文章编号:1672-3961(2014)03-0029-07 DOI:10.6040/j.issn.1672-3961.2.2013.282

# 基于 SimRank 的百度百科词条语义相似度计算

尹坤,尹红风*,杨燕,贾真
(西南交通大学信息科学与技术学院,四川 成都 610031)

**摘要**:提出一种利用百度百科半结构化数据自动获取词语相似度的方法,该方法将百科词条与其相关词条看做有向图的两个节点,且两节点相互之间存在着链接关系,然后利用 SimRank 算法计算百科词条语义相似度。实验表明,该方法优于传统的词语语义相似度测量,能准确地反映词语之间的语义关系。

**关键词**:语义相似度;百科词条;有向图;SimRank
**中图分类号**:TP391 **文献标志码**:A

# Semantic similarity computation of Baidu encyclopedia entries based on SimRank

YIN Kun, YIN Hongfeng*, YANG Yan, JIA Zhen
(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, Sichuan, China)

**Abstract**: The measurement of the semantic similarity using semi-structured data on Baidu encyclopedia was proposed. The encyclopedia entries and related entries were considered as two nodes of a directed graph, of which there was a link between two nodes. Then SimRank algorithm was used to calculate the semantic similarity of encyclopedia entries. Experimental results showed that the proposed measure significantly outperformed the traditional similarity measures, and might accurately reflect the semantic relationship between words.

**Key words**: semantic similarity; encyclopedia entry; oriented graph; SimRank

## 0 引言

词语语义相似度计算在自动问答、情报检索、文本聚类等应用中都是一个非常关键的问题[1]。针对该问题,已有不少学者进行了研究,提出了许多定性和定量方法。目前国内外对词语相似度计算的研究策略大体可分为两类[2]:一类是通过本体(Ontology)计算语义相似度,这种方法简单、有效、直观、易于理解,但是这种方法依赖且受制于语义词典的规模和完备性,而且对于语义词典未登录词(专有名词、网络术语)的语义相似度鲜有研究;另一类是基于大规模语料库的统计信息计算词语语义相似度,这种做法假设凡是语义相近的词,他们的上下文也应该相似。这种方法比较客观、综合地反映了词语的语境信息,但是它比较依赖于训练所用的语料库,且存在数据稀疏的问题。

目前,对于第一类方法,传统语义相似度都以 WordNet、知网、同义词词林等作为通用本体。百科知识与这些通用本体比较,其覆盖的范围更加广泛,

**收稿日期**:2013-06-28 **网络出版时间**:2014-03-17 14:02
**网络出版地址**:http://www.cnki.net/kcms/doi/106040/j.issn.1672396122013282.html
**基金项目**:国家自然科学基金资助项目(61152001,61170111);中国科学院自动化研究所复杂系统管理与控制重点实验室开放课题资助项目(20110102);中央高校基本科研业务费专项资金资助项目(SWJTU11ZT08)
**作者简介**:尹坤(1986-),男,湖北黄冈人,硕士研究生,主要研究方向为自然语言理解. E-mail:yinkun6514@163.com
*通讯作者:尹红风(1964-),男,河南夏邑人,教授,博士,主要研究方向为大数据处理,语义网,搜索引擎.
E-mail: hongfeng_yin2002@yahoo.com

知识描述更加全面,信息内容更新速度更加迅速,所以近年来词语语义相似度研究逐渐转向以百科为基础[3]。百科具有很好的半结构化信息,本研究将百科看成一个巨大的图,图中每个结点表示一个词条,每条边表示一个链接,不同的节点之间通过入链和出链相互连接在一起,如图1所示。

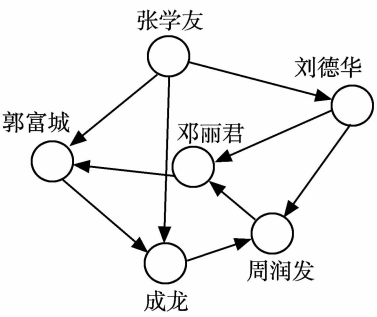


图1 百度百科词条链接图

Fig. 1 The links of Baidu encyclopedia entries

因此,本研究提出基于图论的方法来计算百度百科词条语义相似度。百度百科涉及人物、技术、艺术、地理、体育、科学、文化、历史、生活、社会、自然、经济等领域,包括6 100 293个词条,总的来说内容全面、适用性强。本研究利用百度百科半结构化数据自动获取词语相似度。

