术语相似度计算方法研究

SHUYU XIANGSIDU
JISUAN FANGFA YANJIU

徐　健◎著

中山大学出版社

中山大学“985工程”资助
教育部人文社会科学研究项目（09YJC870031）

术语相似度计算方法研究

SHUYU XIANGSIDU
JISUAN FANGFA YANJIU

徐　健◎著

中山大学出版社
·广州·

图书在版编目（CIP）数据

术语相似度计算方法研究/徐健著. —广州：中山大学出版社，2012.9
ISBN 978-7-306-04307-8

Ⅰ. ①术…　Ⅱ. ①徐…　Ⅲ. ①自然语言处理—研究　Ⅳ. ①TP 391.1

中国版本图书馆 CIP 数据核字（2012）第 217128 号

出 版 人：祁　军
策划编辑：周建华　曹丽云
责任编辑：曹丽云
封面设计：林绵华
责任校对：李　霞
责任技编：何雅涛
出版发行：中山大学出版社
电　　话：编辑部 020-84111996，84113349，84111997，84110779
　　发行部 020-84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传　真：020-84036565
网　　址：http://www.zsup.com.cn　　E-mail:zdcbs@mail.sysu.edu.cn
印 刷 者：广州中大印刷有限公司
规　　格：787mm×1092mm　1/16　14.25 印张　330 千字
版次印次：2012 年 9 月第 1 版　2012 年 9 月第 1 次印刷
印　　数：1～1500 册　定　价：35.00 元

作者简介

徐健（1977－），男，中山大学资讯管理学院讲师，情报学博士。2000年在西安交通大学获学士学位，2003年在中山大学获硕士学位，2010年在中国科学院获情报学博士学位。2003年7月在中山大学硕士毕业后留校任教至今。主要研究方向为：①智能信息处理；②网络信息挖掘；③术语相似度计算及应用技术。已发表研究论文30篇。目前主持1项教育部人文社会科学研究项目“从科技文献中挖掘术语相似性及其在知识发现中的应用”，1项国家社会科学基金项目“用户评论情感分析及其在竞争情报服务中的应用研究”，并参与多项国家、省部级科研项目工作。

内容提要

对术语相似度计算方法的研究，为多种知识发现和自然语言处理任务的开展创造了条件。尽管术语相似度计算相关研究已经或正在探索基于各种特征或知识资源开展术语相似关系挖掘任务的思路和技术方法，但是在相似度指标的计算以及高效集成多种术语相似度指标方面仍存在较多问题。本书在全面介绍当前各种典型术语相似度计算思路的基础上，针对应用中实际存在的问题，提出或改进了基于语词、基于语境以及基于网络资源的术语相似度指标计算方法，设计和实现了多种相似度指标高效集成计算模型，有效地提高了术语相似度计算的综合性能。

本书可作为自然语言处理、知识发现等相关方向的教学、科研人员研究的参考资料。

目　录

图目录

表目录

第1章 绪 论

1.1 术语相似度计算研究缘起

术语代表了研究者之间进行交流的一种方式。当术语出现在文献中时，它代表了特定的领域概念，如实体、过程以及功能等，并且在特定主题内表现出高相关性。领域术语具有动态性特点，出现密集度高，并且在持续不断的变化中[1]。术语是非常有用的语义单位，因为它能够独立地表达某个完整概念。它能够被用在信息检索系统中作为检索式命中返回文档的描述、搜索索引的基础、浏览一个集合的方式，或者文档聚类的重要依据。术语还能够帮助用户快速了解文集内容，提供从文档集中获取文档的快速入口，通过可视化和对重要语词的强调，方便读者浏览文档等。

对于术语的语义相似性概念，不同的学者对此有着不尽相同的解释。维基百科认为，术语语义相似性是一个用来说明术语列表中的术语之间语义内容相似程度的度量[2]。一些文献也强调术语的语义相似性概念和语义相关性概念是不同的。邱松泽[3]认为，语义关联度包括了各种关系，比如，整部关系或成员关系、同义关系、功能关系、联合关系和包容包含关系。语义相似度是关联度的特殊情况，它对应于语义关联度中的同义关系和包含关系。刘群等人[4]认为，词语相似性反映的是词语之间的聚合特点，而词语相关性反映的是词语之间的组合特点。同时，词语相关性和词语相似性又有着密切的联系。如果两个词语非常相似，那么这两个词语与其他词语的相关性也会非常接近；反之，如果两个词语与其他词语的相关性特点很接近，那么这两个词一般相似程度也很高。然而，更多的文献对于语义相似性、语义相关性、语义距离等概念并没有加以严格区分，而是交替地使用这些术语。本书中所指的术语相似度倾向于采用维基百科所给出的定义，即“术语相似度反映出术语 A 与术语 B 有多相关”。

