THE TECHNOLOGY AND APPLICATION OF SEMANTIC COMPUTING BASED ON KNOWLEDGE INTEGRATION IN A BIG DATA ENVIRONMENT

大数据环境下基于知识整合的语义计算技术与应用

蔡圆媛 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书立足于基于知识整合的词汇语义相似度计算技术及其应用，提出了整合两类语义资源的语义相似度计算方法，从语义特征的选择与提取、语义特征融合、语义计算这三方面内容梳理了知识脉络。

本书内容涉及当前主流的技术，如深度学习、文本向量化，附有大量的理论与技术介绍、实验数据、图表以及结果分析，语言通俗易懂，内容结构合理，有助于读者对相关知识概念有一个较为清晰的认识。

本书适宜作为高等院校开设的“信息检索”“自然语言处理”等课程的参考书，也可作为人工智能相关领域科研人员的参考资料。

图书在版编目（CIP）数据

大数据环境下基于知识整合的语义计算技术与应用 / 蔡圆媛著. —北京：北京理工大学出版社，2018.8

ISBN 978-7-5682-6124-1

Ⅰ.①大…　Ⅱ.①蔡…　Ⅲ.①语义分析-自然语言处理-研究　Ⅳ.①TP391

中国版本图书馆 CIP 数据核字（2018）第 185439 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址 / http：//www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 北京地大彩印有限公司
开　　本 / 710 毫米×1000 毫米　1/16
印　　张 / 13.5　　责任编辑 / 梁铜华
字　　数 / 192 千字　　文案编辑 / 曾　仙
版　　次 / 2018 年 8 月第 1 版　2018 年 8 月第 1 次印刷　责任校对 / 黄拾三
定　　价 / 54.00 元　　责任印制 / 王美丽

前　言

近年来，随着社会信息化程度日益提高，特别是移动互联网技术迅速发展，互联网中的数据量呈指数级剧增，我们迎来了大数据时代。大数据具有规模大、种类多、速度快、价值密度低等特点，涵盖了文本、语音、视频、图片等数据类型，其中蕴含的海量信息给我们带来了无穷的价值，改变着我们生活的方方面面，推动着科学决策智能化水平不断提升，并促进现代社会不断进步。在此背景下，智能决策服务对大数据的需求已经从单纯搜集和获取信息，提升为自动化知识推理。人们希望计算机能够理解人的意图，从海量数据中自动提取有价值的信息；实现对大数据的洞察，为决策提供支持；帮助人们提高数据分析效率，降低人力投入。因此，人们对计算机的数据处理与挖掘能力提出了更高的要求。

在大数据中，相对常见的文本数据由词汇、句子、段落、文档等不同粒度的语言单元构成，是一定主旨与意义的表达形式。对于文本的理解依赖于如何分析其中包含的自然语言的语义（即如何计算其语义），具体是指理解、解释自然语言中各个组成单元的具体含义，构建语言单元的语义表示。然而，在语义计算中，语义相似度的计算是其中一个重要内容与难题。语义相似度可以作为挖掘词汇关联的重要依据，在自然语言处理任务中有助于计算机准确理解语句和文档的内容。

根据文本语义资源的来源，典型的语义相似度计算方法包含三类：基于结构化知识库的方法、基于非结构化语料的方法，以及整合利用异构资源的混合方法。基于结构化知识库的方法主要基于领域专家人工建立和维护的语义网络、知识图谱等知识库，词汇覆盖率较低，缺乏可扩展性；基于非结构化语料的方法主要采用统计模型和无监督机器学习技术，从语料库中抽取语义信息，语料库虽然包含丰富的词汇，但是其非结构性导致计算机难以从中提取词汇的有效语义特征信息。针对前两类方法的研究均得到了广泛的关注和应用，而对整合两类资源的混合方法的研究则起步较晚。此外，随

着语义计算方法依赖的语义资源的种类、规模不断发展，从异构数据源中提取语义信息与知识进行有效整合被证明具有较好的效果。因此，近几年有不少研究者关注于将知识工程与大数据机器学习模型的结合，提出异构数据资源的知识整合方案，以及融合不同种类方法的优势的混合计算方法。

