基于本体的语义相似度和相关度计算研究综述

刘宏哲[1,2] 须 德[2]

(北京联合大学信息学院 北京 100101)[1] (北京交通大学计算机研究所 北京 100044)[2]

摘 要 语义相似度和相关度计算广泛应用于自然语言处理中,已有大量语义相似度和相关度算法被提出。分析总结了树和图结构中影响概念相似度或相关度的因素,综述了基于本体的英文语义相似度和相关度计算方法,明确了语义相似度和相关度的区别与联系,系统地对算法进行了分类,最后对每类算法进行了详细的比较。

关键词 语义相似度,语义相关度,本体

Ontology Based Semantic Similarity and Relatedness Measures Review

LIU Hong-zhe[1,2] XU De[2]

(Information College, Beijing Union University, Beijing 100101, China)[1]

(Institute of Computer Science, Beijing Jiaotong University, Beijing 100044, China)[2]

Abstract Measuring semantic similarity and relatedness has many applications in NLP, and many different measures have been proposed. We analysed the factors affecting concept similarity or relatedness in ontology, made a review of research on the ontology based semantic similarity measures, defined the connection and difference between similarity and relatedness, classified the measures, and finally we compared all kinds of the measures.

Keywords Semantic similarity, Semantic relatedness, Ontology

1 语义相似度和相关度

两个对象之间的相似度或相关度计算早已成为数据挖掘和信息提取领域中的基本问题,具体地说,它是文本处理的核心问题[14]。例如,语义相似度或相关度算法[2]已经被应用于词义消歧[3,16]、音频识别错误的检测[10]、信息提取[8]、语音自动摘要[7]、人的姓名解析[6]、文本相似度计算[4]、文本分类和聚类等。

一般来说,语义相关度涵盖语义相似度,语义相似度是指两个概念间的相似程度,通常指两个概念本身之间具有某些共同特性;而语义相关度是指两个概念间的相关程度,这两个概念间可能不存在相似关系,但可以通过某些其它关系相关联形成相关关系。语义相似度是语义相关度的一种特例。Resnik[20]用轿车、汽油和自行车的例子解释了两者之间的区别:"轿车依赖于汽油作为燃料,显然它们之间的相关性比轿车与自行车更为紧密,但人们却普遍认为轿车与自行车之间的相似性大于轿车与汽油。这个例子表明,相关性不能等同于相似性。即使轿车与汽油是紧密相关的,但由于这两者之间没有共同的特性,人们不会认为它们是相似的。而轿车和自行车都是交通工具,都有轮子并且可以载人,因此它们是相似的。"相似性与相关性不是互斥关系,Resnik 认为相似性可以被视为一种特殊的相关性,即对象间基于蕴涵关系的相关性。在本体结构中,通常由"is a"关系关联的两个概念间存在相似关系,由其它关系关联的(例如"part of")两个概念间存在相关关系。需要说明的是,语义相似或相关是基于一定的视角或上下文的,在某个角度相似或相关的概念在另一个角度可能不相似或不相关。

在语义相似度或相关度计算中经常涉及到概念所涉及的本体,在英文语义相似度和相关度计算中,常常涉及到的通用本体是概念语义分类词典 WordNet[5],或者是知识百科 Wikipedia[43]等;另外还有些本体是领域本体。

一般情况下,如果相似度计算过程中只考虑了上下位关系,那么就称该算法为相似度计算算法;如果计算过程中除上下位关系外,还考虑了其它类型的边,例如整体部分关系,那么就称该算法为相关度计算算法。

2 研究状况

国内外已有大量研究者对英文概念语义相似度和相关度进行了研究,已经形成了丰富的研究成果,现分类总结如下。

2.1 基于树状本体结构或以树为主体图结构的语义相似度或相关度计算方法

所谓基于树状本体结构的相似度算法是指在相似度计算过程中主要基于上下位相连的关系,例如 WordNet 中的"is a"关系。而以树为主体的图结构是指上下位关系作为主要关系连接概念节点,同时除了上下位关系,还有少量其它类型的关系编织于概念之间。这些算法在进行语义相关度计算时,

