WSR:一种基于维基百科结构信息的语义关联度计算算法

孙琛琛 申德荣 单 菁 聂铁铮 于 戈

(东北大学信息科学与工程学院 沈阳 110819)

摘 要 该文提出了一种基于维基百科结构信息的语义关联度的计算方法——WikiStruRel(WSR).维基百科作为目前规模最大和增长最快的在线百科系统,其典型包括两个网状结构:文章网络和分类树(以树为主体的图),这两个网状结构包括了丰富的、明确定义的语义知识.WSR充分分析维基百科的文章网络和分类树,进而计算词语间的语义关联度.该方法没有涉及文本处理,算法开销较小,在3个数据集上的实验,取得了较好的准确率和覆盖度.

关键词 语义关联度;维基百科;文章网络;分类树

中图法分类号 TP311 **DOI号**: 10.3724/SP.J.1016.2012.02361

WSR:A Semantic Relatedness Measure Based on Wikipedia Structure

SUN Chen-Chen SHEN De-Rong SHAN Jing NIE Tie-Zheng YU Ge

(*College of Information Science and Engineering, Northeastern University, Shenyang* 110819)

Abstract This paper proposes a semantic relatedness measure based on Wikipedia structure: WikiStruRel(WSR). Nowadays, Wikipedia is the largest and the fastest-growing online encyclopedia, consisting of two net-like structures: an article referenced network and a category tree (actually a tree-like graph), which include lots of explicitly defined semantic information. WSR explicitly analyzes the article referenced network and the category tree from Wikipedia and computes semantic relatedness between words. While WSR achieves effective accuracy and large coverage by testing on three common datasets, the measure doesn't have to deal with text, resulting in low cost.

Keywords semantic relatedness; Wikipedia; article referenced network; category tree

1 引 言

在判断"汽车"与"全球变暖"、"社交网络"与"个人隐私"、"苹果"与"手机"的关系时,通常依赖个人常识和积累的知识,并通过综合判断得出比较准确且满意的结论(取决于个人的见识广博程度和智商).通常,人们认为"苹果"与"手机"的关联度要远大于"苹果"与"飞机"的关联度,因为"苹果电脑公司"的手机产品在全世界范围内广受欢迎;人们不会因为客机上偶尔提供苹果作为餐品而认为两者有很大关联度.然而,对于机器来说,判断不同词语之间的语义关联度是一个复杂而艰难的任务,需要现实世界的有关实体的诸多概念及其关系、常识和某些

收稿日期:2012-06-05;最终修改稿收到日期:2012-08-19.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316201)、国家自然科学基金(60973021,61003060)及中央高校基本科研业务费专项资金(N100704001)资助.**孙琛琛**,男,1987年生,硕士研究生,主要研究方向为信息网络. E-mail: bigchansuns@163.com. **申德荣**,女,1964年生,教授,博士生导师,主要研究领域为Web数据管理和数据空间. **单 菁**,女,1986年生,博士研究生,主要研究方向为Deep Web. **聂铁铮**,男,1980年生,博士,副教授,主要研究方向为数据集成. **于 戈**,男,1962年生,教授,博士生导师,主要研究领域为数据库和云计算.

特定领域的知识[1-2]作为支撑.为计算语义关联度,有些方法是通过对大型语料库进行统计分析来实现[3-4];有些方法则使用经过手工处理得出的语汇结构如同义词典[5-6].无论哪一种情况,背景知识都是一个限制因素.对于前者,无结构和不准确的语料库是难题;对于后者,范围和数量级的限制非常突出.

维基百科是目前最大和增长速度最快的百科知识库,有超过200万的文章数和数以万计的贡献者.其广泛的文章链接构成的相互参考网络、数目庞大的网络入口和层次的分类(以树为主体的图结构)能提供大量的明确定义的语义知识.基于以上原因,研究者热衷基于维基百科来进行语义计算.研究维基百科的最大挑战是,如何使这个庞大的百科知识变得机器可处理.为了克服上文中提到的背景知识的限制因素,本文提出了一种基于维基百科结构信息(文章网络和分类树)的新的语义关联度计算算法WikiStruRel(WSR).

本文的主要贡献如下:

(1)将维基百科的文章网络中的链接分类,并赋予不同经验权重,对目标概念结点的相邻结点进行层次划分,应用Jaccard系数,提出一种基于文章网络的语义关联度计算算法RelArtNet.

(2)利用维基百科的分类树的类本体性,提出一种基于分类树的语义关联度计算算法RelCatTree.

(3)综合基于文章网络和基于分类树的语义关联度计算法,提出一种基于维基百科的结构信息(文章网络和分类树)的语义关联度计算算法WSR,使得语义关联度准确性和计算效率得了提升.

(4)通过实验验证了本文提出的WSR的有效性.

2 相关工作

现有的语义关联度计算方法的主要区别在于背景知识来源的不同.几种关联度算法在测试词集WordSimilarity-353的准确率[7]如表1所示,其中算法准确率通过与人工识别的关联度相比较得到.

表1中的前两个是基于人工产生的语义词典Wordnet和Roget的语义关联度计算算法的准确率.基于Wordnet的算法[8]是通过语义词典的分类体系中的结点信息来判断结点间的关联度;基于Roget的算法[9]是通过计算语义词典的分类体系中结点的语义距离来判断结点间的关联度,距离越近,关联度越大.语义词典全部是人工处理得到的,基于语义词典的关联度计算方法受限于语义词典的词汇量.国内研究者基于HowNet[10]中文语义词典也进行了相关研究.

表1 已有语义关联度算法的准确率

语义关联度计算算法		准确率
基于同义词词典	Roget	0.33～0.35
	Wordnet	0.55
基于语料库	Latent Semantic Analysis (LSA)	0.56
基于维基百科	WikiRelate	0.19～0.48
	Explicit Semantic Analysis (ESA)	0.75
	Wikipedia Link-based Measure(WLM)	0.69

基于语料库的算法是通过对大量文本进行统计分析来得出语义关联度.由Scott Deerwester等人提出的Latent Semantic Analysis(LSA)[11]是最著名和效果最好的基于语料库的语义关联度计算算法.LSA算法核心在于有语义关联的词语被期望出现在同一文本中,算法高度依赖于基于语料库的词汇表.只有较大的语料库才能保证理想的准确度,因而,该算法的语料预处理的工作量非常巨大.

