_{Bi}}{\sqrt{\sum_{i=1}^{k}W_{Ai}^2}\times\sqrt{\sum_{i=1}^{k}W_{Bi}^2}} \tag{1}$$

式中，W_{Ai} 和 W_{Bi} 分别是词项 t_i 在百科名片 A 和 B 中的权值，k 是向量的维度。

3.3 百科词条正文相似度计算

百科词条正文是对词条最完备的信息解释，内容一般比较大。计算两个词条正文之间的相似度，我们的处理方法是先对词条正文分词，再利用停用词列表删除停用词后建立语言网络，选取表征正文的 TOP 比例关键词，以关键词向量相似度作为词条正文的相似度，具体做法如下。

词条正文内容预处理完后，对词项建立语言网络[12]，以词项为节点，为每个句子中跨度为 1 或 2 的节点对建立连边，将各个句子所组成的网络连接起来，合并相同的节点和连边，形成该正文的语言网络。例如，对于句子“词语相似度计算过程”，通过分词产生“词语”、“相似度”、“计算”、“过程” 4 个词语，从而可以建立如图 1 所示的语言网络。对于整个词条正文的语言网络，则可以通过合并各个句子语言网络中的相同节点与连边来产生。

图 1 一个句子的语言网络

语言网络可以用符号表示为：

$G=(V,E)$

式中，$V=\{v_i \mid i=1,2,3\cdots,N\}$ 为顶点的集合，N 为语言网络中节点的个数，$E=\{(v_i,v_j) \mid v_i,v_j \in V\}$ 表示边的集合。

节点 v_i 的度定义为：

$$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}|$$

节点 v_i 的聚集度定义为：

$$K_i=|\{(v_j,v_k):(v_i,v_j)\in E,(v_j,v_k)\in E,v_i,v_j,v_k\in E\}|$$

节点 v_i 的聚集系数定义为：

$$C_i=\frac{K_i}{D_i(D_i-1)/2}=\frac{2K_i}{D_i(D_i-1)}$$

节点的度体现该节点与其它节点的关联情况，节点的聚集度和聚集系数体现在此节点局部范围内的节点相互连接密度。节点的度和聚集系数体现该节点在局部范围内的重要性。

定义语言网络节点的综合特征值 CF_i：

$$CF_i=0.5\times\frac{C_i}{\sum_{i=1}^{n}C_i}+0.5\times\frac{D_i}{n} \tag{2}$$

对词项按 CF 值从大到小进行排序，从中选取 TOP 比例的词项作为关键词项，以此关键词项向量作为词条正文的特征表示。与传统的词频向量相比，一篇正文的关键词向量维度下降了 1－TOP，这在效率上是一个较大的提高。最后利用式(1)通过计算关键词向量之间的相似度得到词条正文的相似度，记为 Sim_2。

3.4 百科词条开放分类相似度计算

百度百科共设 3 级开放分类，所有词条都位于第 3 级开放分类底下，如图 2 所示。

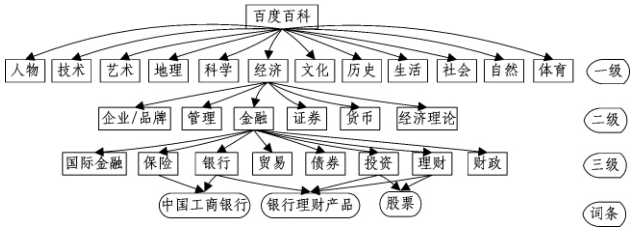

图 2 百度百科开放分类图

通过对百度百科分类关系的分析，可以发现百度百科的分类关系比其它分类词典或语义网络都更为灵活，一个分类节点可以包含任意多个上层分类节点和下层分类节点，而传统分类图一般是多个子节点从属于单一的上位节点。因此在这个分类图中，两词条所在节点之间往往可以找到多条距离相近、却反映不同语义关系的路径。如果两词条之间存在比较多的路径，而这些路径的长度都比较短，并且在分类图中反映的语义关系比较强，那么两词条的语义相似度就比较高。一个典型例子是如果两节点在分类图中距离比较近，并且都属于相同的开放分类，那么认为这两节点的开放分类部分相似度为 1。

