文章编号：2096-1472(2019)-10-36-07
DOI:10.19644/j.cnki.issn2096-1472.2019.10.009

基于维基百科链接特征的词语语义相似度计算

张 波

(贺州学院数学与计算机学院，广西 贺州 542899)

摘 要：针对目前基于维基百科的相似度计算方法预处理过程烦琐、计算量大的问题，本文以维基百科为本体引入基于特征的词语语义计算，提出了一种基于维基百科的快速词语相似度计算方法。根据维基百科页面链接结构的特点，该方法把页面的入链接和出链接作为页面特征值构建特征向量模型，通过计算页面的特征向量相关系数计算对应词语的语义相似度。本文还改进了维基百科消歧处理算法，在一词多义的处理中减少社会认知度低的义项页面的干扰，进一步提高了计算准确度。经Miller & Charles(MC30)和Rubenstein & Goodenough(RG65)测试集的测试，测试结果表明了基于维基百科链接特征的方法在计算相似度方面的可行性，也验证了本文的计算策略和消歧改进算法的合理性。

关键词：语义相似度；维基百科；基于链接；基于特征值
中图分类号：TP391 **文献标识码**：A

A Semantic Similarity Calculation Based on the Features of Wikipedia Links

ZHANG Bo

(*School of Mathematics & Computer Science*,*Hezhou University*,*Hezhou* 542899,*China*)

Abstract: Measuring semantic similarity is a critical basic research in natural language processing.Because Wikipedia has open-editing,huge vocabulary,rapid update and other features,more and more research and applications have been focused on Wikipedia.This paper proposes a page-link approach for calculating word semantic similarity by taking Wikipedia as data resource.This approach improves the Wikipedia Link Vector Model (WLVM) method taking outgoing links as the feature vector,and utilizes page's incoming links and outgoing links as feature values in Wikipedia,then calculates the semantic similarity between words by measuring feature set similarity between the corresponding pages.The method also improves the disambiguation page processing by reducing the interference of the low social recognition pages.Through testing with Miller & Charles (MC30) and Rubenstein & Goodenough (RG65) benchmark,the validity of this method on the measuring word semantic similarity measurement is verified.

Keywords:word similarity;Wikipedia;link-based;feature-based

1 引言(Introduction)

词语之间的语义相似度(SS)测量是自然语言处理领域重要的基础研究之一，在很多领域都有着广泛的应用，比如词义消歧[1]、同义词探测[2]、信息抽取[3]等。传统的词语相似度计算方法主要包括两类：其一是基于语料库的概率统计方法，它的计算策略是使用统计学的方法把词语相似度问题转换为词语在语料库中上下文信息概率分布问题，把词语在语料库中的统计学特征作为语义相似度计算依据；其二是基于某个世界知识本体模型的计算方法，把词语在本体概念模型中特征的共性和差异性作为计算依据[4]。目前英文词语相似度研究所使用的基本参考本体模型主要是WordNet[5]，而中文词语相似度研究主要使用的模型有董振东的“知网”[6]与哈工大的“同义词词林”[7]。由于WordNet具有严谨的词义分类结构和语义关系，基于WordNet的英文词语语义相似度计算已经得到了较为广泛地应用[8-19]。然而，WordNet词汇量覆盖面窄，且版本更新周期长，这些缺点也限制了它的进一步应用范围。进入21世纪后，随着维基百科这一全球性多语言百科全书的推广使用，开始出现了基于维基百科的词语相似度与相关度的

基金项目：广西高校科学研究项目,基于描述逻辑的教育技术标准本体模型研究(项目批准号：ZD2014129).

研究。维基百科是一种带类目结构的网页文本语料库，具有与WordNet相似的分类结构，其包罗万象的网页内容和复杂的页面链接蕴含着丰富的词语语义信息。同时，由于维基百科的内容和结构由世界各地的志愿者合作创建和编辑，其词汇量和更新速度均远超其他语义知识源。因此，根据维基百科的数据特点，探索基于维基百科的语义相似度和相关度计算方法问题，吸引了越来越多研究者的关注。

2　研究背景(Research background)

维基百科是一个允许用户编辑的开放网络百科全书，自2001年推出后得到了迅猛增长，截止到2017年10月2日，共涵盖299种语言，具有46483635个页面[20]，其中包括5486459个英文页面。维基百科每月发布两次数据库备份转储(Database backup dumps)，为基于维基百科数据资源的研究和应用提供便利[21]。

维基百科的基本信息单元是页面(Page)，其中页面分为内容页面(Content Page)、消歧页面(Disambiguation Page)、类目页面(Category Page)。每一个页面的底部都列出该页面所从属的类目(Category)。内容页面通过文本定义概念词条，并显示与该概念词条相关的信息内容，文本中以超链接的形式包括了用于定义该词条的其他页面，比如“United States” 页面定义了“美国”这一概念，并呈现关于“美国”的相关信息内容，其中包括了1632个超链接；消歧页面用于解决一词多义的问题，它以列表的方式呈现该词条的所有义项页面超链接，比如“Apple(disambiguation)”页面呈现了词语“Apple”的59个义项页面；类目页面用于以列表的形式呈现该类别的主页面、子类别、从属页面等信息，比如“Category:Fruit”页面呈现了该类别的一个主页面、四个直接子类别和12个从属页面的信息。