目前有多种基于维基百科的语义关联度计算方法,如WikiRelate、ESA和WLM.WikiRelate[12]是由Strube和Ponzetto提出的基于维基百科的层次分类结构(本文称之为分类树)的语义关联度计算方法.WikiRelate将基于Wordnet的方法"Path based measures"、"Information content based measures"和"Text overlap based measures"进行修改后应用在维基百科上,并且取得了与基于Wordnet相近的准确度.WikiRelate相比于基于Wordnet方法的优势是,维基百科能够提供更广的词汇覆盖度.Explicit Semantic Analysis(ESA)[13]是由Gabrilovich和Markovitch提出的,是迄今为止基于维基百科的准确率最高的语义关联度计算算法.ESA借用了向量空间模型思想,但它没有通过比较语汇权重向量来比较关联度,而是比较维基文章彼此间链接的带权向量,ESA中的向量是由人工定义的概念构成.ESA不仅可以计算词语的语义关联度,而且可以计算文本之间的语义关联度.但是ESA需要比较多的人工参与.Milne和Witten提出了基于维基百科文章链接关系的语义关联度计算算法Wikipedia Link-based Measure(WLM)[7].WLM采用向量空间模型和Normalized Google Distance[14]来处理维基百科中的文章链接,得到词语语义关联度.WLM

比 ESA 算法开销小，但从表 1 可知，准确度只略低于 ESA.

通过分析，我们发现，维基百科中由文章链接构成的相互参考网络和由分类组成的分类树（以树为主体的图结构）能够提供丰富的、明确定义的语义知识. 为此，本文充分挖掘了文章之间、分类之间以及文章与分类之间的关联关系，提出一种基于这两个网络结构的语义关联度计算算法 WikiStruRel (WSR).

3　维基百科和问题定义

本部分介绍维基百科以及相应的两个网络结构、两类链接和两类特殊页面的含义，同时给出了本文问题定义.

3.1　维基百科及其相关概念

维基百科是一个巨大的、协作的、免费的、开放的在线百科全书. 这个大型百科在线系统是目前最大且增长速度最快的百科知识库，其广泛的相互参考网络、数目庞大的网络入口和层次的分类（以树为主体的图结构）能提供大量的明确定义的语义知识. 从 2001 年发布上线到 2012 年 1 月，英文维基百科条目数已有 385 万个条目，全球所有 282 种语言的独立运作版本已突破 2100 万个条目，总登记用户也超越 3200 万人，总编辑次数更是超越 12 亿次. 这些数据说明了它具有大规模语料库的特征. 维基百科中的每一篇文章都描述一个单一的主题，文章题目像传统的同义词典中的语汇或短语一样，简明扼要且表达清晰. 每篇文章至少隶属于一个维基分类. 文章之间的超链接关系描述不同文章之间存在的（如等价、层次或者关联等）语义关系. 维基条目之间的链接构成了一个巨大的语义网络. 维基百科在 2004 年 5 月增加了分类，文章隶属于分类，分类构成了分类树（实际为图结构），这使其增加了语义词典的特征.

针对维基百科的内容组织结构，我们给出如下相关概念.

3.1.1　文章网络和分类树

图 1 揭示了维基百科中的文章网络和分类树的结构以及两个网络结构的关联. 分类树是一个以层次树为主体的有向图结构，其中 C_i 为任意分类；文章网络是一个庞大的包含大量的链接的有向图，其中 A_i 为任意文章，分类树延伸至文章，使得文章网络与分类树产生紧密关联.

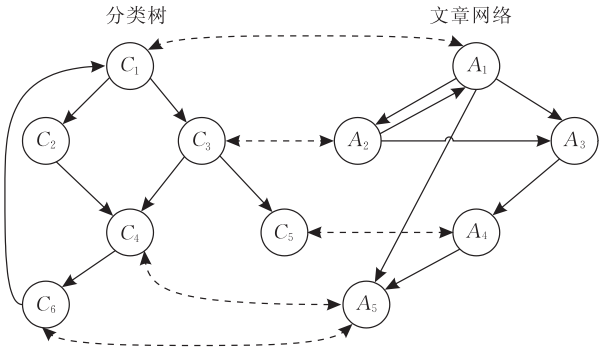

图 1　维基百科分类树和文章网络关系图

3.1.2　文章链接和分类链接

维基百科的文章 A_i、A_j 之间通过超链接相互联系，构成文章网络. 每篇文章都能链向多个其它维基文章. 维基编写者为不同文章中的关联的语汇或短语添加超链接，指向相应的维基页面. 我们视维基文章为结点，文章之间的超链接为从一结点到另一个结点的边，将得到一个有向图（如图 1 中 A_i 组成部分）.

维基百科中，除“Category：Contents”作为所有分类的根结点外，任意文章和分类都隶属于至少一个分类，比如文章“Semantic similarity”隶属于分类“Computational linguistics”和“Statistical distance measures”. 这两个分类可能还有多个双亲分类和多个子分类. 可见，维基百科的分类结构不是一个简单树型结构的分类系统，而是以树为主的有向图. 因此，维基百科的分类具有同义词典的特征.

维基百科中的文章与分类之间的链接是双向的，每篇文章指向一个或多个分类，一个分类指向多篇文章和多个分类. 可以将分类看作是文章的语义标签. 可见，文章网络和分类树（以树为主的图）具有紧密的关联关系. 文章网络中的链接代表文章的关联，分类间的链接则代表上下位关系或部分整体关系.

