第 24 卷第 1 期　　北方工业大学学报　　Vol. 24 No. 1
2012 年 3 月　　J. NORTH CHINA UNIV. OF TECH.　　Mar. 2012

基于维基百科的中文命名实体关联度计算*

刘高军　马砚忠　段建勇
（北方工业大学信息工程学院，100144，北京）

摘　要　利用维基百科中命名实体页面的超链接信息，根据页面共现超链接计算得到命名实体的关联度. 该方法计算得到的命名实体关联度与人工标注的结果比较，二者基本吻合，表明该方法计算得到的命名实体关联度具有较高的准确率.

关键词　命名实体；维基百科；关联度
分类号　TP391.1

实体是指现实世界中存在的特定的事实信息[1]. 命名实体主要包括人名、机构名、地名、时间、专有名词等. 命名实体中包含着句子的关键信息，例如，“乔布斯的生平”的关键信息是“乔布斯”这个命名实体.

命名实体关联度计算在自然语言处理、信息检索、实体分类、自动应答、词义排歧等方面都有很重要的应用价值[2]. 例如，在信息检索中，利用命名实体关联度计算能够提高信息检索的准确率和召回率；在问答系统中，答案和问句的符合程度可以通过计算两者含有命名实体之间的关联度来衡量.

本文研究的基于维基百科的中文命名实体关联度计算方法，是利用中文维基百科的链接信息来计算得到中文命名实体的关联度，是为建立自动问答系统进行的一些基础研究.

1　中文维基百科

中文维基百科正式成立于 2002 年 10 月 24 日，截至到 2011 年 4 月，中文维基百科已拥有 35 万个条目[3]. 中文维基百科的基本条目是网页（又称文章），每个页面都定义和描述一个实体，并通过超链接与维基百科内部或外部的其他网页相连接. 超链接的作用是引导读者获取网页中提到的其它命名实体的信息. 截至到 2010 年 5 月，中文维基百科的内部链接已达 8.0Mb[4]，图 1 显示了近 3 年中文维基百科内部链接的增长情况. 用户浏览维基百科，可以通过这些链接来查看链接指向的命名实体，统计显示每个维基百科页面平均含有 4 个链接[4].

每个维基百科页面都有唯一的标识符，它由词、下划线和空格组成，使用括号“()”解释说明标识符. 例如，“库比蒂诺”对应网页的唯一标识符为“库比蒂诺（加利福尼亚州）”，括号中的内容是解释说明库比蒂诺是加利福尼亚州的一个市.

维基百科页面的超链接是由唯一标识符和锚文本组成，例如，“乔布斯是[苹果公司]的创办人之一，并曾任苹果公司的[董事会主席]、[首席运行官].”是维基百科中的一个句子，句子中含有指向苹果公司、董事会主席和首席运行官的超连接，在维基百科中使用括号“[]”和唯一标识符来表示超链接.

维基百科的条目是由众多的志愿者参与创建和编辑的，对于同一个页面，不同的编辑者使

收稿日期：2011-11-03
* 国家自然科学基金项目（61103112）、国家社会科学基金项目（11CTQ036）.
第一作者简介：刘高军，副教授. 研究方向：数据库技术、数据挖掘等.

用不同的标识符来标识,这就涉及到了维基百科的重定向机制. 例如,奥运会和奥林匹克运动会为两个不同的标识符,但是经过重定向后,奥运会与奥林匹克运动会连接的是同一个页面.

维基百科有消歧机制,例如,对于"博士"这个实体有两个意思:清朝时的"博士"指官位,而现在的"博士"指学位[5]. 维基百科有专门的消歧页面,消歧页面是由那些模棱两可的标识符和指向不同页面的超链接所组成,然后通过使用括号"()"来区分标识符所对应的不同页面. 例如,对于"博士"可以通过"博士(官)"和"博士(学位)"进行区分.

维基百科的条目都是人工编辑的,其质量都很高,因此我们可以利用这些高质量条目含有的链接信息,以及维基百科的重定向和消歧机制来计算命名实体的关联度.