目前，基于维基百科的词语语义相似度计算方法主要包括基于维基百科类目图(Wikipedia Category Graph，WCG)的方法、基于文本语义分析的方法和基于页面链接的方法。

(1)基于WCG的方法以维基百科的类目结构作为语义计算的本体模型基础，把基于语义词典语义相似度计算方法扩展到基于维基百科的相似度计算。该方法可以追溯到Strube等人在2006年提出WikiRelate方法。WikiRelate方法是最早把维基百科作为知识来源的词语相似度计算方法，该方法借鉴了基于语义词典的路径计算方法，在维基百科类目结构的基础上建构维基百科类目树(Category Tree)，根据概念在类目树上的距离和深度计算词语语义相似度[22]。2012年，Taieb等人把基于IC的方法引入维基百科相似度计算中，他们根据WCG的结构特点改进了IC计算方法，并通过限制图的搜索深度提高计算精度[23]。2013年，Taieb等人继续把基于特征值的计算方法扩展到维基百科相似度计算中，他们通过页面词语的词根统计为WCG中的每一个类目构建语义描述特征向量，在此基础上尝试了多种特征向量计算方法，并在计算过程中通过增加页面重定向、页面文本特征等因素改善计算结果[24]。2016年，Aouicha等人改进了基于WCG的IC计算方法，提出了WordNet和维基百科结合的计算方法[25]。由于维基百科由众多用户共建而成，使其类目结构并不严谨，作用主要是词语推荐和网站导航。因此，基于WCG的方法均需要进行复杂的类目数据预处理，计算准确率并不理想。

(2)基于文本语义分析的相似度方法是以维基百科页面文本为计算依据，通过对页面文本进行语义分析和词语统计建立维基百科概念的向量空间模型，根据向量模型的相似度计算维基百科概念的相似度。由Gabrilovich和Markovitch提出的ESA(Explicit Semantic Analysis)方法最具代表性[26]。ESA方法利用维基百科词语覆盖面大、概念语义描述丰富的特点，在使用机器学习技术对页面文本中的词语和词组进行文本解析的基础上，把维基百科概念表示为概念解释向量(interpretation vector)。解释向量由页面文本中包含的文本片段组成，文本片段对于概念的重要程度体现为文本片段在向量中的权重。文本片段的权重计算采用了TF-IDF算法，该算法是信息检索领域与数据挖掘领域的常用加权技术，一个文本片段在某一个页面中的权重等于该文本片段在页面中的统计频率与该页面片段在整个维基百科中逆向页面频率之积。ESA方法不仅可以用于计算词语语义计算，也可以计算任意长度的语义计算，但是该方法需要事先构建解释向量空间，计算量大。

(3)基于页面链接的相似度方法是把维基百科的页面超文本链接结构作为相似度计算数据模型。在维基百科中，页面通过超文本链接相互关联。一个页面的超链接结构既包括在页面文本中链接到其他页面的出链接(Outgoing Link)，也包括从其他页面链接到该页面的入链接(Incoming Link)。Milne和Witten是首先使用维基百科页面链接计算文本相似度的学者，他们在2008年尝试了根据出链接向量余弦夹角计算方法和入链接集合Google距离计算方法，并提出了把两种计算求平均值的方式计算词语相似度的方法[27]。在2009年，Milne在前一种方法的基础上提出了WLVM(Wikipedia Link Vector Model)方法[28]。在WLVM方法中，维基百科页面表示为页面文本中包含的超链接(即出链接)组成的向量，根据向量的余弦夹角计算维基百科概念的相似度。WLVM方法对链接的权重计算提出了改进，以超链接在文本中出现的次数代替TF-IDF算法中的词语统计频率。

本文提出了一种基于链接特征的词语相似度计算方法，该方法以维基百科概念页面之间的链接作为特征，通过计算两个维基百科页面的链接特征集合的相似度得出页面之间的

相似度，进而实现页面对应的词语之间的相似度计算。另外，本文还对维基百科一词多义的处理过程进行了改进，通过对一些社会认知度低的义项页面进行裁剪，进一步提高了计算结果的准确度。

3 计算方法(Computing method)

基于维基百科的词语语义相似度的计算过程主要包括三个步骤。首先在维基百科数据资源中获取词条的定位页面。若定位页面属于内容页面，说明词条义项单一，定位页面即是义项页面；若定位页面属于消歧页面，说明词条是一词多义，需要把页面消歧处理为多个非消歧义项页面，每一个义项页面描述词条的一个义项。最后通过计算义项页面之间的相似度来计算词语之间的相似度，计算基本过程如图1所示。

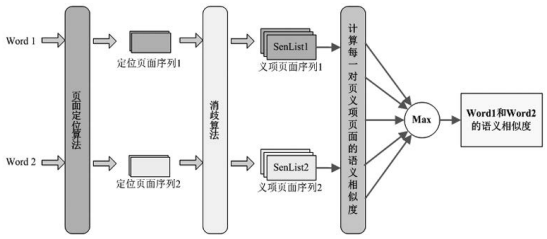

图1 基于维基百科的词语语义相似度计算过程

Fig.1 Wikipedia-based Word Semantic Similarity Computing Process

3.1 页面定位和消歧

计算两个词语相似度的第一步是找到词语在维基百科数据资源中的定位页面，把词语相似度计算转换为页面相似度度量。本文采用查询维基百科词条来实现词语与页面的映射，具体分为三种情况。其一，在维基百科中存在标题与词语完全一致的页面，比如词语“glass”可以定位到标题为“Glass”的页面；其二，在维基百科中词语重定位到的页面，这种情况表示词语与页面的标题属于同义词，比如词语“automobile”重定向为标题为“Car”页面；其三，在维基百科中存在标题为“词语+(disambiguation)”的页面，比如词语“tool”可以定位到标题为“Tool(disambiguation)”的页面。