3.1.3　重定向页面和消歧页面

维基百科中，一个概念只会对应一篇文章来描述它，但一个概念可能存在多个同义词，维基百科为不同的同义词设置了重定向页面，将其重定向到唯一的文章页面. 比如“King”和“Monarch”都有国王的意思，维基百科中只有一篇名为〈Monarch〉的文章来描述这两个词，当我们搜索“King”，系统会自动重定向到〈Monarch〉.

人类语言中的同一个词可能会有多个解释，我们称之为歧义. 维基百科为了解决一词多义的问题，

设置了消歧页面,让用户在消歧页面中选择自己想要了解的意项.消歧页面中意项的格式一般为"歧义概念(注释)",维基会根据被选频率推荐一个首选解释,其它的解释按不同类别排列.比如概念"Ring",推荐解释是"Ring(jewellery)",而又按"Arts and entertainment"、"Music"、"Science and technology"、"People"、"Places"、"Sports"和"Other uses"分别排列.如下所示(不完全示例):

Ring(jewellery)

Arts and entertainment

Ring (film), a 1998 horror film by Hideo Nakata

Rings (short film), a 2005 horror film by Jonathan Liebesman

Music

Ring (The Connells album), 1993 Ring (Miliyah Kato album) Ring (Gary Burton album), 1974

3.1.4 维基概念

维基百科中的每篇文章都被认为是对一个维基概念的描述,而文章的题目就是维基概念的名称.如上文中的文章〈Monarch〉就是对维基概念"Monarch"的描述.这样,维基的数据集就可以看作维基概念的集合.

3.2 问题定义

词语之间的语义关系[15]有语义关联度、语义相似度和语义距离.为了研究完整性,下面给出这些概念的简略定义.

定义 1. 语义关联度.语义关联度描述两个词语之间的任何关联关系(包括如上下位关系、同义关系、反义关系和整体局部关系等),通常是两个词之间数据化的相似性.

定义 2. 语义相似度.语义相似度是关联度的一个特例,只包括词语之间的上位关系和同义关系.

定义 3. 语义距离.语义距离是基于距离的语义关联度的表示方法,两个词之间语义关联度越大,语义距离越小.

本文核心的研究是词语或短语之间的语义关联度的计算.首先,将待比较词语分别映射到维基概念;之后,通过计算维基概念间的语义关联度得到待比较词语的语义关联度.

4 词语映射

词语映射是将要比较的词语映射为维基概念.进行映射时,需要解决两个问题:一词多义(歧义)和同义词.

一词多义是指同一个词汇具有不同的意思,比如 lead 的意思有:(1) v. &n. 领导,引导;(2) v. 领先,占首位;(3) v. 通向,导致,引起;(4) v. 经验,过(生活);(5) n. 铅.我们需要根据上下文语境来推断 lead 的准确意思,金属 lead 是"铅",而"lead a hard life"中的 lead 是"经验,过(私生活)".同义词是指多个词汇具有同一个意思,比如 abandon 的同义词有:(1) desert;(2) forsake;(3) leave;(4) give up.所以在计算某一词与另外一词的语义关联度时,需要能够对它进行同义替代.

本文通过超链接来识别词语的候选概念.维基百科的文档中,跟某一重要主题相关的词汇或短语通常会有超链接指向该主题的文章,因此,维基文章中会有很多的超链接文本,并涉及到歧义和同义,比如,abandon 会根据不同的上下文而指向不同的文章,而 desert、forsake、give up 和 leave 很可能会指向相同的文章.

消歧页面可以解决一词多义,重定向页面可以将同义词重定向到同一文章.因此,本文映射的全过程是:首先,获得所有被计算词语语义相关的维基概念及其相关文章题目;然后,扫描括号中的解释部分实现最佳匹配.在映射过程中,有些文章题目可以直接被映射到,重定向页面可解决同义词的映射,而对于一词多义,可通过分析相应的连接关系消除歧义页面.

可见,在词语映射中,本文充分利用了维基百科的不同类型页面,并且只使用文章题目,而不涉及文本内容.本部分的目标是尽最大努力准确地实现概念映射,且仅需较少代价.

5 基于维基百科结构信息的语义关联度计算算法 WSR

维基百科的结构典型包括如下两个网状结构:文章网络和分类树.WSR 算法是充分利用这两个网状结构中的语义知识而提出的语义关联度计算方法.首先,分析文章网络的多层次结构,结合 Jaccard 系数,得到基于文章网络的语义关联度算法 RelArtNet;然后,根据分类树的类本体结构,结合基于本体的关联度算法思想,得到基于分类树的语义关联度算法 RelCatTree;最后,综合 RelArtNet 和 RelCatTree 得到 WSR 算法.

图 2 为维基百科链接结构示例，其中，概念结点 a 和概念结点 b 之间有多个直接或者间接相邻的其它概念结点，概念结点间的链接类型分为输入链接和输出链接两类. 相邻概念结点之间存在语义关系，有助于语义计算.

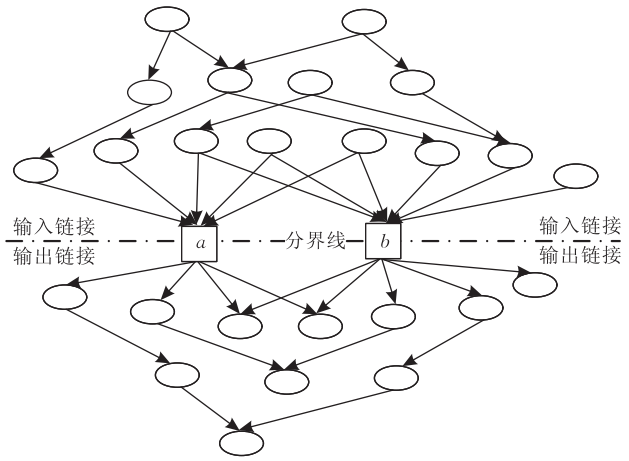

图 2　维基百科链接结构

5.1　基于文章网络的语义关联度计算算法 RelArtNet

5.1.1　Jaccard 系数简介

Jaccard 系数，也叫 Jaccard 指数，来源于统计学，最早由 Paul Jaccard 提出. 它可以用来比较两个样本集合的相似性和差异性. Jaccard 系数通过样本集合的交集除以样本集合的并集得到：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{1}$$

