语义相似度的计算方法研究与分析

魏凯斌,冉延平,余 牛
(天水师范学院,甘肃 天水 741001)

摘 要:语义相似度计算在信息检索、信息抽取、文本分类、词义排歧、基于实例的机器翻译等很多领域中都有广泛的应用。特别是近几十年来随着 Internet 技术的高速发展,语义相似度计算成为自然语言处理和信息检索研究的重要组成部分。介绍了几种典型的语义相似度的计算方法,总结了语义相似度计算的两类策略,其中重点介绍了一种基于树状结构中语义词典 Hownet 的语义相似度计算方法,最后对两类主要策略进行了简单的比较。

关键词:语义相似度;语义距离; 知网; 语料库
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2010)07-0102-04

The Research and Analysis of Computing Methods on Semantic Similarity

WEI Kai-bin, RAN Yan-ping, YU Niu
(Tianshui Normal University, Tianshui 741001, China)

Abstract: Semantic similarity is broadly used in many applications such as information retrieval, information extraction, text classification, word sense disambiguation, example-based machine translation and so on. Especially with the rapid development of Internet technology in recent decades, calculation of semantic similarity has always been an important part of natural language processing and information retrieval research. Introduce several main methods of calculating semantic similarity , then two strategies of semantic similarity measurement are summarized, and focus on the Hownet based on the structure of tree and use them to calculate the semantic similarity , and finally the two strategies are easily compared.

Key words: semantic similarity; semantic distance; CNKI; corpus

0 引 言

语义相似度计算研究的是用什么样的方法来计算或比较两个词语的相似性。自然语言的词语之间有着非常复杂的关系,在实际应用中,有时需要把这种复杂的关系用一种简单的数量来度量,而语义相似度就是其中的一种。

词语的语义相似度计算主要有两种方法:一类是通过语义词典,把有关词语的概念组织在一个树形的结构中来计算;另一类主要是通过词语上下文的信息,运用统计的方法进行求解。对于前一类基于树状层次结构的计算语义相似度方法的研究已经比较成熟,国外的 Dekang Lin[1], Rudi L. Cilibrasi[2]等都给出了自己的比较合理的语义相似度计算公式和方法;国内这方面起步较晚,但发展很快,董振东[3]、刘群、李素建[4]等在这方面的研究做了很多开创性的工作,李峰[5]、李鹏[6]、李熙[7]、杨哲[8]、夏天[9]、张明宝[10]等后来者做了很多补充性和改进性的工作。针对以上研究现状,笔者对当前的语义相似度研究成果进行了简单的归纳和总结,然后对相关方法进行了简单比较,并提出了研究的应用方向,以供相关研究人员参考和应用。

收稿日期:2009-10-30;**修回日期**:2010-02-02
基金项目:甘肃省教育科研项目(0808-07)
作者简介:魏凯斌(1978-),男,甘肃天水人,讲师,硕士,主要研究方向为智能信息系统、人工智能。

1 语义相似度

Dekang Lin 认为任何两个词语的相似度取决于它们的共性(Commonality)和个性(Differences),然后从信息论的角度给出了定义公式:

$$\mathrm{Sim}(A,B)=\frac{\log p(\mathrm{Common}(A,B))}{\log p(\mathrm{description}(A,B))} \quad (1)$$

其中,分子表示描述 A,B 共性所需要的信息量;分母表示完整地描述 A,B 所需要的信息量。

刘群、李素建[4] 以基于实例的机器翻译为背景,认为语义相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语,如果在不同的上下文中可以互相替换且不

改变文本的句法语义结构的可能性越大,二者的相似度就越高,否则相似度就越低。

对于两个词语 W_1, W_2 如果记其相似度为 Sim(W_1, W_2),其词语距离为 Dis(W_1, W_2),根据刘群、李素建的公式:

$$\mathrm{Sim}(W_1, W_2) = \frac{\alpha}{\alpha + \mathrm{Dis}(W_1, W_2)} \quad (2)$$

其中 α 是一个可调节的参数。α 的含义是:当相似度为 0.5 时的词语距离值。

词语距离和词语相似度是一对词语的相同关系特征的不同表现形式,如果两个概念之间的语义距离越近,就认为它们越相似[5],因此二者之间可以给出一个简单对应关系:

$$\mathrm{Sim}(W_1, W_2) = \frac{k}{\mathrm{Dis}(W_1, W_2)} \quad (3)$$

其中,Dis(W_1, W_2)为树中 W_1, W_2 所代表的结点在树中的距离,k 为比例系数。

一般地说,相似度一般被定义为一个 0 到 1 之间的实数。特别地,当两个词语完全一样时,它们的相似度为 1;当两个词语是完全不同的概念时,它们的相似度接近于 0。