本文的页面定位采用JWPL(Java Wikipedia Library)工具，对于任意词语w，通过检索数据资源找到w对应的一到多个页面，并最终选取计算定位页面列表。具体的定位过程包括以下步骤：

(1)若w只能检索到一个页面，则直接确定该页面为w的唯一定位页面。

(2)若w可以检索到多个页面，且存在标题与w完全一致的页面，则把该页面加入w的定位页面序列。比如词语“car”可以检索到标题为“Car”“C.a.R.”“CAR”“Car(disambiguation)”等页面，该步选择“Car”页面加入词语“car”的定位页面序列。

(3)若w可以检索到多个页面，且存在从w重定向的页面，则选取该页面加入w的定位序列。比如词语“gem”可以检索到的页面包括“Gemstone”和“GEM”，由于“gem”重定向到“Gemstone”页面，所以本步选择“Gemstone”页面加入定位页面序列。在维基百科中，满足这一步的条件与步骤(2)互斥，即若存在标题与w一致的页面，则不存在w重定向的页面。

(4)若w可以检索到多个页面，且存在标题为“词语+(disambiguation)”的页面，则把该页面加入定位页面序列。比如步骤(2)所举词语“car”的例子中，该步选择“Car(disambiguation)”页面加入“car”的定位页面序列。

对词语w的维基百科页面定位过程结束后，最终的定位页面序列由1—2个页面组成，页面类型为内容页面或消歧页面。由于消歧页面表示了词w蕴含的多个语义，所以需要对定位页面中的消歧页面消歧处理为多个义项页面，最后形成义项页面序列参与页面计算。需要说明的是，在维基百科中，标题以“(disambiguation)”结尾的页面一定是消歧页面，但标题不以“(disambiguation)”结尾的页面也可能是消歧页面，如页面“Jewel”。因此，判定页面是否是消歧页的方式是判定其页面是否属于“Disambiguation pages”类目，而非标题是否以“(disambiguation)”结尾。

从定位页面序列到义项页面序列的转换过程中，核心是消歧页面的处理算法。若w的义项页面序列不存在消歧页面，则定位页面序列就是义项页面序列。若w的义项页面序列中存在消歧页面，则优先把词语定位页面序列中的内容页面加入义项页面序列，然后通过消歧处理算法，把消歧页面处理为多个内容页面，并依次把它们加入到义项页面序列。若最终w的义项页面序列只包括一个页面，则表示该词语的语义单一，只对应一个义项；若词语有多个义项页面，则表示该词语为一词多义。

目前基于维基百科的研究中，对消歧页面进行消歧处理的方法均使用消歧页的出链接集合作为义项页面集合。由于消歧页面所包含的义项列表是以超链接的方式链接了各个义项页面，所以消歧页面的出链接集合主要是消歧页所列的义项页面集合。这种消歧页处理的方法方便快捷，但是受到消歧页的“See also”“Categories”“References”“External links”等附属内容的干扰，且不能反映超链接在页面中的排序顺序。而本文采用通过解析页面文本的方式获得消歧页面中所有的义项页面。相较于通过数据资源中获取出链接的方法计算义项页面，采用解析页面文本的方式可以有效排除

页面的“See also”“Categories”“References”“External links”等附属内容的干扰；更为重要的是可以形成遵照页面顺序的义项序列，为本文进一步改进消歧算法、提高整体效率和准确度提供基础。

3.2 义项页面的词语相似度计算

经过上述词语的页面定位和消歧处理后，每一个词语可以映射为一个维基百科义项页面序列，即词语之间相似度计算转换为其对应义项页面之间的相似度计算问题。对于任意两个词语w_1和w_2，其相似度取两个义项页面序列中相似度最大的一对义项页面的相似度，其计算方法可以表示为：

$$Sim(w_1,w_2)=\max_{(p_i,p_j)\in SenList(w_1)\times SenList(w_2)} Sim(p_i,p_j) \quad (1)$$

其中，$SenList(w)$表示词语w的义项页面序列，$Sim(p_i,p_j)$表示两个义项页面p_i和p_j的相似度。

在计算词语相似度方法中，基于特征的方法是选取词语在本体模型中的特征属性作为特征集合，通过衡量两个特征集合的共性或差异性来计算词语之间的相似度。基于特征的方法最早可以追溯到1988年，由当时著名的认知心理学家Tversky提出的基于特征的相似度计算方法[15]。Tversky认为词语的特征集的共性越多，两个词语就越相似，反之同理。对于概念c_1和c_2，Tversky计算方法可以表示为：

$$Sim_{Tversky}(c_1,c_2)=\frac{|\psi(c_1)\cap\psi(c_2)|}{|\psi(c_1)\cap\psi(c_2)|+k|\psi(c_1)/\psi(c_2)|+(1-k)|\psi(c_2)/\psi(c_1)|} \quad (2)$$