5.1.2　简单的基于文章网络的语义关联度计算算法 RelArtNetSimple

维基百科中，概念 a 与概念 b 分别有多个直接相邻的概念结点，它们之间的链接关系代表了一定程度的语义关系. 因此，可以通过对相邻概念结点的统计运算来得到概念 a 与概念 b 的语义关联度：

$$Rel(a,b)=\frac{|Neighbor(a)\cap Neighbor(b)|}{|Neighbor(a)\cup Neighbor(b)|} \tag{2}$$

其中，$Neighbor(a)$ 表示与概念 a 相邻的结点，$Neighbor(b)$ 表示与概念 b 相邻的结点；$|Neighbor(a)\cap Neighbor(b)|$ 为概念 a 与概念 b 的相邻概念结点的交集的个数；$|Neighbor(a)\cup Neighbor(b)|$ 为概念 a 与概念 b 的相邻概念结点的并集的个数.

使用 Jaccard 系数可保证语义关联度在 $[0,1]$ 闭区间上，关联度的值越大表示语义关联度越强.

但 RelArtNetSimple 算法存在如下不足：(1) 将所有链接同等对待，没有考虑权重区分度；(2) 只考虑了直接相邻的概念结点，未考虑间接关联的概念结点.

针对上述不足，我们进行了如下改进：(1) 引入带权重的链接，并通过链接的权重来计算概念的语义关联度；(2) 层次划分相邻结点，并按层次计算其语义关联度，各层次的语义关联度加权求和得到最终语义关联度. 详细介绍见下文.

5.1.3　引入带权重的链接

为给链接赋予权重，需要经历链接权值初始化、基于 TF-IDF 的链接权值演化和确定相邻概念结点权值 3 步完成.

(1) 链接权值初始化

维基百科中包括了多种链接，本文只关注特定的几个链接，并且给它们赋予相应的初始权值.

双向链接. 两篇文章分别有链接指向对方，表示两篇文章具有较强的语义关联. 比如说，概念“中国”和“北京”之间就有着很强的关联关系，因为主题为“中国”的页面表示中国的首都是北京，同时主题为“北京”的页面也表示北京是中国的首都. 相反，位于中国山西省平遥县洪善镇的“白家庄村”跟“中国”有较弱的语义关联，因此在“中国”的主题页面上并没有指向“白家庄村”的链接. 这种单向链接代表的语义关联度要比双向链接弱.

See Also 链接. 大部分的维基百科文章都有 See Also 链接. 这些链接指向的文章跟该文章具有很强的语义关联. 因此，See Also 链接对于语义关联度计算是非常重要的. 反向 See Also，即被 See Also 链接所指向的文章接受到的链接，同样具有很强语义关联.

同一分类下的文章之间的链接. 维基百科有一个丰富的分类结构(以树为主体的图结构)，而隶属于同一分类的文章之间应该具有明显的语义关联. 但不可避免的是，有一些分类是比较边缘的，可能会包括一些不关联的文章. 比如“中国大陆演员”包括了超过 1000 篇的文章，但这些演员中大部分(如‘周璇’和‘李胜素’)之间的语义关联并不太强.

维基文章中的普通链接. 代表一定的语义关联，通常经过人工语义判断发现，指向链接(从某篇文章指向另外一篇)的语义作用要强于被指向链接(从别的文章指向本文章).

表 2 给出了各种连接被赋予的经验权重.

表 2　不同类型链接的初始权值

链接类型	初始权值	链接类型	初始权值
See Also 链接	0.7	双向链接	1.0
反向 See Also 链接	0.4	同分类链接	0.6
NavBox 链接	0.6	信息盒链接	0.7
普通指向链接	0.5	普通被指向链接	0.3

(2)基于 TF-IDF 的链接权值演化

在一份给定的文件里,词频(Term Frequency,TF)指的是某一个给定的词语在该文件中出现的次数.逆向文件频率(Inverse Document frequency,IDF)是一个词语普遍重要性的度量.某一特定文件内的高词语频率,以及该词语在整个文件集合中的低文件频率,可以产生出高权重的 TF-IDF.因此,TF-IDF 倾向于过滤掉常见的词语,保留重要的词语.

本文借鉴 TF-IDF 思想,使用概念的相关链接出现的概率来代替 TF-IDF 中的词语出现的概率.设 s,t 分别是源概念和目标概念,那么 $s\rightarrow t$ 的权值描述为

$$w(s\rightarrow t)=w(s\rightarrow t)_0\times\frac{|s\rightarrow t|}{|s\rightarrow x|}\times\log\left(\sum_{y=1}^{all}\frac{|all|}{|y\rightarrow t|}\right) \tag{3}$$

其中,$w(s\rightarrow t)_0$ 是 s 到 t 的初始权值;$|s\rightarrow t|$ 是 s 到 t 的链接数目,$|s\rightarrow x|$ 是从 s 出发的所有链接数目,$|s\rightarrow t|/|s\rightarrow x|$ 是从 s 到 t 的链接占从 s 出发的链接的比例,对应 TF-IDF 中的 TF;$|all|$ 表示维基百科中的链接总数目,$|y\rightarrow t|$ 表示任意概念结点到 t 的链接数目,log 后面的部分是任意一条链接指向目标概念的反比例值,如果不存在该链接,则取 0,对应 TF-IDF 中 IDF.如果同时有很多文章指向了同一目标文章,那么在判断其中两篇文章的关联度时,不能给予这条链接太大权值.比如,如果两篇文章同时指向了"Art",而另外两篇文章同时指向了"Soft rock",显然前者的权值应该小于后者的权值.

对于间接相邻的概念结点,采用递减乘法计算链接权重.如图 3 所示,a 到 e 的权重为 $w(a\rightarrow e)=w(a\rightarrow c)\times(w(c\rightarrow e)\times\varphi)$,其中 φ 是层次递减系数,随着距离的增加,概念之间的关联度则减小.本文中 $\varphi=0.9$.