## 1 相关工作

文献[4]依据概念之间的上下位关系和同义关系,通过计算两个概念在树状概念层次体系中的距离来得到词语间的语义相似度。文献[5]则采用子概念计算概念间相似度,方法是根据每个子概念对相应概念的贡献,赋予每个子概念一个权重,一个概念的所有子概念权重和为1,子概念间相似度权重的线性和,即为相应概念间的语义相似度。文献[6]提出利用上下文词语同现向量来计算词语间语义相似度。文献[7]利用词语关联分布规范化因子,对互信息中目标词和基词的关联度度量方法进行了修正。文献[8]以基于距离的计算模型为基础,把概念的信息和属性作为两个决策因子计算词语相似度。文献[9]提出计算不同本体中概念间语义相似度的方法,该方法通过比较实例间的相似度获得初始概念间语义相似度,结合影响概率间语义相似度的两个系数,计算出最终的概念间语义相似度。文献[10]对概念实例采用联合分布概率统计的方法,确定概念间语义相似度。文献[11-12]除考虑概念树的路径长度之外,还考虑了路径上词汇的差异性,若一个词的深度越低,那么其特异性越低,权重越高。文献[13]基于概念树提出了新的计算方法。文献[14]除考虑了上下位关系之外,还考虑部分-整体关系和同义-反义关系。文献[15-17]则提出利用语义受限的上下文矩阵来计算词语相似度。

近年来词语语义相似度研究逐渐转向基于百科词条之间的链接信息计算词条语义相似度。文献[18]提出基于维基百科类别体系的文本特征表示方法,该方法是将文本中的词映射到维基百科的分类体系。文献[19]采用群体在线合作编辑的WIKI机制,构建语义知识库。文献[20]提出基于维基百科的语义相似度计算方法,该方法利用页面间的链接信息计算维基百科词条语义相似度。文献[21]通过文档间的相互链接关系计算百科词条相似度,并抽取了近40万对语义相关词。文献[22]利用Web页面之间的超链接信息来计算网页间的相似度。

本研究基于Web链接信息计算词语相似度。2002年,JEH G和WIDOM J对PageRank算法的主要思想加以扩展[23],给出两个节点相似的递归定义:"如果与两个节点相连的其他节点相似,那么这两个节点相似",并提出SimRank算法。2008年LIZORKIN D和VELIKHOV P给出了迭代过程中SimRank算法的精确度分析[24],并对SimRank算法进行了优化,以较小的精度损失来获得较大的性能提升。

随着网络知识库日渐成熟,基于百科词条之间的链接信息计算词条语义相似度逐渐成为主流。

## 2 算法设计

### 2.1 动机

图2来源于百度百科网页,是一个百度百科知识条目,该条目可以分为如下6部分:标题部分,标题部分表明该网页所属的百科词条;正文部分,正文部分详细介绍了该百科词条;属性框部分,该部分为百科词条的结构化信息,用于说明该词条的一些基本信息;维基热度部分,该部分用来表明该词条的重要程度,浏览次数和编辑次数越高则代表该词条越重要,利用词条的维基热度可以剔除一些不常用的词条;相关词条部分,该部分用来说明与该百科词条相关的一些词条;开放分类部分,该部分粗略地表示该词条所属类别。

百度百科词条语义相似度计算主要分为3个步骤:

(1) 本研究将百度百科看成一个巨大的图,图中每个节点表示一个百科词条,本研究称其为百科

词条相关图,本研究用 $\boldsymbol{G}=\langle Vq, Eq\rangle$ 代表该图,其中 $q$ 代表节点的个数,$Vq=\{v\}$ 是节点的集合,$Eq=\{e\}$ 是边的集合,如果存在这一条边 $e=\{v\rightarrow w\}$,则说明词条 $v$ 的相关词条包含词条 $w$ 或者词条 $w$ 的相关词条包含词条 $v$。