其中，$\psi(c_1)$和$\psi(c_2)$分别表示c_1和c_2的特征集合，k是非对称调节因子，且$k\in[0,1]$。当$k=0.5$时表示概念c_1和c_2的相似度等于概念c_2和c_1的相似度，Tversky计算方法可以转换为：

$$Sim_{Tversky}(c_1,c_2)=\frac{|\psi(c_1)\cap\psi(c_2)|}{|\psi(c_1)\cup\psi(c_2)|} \quad (3)$$

在基于特征的相似度计算方法中，特征的选取是算法的核心和关键。基于WordNet等语义词典的词语相似度计算方法中，词语的特征主要涉及词性、类别、深度、相邻结点、同义词集合、描述文本等因素。本文提出把维基百科页面的链接(包括入链接和出链接)作为页面的特征值，把页面入链接和出链接构成的特征向量模型作为计算维基百科页面的相似度的依据。

对于页面p，入链接是p作为链接目标页面的超链接，可以表示为：$p'\rightarrow p$，其中p'称为p的入链接页面；出链接是p作为链接起点的超链接，可以表示为：$p\rightarrow p''$，其中p''称为p的出链接页面。页面“Automobile”和“Global Warming”的入链接和出链接示意图如图2所示，其上半部分是两个页面的入链接，下半部分出链接。

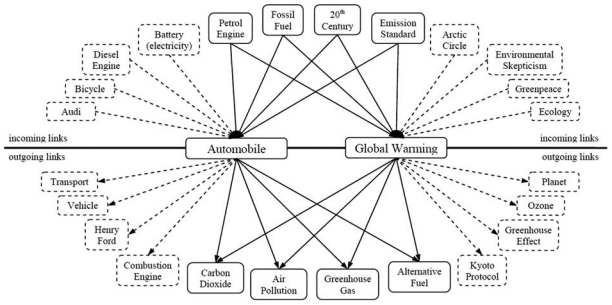

图2 入链接和出链接示意图

Fig.2 Schematic diagram of incoming links and outgoing links

对于页面p，以链接为特征值可以表示为：

$$p_{link}=(\{(p_1'\rightarrow p),(p_2'\rightarrow p),\cdots,(p_n'\rightarrow p)\},\{(p\rightarrow p_2''),(p\rightarrow p_2''),\cdots,(p\rightarrow p_m'')\}) \quad (4)$$

在式(4)中，等式的右边由两个链接集合组成。第一个集合是入链接集合，表示把p作为链接目标页面的超链接；第二个集合是出链接集合，表示从页面p链接到其他页面的超链接。式(4)中，n和m分别表示两个集合的元素个数。

本文认为，维基百科页面之间链接的重叠程度体现了页面相似度，入链接和出链接对页面相似度的影响独立且对等。依据这一思想，本文引入基于特征的相似度计算方法，将入链接和出链接作为维基百科页面特征。对于义项页面p_1和p_2，其页面相似度计算方法为式(5)。

$$Sim_{(p1,p2)}=\left(\frac{\log(|InCL_{p1}\cap InCL_{p2}|)}{\log(|InCL_{p1}\cup InCL_{p2}|)}+\frac{\log(|OutGL_{p1}\cap OutGL_{p2}|)}{\log(|OutGL_{p1}\cup OutGL_{p2}|)}\right)/2 \quad (5)$$

其中，$InCL_{p1}$和$InCL_{p2}$分别表示页面p_1和p_2的入链接页面集合，$OutGL_{p1}$和$OutGL_{p2}$分别为p_1和p_2的出链接页面集合。

在式(5)中，页面p_1和p_2的相似度等于两个页面入链接特征相似度和出链接特征相似度的平均值，体现了两种特征因素在计算页面相似度中发挥对等作用。由于维基百科页面之间入链接和出链接分布极不均衡，为了降低页面链接集合大小的差异性对计算结果的影响，式(5)的链接页面集合元素个数取自然对数以起到降低运算数值变差的作用。

3.3 消歧算法的改进

目前的基于维基百科相似度和相关度研究中，对一词多义的处理大部分采用前文公式(1)的方式，两个词的相似度等于相似度最高的一对义项页面的相似度，即两个词的所有义项页面进行笛卡尔积运算后选择计算结果最大值作为最终词语的相似度。由于维基百科消歧页面的义项数普遍较大，一些研究提出了几种选择合适义项的方法。例如，文献[28]认为义项页面(概念)的入链接数量反映了该义项的普遍程度，入链接数量越多的页面普遍性越高，因此优先选择入链接数量高

的义项页面进行语义计算。而文献[27]和文献[23]认为页面所包含的出链接数量反映了该页面的语义信息的丰富程度，出链接数量越多的页面其语义描述也越充分，因此优先选择入链接数量高的义项页面进行语义计算。