对术语语义相似度计算方法的研究，是开展多种自然语言处理任务和知识挖掘任务的前提和基础。术语语义相似度计算方法可以用于进行文档内和文档间的术语合并和规范化。例如，同一概念的不同表达方式，有些并不符合构词法规则，可以通过语境分析、句法分析等方法判断其相似性，进而对其进行合并，达到术语规范化的目的。此外，一些学者通过术语的语义相似度计算来发现同义词、近义词关系，进而使用这种关系进行聚类或其他处理任务。例如，Turney 等人[5]基于 Web 搜索引擎返回的命中（hits）数量，定义了一个 PMI-IR（point-wise mutual information，information retrieval）测度来识别同义词。Matsuo 等人[6]使用一个相似的方法来测度语词之间的相似性，并将他们的方法应用在基于图的（graph-based）语词聚类算法中。刘群等人[4]在基于实例的机器翻译任务中，基于 WordNet 同义词集计算语词相似度，进而找出相似的翻译实例。通过术语之间的语义相似度计算，还可以发现新的关系、类、实例等，并以此为基

础自动或半自动地使用新关系、新类对本体进行更新，使用术语实例对知识库进行更新[7]。基于术语语义相似度计算的异常值探测也是一种强有力的知识发现方法。例如，通过文本挖掘获取的具有相似性的实体，在特定的实验条件集合下表现非常不同，这可能暗示实验揭示了一些未知的现象，值得进一步调查研究。相似的，从文本中挖掘的关系也能揭示出文献中的一些不一致性或矛盾，或通过建议可能的实体间的关联来识别现有知识中的不足。此外，术语语义相似度计算还为术语分类、文档分类、术语聚类、文档聚类等提供了有力支持。例如，对于公司类内容的聚类，很难将财务、人事这些语词与公司直接关联起来，而缺少这种关联的聚类结果可能无法令人满意；如果能通过相似度计算，得到公司以及与公司类似的词，就能提高聚类准确度。

广泛的应用需求刺激了术语相似度计算相关研究的发展。研究者针对相似术语在各个层面反映出的相似性特征，从不同角度提出了计算术语相似度的思路。典型的思路有：基于语词构成特征的术语相似度计算，基于句法模板的术语相似度计算，基于语境特征的术语相似度计算，基于语词知识体系的术语相似度计算，基于网络知识资源的术语相似度计算。

上述这些计算思路根据某种或某些能够反映术语间相似程度的特征进行相似度计算，在解决某些特定任务时是适用的，但是普遍存在较强的限制性。例如，在基于语词构成特征的术语相似度计算方面，一些语义上相似但语词构成上并不相似的术语对就很难获得良好的计算结果；在基于语境特征的术语相似度计算方面，计算效果很大程度上取决于文集本身的大小和质量。此外，这些计算思路所借助的术语相似特征具有明显的互补性。因此，在对上述计算思路进行改进的基础上，对基于不同思路的术语相似度计算指标进行集成，就成为克服术语相似度计算的应用局限性，提高术语相似度计算性能的必然选择。也有学者提出，通过线性加权的方式将多种术语相似度计算指标进行集成，得到综合反映术语语义相似度的计算结果，但是这类方法在实际操作时难以确定权值分配，而不科学的权值分配会影响到整体计算效果。

本书认为，术语相似度计算可以看做一个分类问题，而借助支持向量机（support vector machine，SVM）分类器能够实现对各种类型术语相似度指标的有效集成。一方面，在仅需提供少量已标引语料的情况下，通过 SVM 自动学习和自动分类操作，能够实现领域术语语义相似度的快速计算；另一方面，可以在充分考察各种术语相似度指标计算特点的前提下，以领域术语相似度网络为基础，合理安排具有条件判断机制的计算流程，在基本不影响计算效果的情况下，大幅提升计算速度。此外，本书还重点针对当前基于语词、语境和搜索引擎的术语相似度计算方法存在的一些问题进行深入研究，力图在术语相似度指标的计算效果上能够有所提升。

1.2 研究的目的及意义

1.2.1 研究目的

通过对相关文献调研和对大量相似术语对的特征分析，本书认为术语之间的相似性

可以通过多方面的特征表现出来。而对这些不同方面相似度特征指标的计算和集成，应该能够更加客观、全面地反映术语语义相似程度。因此，本书把研究和探索基于机器学习的术语语义相似度计算方法作为总体研究目标，以期达到自动、高效地集成多方面术语相似度特征的目的。此外，为了获得更有效的术语相似度特征指标，本书希望从现有术语相似度算法存在的问题着手，结合实际计算过程中所积累的经验和心得，提出新的更为有效的术语相似度指标计算方法，或对已有算法进行改进。