到稿日期:2011-03-09 返修日期:2011-05-28 本文受国家自然科学基金项目(60972145),北京市教委科技面上项目(KM201111417002),北京市属高等学校人才强教计划资助项目(PHR201108419,PHR200907120)资助。

刘宏哲(1971—),女,副教授,主要研究方向为人工智能、数字博物馆,E-mail:xxtliuhongzhe@buu.edu.cn;须 德 男,教授,主要研究方向为数据挖掘。

不仅考虑上下位关系，还考虑了其它类型的关系。例如WordNet中除考虑“is a”关系外，还要考虑“part of”关系等。

图1所示的是基于WordNet名词概念的以树为主体的图结构示例。

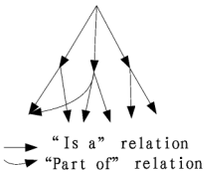

图1　基于WordNet名词概念以树为主体的图结构示例

2.1.1　基于树状本体结构或以树为主体的图结构中影响概念间相似度或相关度的因素

综合考虑相关研究，发现影响概念间相似度包括以下因素：

·被比较概念词在本体层次树中所处的深度[11,25]。层次越高，越抽象；层次越低，越具体。高层次的概念词间的语义相似度一般小于低层次概念词间的语义相似度。例如，图2中$\mathrm{sim}(C_1,C_2)<\mathrm{sim}(C_3,C_4)$。

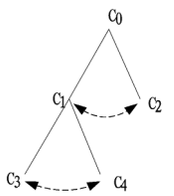

图2　本体深度影响

·被比较概念词在本体层次中所处区域的密度。局部区域密度越大，该区域对节点概念的细化程度就越大。区域内概念间语义相似程度较大[11,25]。例如，图3中左侧的相似度应小于右侧的相似度。

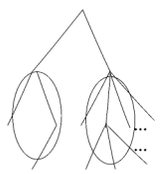

图3　局部密度影响

·被比较概念词连通路径上各个边的类型[25]。在本体中，不同的概念关系所表征的语义相似度和相关度是不同的。例如，“同义关系”所表征的语义相似度和相关度应大于上下位关系与整体和部分关系所表征的语义相似度和相关度。

·被比较概念词相隔路径长度[26]。由于不同的概念关系所表征的语义相似度和相关度是不同的，另外相同类型的边处于不同深度，不同密度区域所代表的语义距离也有所不同，因此可以得出结论：当一对概念的路径包含在另一对概念之中时，这对概念间的相似度大些。例如，图4中$\mathrm{sim}(C_0,C_3)<\mathrm{sim}(C_0,C_1)$。

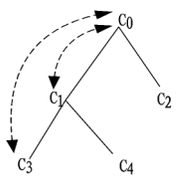

图4　路径长度影响

·被比较概念词连通路径上各个边在本体层次中的关联强度。如图5所示，C_2和C_3被一条part of边相连，而C_2和C_4则没有，虽然C_2和C_3，C_2和C_4两对词在本体中位置是对称的，但是$\mathrm{sim}(C_2,C_3)>(C_2,C_4)$。

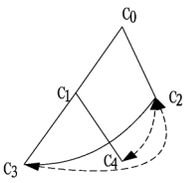

图5　关联强度影响

·被比较概念词连通路径上各个边的两端节点概念词的属性。本体，尤其是领域本体，不仅会对概念及其关系进行准确定义，还会对概念的属性进行详细描述。如果某条边两端的概念词所用的相同属性越多，那么其对语义相似度和相关度的影响也越大。

2.1.2　基于树状本体结构或以树为主体图结构的语义相似度或相关度算法

学者们一般将基于树或以树为主体图结构的语义相似度计算算法划分为4类：基于结构的语义相似度和相关度计算、基于内容的语义相似度和相关度计算、基于属性的语义相似度和相关度计算、混合式语义相似度和相关度计算，以下分别介绍：

2.1.2.1　基于本体结构的语义相似度和相关度计算

基于本体结构的语义相似度和相关度计算分为基于简单结构和基于复杂结构，其中基于简单结构主要是指基于路径距离的语义相似度计算，而另外一些算法基于更为复杂的本体结构做语义相似度和相关度计算。