图 2　百度百科词条

Fig. 2　The Baidu encyclopedia entries

(2) 如果百科词条 $A$ 的相关词条中包含百科词条 $B$,则在百科词条相关图中,存在两条边 $e=\{A\rightarrow B\}$ 和 $e=\{B\rightarrow A\}$。图 3 是词条“张学友”以及“张学友”的相关词条。


图 3　“张学友”百科词条

Fig. 3　The Baidu encyclopedia entries of “ZHANG Xueyou”

如图 3 所示,词条“张学友”的相关词条包括词条“张国荣”、“蔡依林”、“谭咏麟”,所以在百科词条相关图中,存在节点“张学友”、“张国荣”、“蔡依林”、“谭咏麟”,并存在边 $e=\{$“张学友”→“张国荣”$\}$、$e=\{$“张国荣”→“张学友”$\}$、$e=\{$“张学友”→“蔡依林”$\}$、$e=\{$“蔡依林”→”张学友”$\}$、$e=\{$“张学友”→“谭咏麟”$\}$、$e=\{$“谭咏麟”→“张学友”$\}$。

(3) 基于 SimRank 算法计算百科词条语义相似度。

## 2.2　SimRank 算法理论基础

SimRank 算法的理论基础是:“如果与两个节点相连的其他节点相似,那么这两个节点也相似”。它是 1 种用于衡量结构上下文中个体相似度的方法,直观上的含义是利用已有个体的相似度来推算其他有关个体的相似度,这种计算相似度的策略可以挖掘出图中节点深层次的联系,SimRank 算法形式化定义如下:

$$s(a,b)=\frac{C}{|a||b|}\sum_{i=1}^{|a|}\sum_{j=1}^{|b|}s(I_i(a),I_j(b)),\quad(1)$$

$s(a,b)$ 为对象 $a$ 和对象 $b$ 之间的相似度,即图中两个节点 $a,b$ 之间的相似度。$C$ 是衰减因子,它的取值范围为$[0,1]$,$I(a)$ 是节点 $a$ 在图中的入边集合,$I(b)$ 是节点 $b$ 在图中的入边集合。式(1)表明节点 $a$ 和 $b$ 之间的相似度为它们相关节点相似度之和的平均值,这正好符合先前提出的理论基础:“如果与两个节点相连的其他节点相似,那么这两个节点也相似”。同时由式(1)可知:

$$s(a,b)=1,\ \text{if}\ (a=b)。\quad(2)$$

文献[23]对式(1)有如下扩展,并最终得出 SimRank 算法:如果迭代次数 $k=0$,且词条 $a$ 不等于词条 $b$,则词条 $a$ 和 $b$ 之间的相似度为

$$R_0(a,b)=0,\ \text{if}\ (a!\ b)。\quad(3)$$

如果迭代次数 $k=0$,且词条 $a$ 等于词条 $b$,则词条 $a$ 和 $b$ 之间的相似度为

$$R_0(a,b)=1,\ \text{if}\ (a=b)。\quad(4)$$

如果迭代次数 $k\neq 0$,则词条 $a$ 和词条 $b$ 之间相似度为

$$R_{k+1}(a,b)=\frac{C}{|a||b|}\sum_{i=1}^{|a|}\sum_{j=1}^{|b|}R_k(I_i(a),I_j(b))。\quad(5)$$

当迭代次数 $k$ 趋于无穷大,SimRank 算法能准确得到图中两个节点之间的相似度,然而该算法时间复杂度较高,所以有必要对其进行优化。

## 2.3　SimRank 算法的优化

本研究基于两方面对 SimRank 算法优化:一是通过词条的入度实现对不常用词条的过滤,二是通过文献[24]提出的方法对 SimRank 算法优化。

### 2.3.1　基于词条的入度实现对不常用词过滤。

本研究下载了百度百科人物词条作为实验数据集,共得到 100 784 个百科词条,如果本研究用上述方法构建百科词条相关图,图中的节点会有 100 784 个。运用 SimRank 算法计算一个拥有 100 784 个节点的图中任意两个节点相似度,时间复杂度非常高,所以本研究有必要删除那些不重要或者不常用的节点。