本文认为消歧页面所列出的义项页面的排列顺序，体现了对应词义社会认知度的差异性，社会认知度高的词义趋向于排列在消歧页面义项列表的前面，社会认知度低的词义趋向于排列在消歧页面义项列表的后面。这是由于维基百科页面编辑的社会开放性造成的，即人们在编辑消歧页面的时候，会基于自己的认知，倾向于将语义更广泛的义项排放在序列的前列，而把语义更狭窄的义项排放在靠后一些。到目前，英文维基百科每一个页面平均被编辑次数为21.11次以上[20]，故消歧页面中义项的语义排序在一定程度上反映了社会对该词语的认知差异，比如“Apple(disambiguation)”页面有59条语义项，其中前1—15个义项属于植物学范畴的Apple义项页面，第16—19条是名为Apple的四家公司，第20—22条是3个影视作品，第23—28条是音乐作品，29—34条是地名，35—39条是技术，第40—59条是Apple相关的其他语义项，页面截图如图3所示。词语相似度和相关度计算本质上是模拟人类对词语之间语义的相似度和相关度的判定。由于语义序列排列靠前的义项社会认知度较高，语义计算应优先选择排列在语义序列前列的义项页面进行语义计算。由于靠后的义项社会认知度较低，这些义项页面参与计算通常会对计算结果产生一定程度的干扰。

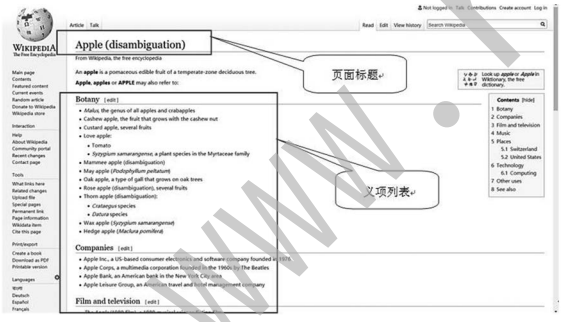

图3 Apple(disambiguation)页面截图

Fig.3 Screenshot of *Apple* (*disambiguation*) page

为了减少社会认知度低的义项参与运算导致的干扰，本文在处理消歧页面时通过设置阈值参数θ来控制义项的选取个数，选取消歧页义项列表前列的词义参与相似度计算，排除义项列表后列义项的干扰，从而提高计算的准确度。阈值参数$\theta\in[0,1]$，表示在义项页面序列中选取义项比例，选取顺序依照消歧页面中义项列表的文本排列顺序。词语w的带阈值参数的义项序列可以表示为：

$$SenList^{\theta}(w)=\{p_i\in SenList(w)|i\in[1,\theta*N]\} \quad (6)$$

其中，$SenList(w)$是w的全部义项页面列表，$N=|SenList(w)|$。阈值参数θ表示选取比例，$\theta\in[0,1]$，当θ为0时表示不选取任何义项，当θ为1时表示选取所有的义项。

两个词语w_1和w_2带阈值参数的相似度计算可以表示为：

$$Sim^{\theta}(w_1,w_2)=\max_{(p_i,p_j)\in SenList^{\theta}(w_1)\times SenList^{\theta}(w_2)}Sim(p_i,p_j) \quad (7)$$

为了确保阈值参数θ取值的客观性，我们选择了AG203相似度基准测试集作为训练集来训练计算参数。AG203[29]是Agirre等在2009年创建，由203对词组成，词语的词性包括名词、动词和形容词，人工判定值在0—10。在训练集AG203上使用式(4)、式(5)、式(6)和式(7)，通过不断改变阈值参数θ的取值，计算相似度值与人工判定值的Pearson相关系数值，当Pearson相关系数达到峰值时阈值参数θ的取值确定为最终参数值。在训练集AG203上，本文方法计算相似度与人工判定值的Pearson系数随阈值参数θ取值变化情况如图4所示。

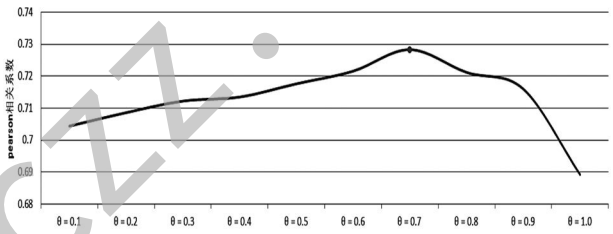

图4 Pearson相关系数随阈值参数θ取值的变化图

Fig.4 Pearson correlation coefficient versus threshold parameter θ

如图4所示，设置阈值参数的测试结果整体优于不设阈值的结果，且阈值θ=0.7时，训练数据集计算结果与人工判定值的Pearson相关系数达到最优状态，在本文中，阈值参数最终确定为0.7。在下文章节所采用的测试数据集在计算过程中，阈值θ=0.7为计算的最优状态这个现象依然表现明显。

4 实验与分析(Experiments and analysis)

使用基准测试数据集对算法进行测试和对比是目前最主要的相似度计算方法的评测方式，本文选取两种国际通行计算词语相似度基准测试集：Miller & Charles(MC30)和Rubenstein & Goodenough(RG65)。其中，MC30数据集是由Miller和Charles在1991年发布的，包括30对名词，人工判定值的取值在0—4[30]。RG65是Rubenstein和Goodenough在1965年创建，由65对名词测试数据组成，人工判定值在0—4[31]。为了便于计算结果对比，本文对测试集进行了归一化处理，即MC30和RG65的人工判定值除以4，使其值域统一为[0,1]。

首先，我们对本文改进的语义相似度方法进行测试，消岐算法采用传统的取消歧页面所有的义项页面的方法，相关公式包括式(1)、式(4)和式(5)。测试采用的维基百科数据资源