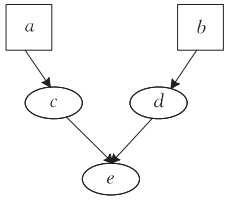

图 3 间接相邻的概念结点

(3)相邻概念结点的权值

为适应 Jaccard 系数公式,需要将链接的权值转换为概念结点的权值.因此,以 a,b 为源概念结点,则目标概念结点 x 的权值是

$$w(x)=\begin{cases}(w(a\rightarrow x)+w(b\rightarrow x))/2, & x\in(a\cap b)\\ w(a\rightarrow x)\text{ 或 }w(b\rightarrow x), & x\in(a\cup b-a\cap b)\\ 0, & \text{否则}\end{cases} \tag{4}$$

如果 x 是 a 和 b 的共同邻结点(直接或间接),那么 x 的权值是 $w(a\rightarrow x)$ 和 $w(b\rightarrow x)$ 的平均数;如果 x 只是 a 或 b 中的一个结点的单一邻结点,那么 x 权值是 $w(a\rightarrow x)$ 或 $w(b\rightarrow x)$;其它情况下,x 被认为与源概念结点无关,权值是 0.

5.1.4 层次划分相邻结点

将结点按层次划分,与源概念结点直接相邻的为第一层相邻结点,依次类推.对于某一层的邻结点,源概念的关联度描述为

$$Rel(a,b)=\frac{\sum w(x)}{\sum w(y)}, \tag{5}$$

$$x\in Neighbor(a\cap b),\ y\in Neighbor(a\cup b)$$

5.1.5 语义关联度计算算法 RelArtNet

结合结点层次结构和链接权重,基于文章网络的语义关联度描述为

$$Rel(a,b)=\alpha\times Rel_1+\beta\times Rel_2+\cdots+\omega\times Rel_N \tag{6}$$

其中,$Rel_1,Rel_2,\cdots,Rel_N$ 是概念 a,b 的相应层次的关联度,$\alpha,\beta,\cdots,\omega$ 是相应层次的权重,且 $\alpha+\beta+\cdots+\omega=1$.本文 $N=3,\alpha=0.6,\beta=0.3,\gamma=0.1$(本文的最外层权重).

5.2 基于分类树的语义关联度计算算法 RelCatTree

维基百科中每篇文章都至少隶属于一个维基分类,并且文章与分类之间存在紧密的语义关系.维基分类是以树为主体的图结构,具有分类系统的特征,因此,我们将分类系统的关联度计算方法用在分类树上.

Lin[16]提出了基于分类系统中的结点的关联度计算方法.比较分类系统中的两个分类 C_1 和 C_2(具体实例)的关联度时,并不比较它们自身.例如,计算工具汽车跟火车的关联度时,不是将工具汽车的集合与火车的集合(具体实例)进行比较,而是比较工具汽车类与火车类(抽象类).即用 $Rel(C_1,C_2)$ 来表示 x_1,x_2 的关联度,其中,x_1 和 x_2 是实例,C_1 和 C_2 是类,$x_1\in C_1,x_2\in C_2$."$x_1\in C_1$"和"$x_2\in C_2$"是独立的,因为从 C_1 中任选一个实例 x_1 和从 C_2 中任选一个实例 x_2 是完全没关系的."$x_1\in C_1$, $x_2\in C_2$"的信息量是 $-\log(P(C_1))-\log(P(C_2))$,其中,$P(C_1)$ 和

$P(C_2)$是从 C_1 和 C_2 中分别随机选取实例的概率.

在树型的分类系统中，如果 C_1，C_2 是分类系统中的类或概念，$x_1 \in C_1$，$x_2 \in C_2$，C_0 是 C_1，C_2 的最小共同蕴含(least common subsume)，那么，x_1，x_2 的关联度为

$$Rel(x_1,x_2)=\frac{2\times\log(P(C_0))}{\log(P(C_1))+\log(P(C_2))} \quad (7)$$

例如，图 4 为维基百科分类树中的有关“Train”和“Truck”的部分示例，分类旁边数字是信息量.

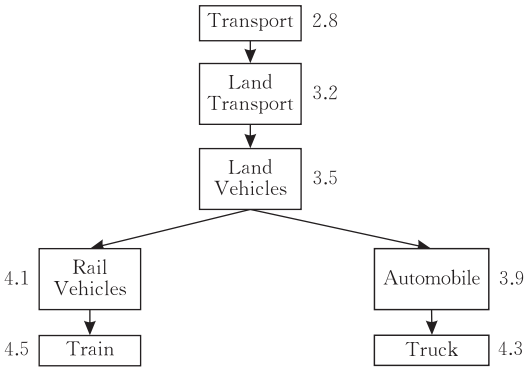

图 4　维基百科分类系统示例

那么，“Train”和“Truck”的关联度为

$$Rel(\text{Train},\text{Truck})=\frac{2\times\log(P(\text{Land Vehicles}))}{\log(P(\text{Train}))+\log(P(\text{Truck}))}=0.423.$$

5.3　语义关联度计算方法 WSR

综合基于文章网络的语义关联度计算方法 RelArtNet 和基于分类树的语义关联度计算方法 RelCatTree，得到基于维基百科结构信息语义关联度计算方法(WSR). WSR 的处理流程如图 5.

概念结点 a 和 b 的 WSR 语义关联度是

$$WSR(a,b)=\sum_{i=1}^{N}RelArtNet_i\times\alpha_i+\sum_{j=1,k=1,l=1}^{M,P,Q}RelCatTree_l(C_{a_j},C_{b_k})\times\beta_l,\quad \sum_{i=1}^{N}\alpha_i+\sum_{l=1}^{Q}\beta_l=1,\ M\times P=Q \quad (8)$$

其中，式(8)中加号前半部分表示基于维基百科文章网络的语义关联度，α_i 是不同层次概念结点的权重，本文中 $N=3$；式(8)中加号后半部分表示基于分类树的语义关联度，根据图 1 概念结点与分类的隶属关系，概念结点 a 和 b(对应图 1 中的 A_i)，至少分别隶属于一个分类 a_j 和 b_k(对应图 1 中的 C_i)，$RelCatTree(C_{a_j},C_{b_k})$ 是 a_j 和 b_k 的语义关联度，β_l 是其权重，M 和 P 分别是 a 和 b 分类的数目.