PageRank 算法表明:如果一个网页的链接数越多,则说明该网页越重要。类似,本研究可以做出这样一个假设:一个词条的相关词条越多、入度越大,则说明该词条越重要。

假设 $Vq=\{v\}$ 是词条相关图节点的集合,本研究定义这样一个集合 $A_q=\{a\}$,其中 $a_i$ 是百科词条

$v_i$ 入度。

图4为入度与次数关系图。图4表明不重要或者不常用的百科词条占多数。所以本研究设定一个阈值,如果词条 $v_i$ 的入度 $\alpha_i > \beta$,本研究则认为词条 $v_i$ 是重要的,应保留。如果本研究取 $\beta = 8$,最后会得到27 893个词条即百科词条相关图有27 893个节点,很好地降低了算法的运行时间。

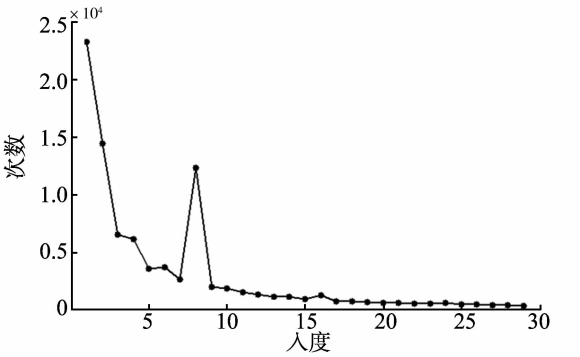


图4 入度与次数关系图

Fig. 4 The relationships between in-degree and frequency

为了验证本方法的合理性,本研究从入度≤8的节点集合中随机选择1 000个词条作为测试词条,这1 000个词条可能不全是人物词条,因为本研究是通过开放分类树判断该词条是否为人物词条,而开放分类树只能粗略地表示一个词条所属类别,其中有部分错误。然后利用词条的维基热度和人的经验判断该词条是否为重要词条,通过实验发现,这1 000个词条中不重要或者不是人物的词条所占比率为93.5%。

**2.3.2** SimRank近似估计

文献[24]给出了迭代过程中SimRank算法的精确度分析,提出了以较小精度损失来获得较大性能提升的方法,并定义了如下公式:

$$s(a,b) - R_k(a,b) \leqslant C^{(K+1)}, \tag{6}$$

其中 $s(a,b)$ 为词条 $a$ 和词条 $b$ 的相似度,$R_k(a,b)$ 为词条 $a$ 和词条 $b$ 在第 $k$ 次迭代时SimRank值,$C$ 为衰减因子,$C^{(k+1)}$ 为在第 $k$ 次迭代时SimRank算法的最大误差。所以如果设 $C = 0.8$,最大误差为0.000 1,则只需49次迭代就能得到相当精确的结果。

## 3 实验结果与分析

### 3.1 实验数据

本实验中的数据采用百度百科2012-02-10的网页数据文件,该数据记录了所有在线百度百科的词条信息。

实验中会用到百度百科网页的标题和相关词条,所以需对下载的网页进行网页信息抽取。

### 3.2 实验评测与结果

由于词语相似度计算,还没有一个优劣评价标准,所以本研究采用3种方法来衡量算法的性能:采用人工判别的方法,采用统计准确率 $P$ 和查全率 $R$ 的方法,采用对比的方法。

**3.2.1** 采用人工判别的方法

表1列举了部分词条在迭代次数 $K$ 变化时的词语相似度归一化结果。

表1 词条语义相似度

Table 1 Semantic similarity of entries

| 词条一 | 词条二 | $K=1$ | $K=10$ | $K=20$ | $K=30$ |
|---|---|---|---|---|---|
| 刘德华 | 张学友 | 0.000 | 0.809 | 0.856 | 0.857 |
| 刘德华 | rain | 0.952 | 0.960 | 0.968 | 0.969 |
| 刘德华 | 汤唯 | 0.000 | 0.833 | 0.873 | 0.874 |
| 刘德华 | 李安 | 0.952 | 0.984 | 0.987 | 0.987 |
| 刘德华 | 郑伊健 | 1.000 | 0.992 | 0.994 | 0.994 |
| 刘德华 | 郑中基 | 0.000 | 0.698 | 0.778 | 0.775 |
| 老舍 | 阿Q正传 | 0.000 | 0.590 | 0.680 | 0.702 |
| 老舍 | 钱钟书 | 0.000 | 0.690 | 0.760 | 0.770 |