是2017-8-1版维基百科数据库备份转储。在两个测试集上的计算结果与人工判定值之间的Pearson相关系数如表1所示。

表1 本文方法取全部义项的计算结果与人工判定值的相关系数

Tab.1 The correlation coefficients of taking all meanings

数据集	MC30	RG65
Pearson相关系数	0.865184769	0.779429063

接下来，我们在上述页面相似度算法的基础上，采用本文改进的消岐算法处理一词多义，使用两个测试数据集对算法重新测试，相关公式包括式(4)、式(5)、式(6)和式(7)，阈值参数θ设为0.7，计算结果如表2所示。

表2 本文方法阈值参数取0.7时计算结果与人工判定值的相关系数

Tab.2 The correlation coefficients when the threshold parameter is 0.7

数据集	MC30	RG65
Pearson相关系数	0.906513536	0.82858353

MC30测试集测试结果与人工判定值数值对比如表3所示。

表3 MC30测试集相似度测量结果

Tab.3 Similarity measurement results of MC30 test set

词1	词2	人工值	取全部义项	θ=0.7	词1	词2	人工值	取全部义项	θ=0.7
car	automobile	0.98	1	1	lad	brother	0.415	0.454138	0.454138
gem	jewel	0.96	1	1	journey	car	0.29	0.525163	0.525163
journey	voyage	0.96	1	1	monk	oracle	0.275	0.391677	0.391677
boy	lad	0.94	1	1	cemetery	woodland	0.2375	0.342971	0.175806
coast	shore	0.925	0.779801	0.779801	food	rooster	0.2225	0.417893	0.417893
asylum	madhouse	0.9025	0.714846	0.714846	coast	hill	0.2175	0.343104	0.343104
magician	wizard	0.875	1	1	forest	graveyard	0.21	0.31132	0.31132
midday	noon	0.855	1	1	shore	woodland	0.1575	0.404203	0.404203
furnace	stove	0.7775	0.83284	0.83284	monk	slave	0.1375	0.377536	0.377536
food	fruit	0.77	0.703726	0.703726	coast	forest	0.105	0.414149	0.414149
bird	cock	0.7625	0.408315	0.408315	lad	wizard	0.105	0.67638	0.255194
bird	crane	0.7425	0.614599	0.614599	chord	smile	0.0325	0.229387	0.166956
tool	implement	0.7375	1	1	glass	magician	0.0275	0.260555	0.260555
brother	monk	0.705	0.744799	0.744799	noon	string	0.02	0.389263	0
crane	implement	0.42	0.413047	0.413047	rooster	voyage	0.02	0.345394	0.271839

为了比较本文相似度算法与现有相似度算法的优劣，本文使用2017-8-1版维基百科数据库对前文学者提出的出链接余弦夹角计算方法、Google距离计算方法和WLVM进行重新计算，并与文献中主要的词语语义相似度算法测试结果进行对比，如表4所示。

表4 不同方法计算结果与人工值之间的Pearson 相关系数

Tab.4 Pearson correlation coefficients of calculated results of different methods

方法	类型	知识来源	MC30	RG65	计算文献
本文方法	基于链接	维基百科	0.91	0.83	本文
出链接余弦夹角方法	基于链接	维基百科	−0.11	0.06	本文
出链接余弦夹角方法(改进消歧算法)	基于链接	维基百科	−0.11	0.22	本文
Google距离	基于链接	维基百科	0.21	−0.01	本文
Google距离(改进消歧算法)	基于链接	维基百科	0.33	0.02	本文
余弦夹角与Google距离平均值	基于链接	维基百科	0.08	0.05	本文
余弦夹角与Google距离平均值(改进消歧算法)	基于链接	维基百科	0.14	0.12	本文
WVLM	基于链接	维基百科	0.66	0.63	本文
WVLM(改进消歧算法)	基于链接	维基百科	0.77	0.71	本文
WikiRelate方法	基于WCG	维基百科	0.45	0.52	[27]
Taieb等(1)	基于WCG+IC	维基百科	0.73	0.65	[23]
Taieb等(2)	基于WCG+特征	维基百科	0.83	0.78	[24]
Aouicha等(1)	基于WCG+IC	维基百科	0.57	0.53	[25]
Aouicha等(2)	杂合	维基百科+WordNet	0.84	0.84	[25]
Gabrilovich和Markovitch	基于ESA	维基百科	0.73	0.82	[27]
Rada	基于路径	WordNet	0.66	0.75	[8]
Wu和Palmer	基于路径+深度	WordNet	0.75	0.79	[9]
Leacock和Chodorow	基于路径+深度	WordNet	0.75	0.79	[10]
Hao等	基于路径+深度	WordNet	0.82	0.84	[11]
Resnik	基于信息内容	WordNet	0.72	0.72	[12]
Lin	基于信息内容	WordNet	0.70	0.72	[13]
Jiang和Conrath	基于信息内容	WordNet	0.73	0.75	[14]
Tversky	基于特征	WordNet	0.63	0.68	[15]
Rodriguez和Egenhofer	基于特征	WordNet	0.71	0.78	[16]
Petrakis等	基于特征	WordNet	0.68	0.78	[17]
Taieb等(3)	杂合方法	WordNet	0.85	0.88	[18]