式(8)中的 α_i 和 β_l 通过实验对比得到. 从定性角度分析，文章节点是千万级而分类节点是百万级，前者包含的语义知识和关联要大于与后者. 本文通过 α_i 和 β_l 分别取多组不同的值，对相应的准确率进行比较，发现 $\sum\alpha_i=0.73$，$\sum\beta_l=0.27$ 的时候，准确率最大，其中，$\alpha_1=0.49$，$\alpha_2=0.15$，$\alpha_3=0.09$；β_l 根据目标节点所属分类的具体情况而定. 文章网络的层次 N 也是通过实验对比取得. 通过实验发现，随着 N 的增加，准确率不断增长，然而，当 $N>3$ 以后，准确率增加变的不明显，而算法开销却以指数增加，因此，综合考虑准确率和开销，本文 $N=3$.

6　实验与分析

6.1　数据源与数据集

本文使用的维基百科数据版本于 2011 年 1 月

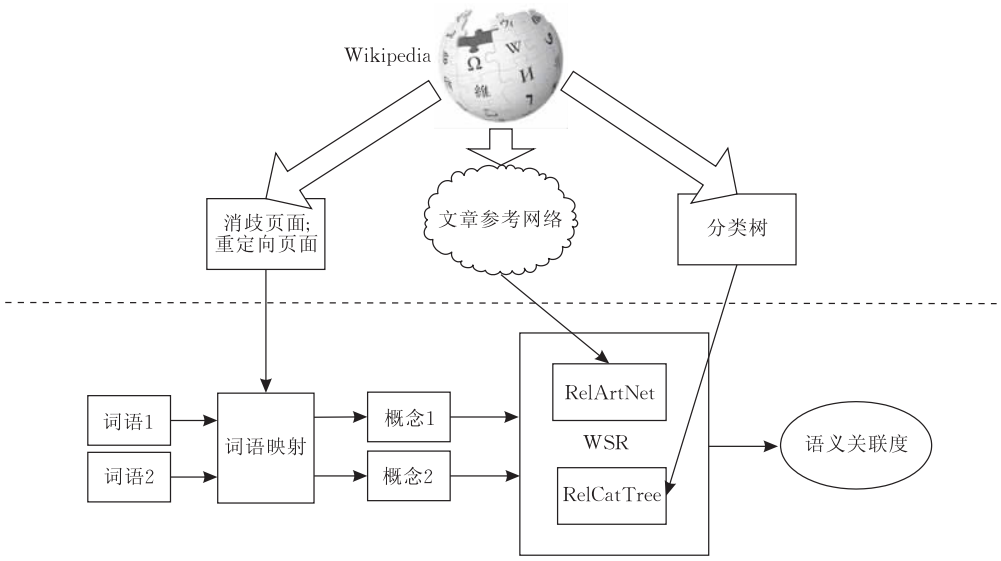

图 5　WSR 处理流程

15 日发布.我们选择了数据的 sql 备份版本,page.sql.gz(716.6MB)是页面信息(包括 id、标题等,不包括文章内容)sql 脚本,category.sql.gz(14.1MB)是分类信息(包括 id、分类题目等),categorylinks.sql.gz(698.9MB)是分类链接信息(包括分类之间、分类与文章之间的链接信息),pagelinks.sql.gz(3.5GB)是页面链接信息(包括页面间的多对多的链接信息).

本文的测试集是语义关联度研究领域常用的 3 个数据集:Miller and Charles (1991)(含 30 对词语)、Rubenstein and Goodenough(1965)(含 65 对词语)和 WordSim-353 datasets (Finkelstein et al., 2002)(含 353 对词语).

6.2 算法评价方法

定义 4. 准确率.采用 Spearman 等级相关系数来衡量目标算法与人工识别的结果的相关程度,称为准确率,即语义关联度.

Spearman 等级相关系数公式如下

$$\rho = 1 - \frac{6 \times \sum_i d_i^2}{n(n^2-1)} \tag{9}$$

其中,d_i是第 i 个元素的等级差,对应第 i 个词对的关联度算法结果和人工识别的结果在各自排序表中位置的差值,n 表示测试集的规模.

6.3 实验结果与分析

在 3 个测试数据集上,将本文提出的 3 个算法(RelArtNet、RelCatTree、WSR)进行实验,并对 RelArtNet 进行单因素改变实验,实验结果如下.

6.3.1 WSR 与传统方法的对比

图 6 是 WSR 算法与传统的语义关联度算法的准确率对比,由于 WSR 克服了传统方法的缺陷,因此准确率取得了显著的提高.算法时间开销方面,WSR 与传统方法相当.

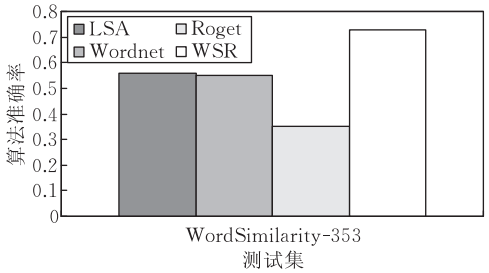

图 6 WSR 与传统方法的准确率对比

6.3.2 WSR 与 WikiRelate、WLM 和 ESA 的对比

观察图 7 可以发现:WSR 算法准确率明显优于 WikiRelate,也优于 WLM.WikiRelate 利用了维基百科的分类树中的语义知识,WLM 利用了维基百科文章网络中的页面链接信息(直接关链的页面),而 WSR 算法充分利用了维基百科文章网络中包括直接和间接的多层次的链接信息、文章与分类和分类之间的语义关联关系.ESA 是基于维基百科页面文本的算法,使用了大量的文本信息、分类信息和标题信息,它不仅可以计算词语之间的关联度,还可以计算文本之间的关联度.ESA 由于背景知识量庞大,需要大量的预处理,算法开销巨大.而 WSR 仅基于维基百科的结构信息,背景知识量远小于 ESA,因此,WSR 的算法的开销远小于 ESA.由于背景知识的差距,WSR 的准确率略小于 ESA.