基于距离的语义相似度计算

其基本思想是通过两个概念词在本体树状分类体系中的路径长度来量化它们之间的语义距离。语义相似度和语义距离之间存在着密切的关系：两个词语的语义距离越大，其相似度越低；反之，两个词语的语义距离越小，其相似度越大。代表算法有：Shortest Path法[19]、Weighted Links法[25]、Wu and Palmer法[23]、Leacock and Chodorow法[13]等。Rada等[19]认为概念词间的相似度与其在本体分类体系树中的距离有关，距离越大，相似度越小。该算法在现有算法中计算复杂性最小，不过其主要缺陷是需假设本体分类体系中所有边的距离同等重要。显然，该假设并不成立，边的重要性受其位置信息、自身的类型和所表征的关联强度等因素影响。Weighted Links法[25]基于权重的思想对Shortest Path法进行了扩展，考虑到概念词在本体层次树中的位置信息（所在深度和所处区域的密度）和边所表征的关联强度，通过将组成概念对之间连通路径的各个边的权值相加，而不是简单统计两个概念词间边的数量，来计算两个概念词的距离。Wu and Palmer法[23]，与Shortest Path法和WeightedLinks法不同，其并不是通过直接计算概念对间的路径长度来计算它们之间的相似度，而是基于它们与其最近公共父节点概念（Most Common Parent Concept）的位置关系来计算的。Leacock和Chodorow[13]则还考虑了本体分类体系树自身的深度对被比较概念词相似度的影响。

Hirst-Stonge[9]认为，如果概念对间存在一条较短的路径，且在遍历过程中改变路径方向的次数较少，那么这两个概

念词语义相关。概念间路径涉及多种类型的关系。Hirst-Stonge 法开辟了一个新的视角，但由于其在很大程度上取决于“方向”问题而不是概念关系，因此表现似乎不是很好。

Yang 和 Power[24] 在 2005 年提出一个基于本体中关系边数的相关度算法，算法中除了利用之前算法主要使用的“is a”关系外，还使用 “equivalence”和“part of”关系。此算法获得了当时最好的关联度，但是最大缺点是它一共涉及到 7 个可以自由调节的参数[29]。

基于复杂本体结构的语义相似度和相关度计算方法

经过研究发现，除了路径距离外，诸多信息隐藏在本体结构中，充分利用它们来进行语义相似度和相关度计算将有利于提高计算效果。与基于简单本体结构相似度算法不同基于复杂本体结构相似度算法不一定考虑了更多的边的类型，而是利用了更多的本体结构信息而非只是基于边的个数的语义距离。J. W. Kim[12] 等人提出一个 CP/CV 的概念传播方法，该方法依赖于概念之间的语义关系，而语义关系由概念节点在本体结构中所处的层次结构来决定，CP /CV 方法比其它只利用简单本体结构的方法有更好的关联度。另研究发现，概念的局部密度信息及概念的相关概念信息隐藏在本体层次结构里。于是在文献[26]提出了一个基于相关概念节点局部密度的概念向量模型来计算概念间语义相似度和相关度(简称 RNCVM 方法)。具体做法是首先定义在一个树状结构中，某一概念节点的祖先概念节点和后代概念节点为它的相关概念节点，然后再根据树的结构特点赋予相关概念节点不同权重，由相关概念节点的局部密度形成概念向量。实验证明该方法比其它方法有较为显著的改进。

