基于不同语义资源的词语相似度算法综述

蔡辉虎
（厦门大学信息科学与技术学院，福建 厦门 361005）

摘　要：词语相似度研究作为人工智能领域中一项重要研究，被广泛应用于信息检索，词义消歧，机器翻译，语音自动摘要，分类和聚类等方面。现有的词语相似度算法主要分为基于语义资源和基于统计两类方法，第一种也被称为基于本体的词语相似度算法，主要根据词语所处的语境来反应词语的词义，即根据不同的层次结构组织中词所处的上下位与同位关系来计算词语的相似度。另一种也被称为基于大规模语料库的算法，研究上下文环境中各个词语之间出现的某种规律，利用统计技术计算的一种无监督机器学习的方法。本文重点介绍基于不同的语义资源的词语相似度算法，对词语相似度算法的未来做了展望。

关键词：词语相似度；语义资源；维基百科
DOI:10.16640/j.cnki.37-1222/t.2016.05.211

1　引言

随着云时代的来临，大数据越来越受人们关注。伴随着办公室无纸化推行，人们逐渐习惯于利用计算机进行数字化处理数据，自然语言处理的研究也飞速发展。词语是自然语言处理的最小单位，词语相似度的计算在自然语言处理的各个领域占有很重要的地位。词语相似度计算研究的是计算两个词语相似度的方法，词语之间有着非常复杂的关系，应用中常常将这种复杂的关系用简单的数量来度量。可见词语相似度研究有广阔的应用前景和重大研究价值。本文综合介绍了近年来基于几种常见语义资源的词语相似度算法和最新研究成果，对该领域的发展前景做出了展望。

2　基于 Wordnet 的方法

Wordnet 是由普林斯顿大学的心理学家，语言学家和计算机工程师联合设计的一个在线词典参考系统，在认知语言学理论下推动形成的覆盖范围非常广阔的词汇语义网。Wordnet 不像传统的在线词典按照字母排序构造而成，这个系统中的词语根据同义关系，反义关系，部分关系聚类分为代表某一类词汇概念的相关集合。并在这些聚类后形成的集合之间建立起不同关系。

Wordnet 主要代表算法是通过计算两个词语在本体结构分类的路径长度，本体库的统计特征，概念层次树上下位关系和同位关系或对词语涉及的边进行处理。例吴思颖等 [1] 利用语义网同义词集上下位关系图中，引入了距离，密度，深度 3 个因素来估计同义词集之间的相似度，采用一个自适应的方案来解决候选同义词集组合的权重和取舍问题。实现了一个可以计算英 - 英，汉 - 英，汉 - 汉词语之间相似度的算法。基于 wordNet 算法的主要优点是覆盖范围宽广，数据足够密集，减少数据中无法解释的数据变动的干扰。主要缺点受个人偏见或局限性影响较大，对客观现实的反应不够准确。

3　基于知网的方法

< 知网 >(英文名称 HowNet) 是著名机器翻译专家董振东 [2] 先生创建的相对丰富的语义知识词典，它所所描述的对象是以词语为代表的概念，概念之间的关系用关系义原或者关系符号来表示，并且描述了多种概念的属性与属性之间的关系，具有种类多，数量广并且多样化的关系层次词汇语义知识。

“ 概念 “ 和 ” 义原 “ 是《知网》结构中有两个最主要的概念，” 义原 “ 是用来描述 ” 概念 “ 的 ” 知识表示语言 “ ，” 义原 “ 还是描述 ” 概念 ” 不可分割的基本单位。一个概念可以描述一个词，或者多个概念组合描述一个词，利用基本义原，语法义原和关系义原来描述概念，也是词语的某一部分特性，计算出义原的相似度就可以求出词语的相似度。例王斌 [3]，刘群等 [4]，Li 等 [5] 根据《知网》中树形图由义原上下位关系构成，分别计算出其中节点之间路径的方法，或者利用集合，特征结构整体计算得到语义距离并进行转换的方法，提出各种基于《知网》义原关系计算的词语相似度算法。《知网》提供了更加直观，结构化的词汇语义信息，但是随着知识语言发展，未登录词语越来越多，暴露了覆盖的词汇有限的局限性。

4　基于同义词词林的方法

1983 年梅家驹等 [6] 人为了加速创作和翻译工作，对同义词语进行收集汇编分类，由此编纂了 《同义词词林》。这本词典最主要的是包括大部分词的同义词，当然也包含了一定数量的广义相关词。依照树状层次结构把所有收录的词条组织到一起，把词汇分成大中小三类，采用层级体系，具有五层结构。

基于《同义词词林》的词语相似度算法主要采用概念切分法，节点路径计算法，或者综合算法。例天久乐等 [7] 从词语的语义出发，根据两个词语的义项在同义词词林中的位置，算出相对距离，用具体的实数值表示，并且结合两个词语在相类似语境中相互替换或者共现的可能性计算出相关性，具有高相关性的词语具有相似性的程度也相应较高，导入一定的测试函数计算出词语的合理相似度。吕立辉等 [[]] 通过两个单词在词林书中相距的路径长，以及所在分支词义密度来计算两个中文单词间的相似度，利用皮尔逊线性相关系数来评价该方法。基于同义词词林词语相似度算法的优缺点与基于 Wordnet，并且同义词词林数据更新缓慢。