本文利用分类图节点的信息内容来计算开放分类之间的相似度[13]。信息内容的计算依赖于一个节点下辖的子节点的数量以及整个分类图的规模，计算公式为：

$$ic(w)=1-\frac{\log(totalNum(w))}{\log(|W|)}$$

式中，$totalNum(w)$ 表示节点 w 下辖的所有子节点的数量，$|W|$ 表示百度百科分类图中的总节点数。

综合多条路径的长度和权重，基于开放分类的相似度计算公式为：

$$\mathrm{Sim}_3=\sum_{k=1}^{n}\frac{ic_k(A,B)}{length_k(A,B)} \tag{3}$$

式中，$ic_k(A,B)$ 表示两词条 A 和 B 最近公共节点的信息内容，$lenghth_k(A,B)$ 表示两节点经过分类图最近公共节点的路径长度，n 为近距离路径条数。

3.5 相关词条相似度计算

百度百科的相关词条是与该词条具有较为紧密联系的横向关联的词条列表。将词条列表视为一维向量，每个词项权重为 1，利用式(1)计算相关词条之间的相似度，结果记为 Sim_4。

3.6 词条相似度计算

通过上述 4 部分的相似度计算，得到词条的整体相似度，记为：

$$\mathrm{baikeSim}(A,B)=\sum_{i=1}^{4}\theta_i\mathrm{Sim}_i(A,B) \tag{4}$$

式中，θ_i 是可调节的参数，且有

$$\theta_1+\theta_2+\theta_3+\theta_4=1,\theta_1\geqslant\theta_2\geqslant\theta_3\geqslant\theta_4$$

后者反映了 Sim_1 到 Sim_4 对于总体相似度所起的作用依次递减。由于百科名片和百科词条正文反映了一个词条最主要的特征，因此应该将其权值定义得比较大。

4 实验

由于目前对中文词语相似度的研究还没有形成统一的规范，也没有相关标注语料提供实验平台，因此中文词语相似度计算的实验设计与数据筛选比较困难，如果随机选取一些词语，很难说明问题。本文通过综合选择，将英文的 WordSimilarity-353 词对翻译成中文作为实验标准数据集。该数据集包括 353 对英文词对，这些词对的相似度值是通过人工判断统计得到的，可以用来收集、训练或测试计算机执行语义相似度的方法。随机选取的部分标准数据词对集如表 2 所列。

表 2 实验数据摘要

英文词对	标准值	中文词对
journey-voyage	0.929	旅程-航程
money-cash	0.908	货币-现金
computer-software	0.85	计算机-软件
network-hardware	0.831	网络-硬件
nature-environment	0.831	自然-环境
psychology-Freud	0.821	心理学-弗洛伊德
news-report	0.816	新闻-报告
war-troops	0.813	战争-部队
bank-money	0.812	银行-货币
stock-market	0.808	股票-市场
century-nation	0.316	世纪-国家
volunteer-motto	0.256	志愿者-座右铭
reason-hypertension	0.231	原因-高血压
energy-secretary	0.181	能源-秘书
stock-phone	0.162	股票-手机

实验首先采用中科院分词软件 ICTCLAS 对百科名片和词条正文分词，再利用停用词列表删除停用词。词条各部分的相似度计算按以下步骤进行：

(1)对百科名片统计分词后各词项的词频，以词频作为词项权重，建立向量空间，再利用式(1)计算百科名片之间的相似度，得到 Sim_1；

(2)对词条正文分词后的词项建立语言网络,再利用式(2)计算各个词项的综合特征值,按综合特征值从大到小排序后选择 TOP 比例的词项作为关键词建立向量空间,再利用式(1)计算词条正文之间的相似度,得到 Sim_2;

(3)对开放分类,利用式(3)计算词条之间的相似度,得到 Sim_3;

(4)对相关词条,直接建立向量空间后利用式(1)计算相似度,得到 Sim_4。

在得到词条各部分的相似度后,本文经过多次实验比较,根据关键词抽取规则[15]以及词条中百科名片和词条正文对词条意义贡献比较大、开放分类和相关词条对词条意义贡献较小的情况,选取的参数为 TOP=30%,$\theta_1=0.4$,$\theta_2=0.4$,$\theta_3=0.1$,$\theta_4=0.1$,利用式(4)计算词条之间的相似度。本文使用文献[7]基于《知网》的方法和文献[8]基于《同义词词林》的方法作对比,得到的实验结果数据集如表 3 所列。

表 3 实验结果数据

中文词对	标准值	基于知网	基于同义词词林	本文算法
旅程-航程	0.929	0.04	0	0.63
货币-现金	0.908	1	0.43	0.57
计算机-软件	0.85	0.44	0.22	0.78
网络-硬件	0.831	0.29	0.22	0.54
自然-环境	0.831	0.05	0	0.62
心理学-弗洛伊德	0.821	0	0	0.61
新闻-报告	0.816	0.62	0.22	0.65
战争-部队	0.813	0.15	0.61	0.75
银行-货币	0.812	0.11	0.21	0.72
股票-市场	0.808	0.11	0.21	0.45
世纪-国家	0.316	0.11	0	0.21
志愿者-座右铭	0.256	0.1	0	0.15
原因-高血压	0.231	0.29	0.21	0.13
能源-秘书	0.181	0.10	0	0.08
股票-手机	0.162	0.26	0	0.11