2.1.2.2 基于信息内容的语义相似度计算

基于信息内容的语义相似度计算算法的基本原理是：如果两个概念词共享的信息越多，它们之间的语义相似度也就越大；反之，共享的信息越少，相似度也就越小。在本体分类体系树中，每个概念的子节点都是对其祖先节点概念的一次细分和具体化，因此，可以通过被比较概念词的公共父节点概念词所包含的信息内容来衡量它们之间的相似度。根据信息论可知，概念词所包含的信息内容可通过其在给定的文献集中出现的频率来衡量，频率越高，信息内容就越贫乏；反之，所含的信息内容也就越丰富。在一个树状结构中，概念节点的频率是指此概念实例所出现的频率，任何一个非叶子节点所对应的概念出现的频率是它所对应的所有子孙节点出现的频率之和，显然根节点实例出现的频率为 1，其所含的信息含量最低；随着节点的下移，其所对应的概念就越具体，对应实例出现的频率就越低，所含的信息含量就越高，叶子节点信息含量最高；尽管树状结构中的理论是如此，但依靠大规模语料库统计的频率并不一定完全符合以上规律。所有基于信息内容的语义相似度计算算法都建立在被比较概念词对共享父节点所含信息内容基础上。Lord 等人[27] 提出使用共享父节点所包含的信息内容来计算概念词间的语义相似度。他们直接使用最近公共父节点概念词的信息量来计算被比较概念词对间的相似度。Resnik[21] 使用共享父节点信息内容来计算概念词间的语义相似度，该算法与 Lord 等人的算法不同之处在于它并不是基于最近公共父节点的信息内容，而是基于公共父节点概念词中信息量最大的父节点的信息内容。上述两种算法都只考虑了被比较概念词对的共享信息内容，Lin[15] 认为还应该考虑被比较概念词对各自所包含的信息内容。当被比较概念词集属于同一个本体时，使用 Lin 法可以获得比 Resnik 法更好的基于相似度的排列结果。Jiang 和 Conrath 法[11] 与上述 3 种算法不同的是：直接通过对语义距离的计算来表征被比较概念词间的相似度。和 Lin 法一样，该算法也同时使用了共享父节点和被比较概念词所包含的信息内容。

2.1.2.3 基于属性的语义相似度或相关度计算

基于属性的语义相似度和相关度计算的基本思想：事物由其属性特征反映其本身，人们用以辨识或区分该事物的标志就是属性特征。事物之间的关联程度与其所具有的公共属性数相关。基于此提出了基于属性的语义相似度和相关度计算。对于两个被比较概念词而言，公共属性项越多，相似度越大。Tversky 算法[1] 是该类算法的典型，该算法既没有考虑被比较概念词在分类体系树中的位置信息，也没有考虑其祖先概念节点和自身所包含的信息内容，只利用了相应本体的属性集信息。

Banerjee 和 Pedersen[28] 以及 Patwardhan 等人[18,30] 提出的基于概念释词的方法，其思想是概念相似度或相关度值通过两个概念在本体中的释词(gloss)重叠程度来获得。由于释词的个数可能很少，因此以上办法是将释词的个数扩展至与目标概念直接相连的概念[18] 或目标概念上下文概念[30] 释词的重合程度。这不仅涉及上下位关系，也涉及其它关系。通过在释词集合中提取共同属性或判断属性的相似程度来判断两个概念的语义相似程度，也属于基于概念属性的方法。

2.1.2.4 混合式语义相似度和相关度计算

混合式语义相似度和相关度计算算法实际上是对上述 3 种算法的综合考虑，即同时考虑了概念词的位置信息、边的类型、概念词的属性信息等。代表算法有 Li 等人[14] 法，该算法同时考虑了被比较的词语队间的最短路径 L 和最近公共父节点在分类体系树中所处的深度，以及词语所处位置的局部密度信息，是个非线性函数。Marco[29] 等人提出的基于图模型的语义相似度和相关度计算方法(简称 SSA)是基于以树为主体的本体结构和释词方法的混合，也取得了不错的关联度和值连续性。

另外，近年来国内外出现的大量相关研究成果[34-41] 大都属于基于以上计算方法的改进方法、混合方法和具体应用。例如 Shi bin 等[41] 人提出了 OHIIC 基于局部密度、信息量和概念深度的混合算法，只不过算法中的局部密度扩展成由当前节点与其儿子节点连接边的个数决定。