本书立足于基于知识整合的词汇语义相似度计算技术及其应用，以异构数据源为对象，从语义特征的选择与提取、语义特征融合、语义计算这三方面内容展开知识脉络的详细描述。书中附有大量的理论与技术介绍、实验数据、图表以及结果分析，有助于读者对相关知识概念有较为清晰的认识，能够正确、直观地理解语义计算的内容。

本书内容涉及当前主流的技术，如深度学习、文本向量化、语义相似度计算。全书共分为六章，前五章提出多种整合异构数据的概念/词汇相似度计算方法，并且给出了各计算方法对应的应用案例(如基于语义的 Web 服务发现)，介绍了语义计算理论（包括国内外现状、发展趋势以及存在的主要问题)，重点描述了知识图结构、向量空间模型等概念，提出了在选择语义计算表示方法时应遵循的原则以及进行词汇语义计算的几种方法。最后一章主要介绍了知识库与深度学习技术的结合，分析了词汇语义计算的重要性及未来研究方向。本书重点介绍了基于 WordNet 的图结构和词汇的低维向量表示，分别从概念信息含量的量化模型、语义增强的词向量、度量方法的优化组合三方面，提出了在知识库和语料库中对语义知识的整合方法，进行了详尽的对比实验，并且在具体的 Web 服务发现应用上验证了相关方法的有效性。

本书包含了大量的语言统计模型、模型的验证标准等知识点，因此适宜作为高等院校开设的“自然语言处理”“信息检索”等本科生、研究生相关课程的参考教材。本书中的理论部分内容是课程知识的补充和延伸。此外，本书所涉及的内容可以作为基于深度学习的文本理解研究的基础，尤其是关于知识整合如何应用于语义计

算，能够给开展相关研究的读者提供新思路。

本书涉及的内容及工作依托农产品质量安全追溯技术及应用国家工程实验室，是在北京市自然科学基金青年项目“面向机器阅读理解多粒度文本的多维跨层级注意力机制研究”（4184084）、北京市自然科学基金面上项目“基于异质数据融合的电信欺诈检测技术研究”（4172014）、教育部人文社会科学研究青年基金项目“基于深度学习的视频直播弹幕违规内容识别研究”（17YJCZH007）、北京工商大学青年教师科研启动基金项目“地质大数据中基于深度学习的实体抽取和表示”（QNJJ2017－17）、国家重点研发计划项目“主要食品全产业链品质质量控制关键技术开发研究”（2016YFD0401205）等资助下的综合成果，也包含了笔者近年来的主要研究成果。在此，特别感谢给予本书修改意见的阅评专家，尤其是北京交通大学的卢苇教授、北京工商大学的姜同强教授、左敏教授，他们认真阅读了本书全稿，提出了许多有价值的宝贵意见。此外，本书的出版得到了北京工商大学和北京理工大学出版社的大力支持，在此一并致以真诚的感谢。

本书在编写过程中借鉴、引用了众多学者的相关研究成果，在此表示诚挚的敬意和感谢，若有遗漏未标注之内容，敬请谅解。由于作者水平有限，缺憾与不足之处在所难免，欢迎读者批评指正。