表1表明利用本方法计算的结果,随着迭代次数的增加,SimRank值趋于稳定,当迭代次数超过20,词条相似度变化幅度趋近于0,并且词条相似度基本符合客观事实,但也有小部分错误,比如老舍-阿Q正传、刘德华-李安等。

**3.2.2** 采用准确率和查全率的方法

本研究利用准确率 $P$ 和查全率 $R$ 两项指标来衡量算法的性能。

$$P = \frac{C_1}{C_2}, \tag{7}$$

$$R = \frac{C_1}{C_3}, \tag{8}$$

其中,$C_1$ 代表抽取出来的正确的同义词词对数量,$C_2$ 代表抽取出来的同义词词对数量,$C_3$ 表示语料含有同义词词对数量。图5给出了查全率、准确率随迭代次数变化的趋势。

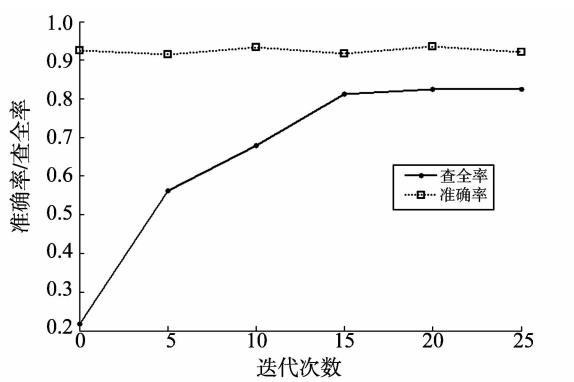


图5 SimRank迭代结果

Fig. 5 The iteration results of SimRank

如图5所示,当迭代次数较低的时候,抽取的同

义词词对较少,但正确率较高;当迭代次数增加,在保证准确率的同时,查全率也得到相应的提升;当迭代超过 20 次,simRank 值趋于稳定,查全率和准确率也趋于稳定。

图 5 表明本方法即使在迭代次数较低的情况下,也能保证抽取同义词词对的准确率。因为百科建设者在不停地采用各种方法提高百科知识的准确度和覆盖度,所以百科词条的相关词条准确度也得到很大提升。而在迭代次数较低的情况下查全率也较低,因为 SimRank 算法的迭代次数和遍历图的深度有关,当迭代次数较低时,遍历的深度也较低,所以如果图中两个节点距离较远,则不能准确地计算出它们之间的相似度。当迭代次次数增加时,两个词条的 SimRank 值逐渐趋于稳定,并接近真实值,因为这种计算相似度的策略可以挖掘出图中节点深层次的联系,可以考虑全图所有节点的特性。

**3.2.3**　采用对比的方法

本方法既不受通用本体规模的限制,也解决了基于大型语料库计算词语相似度数据稀疏、计算量大、准确率低等问题。为了说明本方法的优越性,本研究进行了两个对比实验:同基于通用本体计算词语相似度的方法进行对比,同基于大型语料库计算词语相似度的方法进行对比。

(1)同通用基于本体计算词语相似度的方法对比

本研究实现并对比了 3 种基于本体计算词语相似度的方法:文献[1-2]提出的基于知网的计算方法,文献[25]提出的基于同义词词林的计算方法。

方法一为文献[2]提出的方法,方法二为文献[1]提出的方法,方法三为文献[25]提出的方法,方法四为本研究提出的方法。表 2 中,通用本体的登录词,用下划线表示,比如:北京、武汉 、爸爸、妈妈。表 2 列举了部分词条对比结果,其中方法四(本研究方法)的结果为 SimRank 迭代 49 次后归一化的结果。四种方法的词语相似度对比结果见表 2。