2.2 基于有向图结构的语义相关度计算方法

传统语义相似度和相关度计算都以 WordNet 为通用本体，Wikipedia 与 WordNet 比较，其覆盖的范围更加广泛，知识描述更加全面，信息内容更新速度更加迅速。所以近年来语义相关度逐渐转向以 Wikipedia 为基础。维基百科具有很好的结构化信息，可以将维基百科看成两个巨大的网络：(1)由页面组成的网络(简称页面网)，如图 6 所示，其中每个节点表示一个页面，每根线条表示一个连接。不同的页面之间通过入链和出链相互连接在一起。

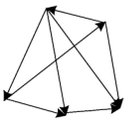

图 6 Wikipedia 页面网示例

(2)由类别组成的网络[4](简称类别网),如图 7 所示,其中每个矩形框代表一个维基百科的一个类。不同的类别通过子类和父类的关系相互连接在一起。

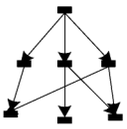

图 7　Wikipedia 类别网示例

这里的类别不像 WordNet 中的类别关系是树或者以树为主体的图结构,而是有向图结构。因此维基百科这两个网都可以抽象成有向图(DAG),对 Wikipedia 类别网和页面网的处理也就抽象成对有向图的处理。

2.2.1　图结构中影响概念间相似度和相关度的因素

图结构中影响概念间相似度包括以下因素:

·被比较概念词在图中连接边的个数。如图 8 所示,被比较概念词在图中连接边的条数决定概念词间的语义相关度。概念间的直接或间接连接边数越多,其间的语义相关度越大。例如图 8 中,C_1,C_2 间有 3 条边相连,而 C_1,C_3 间有一条边相连,所以 $\mathrm{sim}(C_1,C_3)<\mathrm{sim}C_1,C_2)$。

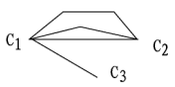

图 8　连接边数

·被比较概念词相隔路径长度。由于以 Wikipedia 为代表的图结构中边的类型和深度一般不加以区分,在局部密度信息一定的情况下,笼统地讲,被比较概念词相隔路径越长,其所表征的语义距离越大,概念间的语义相似度就越小。例如图 9 中,C_1,C_2 间相隔 3 条边,而 C_1,C_3 间隔一条边,所以 $\mathrm{sim}(C_1,C_2)<\mathrm{sim}(C_1,C_3)$。

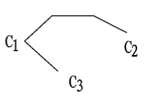

图 9　路径长度影响

·被比较概念词在图中所处区域的密度。局部区域密度越大,该区域对节点概念的细化程度也越大。区域内概念间语义相似程度较大。

·被比较概念词的属性。如果一对概念词所拥有的共同属性越多,概念间的相似度也就越大。很多基于概念解释页面文本的算法就是通过提炼概念间的共同属性来决定概念间相似度或相关度大小的。

2.2.2　基于图结构的算法

以基于 Wikipedia 为代表的算法有 WikiRelate![31]、Explicit Semantic Analysis(简称 ESA)[32] 和 Wikipedia Link Vector Model(简称 WLVM)[33]。Strube 等人提出的 WikiRelate! 方法利用 Wikipedia 的文档类型结构代替 WordNet 的概念层次结构,利用 Wikipedia 的文档内容代替 WordNet 的词汇定义,模仿基于 WordNet 的度量方法进行语义相关性的计算。Gabrilovich 等人提出了类似于信息检索中向量空间模型的方法 ESA,它以 Wikipedia 中的所有文档为向量元素构造词汇的属性向量,通过比较词汇向量之间的相似性来判断词汇的相关性。David 提出一种基于向量模型计算语义相关性的方法 WLVM,它只利用了 Wikipedia 文档之间的链接信息,没有涉及文本内容。

3　算法评价标准

关连度

早在 1965 年,作为对"文本相似度与意义(同义词)的相似度之间的关系"的研究的一部分,Rubenstein 和 Goodenough[22] 让 51 个受试者对 65 对词汇做出"同义判断"。这些词汇对经过精心挑选,从"高度同义"到"语义不相关",并且这些实验对象被要求依照他们自己的"意义的相似度"在 0.0 至 4.0 的范围内对它们估值。还有一个类似的研究:Miller 和 Charles[17] 从 Rubenstein 和 Goodenough 的 65 对词汇中选取 30 对,其中 10 对高层(3-4),10 对中层(1-3),10 对低层(0-1),然后让 38 个对象对这 30 对进行语义相关度判断。Miller 和 Charles 与 Rubenstein 和 Goodenough 的结果关联度接近 0.97,说明了两个数据集的有效性。