根据表4数据分析，由于维基百科数据资源页面规模和义

项数的增加，使用新版本维基百科数据资源对原有的基于链接的维基百科相似度方法的重算结果均差于原文献的计算结果，本文提出的页面链接为特征值的维基百科相似度方法总体上优于前文学者提出的出链接余弦夹角计算方法、Google距离计算方法和WLVM等方法。对比这些方法的计算过程中使用本文改进的消歧算法，可以看出本文的消歧方法在这些方法中依然起到改善计算结果的作用。本文方法在总体上优于文献中其他基于维基百科的相似度计算结果，也优于目前大部分基于WordNet的语义相似度计算方法。虽然本算法的RG65测试结果略低于目前基于WordNet方法的最高值[11,18]，但相较于WordNet(3.1版本)[5]中数据集中只有155287个词、117659个同义词集和206941个义项，维基百科具有超过其30倍的词汇量，且在快速更新，因此本文方法具有更为广泛的应用前景。

5 结论(Conclusion)

本文使用维基百科链接作为数据资源计算词语语义相似度，提出了一种基于维基百科页面链接特征的计算词语语义相似度方法。首先，通过页面定位和消歧，把词语定位到维基百科页面，把词语相似度问题转换为页面相似度的计算问题。然后把页面表示为由入链接和出链接组成的特征集合，通过计算特征集合的相似度计算对应词语的相似度。本文提出了对一词多义问题的消歧算法的改进，通过阈值参数限制消歧义项页面的方式减少社会认知度较低义项对测试过程的干扰，以达到提高计算准确度目的。

本文使用2017-8-1版维基百科数据资源，通过基准测试数据集对本文方法今次测试。MC30数据集的计算结果与人工值之间的Pearson相关度达到0.906513536，RG65达到0.82858353。经实验数据与其他方法的对比，本文提出的方法总体优于现有的基于维基百科链接的词语相似度计算方法，且与其他维基百科方法和基于WordNet方法对比也有较好的表现，说明本文提出的基于维基百科链接的词语语义相似度计算方法是合理有效的。

参考文献(References)

[1] Patwardhan S,Banerjee S,Pedersen T.Using measures of semantic relatedness for word sense disambiguation[C].International Conference on Computational Linguistics & Intelligent Text Processing.Springer-Verlag,2003:241-257.

[2] Lin D.An Information-theoretic Definition of similarity[C]. The Fifteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.,1998:296-304.

[3] Atkinson J,Ferreira A,Aravena E.Discovering Implicit Intention-Level Knowledge from Natural-Language Texts[J].Knowledge-Based Systems,2009,22(7):502-508.

[4] 石静,吴云芳,邱立坤.基于大规模语料库的汉语词义相似度计算方法[J].中文信息学报,2013,27(01):1-6.

[5] Princeton University.WordNet[EB/OL].http://wordnet.princeton.edu,2019-8-1.

[6] 董振东,董强.知网[EB/OL].http://www.keenage.com,2015-1-8.

[7] 哈工大社会计算与信息检索研究中心.同义词词林扩展版[EB/OL].http://ir.hit.edu.cn/demos,2019-7-20.

[8] Rada R,Mili H,Bicknell E,et al.Development and application of a metric on semantic nets[J].IEEE Transactions on Systems,Man and Cybernetics,1989,19(1):17-30.

[9] Wu Z,Palmer M.Verbs Semantics and Lexical Selection[C]. Proceedings of the 32nd annual meeting on Association for Computational Linguistics(COLING-94).Association for Computational Linguistics,1994:133-138.

[10] Leacock C,ChodorowM.Combining Local Context and WordNet Similarity for Word Sense Identification[M]. WordNet:An Electronic Lexical Database,1998:265-283.

[11] Hao D,Zuo W,Peng T,et al.An Approach for Calculating Semantic Similarity between Words Using WordNet[C]. Second International Conference on Digital Manufacturing & Automation.IEEE,2011:177-180.

[12] Resnik P.Using Information Content to Evaluate Semantic Similarity in a Taxonomy[C].the 14th International Joint Conference on Artificial Intelligence,Morgan Kaufmann Publishers Inc.,1995:448-453.

[13] D Lin.An Information-Theoretic Definition of Similarity[J]. Fifteenth International Conference on Machine Learning,1998,1(2):296-304.

[14] JJ Jiang,DWConrath.Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[J].International Conference Research on Computational Linguistics,1997:19-33.

[15] A Tversky.Features of similarity[J].Readings in Cognitive Science,1988,84(4):290-302.

[16] MA Rodríguez,MJ Egenhofer.Determining Semantic Similarity among Entity Classes from Different Ontologies[J].Knowledge & Data Engineering IEEE Transactions on,2003,15(2):442-456.

[17] EGM Petrakis,G Varelas,A Hliaoutakis,et al.X-similarity:computing semantic similarity between concepts from different ontologies[J].Journal of Digital Information Management,2006,4(4):233-237.

[18] Hadj Taieb M A,Ben Aouicha M,Ben Hamadou A.Ontology-based approach for measuring semantic similarity[J].Engineering Applications of Artificial Intelligence,2014,36:238-261.

[19] Zhu X,Li F,Chen H,et al.A density compensation-based path computing model for measuring semantic similarity[J]. Computer Science,2015:17-25.

[20] Wikipedia.List of Wikipedias[EB/OL].https://meta.wikimedia.org/wiki/List_of_Wikipedias,2019-8-1.

[21] Wikipedia.Wikimedia Downloads[EB/OL].http://dumps.wikimedia.your.org,2018-8-2.