表 2　词语相似度对比结果

Table 2　The results of words similarity comparison

| 词语一 | 词语二 | 方法一 | 方法二 | 方法三 | 方法四 |
|---|---|---|---|---|---|
| 妈妈 | 母亲 | 1.000 | 1.000 | 1.000 | 0.864 |
| 母亲 | 爹 | 1.000 | 1.000 | 1.000 | 0.701 |
| 母亲 | 爸爸 | 1.000 | 1.000 | 1.000 | 0.473 |
| 餐馆 | 饭店 | 1.000 | 1.000 | 1.000 | 0.943 |
| 北京 | 武汉 | 0.860 | 0.900 | 0.940 | 0.952 |
| 北京国安 | 武汉 | 0.000 | 0.430 | 0.000 | 0.000 |
| 足球俱乐部 | 足球 | 0.000 | 0.260 | 0.000 | 0.000 |
| 阿森纳 | 利物浦 | 0.000 | 0.000 | 0.000 | 0.954 |
| 齐达内 | 卡卡 | 0.000 | 0.000 | 0.000 | 0.962 |
| Java | C + + | 0.000 | 0.000 | 0.000 | 0.787 |
| CPU | 硬盘 | 0.000 | 0.000 | 0.000 | 0.742 |

表 2 表明:方法一、方法二、方法三对于通用本体登录词的计算相对合理,其结果也稍优于本方法。因为通用本体由人工编撰,有着良好的层次化结构,但人工编撰耗时耗力,所以其规模相对较小,据统计知网仅收录 66 191 个词条,且不包括专有领域词汇,比如:阿森纳、利物浦、Java、C ++ 等。方法一和方法三对未登录词无法处理,方法二虽有研究未登录词语义相似度计算,但也只是利用词素切割的方法,结果不是很理想。而本研究基于百度百科知识,百度百科知识拥有 6 100 293 个条目,词语的覆盖比较全面,且实时更新。所以本研究提出的方法虽然对于通用本体的登录词计算结果稍差,但覆盖范围广,能较为准确的计算专有领域词的相似度。

(2)同基于大型语料库计算词语相似度的方法对比

由于资源的限制,本研究只对文献[7]提出的方法进行对比。本研究利用准确率 $P$、查全率 $R$ 的方法来评价这两个算法的性能。

图 6 表明本方法的查全率、准确率均高于文献[7]的查全率、准确率。但由于本研究与文献[7]阈值选取标准不同,直接进行性能指标值对比没有多大意义。为了更好地说明本方法的性能,本研究选取文献[7]性能指标达到的最优值同本文方法在不同阈值下的性能值作对比。由图 6 可知,当文献[7]选择阈值 $a=0.02$ 时,文献[7]的性能最优,所以本研究将文献[7]的阈值 $a=0.02$、本方法阈值 $a=0.80, K=10$、本方法阈值 $a=0.80, K=20$、本方法阈值 $a=0.80, K=30$、本方法阈值 $a=0.80, K=40$ 等 5 个阈值下的综合性能值做对比,对比结果见表 3。

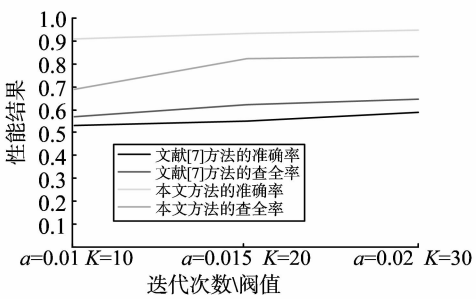


图 6　准确率和查全率结果

Fig. 6　The recall and precision results

表 3　对比结果

Table 3　The results comparison

| 方法 | 阈值 | 同义词类 | |
|---|---|---|---|
| | | 准确率 $P$ | 查全率 $R$ |
| 方法一 | $a=0.02$ | 0.700 | 0.639 |
| 方法二 | $a=0.80$ $k=10$ | 0.910 | 0.680 |
| 方法二 | $a=0.80$ $k=20$ | 0.933 | 0.830 |
| 方法二 | $a=0.80$ $k=30$ | 0.926 | 0.840 |
| 方法二 | $a=0.80$ $k=40$ | 0.934 | 0.860 |