5　基于维基百科的方法

维基百科是一个基于 Web2.0 技术的全球性多语言合作型语料库，同时也是作为词语相似度计算的网络百科全书，其目标及宗旨是由全人类提供的自由的百科全书，维基百科中使用语义解释丰富的词条来表示主题，每篇文章都可以归类于某一类主题。词条之间具有上下位关系，这种独特的结构方式使维基百科成为最新词汇语义信息的重要来源。

基于维基百科的词语相似度算法主要利用维基百科中词条丰富的语义解释，层次的上下位关系，文章之间借助内容的超链接相互关联反映的词汇间词义关系进行相似度计算。例 Strube 等 [9] 最早提出 Wikerelate! 算法，比较不同词性的词语之间的语义相似度，随后 Gabrilovich 等 [10] 提出了基于维基百科文章内容的显性语义分析法 . 把文本内容的词语含义通过机器学习技术表达为维基百科概念的加权向量。MiLine[11] 提出了利用维基百科文章集合中内容的超链接信息计算词语相似度的方法 WLVM，该方法只利用了文章中内容超链接

气相液相色谱技术在食品安全检测中的应用研究

田利利，蒋绍金
（光明乳业股份有限公司，上海 201111）

摘　要：食品安全问题的出现，不仅会影响人们的身体健康，还会扰乱社会秩序，继而给社会发展带来不良的影响。而使用气相液相色谱技术进行食品检测，可以为食品安全提供保障。因此，本文对气相液相色谱技术在食品安全检测中的应用问题展开了研究，从而为关注这一话题的人们提供参考。

关键词：气相液相色谱技术；食品安全检测；应用
DOI：10.16640/j.cnki.37-1222/t.2016.05.212

0　引言

为了谋取暴利，一些不法商贩在食品生产过程中大量使用了苏丹红和三氯氰胺等非法添加剂，从而给人体健康带来了严重威胁。而使用传统的检测方法，已经很难将这些有毒有害物质检测出来。在这种情况下，食品安全检测部门对气相液相色谱技术展开了研究，并将其应用在了食品安全检测中，从而为食品安全提供了一定的技术保障。

1　气相液相色谱技术及其在食品安全检测中的应用范围

1.1　气相液相色谱技术

气相液相色谱技术又可以划分成气相色谱技术和液相色谱技术，两种技术拥有不同的载体。其中，气相色谱技术是利用惰性气体为载体，需要使用气相色谱仪进行检测样本的色谱分析。而液相色谱技术是以液体的流动相为载体，需要使用高压输液系统将不同极性的溶液流动相泵入到色谱柱中，然后利用装有固定相的色谱柱完成样品分离，并利用不同检测仪器进行检测分析。就目前来看，在检测气体混合物或容易挥发的液体和固体时，通常会使用气相色谱技术。在检测单一流动溶液或混合溶液时，可以使用液相色谱技术进行分离检测[1]。而在使用气相液相色谱技术时，需要使用高效色谱仪器，并且按照各种项规定进行色谱柱填料和流动相分组。

结构和文章维基类别等信息来计算相似度，而没有利用维基百科中所有的文本内容，计算方式简便，速度也提高了，但却已牺牲了准确性高为代价。基于维基百科的词语相似度算法主要的优点是维基百科提供了最新的语义信息和独特的信息结构。主要缺点是维基百科是并不像前面提到的三种语义资源由专业的人士或者团队收集而来，缺乏专业性质

6　总结

词语相似度的计算在自然语言处理领域有着非常重要的意义，是信息检索，文本分类等相关领域的基础。综上对基于四种不同语义资源的算法，前三种均存在更新缓慢的缺点，维基百科的出现弥补了这一点。与传统的语义词典相比，维基百科含有丰富的语义信息，涉及的知识面广阔，独特的信息组织方式的优点，同时其语义资源更新频率高，有效地提高了词语相似度计算的准确率。有机地融合维基百科和其他背景信息，能够在多种不同类别的词汇语义信息来源中取长补短，提高计算的准确性。所以，针对维基百科和通过融合方法的词语相似度算法将成为词语相似度研究今后的发展趋势。

参考文献：

[1] 吴思颖，吴扬扬．基于中文 WordNet 的中英文词语相似度计算 [J]．郑州大学学报，2010(06):42-2.
[2] 董振东，董强．知网 [EB/OL].[2012-03-20].www.keenage.com
[3] 王斌．汉英双语语料库自动对齐研究 [D]．北京：中国科学院计算技术研究所，1999.
[4] 刘群，赵捧未，刘怀亮．词语相似度计算研究 [J]．情报理论与实践，2007,30(01):105-108
[5]Li S J,Huang X,et al.Semantic Computation in Chinese Quertion-Answering Systm[J]. Journal of Computer science and Technology,2002,17(6):933-939.
[6] 梅家驹，竺一鸣，高蕴琦等．同义词词林 [M]．上海：上海辞书出版社，1983.
[7] 田久乐，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，2010(11):28-6.
[8] 吕立辉，梁维薇，冉蜀阳．基于词林的词语相似度的度量 [J]．研究与开发，2013(01).
[9]Strube M,Ponzetto S P。WikiRelate! Computing Semantic Related Using Wikipedia [C].In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI06).AAAI Press,2006:1419-1424.
[10]David MiLne.Computing semantic relatedness using Wikipedia link structure[C].In Proceedings of the New Zealand Computer Science Research Student Conference(NZCSRSC07),2007.

作者简介：蔡辉虎（1988-），男，福建泉州人，硕士研究生，研究方向：数据挖掘。