目　录

第 1 章　引论 …… 001

1.1　语义计算 …… 001

1.1.1　语义网络 …… 003

1.1.2　本体 …… 004

1.1.3　语义网络与本体的联系 …… 005

1.2　语义相似度 …… 006

1.3　相关理论的演变与现状 …… 010

1.3.1　基于知识库的概念语义相似度计算 …… 010

1.3.2　基于语料库的单词语义相似度计算 …… 011

1.4　本书主要内容与创新 …… 014

1.4.1　基于 IC 模型的异构数据整合 …… 015

1.4.2　基于语义特征的异构数据整合 …… 016

1.4.3　基于度量方法的异构数据整合 …… 016

第 2 章　语义相似度计算的相关理论 …… 018

2.1　语义资源 …… 018

2.1.1　知识库 …… 019

2.1.2　语料库 …… 026

2.2　基于图结构的概念语义相似度 …… 027

2.2.1　概念语义相似度计算方法的分类 …… 028

2.2.2　不同概念语义相似度计算方法的特征比较 …… 031

2.3　文本的表示学习 …… 033

2.3.1　独热表示 …… 035

2.3.2　基于共现矩阵的向量空间模型 …… 035

2.3.3　基于神经网络训练的稠密向量空间模型 …… 038

2.4　基于向量空间的单词语义相似度 …… 044

2.4.1　欧氏向量空间 …… 044

2.4.2 概率向量空间 …… 047
2.5 本章小结 …… 048
第3章 基于IC加权最短路径的概念语义相似度计算 …… 049
3.1 计算方法的分类 …… 050
3.1.1 基于路径距离 …… 050
3.1.2 基于信息含量 …… 052
3.1.3 基于特征属性 …… 054
3.1.4 混合式 …… 056
3.2 概念的语义继承关系和结构属性 …… 056
3.2.1 语义继承关系 …… 057
3.2.2 知识库的图结构 …… 059
3.3 结合IC与路径距离的混合式计算方法 …… 061
3.3.1 计算元素的定义 …… 061
3.3.2 固有IC混合模型 …… 064
3.3.3 语义相似度计算策略 …… 067
3.4 应用案例 …… 069
3.4.1 基准数据集与评价标准 …… 069
3.4.2 IC计算模型的设计选择 …… 071
3.4.3 相似度计算策略中权重参数的有监督学习 …… 074
3.4.4 相似度计算策略中权重参数的无监督选择 …… 075
3.5 本章小结 …… 077
第4章 基于多语义融合的单词语义相似度计算 …… 079
4.1 概念向量的构建 …… 079
4.1.1 利用知识库进行语料语义标记 …… 080
4.1.2 利用聚类算法对相似词汇进行归类 …… 080
4.2 向量的特征融合 …… 083
4.2.1 相关背景知识 …… 083

4.2.2 向量组合模型 …… 084
4.3 多语义属性的融合模型 …… 086
4.3.1 多语义融合（MSF）模型 …… 087
4.3.2 基于语料的词向量训练 …… 091
4.3.3 概念实体的定义与关系抽取 …… 094
4.3.4 语义融合策略定义 …… 095
4.3.5 构造语义增强词向量的多种策略 …… 097
4.4 应用案例 …… 101
4.4.1 词对相似度评测 …… 103
4.4.2 语义 Web 服务匹配 …… 112
4.5 本章小结 …… 117
第 5 章 基于差分进化算法的单词语义相似度计算 …… 119
5.1 差分进化算法 …… 120
5.1.1 养分进化算法的原理 …… 120
5.1.2 差分进化算法的流程 …… 121
5.1.3 差分进化算法的组成 …… 123
5.1.4 六种变异策略 …… 124
5.1.5 常用的交叉策略 …… 125
5.1.6 相关研究及应用 …… 127
5.2 基于特征的有监督学习模型 …… 129
5.2.1 回归学习 …… 130
5.2.2 排序学习 …… 132
5.3 基于差分进化算法的语义计算 …… 134
5.3.1 方法组合策略 …… 135
5.3.2 问题定义 …… 136
5.3.3 向量相似度计算函数 …… 136
5.3.4 算法描述 …… 137
5.4 应用案例 …… 140
5.4.1 基于无监督差分进化算法的

相似度计算 …… 141
5.4.2 基于有监督机器学习算法的
相似度计算 …… 145
5.4.3 原始词向量和语义增强词向量的
空间探索 …… 147
5.5 本章小结 …… 151
第 6 章 知识整合的前世今生 …… 152
6.1 知识图谱与深度学习的研究与应用 …… 153
6.1.1 知识图谱 …… 154
6.1.2 深度学习 …… 160
6.1.3 典型应用 …… 161
6.2 知识作为神经网络的输入 …… 167
6.2.1 知识图谱的表示学习与推理 …… 167
6.2.2 知识图谱向量化表示的应用 …… 170
6.3 知识作为神经网络的约束 …… 172
6.4 本章小结 …… 173
附录 术语 …… 174
参考文献 …… 180