2　语义相似度的计算方法

词语距离有两类常见的计算方法,一种是根据某种世界知识(Ontology)或分类体系(Taxonomy)来计算,一种利用大规模的语料库进行统计。

2.1　根据分类体系计算词语语义距离的方法

该方法又称基于树的语义相似度研究方法,基于树的语义相似度计算的算法大体上分为两种:一是基于距离的语义相似性测度;二是基于信息内容的语义相似性测度。一般是利用一部语义词典(如 Wordnet, Hownet),语义词典都是将所有的词组织在一棵或几棵树状的层次结构中[6]。在一棵树状图中,任何两个结点之间有且只有一条路径。于是,这条路径的长度就可以作为这两个词语概念间语义距离的一种度量;而且随着概念所处结点越深,其所包含的语义信息越丰富,越能准确地决定概念的性质,它们对语义相似度起着决定作用。

2.1.1　基于树状层次计算语义相似度的基本思想

根据公式(3)给出的定义,计算思想是以边为距离来计算语义相似度。如果树状语义网中所有的边即树的分支是等长的,那么边的数目可以作为距离的测度。假定要确定词语 W_1, W_2 之间的语义相似度,可以在该语义网中首先找到包含待比较词的那些子概念(或义原)。在此情况下, W_1, W_2 之间的语义相似性可以用连接这两个概念之间的最短路径来表示。例如,在图 1(取自 Wordnet 本体中的一小部分)中,boy 和 girl 之间的最短路径是 boy – male – person – female – girl,最小路径长度为 4。而 teacher 和 boy 之间的最小路径长度为 6。因此,girl 比 teacher 在语义上更接近于 boy。该测度算法在基于 Wordnet 的语义网中获得了较好的计算结果。

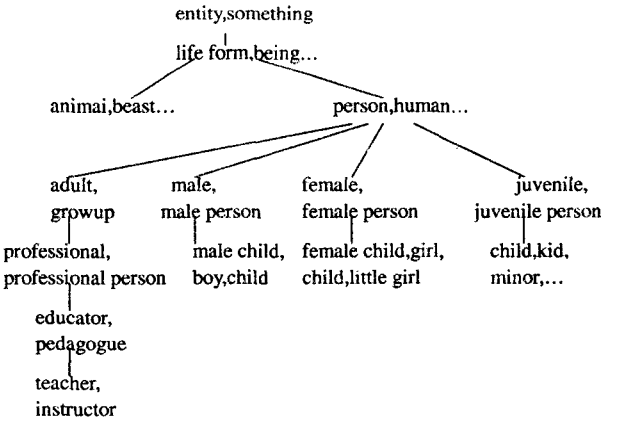

图 1　部分词义网络

2.1.2　基于《知网 Hownet》的语义相似度计算

《知网》中有两个主要的概念:“概念”与“义原”。“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念[4]。“概念”是用一种“知识表示语言”来描述的,这种“知识表示语言”所用的“词汇”叫做“义原”。“义原”是用于描述一个“概念”的最小意义单位。与一般的语义词典 Wordnet 不同,《知网》并不是简单地将所有的“概念”归结到一个树状的概念层次体系中,而是试图用一系列的“义原”来对每一个“概念”进行描述。

由于《知网 Hownet》中词语不是组织在一个树状的层次体系中,而是一种网状结构;同时借助义原和符号对概念进行描述。对于两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个义项(概念):$S_{11}, S_{12}, \cdots, S_{1n}$, W_2 有 m 个义项(概念):$S_{21}, S_{22}, \cdots, S_{2m}$,刘群、李素建认为 W_1 和 W_2 的相似度是各个概念的相似度之最大值,也就是说:

$$\mathrm{Sim}(W_1, W_2) = \max_{i=1\cdots n, j=1\cdots m} \mathrm{Sim}(S_{1i}, S_{2j}) \quad (4)$$

为了更加精确地计算出词语的语义相似度,在《知网》中对一个实词的描述可以表示为一个特征结构,该特征结构含有以下四个特征:

* 第一基本义原描述:其值为一个基本义原,将两个概念的这一部分的相似度记为 $\mathrm{Sim}_1(S_1, S_2)$;

* 其它基本义原描述:对应于语义表达式中除第一基本义原描述式以外的所有基本义原描述式,其值为一个基本义原的集合,将两个概念的这一部分的相

似度记为 $Sim_2(S_1,S_2)$;

* 关系义原描述:对应于语义表达式中所有的关系义原描述式,其值是一个特征结构,对于该特征结构的每一个特征,其属性是一个关系义原,其值是一个基本义原,或一个具体词[6]。将两个概念的这一部分的相似度记为 $Sim_3(S_1,S_2)$;