本书所要达到的具体研究目的有以下五个方面：

（1）对当前术语相似度计算的典型方法和思路进行全面、系统的调查研究、分析和总结，为进一步的术语相似度计算研究提供借鉴。

（2）针对目前已有的语词相似度计算方法对术语中主词的语义相似匹配考虑不足，以及未能对术语中的修饰词相似情况进行权重区分等问题，提出基于术语中主词软匹配的 SSHW 术语相似度计算方法。通过将术语中主词匹配由原来的精确匹配改为基于词干和编辑距离的模糊匹配，以及根据修饰词与术语中主词之间的距离长度为修饰词分配不同相似度权重等方法，以期能够有效提升基于语词的术语相似度指标计算效果。

（3）针对目前术语语境相似度算法采用 n-Gram 语境模板质量不高这一问题，从语句中语词间的依赖关系入手，提出基于语境依赖关系模式的相似度算法（dependency relation context pattern，DRCP），希望能够通过高质量术语语境模板的构造和获取，在术语语境相似度指标方面能有所提升。

（4）针对目前基于搜索引擎的术语相似度算法在领域限定特征方面体现不足这一问题，在 Web-PMI 术语相似度计算方法的基础上，引入特定领域限定因素，以期提高 Web-PMI 算法在进行特定领域术语相似度计算时的效果。

（5）针对当前术语相似度指标集成过程中，各项相似度指标权重难以确定的问题，构建基于机器学习实现术语语义相似度指标集成的计算模型。该模型将术语相似度指标的集成看做一个分类问题，通过在少量训练集合上的训练获得术语相似度判断模型，并在此基础上开展以相似度指标向量为输入的相似度集成计算。此外，根据各种相似度指标的计算特点，合理组织具有条件判断机制的计算流程，希望能够在相似度计算的效果和效率方面得到有效提升。

1.2.2 研究意义

术语的相似性特点可以通过语词构成层面、句法构成层面、语境特征层面等多个方面特征反映出来。如果能够采用合理有效的方法，将这些反映术语相似关系的特征析取出来并进行量化表示，进而通过有效的集成方法将这些片面的指标集成在一起，就能够较为全面、客观地反映术语间的相似程度，为各种以术语相似度计算为基础的任务开展提供良好支持。

本研究具有以下三方面的意义：

（1）有效地计算出术语之间的语义相似度，为自然语言处理多项任务的开展奠定基础。术语规范化、识别同义/近义词、基于实例的机器翻译等任务都将术语相似度计算作为任务过程的核心模块。

（2）术语语义相似度计算为挖掘文本中的潜在知识提供支持。例如，在本体构建和知识库构建任务中，通过计算术语间的语义相似度，能够发现术语之间存在的新关系，并在此基础上对本体进行完善；判断为相似的术语可以作为相似类的实例存入知识库。此外，术语语义相似度计算还能够对领域内术语的异常值（outlier）探测、社群网络挖掘（community mining）等知识挖掘任务提供支持。

（3）术语语义相似度计算是领域结构分析的基础。领域中的概念主要以术语的形式表现出来。通过对术语间语义相似度的计算，能够发现领域中术语间的相似关系，在此基础上可以进行领域术语聚类，从中发现领域结构以及领域动态变化情况。

1.3 相关概念界定

1.3.1 术语

术语又称为技术名词、科学术语、科技术语或技术术语，是科学工作者开展交流的一种方式。从形式方面看，术语可以是简单术语（只有一个语词的术语，例如，声、光、电等），也可以是复合术语（由两个或更多个语词构成的术语，例如，声波、光束、电压等）。从内涵方面看，术语具有单名单义性、简明性、稳定性、合乎语言习惯等特征。“汉典”将术语定义为：专门学科的专门用语[8]。维基百科将术语定义为：是在特定专业领域中一般概念的词语指称，一个术语表示一个概念[9]。

为了更好地说明问题和开展相关实验，如不作特别说明，本书中所提及的“术语”一词特指通过自动或半自动的方式从英文电子期刊文摘中抽取的、能够独立表达领域内某个完整概念的英文名词性语词或短语。“术语”不限于特定领域术语词典中出现的语词或词组，一些新出现的或表述并不完全符合现有术语规范的语词或词组也在研究范围内。