第1章 引 论

近年来，随着网络用户的规模呈爆炸式增长以及移动互联、社交网络、云计算等技术的日益成熟，互联网进入了“大数据”时代，大量无结构化的信息（包括文档、多媒体内容）不断出现，其规模呈指数级增长。对于计算机而言，这些无结构化的数据难以理解和使用。语义计算通过理解、提炼数据中的语义可以解决这一问题，帮助计算机执行复杂的数据计算与分析操作，实现其过程的自动化和实时性。

1.1 语 义 计 算

人们在进行有意义的交流时，往往需要借助被讨论对象的语义指代。语义（Semantics）指数据信息的含义，表示数据在某个领域的解释和逻辑。在现实世界中，某一事物或事实所代表的概念以及与其他概念之间的关系，均被认为是语义。例如，“胡萝卜是一种蔬菜，富含多种维生素”，就解释了“胡萝卜”这个概念（符号）的含义。

语义的典型特征包括：语义的客观性和主观性；语义的概括性和具体性；语义的清晰性和模糊性；语义的领域性；等等。其中，语义的客观性和主观性是指词义本身能够反映出客观的存在，在此基础上，语义还是人们对于客观存在的认识，能在一定程度上反映客观事物的特征。语义的清晰性和模糊性主要体现在语义边界上，例如，形容词“美丽”

具有的语义是一个模糊概念，边界并不清晰，无法用简单的判断逻辑来定义。语义的领域性是指一些词语含义的理解和表述需要结合某一个具体的领域，这是由于同一事物在不同领域的意思不同。例如，“苹果”在水果领域和手机领域具有不同的含义。

作为自然语言最基本的语义单元，词由外在表现和内在含义（概念）两部分构成。由于存在多义词和同义词，词汇与概念便存在多对多的关系，即：同一概念可以用不同的单词来表达；一个单词可能对应多个概念，具体表示何种意思，依赖单词所处的语境（上下文）。例如，在下列两个英文句子中，单词“bank”分别对应着“银行”和“河岸”的含义。

句1：She went to bank to deposit cash。

句2：She went to bank to catch fishes。

因为语义具有主观性，所以自然语言天然具有多义属性，这给理解自然语言、分析语义过程中的语义计算（Semantic Computing）带来了极大的挑战。

语义计算指通过构建计算对象的语义表示来分析、理解数据或者语言单元的含义，并从数据中凝练、提取出语义知识。语义计算涉及语言学、认知科学、心理学、自然语言处理、信息检索等相关学科领域与技术。语义计算的基础包括语义知识的表示与组织、相似度计算模型。其中，语义知识的表示与组织方式主要分为三大类：语义网络（Semantic Network）、特征模型和语义空间[1]。

语义网络将人类已有的知识组织为一个有向图，图结构中的节点表示语义概念，节点之间的连接边表示概念之间的语义关系。基于语义网络，一个单词的词义可以由其对应的概念节点以及相关节点的连接边表示。

特征模型基于词义可以表象为语义特征的集合这一思想，将词义表示为特征的概念分布。例如，利用“红色”“圆形”“美味”“有光泽”等一系列的感官特征表示“苹果”。与语义网络相同，大多数特征模型的构建依赖于构造者能够基于先验知识选择出最能够表达词义的特征。

语义空间则建立在一个假设上：假设词汇的语义可以由其所在的语

境决定，具有相同语境的两个词汇在很大程度上具有相似的词义。向量空间模型（VSM）是应用得最为广泛的一个语义空间模型。

1.1.1 语义网络

语义网络是美国心理学家奎廉（M. R. Quilian）于1968年提出的心理学模型，用于构建概念之间的关联，以便帮助人们记忆。美国人工智能专家西蒙（R. F. Simmons）和斯乐康（J. Slocum）于1972年将语义网络用于自然语言理解任务。认知心理学家和计算语言学家于1985年开始以“语义网络”的形式来描述词汇的含义，为人类提供从“关系”的角度来分析问题的能力。