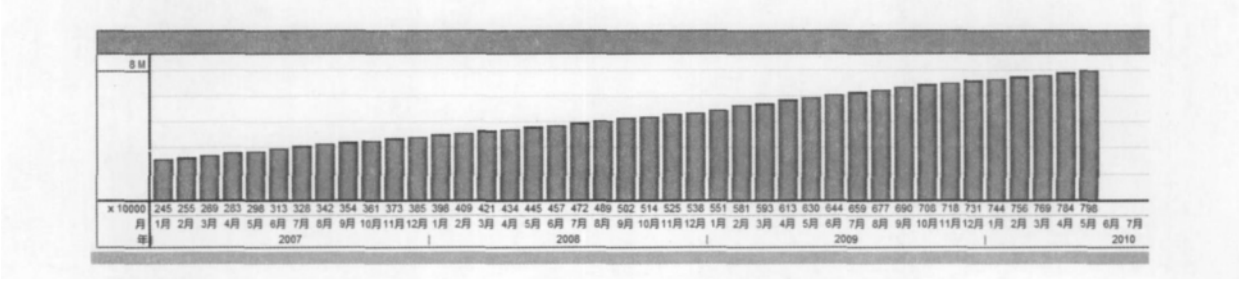

图 1　中文维基百科内部链接总数

2　基于维基百科的命名实体关联度计算

2.1　命名实体关联度

命名实体关联度是指命名实体的相关性. 关联度是一个数值,取值范围为(0,1). 一个命名实体与本身的关联度为 1,如果两个命名实体的相关性为 0,则它们的关联度为 0.[6]

2.2　维基百科链接分析

如表 1 所示,对于"乔布斯"和"苹果公司"这两个命名实体,表中列出了它们在维基页面中含有的部分链接,其中"Macintosh"和"麦金塔"这两个标识符虽然不同,但经过重定向后,它们都连接到"麦金塔电脑"这个命名实体页面. 对于"财富"和"时代"这两个标识符,"财富"有两个意思分别是:1)经济学上,财富是指物品按价值计算的富裕程度,或对这些物品的控制和处理的状况;[7] 2)《财富》杂志. "时代"也有两个意思分别是:1)历史上以经济、政治、文化等状况为依据而划分的时期;2)《时代》杂志. 这里通过对"财富(杂志)"和"时代(杂志)"进行消歧,二者分别连接到财富(杂志)和时代(杂志)这两个命名实体的页面.

计算结果得"苹果公司"这个命名实体页面含有 265 个链接,"乔布斯"这个命名实体页面含有 140 个链接,经过维基百科的重定向和消歧以后,二者所含链接指向命名实体相同的链接为 84 个. 如表 1 所示,斜体部分就是"苹果公司"和"乔布斯"这两个页面所含链接指向的命名实体相同部分.

2.3　命名实体关联度计算方法

根据上述分析,我们发现"苹果公司"和"乔布斯"这两个命名实体在维基百科中的页面含有很多共现链接,这样我们根据共现链接的数量可以计算命名实体的关联度[8-9]. 这里所说的共现链接是指链接指向的命名实体相同的链接,计算公式如下:

$$\mathrm{sim}(EN1,EN2)=\frac{\log(\max(a,b))-}{\log(w)-}\rightarrow \leftarrow\frac{\log(c)}{\log(\min(a,b))},$$

$EN1$、$EN2$ 表示 2 个命名实体;a 表示命名实体 $EN1$ 页面所含链接数;b 表示命名实体 $EN2$ 页面所含链接数;c 表示 $EN1$ 和 $EN2$ 页面共现链接数;w 为全部的维基页面. 计算方法如下:

Sim(A,B)

A:$EN1$ 页面含有的超链接向量集合

B:$EN2$ 页面含有的超链接向量集合

$Redirect(A)$:重定向 A

$Redirect\ (B)$:重定向 B

For $i=1,2,\cdots,|A|$

对 i 执行 R(重定向)操作,修改 $A(i)$;

End

For $j=1,2,\cdots,|B|$

对 j 执行 R(重定向)操作,修改 $B(j)$;

End

$C=Co\text{-}occurrence(A,B)$:计算 A 和 B 的共现超链接存入 C

$S=(\log(\max(|A|,|B|))-\log(|C|))/(\log(w)-\min(|A|,|B|))$:计算相关度

Return S

表1 苹果公司和乔布斯两个命名实体页面含有的部分链接

命名实体	页面含有的链接	链接指向的实体	命名实体	页面含有的链接	链接指向的实体
苹果公司	美国	美国	乔布斯	董事会主席	董事长
	库比蒂诺	库比蒂诺(加利福尼亚州)		2006年	2006年
	Apple II	Apple II		迪士尼、迪士尼公司	华特迪士尼公司
	Macintosh	*麦金塔电脑*		*麦金塔*	*麦金塔电脑*
	iPod	*iPod*		*iPod*	*iPod*
	iTunes商店	*iTunes Store*		*iTunes Store*	*iTunes Store*
	iPhone	*iPhone*		*iPhone*	*iPhone*
	iPad	*iPad*		*iPad*	*iPad*
	CNN	CNN		财富	财富(杂志)
	中石油	中国石油		时代	时代(杂志)

3 实验与评价

3.1 实验设计

挑选100对(200个)常见的命名实体,人工对其关联度进行了标注.标注方法是将100对命名实体分别放入3个搜索引擎Google、百度和雅虎进行检索.通过下面公式计算出它们的关联度:

$$\mathrm{sim}(a,b)=\frac{|a\cap b|}{|a|\cup|b|}.$$

a、b 表示2个命名实体;$|a|$、$|b|$ 表示 a、b 这2个命名实体分别在搜索引擎中检索到的结果数量,$|a\cap b|$ 表示将 a,b 这两个命名实体共同放入搜索引擎中检索到结果数量.采用同样的方法将这些命名实体分别放入Google、百度和雅虎进行检索,计算它们的关联度的平均值.其关联度的分布图如图2所示,横轴表示关联度范围,竖轴表示实体对个数.