1.3.2 术语语义相似度

本书中所提及的“术语语义相似度”概念倾向于采用维基百科对“术语语义相似度”的定义，即：术语语义相似性是一个用来说明术语列表中的术语之间语义内容相似程度的度量[2]。这里需要特别强调，术语语义相似度和术语语词相似度是不同的。前者侧重于分析和计算术语本身所代表的概念在语义内涵上的相似程度，而后者则侧重于仅从术语的语词构成角度来计算术语之间的“类似”程度。例如，对于“classification”和“assortment”这对术语，采用术语语义相似度计算方法来计算，两者应该是相似的，但采用术语语词相似度计算方法来计算，则结果很可能是两者不相似。

尽管本书所提出或应用的一些术语相似度计算方法能够计算术语对之间的量化相似度值（一般相似度计算结果的范围在[0,1]之间），出于实际应用需求的考虑（例如，基于术语相似度计算的术语聚类任务），本书中对术语相似度最终计算结果进行二类划分，即“相似”或“不相似”两种计算结果。这在最终形成的术语相似度网络中以术语之间的连线表示术语对是相似的。此外，本书研究的目标是挖掘术语对潜在的语义相

似性关联，因而本书并不刻意去识别存在于术语之间的关系类型，而仅仅是发现它们之间的相似性关联。

1.4 研究思路与方法

1.4.1 研究思路

本书的总体思路框架如图1-1所示。首先，阐述了本书研究背景，包括研究的必要性和现有方法的不足。其次，提出了要研究的两个问题，对本书关键问题进行综述，

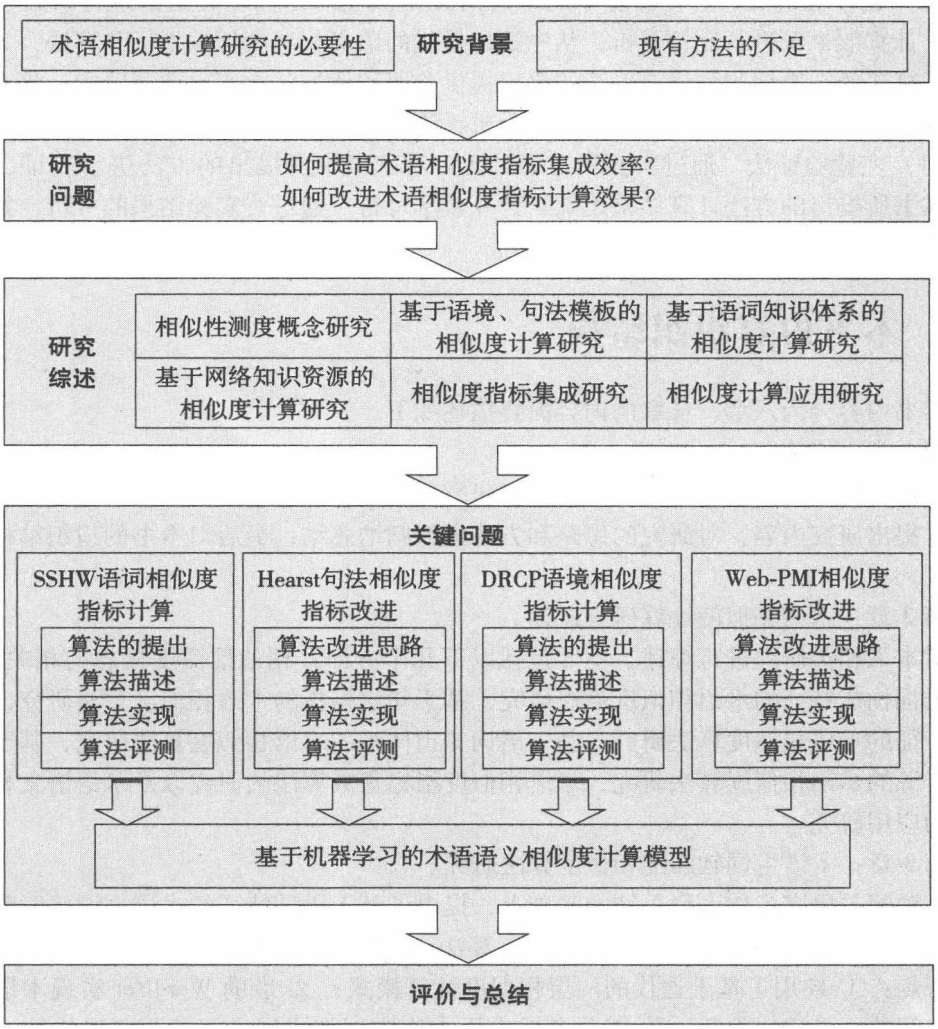

图1-1 研究思路框架