随着研究的深入,有研究者提出,Miller 和 Charles 以及 Rubenstein 和 Goodenough 的数据集涵盖的数据太少,于是 Finkelstein 等[42] 给出了一个包含 353 对词汇的大的语义相似度和相关度测试集作为检验语义相似度和相关度算法测试训练集,353 对词汇涵盖了 Miller 和 Charles 的 30 对名词。

在相似度或相关度计算中,如果两个概念有多个意思,那么一般取多个意思中相似度最大的那对意思的相似度/相关度值作为两个词的相似度/相关度值。例如概念 1 有意思 S11,S12,S13,而概念 2 有意思 S21,S22,S23,那么相似度或相关度计算一般采用图 10 方法来计算每一对可能的意思对的相似度或相关度,然后取值最大的一对作为这对概念相似度或者相关度的值。

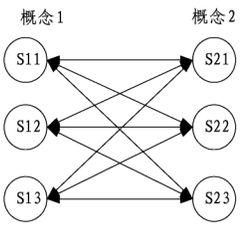

图 10　概念多个意思间的相似度或相关度计算

为评价相关度计算算法的好坏,国际上通常根据算法计算得到的英语概念相似度值与人的判断结果(以上 3 个基准数据集为代表)计算皮尔森关联(Correlation),关联度越高说明计算得到的相似度值与人的判断吻合度就越高。

相似度结果的连续性

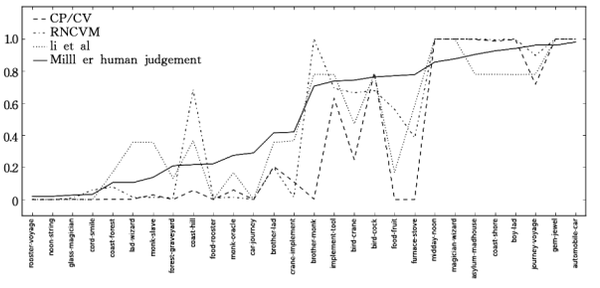

图 11　相似度或相关度结果的连续性

在关联度近似的情况下，相似度或相关度值波动幅度可能大不相同，有的则连续性较好，而有的波动幅度较大。如图11将li等人[14]的算法、RNCVM[26]算法和CP/CV[12]算法按照Miller和Charles测试名词对相似度值从小到大的顺序连线，虽然它们都具有较高的关联度值，但是它们围绕基准数据波动的幅度不同，而波动幅度小的值连续性较好，不会出现个别结果偏差太大。

调节参数个数

相似度计算算法中，设置的可自由调节的参数个数决定着算法应用时的稳定性，如果参数不能得到很好的调节，可能带来计算结果偏差。例如提到的文献[24]中Yang和Powers的算法，虽然获得了很高的关联度(0.921)，但其算法中一共涉及7个可自由调节的参数，使得该算法有很大的灵活度向关联度大的结果靠拢[29]。

在评估语义相似度或相关度计算算法性能时，不但要从理论上分析算法来衡量算法的性能好坏，更重要的是将算法放在具体应用中来评估它的性能[2]。

4 算法分析比较

综合以上影响语义相似度和相关度的因素和相似度算法的评价标准，具体分析如下：

基于树或以树为主体图的算法中，基于结构的方法分为基于简单结构和基于复杂结构的算法。基于简单结构的算法也就是基于路径的方法简单，易于实施，不依赖附加信息。但是不能体现概念对所在位置深度，使得该方法计算出的树中位置深度较深的概念间的相似度偏小。也不能体现概念对所在位置局部密度，使得方法计算出的本体中局部密度较大的概念对的相似度偏小。基于复杂结构的方法充分地利用了隐藏在本体内固有的、丰富的结构信息，提高了语义相关度计算效果，同时相对简单，且不依赖附加信息，普遍具有较好的关联度。