* 关系符号描述:对应于语义表达式中所有的关系符号描述式,其值也是一个特征结构,对于该特征结构的每一个特征,其属性是一个关系义原,其值是一个集合,该集合的元素是一个基本义原,或一个具体词。将两个概念的这一部分的相似度记为 $Sim_4(S_1,S_2)$;

通过以上分析,可知在实际的《知网》结构中,由于各个义原所处的层次不一样,因而它们对词语相似度的影响程度也不一样,也就是说部分相似性在整体相似性中所占的权重是不一样的,权重(百分比)用 β_i 表示,于是,在知网中,概念的整体相似度可以记为:

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \quad (5)$$

其中,$\beta_i(1\leqslant i\leqslant 4)$ 是可调节的参数,且有:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。后者反映了 $Sim_1(S_1,S_2)$ 到 $Sim_4(S_1,S_2)$ 对于总体相似度所起到的作用依次递减。由于第一独立义原描述式反映了一个概念最主要的特征,所以应该将其权值定义得比较大,一般应在 0.5 以上。在以上计算中,最后求加权平均时,各部分取相等的权值。这样,就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。

2.1.3 实验及结果

根据以上方法,刘群、李素建实现了一个基于《知网》的语义相似度计算程序模块,这里选取其中的一个实验结果片段来分析:

方法 1:仅使用《知网》语义表达式中第一基本义原来计算词语相似度;

方法 2:刘群、李素建[4]的语义相似度计算方法;

实验结果如表 1 所示。

实验结果分析:考察方法 1 的结果,可以看到,"男人"(取义原"人,家,男")和其它各个词的相似度与人的直觉是比较相符合的。将方法 1、方法 2 的结果相比较,可以看到:方法 1 的结果比较粗糙,只要是人,相似度都为 1,显然不够合理;而方法 2 的结果中,这两个相似度的差距更合理一些。

2.2 利用大规模的语料库进行统计

基于语料库的词语相似度研究大都采用了上下文语境的统计描述方法,即认同这样一个论断:词语的上下文可以为词语定义提供足够信息[8]。词语向量空间模型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,算法复杂度也能够实现的模型[9]。该模型事先选择一组特征词,然后计算这一组特征词与每一个词的相关性(一般用这组词在实际的大规模语料中以该词在上下文中出现的频率来度量),于是对于每一个词都可以得到一个相关性的特征词向量,然后利用这些向量之间的相似度作为这两个词的相似度。

表 1 语义相似度实验结果片段

词语 1	词语 2	词语 2 的义原	方法 1	方法 2
男人	女人	人,家,女	1.000	0.861
男人	父亲	人,家,男	1.000	1.000
男人	母亲	人,家,女	1.000	0.861
男人	和尚	人,宗教,男	1.000	0.861
男人	经理	人,#职位,官,商	1.000	0.630
男人	高兴	属性值,境况,福,良	0.016	0.048
男人	收音机	机器,*传播	0.186	0.112
男人	鲤鱼	鱼	0.347	0.209
男人	苹果	水果	0.285	0.171
男人	工作	事物,$担任	0.186	0.112
男人	责任	责任	0.016	0.126

3 其他方法:谷歌相似性距离

这里另外所介绍的两种方法,第一种方法主要是基于树状结构中两个结点所含的信息量的大小来计算语义相似度[11],其基本思想是利用信息理论来进行研究。如以下 Dekang Lin 给出的公式:

$$Sim(S_1,S_2)=\frac{2\times\log p(S_p)}{\log p(S_1)+\log p(S_2)} \quad (6)$$

其中,S_1,S_2 表示两个义原,S_p 表示离它们最近的共同祖先,$p(S)$ 是该结点的子结点个数(包括自己)与树中的所有结点个数的比值。

第二种方法由计算机自然语言处理专家 Rudi L. Cilibrasi 和 Paul M. B. Vita′ nyi[2007.12]提出的语义相似度计算方法,该方法理论基础涉及信息论、压缩原理、柯尔莫哥洛夫复杂性、语义 WEB、语义学等,基本思想是把 Internet 作为一个大型的语料库,以 Google(对其它的搜索引擎如百度同样适用)作为搜索引擎,搜索返回的结果数作为计算的数值依据,其计算公式如下:

$$NGD(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}} \quad (7)$$

其中,NGD (Normalized Google Distance,介于 0 与 1 之间) 表示标准谷歌距离(以此衡量语义相似性大小),$f(x)$,$f(y)$ 分别表示含概念 x,y 的网页数,$f(x,y)$ 表示同时含有概念的网页数,N 表示 Google 引用的