表3中,方法一为文献[7]提出的方法,方法二为本研究提出的方法。$a$ 为相似度阈值,$K$ 为 SimRank 算法迭代次数。由表3可知,本研究提出的方法即使在迭代次数较低时,准确率和查全率也优于基于大型语料库计算词语相似度的方法最优值,而随着迭代次数逐渐增加,其优势更加明显。因为,基于大型语料库计算词语相似度的方法是根据自由文本的上下文信息计算词语相似度,没有其他额外信息,所以会存在数据稀疏、计算量大、准确率偏低等问题。而本方法基于百度百科相关词条计算词语相似度,百度百科相关词条由百科建设者们共同编辑,百科建设者们编辑这些相关词条给了我们一定的指导信息,所以利用这些指导信息,本研究能有效地解决数据稀疏、准确率低等问题。

## 4 结论

本研究提出了基于 SimRank 算法的百科词条语义相似度计算方法。从实验结果来看,基于百科计算词语相似度的方法既能有效的解决基于大规模语料库统计方法所存在的数据稀疏、计算量大等问题又能有效的解决基于通用本体方法所存在的词汇覆盖面小等问题。然而,本研究所提出的方法适合于相关词条较多的百科词条,对于相关词条较少的百科词条计算结果还不是很准确,这些问题将在以后的研究中改进。由于百度百科知识条目还存在着其他特征,下一步将可以利用百度百科网页的多特征信息计算百科词条语义相似度。

**参考文献:**


[1] 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.
XIA Tian. Study on Chinese words semantic similarity computation[J]. Compute Engineering, 2007, 33(6): 191-194.

[2] LIU Qun, LI Sujian. Word similarity computing based on Hownet[C]//Proceedings of Computational Linguistics and Chines Language Processing. Taipei, China: [s.n], 2002:59-76.

[3] 刘宏哲, 须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学, 2012, 39(2):8-13.
LIU Hongzhe, XU De. Ontology based semantic similarity and relatedness measures review[J]. Compute Science, 2012, 39(2):8-13.

[4] 吴健, 吴朝晖, 李莹, 等. 基于本体论的词汇语义相似度 Web 服务发现[J]. 计算机学报, 2005, 28(4):595-602.
WU Jian, WU Zhaohui, LI Ying, et al. Web service discovery based on ontology and similarity of words[J]. Chines Journal of Computers, 2005, 28(4):595-602.

[5] WONG A K Y, RAY P, PARAMESWARAN N, et al. Ontology mapping for the interoperability problem in network management[J]. IEEE Journal on Selected Areas in Communication, 2005, 23(10):2058-2068.

[6] 张涛, 杨尔弘. 基于上下文词语同现向量的词语相似度计算[J]. 电脑开发与应用, 2005, 18(3):41-43.
ZHANG Tao, YANG Erhong. Study on word similarity base on contextual word co-occurrence vector[J]. Computer Development & Applications, 2005, 18(3):41-43.

[7] 赵军, 胡栓柱, 樊兴华, 等. 一种新的词语相似度计算方法[J]. 重庆邮电大学学报:自然科学版, 2009, 21(4):528-532.
ZHAO Jun, HU Shuanzhu, FAN Xinghua, et al. Word similarity computation based on word link distribution [J]. Journal of Chongqing University of Posts and Telecommunications:Natural Science Edition, 2009, 21(4): 528-532.

[8] 黄果, 周竹荣, 周亭, 等. 基于领域本体的语义相似度计算研究[J]. 计算机工程与科学, 2007, 29(5):112-118.
HUANG Guo, ZHOU Zhurong, ZHOU Ting, et al. Research on the domain ontology based semantic similarity computing[J]. Computer Engineering & Science, 2007, 29(5):112-118.

[9] 王家琴, 李仁发, 李仲生, 等. 一种基于本体的概念语义相似度方法的研究[J]. 计算机工程, 2007, 33(11): 201-204.
WANG Jiaqin, LI Renfa, LI Zhongsheng, et al. Research on method of concept semantic similarity based on ontology[J]. Computer Engineering, 2007, 33(11): 201-204.