基于内容的方法相对比较客观，能综合反映概念在句法、语义、语用等方面的相似性和差异，但也存在一些问题：例如，Resnik[21,27]单纯依据信息内容的方法不能体现概念间的距离，而且忽略了密度深度等结构信息，即使有的办法考虑了概念间的部分结构因素[11,15]，得出较粗略的结果，例如一个节点与同一子树中任意节点的相似度值都相同。另外这类算法还比较依赖于训练所用的语料库，受数据稀疏和数据噪声的干扰较大，有时会出现明显的错误。另外，当建立一个新的应用时，尤其是应用到某些领域本体时，针对领域本体的语料库不全或者尚未建立，使得此类算法很难实施。

基于属性的方法因各个方法差异，优缺点也各有不同，其共同局限性是：此类办法必须依赖于概念具备完备的属性集，对于不存在针对概念完备属性集的情况，此类办法则无法实施。

混合方法不能减少对附加信息的依赖，没有从根本上克服它基于方法的局限性。

以上分析的是基于WordNet为代表的树或以树为主体图结构的相似度或相关度计算方法，与基于Wikipedia为代表的基于有向图的算法相比，其普遍具有较高的关联度。原因是Wikipedia虽然含有丰富的语义信息，但其数据噪声较大，且与WordNet相比，其数据的结构性不强，因此计算效果普遍较差。

结束语 任何一种相似度计算算法都不能解决所有问题，算法也没有绝对的好与坏，可能因为应用场合不同而表现各异。Wikipedia的覆盖范围更加广泛，知识描述更加全面，信息内容更新速度更加迅速。另外，以WordNet为基础的算法得到较充分的研究，结果已经取得了较好的关联度，而以Wikipedia为基础的算法还比较少，关联度还有较大的提升空间，所以不失为未来语义相似度和相关度计算研究的趋势。

参 考 文 献

[1] Tversky A. Features of Similarity [J]. Psychological Review, 1977, 84(4): 327-352

[2] Budanitsky A, Hirst G. Evaluating wordnet-based measures of lexical semantic relatedness [J]. Computational Linguistics, 2006, 32(1): 13-47

[3] Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network [C]// Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93). Arlington, Virginia, 1993: 67-74

[4] Corley C, Mihalcea R. Measuring the semantic similarity of texts [C]// Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. Ann Arbor, MI, US, June 2005: 13-18

[5] Fellbaum C. WordNet: An Electronic Lexical Database [M]. MIT Press, 1998

[6] Fleischman M, Hovy E. Multi-document person name resolution [C]// Harabagiu S, Farwell D, eds. Proceedings of the Workshop on Reference Resolution and its Applications. Barcelona, Spain, July 2004: 1-8

[7] Gurevych I, Strube M. Semantic similarity applied to spoken dialogue summarization [C]// Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004: 764-770

[8] Hassan H, Hassan A, Emam O. Unsupervised information extraction approach using graph mutual reinforcement [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, July 2006: 501-508

[9] Hirst G, Onge D S. Lexical chains as representations of context for the detection and correction of malapropisms [C]// Fellbaum C, ed. WordNet: An Electronic LexicalDatabase. MIT Press, 1998: 305-332

[10] Inkpen D, Esilets A. Semantic similarity for detecting recognition errors in automatic speech transcripts [C]// Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, October 2005: 49-56

[11] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy [C]// Proceedings of the 10th International Conference of Research on Computational Linguistics. Taiwan, August 1997

[12] Kim J W, Candan K S. Cp/cv: concept similarity mining without

frequency information from domain describing taxonomies[C]// Proceedings of the 15th ACM International Conference on Information and Knowledge Management. New York, NY, USA, ACM Press, 2006: 483-492

[13] Leacock C, Chodorow M. Combining local context and wordnet similarity for word sense identification[C]// Fellbaum C, ed. WordNet: An Electronic Lexical Database. MIT Press, 1998: 265-283

[14] Li Y, Bandar Z, McLean D. An approach for measuring semantic similarity between words using multiple information sources [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 871-882