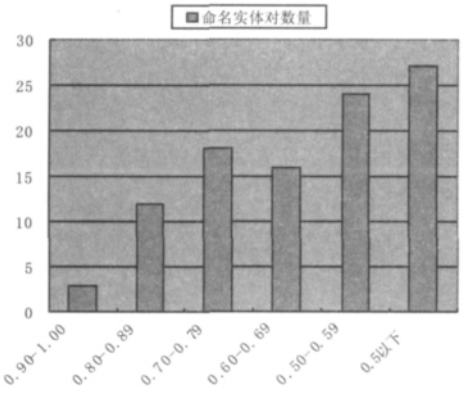

图2 人工标注的命名实体关联度分布图

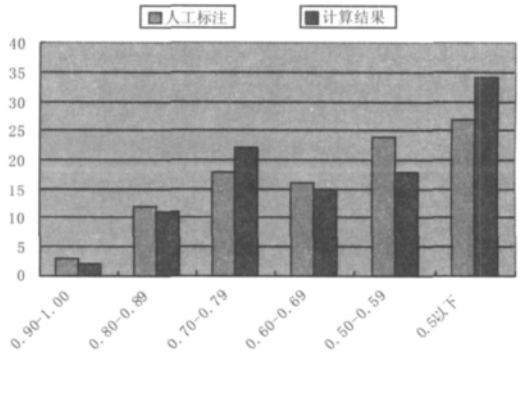

图3 计算得到的命名实体的关联度与人工标注分布比较图

3.2　结果评价

我们计算了几个常见的命名实体,如苹果公司、乔布斯、中华人民共和国、北京、内蒙古、牛奶、面包的关联度,计算结果如表 2 所示.从表 2 可以看出,计算结果与人们认知结果相吻合.如乔布斯是苹果公司的创建人,因此它们有很高的关联度;因苹果公司是一家美国公司,所以它与中华人民共和国的关联度不高.

表 2　采用本方法计算得到的命名实体关联度

命名实体 1	命名实体 2	关联度值
苹果公司	乔布斯	0.86
苹果公司	中华人民共和国	0.5
北京	中华人民共和国	0.8
北京	内蒙古	0.75
牛奶	面包	0.51
牛奶	乔布斯	0

图 3 给出了用本文方法计算得到的 100 对命名实体的关联度与人工标注分布比较图.从图中可以看出,本方法与人工标注的结果基本吻合,有 72 对命名实体的关联度所在区间与人工标注完全一致,在不一致的 28 对中有 13 对命名实体与人工标注结果的差值小于 0.1,因此本试验的准确率可达到 72%以上.

4　结语

本文提出了一种基于维基百科的计算命名实体关联度的方法,该方法充分利用了维基百科中的超链接信息,通过命名实体的维基百科页面含有的共现超链接计算得到命名实体的关联度.通过与人工标注的命名实体关联度进行对比,该方法计算得到的命名实体的关联度具有较高的准确率,计算结果与人们认知结果基本吻合.

参　考　文　献

[1] 牟晋娟,包宏.中文实体关系抽取研究[J].计算机工程与设计,2009,30(15):3587-3590

[2] 田久乐,赵蔚.基于同义词词林的词语关联度计算方法[J].吉林大学学报:信息科学版,2010,28(6):602-608

[3] http://wikipedia.jaylee.cn

[4] http://stats.wikimedia.org/ZH/TablesDatabaseLinks.htm

[5] http://zh.wikipedia.org/zh-cn/Wikipedia_talk:%E6%B6%88%E6%AD%A7%E4%B9%89

[6] 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,中国台北:2002

[7] http://wiki.mbalib.com/wiki/%E8%B4%A2%E5%AF%8C

[8] Rada Mihalcea, Andras Csomai. Wikify! Linking Documents to Encyclopedic Knowledge [C]. CIKM, 2007

[9] Olena Medelyan, Ian H Witten, David Milne. Topic Indexing with Wikipedia[C]. AAAI, 2008

Calculating Correlation of Chinese Named Entity Based on Wikipedia

Liu Gaojun　Ma Yanzhong　Duan Jianyong

(Col. of Information Engineering, North China Univ. of Tech., 100144, Beijing, China)

Abstract　In this paper, the correlation of the named entity is obtained by calculating the common hyperlinks in Wikipedia pages. The result calculated by this method is consistent with the manual coding, which shows that this method achieves higher accuracy.

Key Words　named entity; Wikipedia; correlation