从以上实验结果可以看到:(1)基于《知网》的相似度计算方法存在一些未登录词,对于一些新词以及一些不常用的词无法计算相似度。基于《同义词词林》的方法也同样存在一些未登录词,同时受到同义词词林层次关系的限制,得到的结果值集中于 5 个数值,不能很好地反映词间的语义关系。本文算法在这一点得到了很好的改进,能够计算出任意两个词语间的相似度值。(2)个别存在比较明显的概念关系的词语,其它两种方法的结果更优,但在整体效果与标准数据集相比方面,本文算法的结果显得更加合理有效。

结束语 本文提出了一种新的基于百度百科的词语相似度计算方法。与传统的基于语义词典和大规模语料库的方法不同,本文通过计算表征百科词条各个部分内容的相似度加权得到词条相似度。具体讨论了百科名片、词条正文、开放分类和相关词条部分的相似度计算方法,对其再加权就得到整体的相似度结果。从实验结果看,这个新方法产生的结果优于已有的方法。

本文后续的研究将在现有探讨词条相似度的基础上,进一步深入分析词条信息所蕴含的语义相似性特征,考虑百科名片、词条正文等语义结构信息,更好地提高词语相似度效果。

参 考 文 献

[1] 章志凌,虞立群,陈奕秋,等. 基于 Corpus 库的词语相似度计算方法[J]. 计算机应用,2006,26(3):638-640,644

[2] Salton G, Lesk M E. Computer evaluation of indexing and text processing[J]. Journal of the ACM, 1968, 15(1): 8-36

[3] Rada R. Development and application of a metric on semantic nets[J]. IEEE Transactions on System. Man and Cybernetics, 1989, 19(1): 17-30

[4] Lee J H. Information retrieval based on conceptual distance in ISA hierarchies [J]. Journal of Documentation, 1993, 49(2): 188-207

[5] Agirre E, Rigau G. A Proposal for word sense disambiguation using conceptual distance [C]// International Conference/Recent Advances in Natural Language Recessing RANLP. 95. Tzigov Chark, Bulgaria, 1995: 91-98

[6] Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network[C]// Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM'93). Washington, DC, US, 1993: 67-74

[7] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]// 台北第三届汉语词汇语义学研讨会

[8] 王斌. 汉英双语语料库自动对齐研究[D]. 北京:中国科学院计算技术研究所,1999

[9] Li Su-jian, et al. Semantic computation in Chinese question-answering system [J]. Journal of Computer Science and Technology, 2002, 17(6): 933-939

[10] Brown P. Word sense disambiguation using tactical methods[C]// Proceedings of 29th Meeting of the Association For Computational Linguistics (ACL291). 1991: 210-207

[11] 胡俊峰,俞士汶. 唐宋诗词汇间语义相似度计算[J]. 中文信息学报,2002(4):40-45

[12] Ferreri Cancho R, Sole R V. The small world of human language [J]. Biological Sciences, 2001, 268(1482): 2261-2265

[13] Seco N, Veale T, Hayes J. An Intrinsic Information Content Metric for Semantic Similarity in WordNet[C]// Proc of ECAI. 2004

[14] 黄承慧,印鉴,候昉,等. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011(5):856-864

[15] 郑家恒,卢娇丽,等. 关键词抽取方法的研究[J]. 计算机工程,2005(9):194-196

(上接第 191 页)

[14] O'Callaghan L, Mishra N, Meyerson A. Streaming-data algorithms for high-quality clustering[C]// Proceedings of 18th International Conference on Data Engineering. Los Alamitos, CA, USA: IEEE, 2002: 685-94

[15] Gorawski M, Pluciennik-Psota E. Distributed data mining methodology for clustering and classification model[C]// Proceedings of 10th International Conference on Artificial Intelligence and Soft Computing. Berlin, Germany: The Institution of Engineering and Technology, 2010: 323-30

[16] 孙岳,毛国君,刘旭. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报,2008,34(1):93-97

[17] 吴枫,仲妍,吴泉源. 基于时间衰减模型的数据流频繁模式挖掘[J]. 自动化学报,2010,36(5):674-684