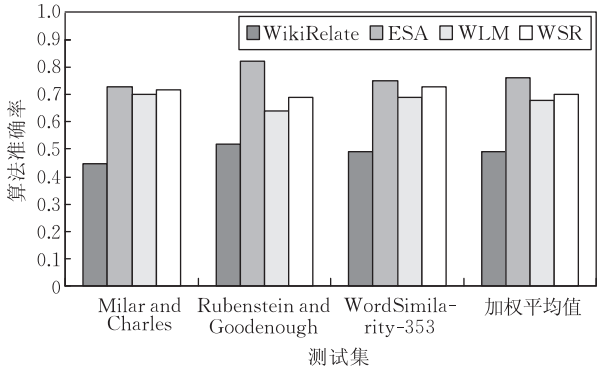

图 7 不同测试集上 WSR 与其它 3 种方法的准确率对比

从算法时间角度分析,WikiRelate 只涉及分类树,数量级是百万,WikiRelate 算法时间开销最小;WSR 和 WLM 都涉及到分类树和文章网络,数量级是千万,两者的时间开销是相当的;ESA 基于维基百科页面文本内容,因此,它的时间开销要远大于前三者.

权衡算法准确率和开销,WSR 是一个具有优势的词语或短语之间的语义关联度计算算法.

6.3.3 RelArtNet、RelCatTree 和 WSR 对比

本文提出的 3 个算法(RelArtNet、RelCatTree 和 WSR)分别运行在 3 个测试数据集上的测试结果见图 8.

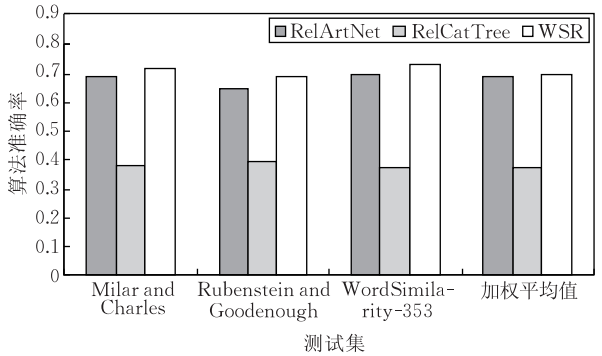

图 8 本文中提出的 3 种算法的准确率

通过对比图 7 和图 8 中测试结果发现，基于文章网络的 RelArtNet 算法的准确率略高于基于文章网络的 WLM 算法的准确率，而基于分类树的 RelCatTree 算法的准确率与基于分类树的 WikiRelate 算法准确率基本相当. 因为，RelArtNet 算法利用了目标概念节点的直接的和间接的相邻概念节点，而 WLM 只使用了目标概念节点的直接相邻的概念节点.

观察图 8，RelArtNet 的准确率明显优于 RelCatTree，因为维基百科的文章网络比分类树蕴含了更多的、更明确的语义知识. WSR 的准确率都高于同一测试集下的 RelArtNet 和 RelCatTree.

可见，同时利用文章网络与分类树蕴含的语义知识，可以取得比单独使用其中一类具有更高的算法准确率.

RelCatTree 和 WSR 的算法时间开销在相同数量级，而 RelArtNet 的算法时间开销远小于前两者.

6.3.4　RelArtNet 算法分析

(1) RelArtNetSimple 与单层 RelArtNet 对比

单层 RelArtNet 算法是指只考虑目标概念节点直接相邻的概念节点时的 RelArtNet 算法. 由图 9 对比可知，RelArtNetSimple 算法的准确率较低，经过赋予链接权值和演化、生成概念节点权值等处理步骤之后，得到了单层 RelArtNet 算法，取得了更理想的算法准确率. 单层 RelArtNet 算法的时间开销明显大于 RelArtNetSimple 算法.

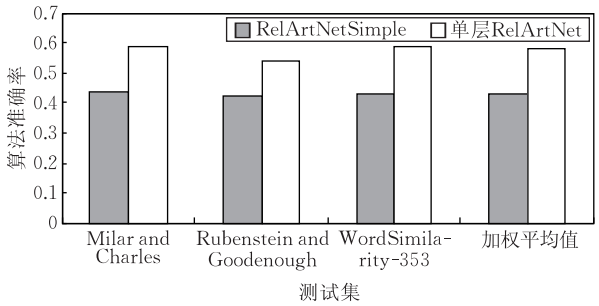

图 9　RelArtNetSimple 算法与单层 RelArtNet

(2) 层次划分节点对算法 RelArtNet 的影响

本小节研究对节点划分层次对算法 RelArtNet 的影响，实验数据如图 10. 图 10 中单层 RelArtNet 是指只考虑目标概念节点的直接相邻概念节点的 RelArtNet 算法，双层 RelArtNet 是指考虑直接相邻和次相邻的概念节点的 RelArtNet 算法，依次类推. 算法 RelArtNet 中，层次划分节点具有重要的意义，它将相邻概念节点的范围从直接相邻拓展到了间接相邻. 观察图 10，随着层次的增加，算法的准确率明显地提高；关注每个测试集上第 1 列的值、第 1、2 列与第 2、3 列的差值可以发现，随着层次的增加，越靠近外层的概念节点，对目标概念节点的语义关联度影响越小. 本文的实验中，随着 N 的增加，准确率不断增长，然而，当 $N>3$ 以后，准确率增加变得不明显，而算法开销却以指数增加，因此，综合考虑准确率和开销，本文 $N=3$.