[15] Lin D. An information-theoretic definition of similarity [C]// Proceedings of the 15th International Conference on Machine Learning. Wisconsin, USA, July 1998: 296-304

[16] McCarthy D. Relating wordnet senses for word sense disambiguation[C]// Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together. Trento, Italy, 2006: 17-24

[17] Miller G A, Charles W G. Contextual correlates of semantic similarity[J]. Language and Cognitive Processes, 1991, 6(1): 1-28

[18] Patwardhan S, Pedersen T. UsingWordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts[C]// Proceedings of the EACL Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy, April 2006: 1-8

[19] 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

[20] Resnik P. Using information content to evaluate semantic similarity in a taxonomy[C]// Proceedings of the 14th International Joint Conference on Artificial Intelligence. volume 1, Montreal, Canada, August 1995: 448-453

[21] Resnik P. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language[J]. Journal of Artificial Intelligence Research, 1999, 11: 95-130

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

[23] Wu Z, Palmer M. Verbs semantics and lexical selection[C]// Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, USA, 1994: 133-138

[24] Yang D, Powers D M W. Measuring semantic similarity in the taxonomy of WordNet[C]// Proceedings of the 28th Australasian Computer Science Conference. Newcastle, Australia, Jan/Feb 2005: 315-322

[25] Richardson R, Smeaton A F. Using WordNet in a Knowledge-Based Approach to Information Retrieval[Z]. Working Paper, CA-0395. School of Computer Applications, Dublin City University, Ireland, 1995

[26] Liu Hong-zhe, Bao Hong, Xu de. Concept Vector for Similarity Measurement based on Hierarchical Domain Structure[J]. Computing and informatics, 2011(30): 1001-1021

[27] Lord P W, Stevens R D, Brass A, et al. Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Sequence and Annotation [J]. Bioinformatics, 2003, 19(10): 1275-1283

[28] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness[C]// Proceedings of IJCAI. Mexico 2003: 805-810

[29] Marco A A, SeungJin L. A Graph Modeling of Semantic Similarity between Words[C]// International Conference on Semantic Computing(ICSC 2007). 2007: 355-362

[30] Wan S, Angryk R A. Measuring semantic similarity using wordnet-based context vectors[C]// Systems, Man and Cybernetics. 2007: 908-913

[31] Strube M, Ponzetto S P. WikiRelate! Computing Semantic Relatedness Using Wikipedia[C]// Proc. of AAAI. 2006

[32] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis[C]// IJCAI. 2007: 1606-1611

[33] Milne D. Computing semantic relatedness using Wikipedia link structure[C]// NZCSRSC'07. 2007

[34] Jike G, Yuhui Q. Concept Similarity Matching Based on Semantic Distance[C]// SKG: 380-383

[35] Anna F. Concept similarity by evaluating information contents and feature vectors: a combined approach[J]. Communications of the ACM, 2009, 52(3): 145-149

[36] Gerasimos S, Georgios S, Andreas S. A hybrid Web-based measure for computing semantic relatedness between words[C]// 2009 21st IEEE International Conference on Tools with Artificial Intelligence, ICTAI. 2009: 441-448

[37] Zhao Zhong-cheng, Yan Jian-zhuo, Fang Li-ying, et al. Measuring Semantic Similarity Based On WordNet[C]// Web information system and application conference. 2009: 89-92

[38] Cai Song-mei, Lu Zhao. An Improved Semantic Similarity Measure for Word Pairs[C]// International Conference on e-Education, e-Business, e-Management and e-Learning. 2010: 212-216

[39] Qin Peng, Lu Zhao, Yan Yu, et al. A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory[C]// International Conference on Web Information Systems and Mi-ning. 2009: 181-185

[40] Sheng Yan, Li Yun, Luan Luan. A Concept Similarity Method in Structural and Semantic Levels[C]// Second International Symposium on Information Science and Engineering: 620-623

[41] Shi Bin, Fang Li-ying, Yan Jian-zhuo, et al. Ontology-Based Measure of Semantic Similarity between Concepts[C]// World Congress on Software Engineering. 2009, 2: 109-112

[42] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

[43] http://www.wikipedia.org/