[22] M.WikiRelate!Computing Semantic Relatedness Using Wikipedia[J].National Conference on Artificial Intelligence,2006,2(6):1419-1424.

[23] MAH Taieb,MB Aouicha,M Tmar,AB Hamadou.Wikipedia Category Graph and NewIntrinsic Information Content Metric for Word Semantic Relatedness Measuring[C].

International Conference on Data and Knowledge Engineering,Springer Berlin Heidelberg,2012:128-140.
[24] MAH Taieb,MB Aouicha,AB Hamadou.Computing semantic relatedness using Wikipedia features[J].Knowledge-Based Systems,2013,50(50):260-278.
[25] MB Aouicha,MAH Taieb,AB Hamadou.Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness[J].Applied Intelligence,2016,45(2):1-37.
[26] E Gabrilovich,S Markovitch.Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis[C].International Joint Conference on Artifical Intelligence,2007:1606-1611.
[27] D Milne,IH Witten.An Effective,Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links[J]. Proceedings of AAAI,2008:35-42.
[28] D Milne.Computing Semantic Relatedness using Wikipedia Link Structure[C].NewZealand Computer Science Research Student Conference,2009.
[29] E Agirre,E Alfonseca,K Hall,et al.A Study on Similarity and Relatedness Using Distributional and Word Net-based Approaches[C].The Conference of the North American Chapter of the Association for Computational Linguistics,2009:19-27.
[30] George A.Miller,Walter G.Charles.Contextual Correlates of Semantic Similarity[J].Language Cognition & Neuroscience,1991,6(1):1-28.
[31] H Rubenstein,JB Goodenough.Contextual Correlates of Synonymy[J].Communications of the ACM,1965,8(10):627-633.

作者简介：

张 波(1983-)，男，硕士，讲师.研究领域：自然语言处理，远程教育技术.

(上接第46页)

参考文献(References)

[1] Kenthapadi K,Mironov I,Thakurta AG.Privacy-preserving data mining in industry[C].In:Proc.of the Web Conference 2019-Companion of the World Wide Web Conference (WWW '19),2019:1308-1310.
[2] Li YL,Miao,CL,Su L,et al.An efficient two-layer mechanism for privacy-preserving truth discovery[C].In:Proc.of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18).ACM,New York,NY,USA,2018:1705-1714.
[3] Teo SG,Cao JN,Lee VCS.DAG:A general model for privacy-preserving data mining[J].IEEE Transactions on Knowledge and Data Engineering,2018,Article in Press.
[4] Bullek B,Garboski S,Darakhshan JM,et al.Towards understanding differential privacy:when do people trust randomized response technique?[C].In:Proc.of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).ACM,New York,NY,USA,2017:3833-3837.
[5] Aldà F and Simon HU.Randomized response schemes,privacy and usefulness[C].In:Proc.of the 2014 Workshop on Artificial Intelligent and Security Workshop (AISec '14).ACM,New York,NY,USA,2014:15-26.
[6] Warner SL.Randomized response:A survey technique for eliminating evasive answer bias[J].The American Statistical Association,1965,60(309):63-69.
[7] 陈光慧,韩兆洲.基于随机化回答模型的最低工资敏感性问题研究[J].统计与信息论坛,2012,27(09):3-7.
[8] 郭宇红,童云海,唐世渭,等.带学习的同步隐私保护频繁模式挖掘[J].软件学报,2011,22(08):1749-1760.
[9] Sun CJ,Fu Y,Zhou JL,et al.Personalized privacy-preserving frequent itemset mining using randomized response[J].The Scientific World Journal,2014.
[10] 许胜之.满足差分隐私保护的频繁模式挖掘关键技术研究[D].北京邮电大学,2016.
[11] 蒋辰,杨庚,白云璐,等.面向隐私保护的频繁项集挖掘算法[J].信息网络安全,2019(04):73-81.
[12] 邢欢.基于隐私保护的关联规则挖掘研究[D].南京邮电大学,2016.
[13] Rizvi SJ,Haritsa JR.Maintaining data privacy in association rule mining[C].In:Proc.of the 28th Int' l Conf.on Very Large Data Bases (VLDB '02).Morgan Kaufmann,2002:682-698.
[14] Agrawal S,Krishnan V,Haritsa J.On addressing efficiency concerns in privacy preserving mining[C].In:Proc.of the 9th Int' l Conf.on Database Systems for Advanced Applications (DASFAA' 04).LNCS 2973,Springer-Verlag,2004:113-124.
[15] Xia Y,Yang Y,Chi Y.Mining association rules with non-uniform privacy concerns[C].In:Proc.of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery(DMKD '04).ACM Press,2004:27-34.
[16] Andruszkiewicz P.Optimization for mask scheme in privacy preserving data mining for association rules[C]. In:Proc.of Int' l Conf.Rough Sets and Emerging Intelligent Systems Paradigms(RSEISP '07).LNAI 4585,Springer-Verlag,2007:465-474.
[17] Huang ZL,Du WL,Teng ZX.Searching for better randomized response schemes for privacy-preserving data mining[C]. In:Proc.of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '07). LNCS 4702,Heidelberg:Springer-Verlag,2007:487-497.

作者简介：

郭宇红(1979-)，女，博士，副教授.研究领域：数据挖掘，推荐系统.
童云海(1971-)，男，博士，教授.研究领域：数据挖掘，联机分析处理.