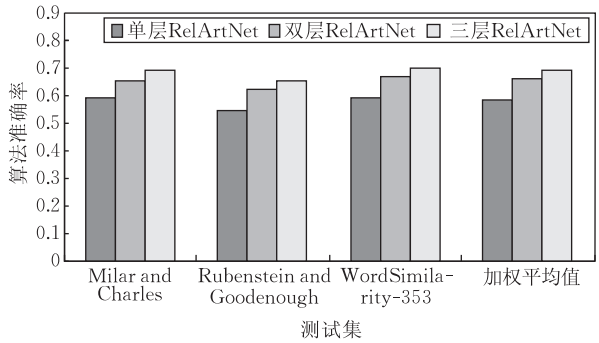

图 10　层次划分节点对算法 RelArtNet 的影响

7　结束语

语义关联度计算是基础性研究课题，它在信息检索、自然语言处理和人工智能等研究领域有着重要地位. 语义关联度计算的准确性会直接影响计算机处理信息的准确性. 本文主要从维基百科结构信息出发，分析了结构信息中包含的语义知识及关联，提出了一种基于维基百科结构信息的语义关联度计算算法 WSR. 在后续的工作中，我们将进一步研究如下方面的问题：

(1) 维基百科包含最丰富语义知识的是维基文章，即维基百科的页面内容，在本文研究的基础上，对页面内容进行语义分析和利用，必定会提高语义关联度计算的准确率.

(2) 本文研究的链接全部是维基百科的内部链接，除内部链接外，维基百科还有大量指向维基百科以外的链接，通过这些链接，使得它与更广阔的英特网联系起来，英特网拥有海量的信息. 通过对维基百科结构化的、明确的语义知识和英特网中的海量信息进行分析，可能会提高语义关联度计算的覆盖度和准确率.

(3) 在本文的研究基础上，希望将 WSR 应用在信息检索(如检索推荐)、知识关联和自然语言处理等研究任务中.

参　考　文　献

[1]　Buchanan B G, Feigenbaum E A. Forward//Davis R, Lenat D B.

Knowledge-Based Systems in Artificial Intelligence. New York: McGraw-Hill, 1982: 39-51
[2] Lenat D, Guha R. Building Large Knowledge Based Systems. New York: Addison Wesley, 1990
[3] Ricardo B Y, Berthier R N. Modern Information Retrieval. New York: Addison Wesley, 1999
[4] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391-407
[5] Alexander B, Graeme H. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 2006, 32(1): 13-47
[6] Mario J. Roget's thesaurus as a lexical resource for natural language processing [Ph. D. dissertation]. University of Ottawa, Ottawa, 2003
[7] Milne D, Witten I H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links// Proceedings of the 23th Association for the Advancement of Artificial Intelligence. Chicago, US, 2008: 25-30
[8] Philip R. Using information content to evaluate semantic similarity in a taxonomy//Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada, 1995: 448-453
[9] Mario J, Stan S. Roget's thesaurus and semantic similarity// Proceedings of Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria, 2003: 212-219
[10] Li Yun. Mining semantic knowledge from Chinese Wikipedia [Ph. D. dissertation]. Beijing University of Posts and Telecommunications, Beijing, 2009
[11] Thomas K L, Peter W F, Darrell L. An introduction to latent semantic analysis. Discourse Processes, 1998, 25(2-3): 259-284
[12] Strube M, Ponzetto S P. WikiRelate! computing semantic relatedness using Wikipedia//Proceedings of the 21st National Conference on Artificial Intelligence. Boston, US, 2006: 1419-1424
[13] Gabrilovich E and Markovich S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis// Proceedings of the 20th International Joint Conference on Artificial Intelligence. International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 163-168
[14] Cilibrasi R L, Vitanyi P M B. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383
[15] Jorge G, Eduardo M. Web-based measure of semantic relatedness//Proceedings of the 9th International Conference on Web Information Systems Engineering. Berlin, Germany, 2008: 136-150
[16] Lin De-Kang. An information-theoretic definition of similarity //Proceedings of the 15th International Conference on Machine Learning. San Francisco, USA, 1998: 296-304

SUN Chen-Chen, born in 1987, M. S. candidate. His research interest is in information network.

SHEN De-Rong, born in 1964, pfofeesor, Ph. D. supervisor. Her research interests include Web data management and data space.

SHAN Jing, born in 1986, Ph. D. candidate. Her research interest is Deep Web.

NIE Tie-Zheng, born in 1980, Ph. D., associate professor. His research interest is data integration.

YU Ge, born in 1962, professor, Ph. D. supervisor. His research interests include database and cloud computing.

Background

This work is supported by the National Basic Research Program (2012CB316201), the National Natural Science Foundation of China (60973021, 61003060) and the Fundamental Research Funds for the Central Universities (N100704001). Semantic relatedness is a fundamental branch of computer science, which is popular among information retrieval, artificial intelligence, natural language processing and so on. Semantic relatedness between words helps lots in intelligently dealing with nowadays' fast growing information produced by internet and mobile internet. Some researchers use statistical analysis of large corpora to compute semantic relatedness such as Latent Semantic Analysis (LSA) while others deal with knowledge bases and get lexical structures like Taxonomies and Thesauri to compute semantic relatedness, such as Wordnet and Roget. However, both are limited by background knowledge; the former is bad structure and imprecise, and scalability and scope limit the latter.

Wikipedia is an excellent semantic knowledge base, consisting of the article referenced network and the category tree, which are two structures like networks, with quite amounts of explicit semantic knowledge in good structures. Researchers try to break the bottleneck above by Wikipedia, such as WikiRelate, ESA, WikiWalk and WLM, which achieve some breakthroughs. This work analyses article referenced network and category tree in Wikipedia. Then, authors propose a new semantic relatedness algorithm named RelArtNet, which bases on hierarchically divided wiki-concepts with weights in the Wikipedia article referenced network, and a new semantic relatedness algorithm named RelCatTree based on category tree of Wikipedia. Finally, an integrative algorithm named WSR is given, which combines both RelArtNet and RelCatTree and take advantages of both. WSR is better than traditional methods and other Wikipedia based methods, considering cost and correlation between method's results and humans' judgments.