互联网上的网页总数。

可以以一次实验来说明,假设用 Google 搜索词语"horse"返回 46 700 000(记为 $f(x)$)条结果,搜索词语"rider"返回结果数为 12 200 000(记为 $f(y)$),搜索同时含"horse,rider"的网页数是 2 630 000(记为 $f(x,y)$),Google 共引用的网页数是 $N=8\,058\,044\,651$,代入上述公式(7)求得:

$NGD(horse,rider)\approx 0.443$

4　两类主要语义相似度计算方法的比较

下面对基于语义词典和基于语料库的词语相似度计算这两类策略的方法、前提条件、所用工具等 6 个方面进行比较,见表 2。

表 2　两类主要语义相似度计算方法比较

	基于语义词典的词语相似度计算	基于语料库的词语相似度计算
方法	客观计算	经验法
前提条件	两个词汇具有一定的语义相关性,当且仅当它们在概念间的结构中有且仅有一条路径	词语的上下文可以为词语定义提供足够信息,两个词语语义相似当且仅当它们处于相似的上下文环境中
所用工具	语义词典	大规模语料库
理论依据	树论,图论	向量空间
优点比较	直观而且简单有效,可以计算出字面上不相似的词汇间的相似度	能够客观地反映词语的形态、句法、语义等特点
缺点比较	受人的主观影响比较大,有时不能反映客观现实性能	依赖于语料库的优劣,存在数据稀疏的问题,也有噪声干扰

5　结束语

鉴于语义相似度在现代科学领域中的广泛应用,在该文中,比较系统介绍了当前语义相似度计算的一些理论及方法,并简单比较了两种主要方法的特点及区别,重点描述了基于中文语义词典《知网 Hownet》的相似度计算方法;最后简单介绍了国外基于搜索引擎的相似度计算法。除了完善语义词典的全面性和准确性之外,选择或找到一种相对比较简捷地准确计算出语义相似度的方法,以确定出相似度,然后将此方法应用于信息检索等领域,改进当前仅仅依靠寻找匹配词查询所需信息的局限性。此外,在研究文本的相似性问题时,可以通过计算词与词之间、句与句的相似度得到整个文本的相似度,当相似值达到所设定的标准值时,即可认定所检验的两篇文章有抄袭其中之一的嫌疑。

关于文本相似性的研究对于当前的考试作弊雷同卷、论文抄袭等的鉴定性工作方面起到很大的改进作用,同时节省大量的人力和物力。

参考文献:

[1] Lin D. An Information Theoretic Definition of Similarity Semantic Distance in WordNet[C]//Proceedings of the Fifteenth International Conference on Machine Learning. [s.l.]: [s.n.],1998.

[2] Cilibrasi R L, Vita′ nyi P M B. The Google Similarity Distance [J]. IEEE Transactions on Knowledge and Data Engineering, 2007,19(3):370-383.

[3] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用,1998(3):79-85.

[4] 刘　群,李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 台北:出版者不详,2002.

[5] 李　峰,李　芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007(3):99-105.

[6] 李　鹏,陶　兰,王弼佐. 一种改进的本体语义相似度计算及其应用[J]. 计算机工程与设计,2007(1):227-229.

[7] 李　熙,徐德智. 基于 WordNet 的概念语义相似度研究[J]. 湖南科技学院学报,2008,29(12):115-116.

[8] 杨　哲. 基于启发式规则的本体概念语义相似度匹配[J]. 计算机应用,2007,27(12):2919-2921.

[9] 夏　天. 汉语词语语义相似度计算研究[J]. 计算机工程,2007(6):191-194.

[10] 张明宝,马　静. 一种基于知网的中文词义消歧算法[J]. 计算机技术与发展,2009,19(2):9-11.

[11] Doan A, Madhavan J. Learning to Match Ontologies on the Semantic Web[J]. The VLDB Journal, 2003,12(4):116-120.

(上接第 101 页)

In proceedings of the 19th ACM SOSP. New York, NY, USA: ACM press,2003:298-313.

[14] Birrer S, Bustamante F E. Resilient peer-to-peer multicast without the cost[C]//In Proceedings of MMCN. Berkeley: University of California,2005:113-120.

[15] Birrer S, Bustamante F E. Magellan: Performance-based, cooperative multicast[C]//In Proceedings of IWCW. Evanston, IL, USA:[s.n.],2005:133-143.

[16] Li Zhenyu, Zhu Zengyang, Xie Gaogang, et al. Fast and proximity-aware multi-source overlay multicast under heterogeneous environment[C]//In computer communications. [s.l.]:[s.n.],2008:257-267.