[10] DOAN A H, MADHAVAN J, DOMINGOS P, et al. Learning to map between ontologies on the semantic Web[C]//Proceedings of the 11th International Conference on World Wide Web. New York, USA: ACM, 2002:662-673.

[11] LEACOCK C, CHODOROW M. Combining local context and WordNet similarity for word sense identification [M]. [S.l]:[s.n.], 1998:265-283.

[12] WU Zhibiao, MARTHA Palmer. Verbs semantics and lexical selection[C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA:IEEE, 1994:133-138.

[13] LI Y, BANDAR Z A, MCLEAN D. An approach for measuring semantic similarity between words using multiple information sources[J]. Knowledge and Data Engineering, IEEE Transactions on, 2003, 15(4): 871-

882.

[14] YANG Dongqiang, DAVID M W Powers. Measuring semantic similarity int the taxonomy of WordNet[C]// Proceedings of the 28th Australasian Conference on Computer Science. Darlinghurst, Australia: ACM, 2005, 102:315-332.

[15] 王石, 曹存根, 裴亚军,等. 一种基于搭配的中文词汇语义相似度计算方法[J] 中文信息学报, 2013, 27(1):7-15.
WANG Shi, CAO Cungen, PEI Yajun, et al. A collocation-based method for semantic measure for Chinese words[J]. Journal of Chinese Information Processing, 2013, 27(1): 7-15.

[16] WANG Shi, CAO Cungen, CAO Yanan, et al. Measuring taxonomic similarity between words using restrictive context matrices[C]//Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2008). NJ, USA: IEEE, 2008:193-197.

[17] Herbert Rubenstein, John B Goodenough. Contextual correlates of synonymy [J]. Communications of the ACM, 1965, 8(10): 627-633.

[18] 王锦, 王会珍, 张俐, 等. 基于维基百科类别的文本特征表示[J]. 中文信息学报, 2011, 25(2):27-32.
WANG Jin, WANG Huizhen, ZHANG Li, et al. Text representation by the Wikipedia category[J]. Journal of the Chinese Information Processing, 2011, 25(2):27-32.

[19] 张海粟, 马大明, 邓智龙, 等. 基于维基百科的语义知识库及其构建方法的研究[J]. 计算机应用研究, 2011, 28(8):2807-2812.
ZHANG Haisu, MA Daming, DENG Zhilong, et al. Semantic knowledge bases construction based on Wikipedia[J]. Application Research of Computes, 2011, 28(8):2807-2812.

[20] 盛志超, 陶晓鹏. 基于维基百科的语义相似度计算方法[J]. 计算机工程, 2011, 37(7):193-196.
SHENG Zhichao, TAO Xiaopeng. Semantic similarity computing method based on Wikipedia[J]. Computer Engineering, 2011, 37(7):193-196.

[21] 李赟, 黄开研, 任福继, 等. 维基百科的中文语义相关词获取及相关度分析计算[J]. 北京邮电大学学报, 2009, 32(3):109-112.
LI Yun, HUANG Kaiyan, REN Fuji, et al. Wikipedia based semantic related Chinese words exploring and relatedness computing[J]. Journal of Beijing University of Posts and Telecommunication, 2009, 32(2):109-112.

[22] DEAN J, HENZINGER M R. Finding related pages in the World Wide Web[J]. Computer Networks, 1999, 31:1467-1479.

[23] JEH G, WIDOM J. SimRank: a measure of structural context similarity[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2002:538-543.

[24] LIZORKIN D, VELIKHOV P, GRINEV M, et al. Accuracy estimate and optimization techniques for SimRank computation[J]. Proceedings of the VLDB Endowment, 2008, 1(1):422-433.

[25] 田乐九, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报:信息科学版, 2010, 28(6): 602-608.
TIAN Lejiu, ZHAO Wei. Words similarity algorithm based on Tongyici Cilin in semantic Web adaptive learning system[J]. Journal of Jilin University: Information Science Edition, 2010, 28(6):602-608.

(编辑:胡春霞)