语义网络在本质上是利用概念及其语义关系来表达知识的一种有向网络图，是一种基于图的数据结构，由节点（Vertex）和边（Edge）组成。在图结构里，每个节点表示现实世界中存在的“实体”或“语义概念”，每条边对应实体之间或概念之间的“关系”。语义网络能够把多元信息连接在一起，表示复杂的概念、事物以及语义联系，如类属关系（is－a）、从属关系（part－of）、位置关系（located－on）等。如图1－1所示，在动物的概念以及不同动物的相互关系构成的语义网络结构中，“熊猫”和“熊科动物”各自代表一个节点，这两个节点之间由一个表示“是”的箭头来表示命题“熊猫是一种熊科动物”，即两者之间存在继承类型的语义关系；“熊猫”节点又与“毛”节点存在着包含关系，由表示“有”的箭头来表示。

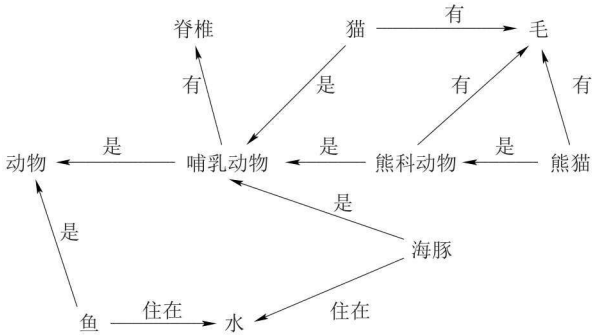

图1－1 语义网络的图结构

构建语义网络的目的在于以明确的规范为约束来获取、表示和组织知识，构建出复杂的知识结构和体系，并以此为依据进行知识推理，从而在求解任务问题时不必遍历整个知识结构。语义网络有强大的表达能力和灵活性，能通过多种机制来表达概念、规则及其之间的关联知识，因而在各个领域都得到了广泛的应用。随着语义网络研究的增多，越来越多的人意识到：除了利用语义成分（义素分析法）以外，还可以利用语义关系来表示词汇。因此，由专家建立的一大批领域知识库应运而生。

1.1.2 本体

本体（Ontology）来源于哲学术语，用于描述事务的本质，是客观存在的一个系统的解释或说明。21 世纪初，“知识本体”的研究开始成为计算机科学的一个重要领域。1993 年，格鲁伯（Gruber）将本体定义为“概念体系的明确规范说明”。本体能够将领域中的各种概念及相互关系明确地、形式化地表达出来，因而在语义研究方面发挥着重要的作用。

本体可以有效地进行知识表达、知识查询，或对不同领域的知识进行语义消解，是语义网（Semantic Web）的基础。本体可以支持基于语义的服务发现、匹配和组合，提高自动化程度。语义网是以资源描述框架（RDF）为基础、以 XML 为语法、以 URI 为命名机制，将各种不同的应用集成在一起，对互联网中的数据所进行的一种抽象表示。语义网所指的“语义”是“机器可处理的”语义，而不是人在自然语言中表达的语义等目前计算机所不能够处理的信息。

知识本体是领域概念及概念之间关系的规范化描述，这种描述是规范的、明确的、形式化的、可共享的。“明确”意味着对所采用概念的类型和它们应用的约束实行明确的定义。“形式化”指知识本体是计算机可读的，能被计算机处理。“共享”反映知识本体应捕捉该领域中一致公认的知识，反映的是相关领域中公认的概念集，即知识本体针对的是团体而非个体的共识。

知识本体的目标是捕获相关领域的知识，提供对该领域知识的共同

理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。对本体的知识进行管理可以实现语义级知识服务，提高知识利用的深度，还可以支持对隐性知识进行推理，方便异构知识服务之间实现互操作，方便融入领域专家知识及经验知识结构化等。对本体的知识进行管理一般要求满足以下基本功能：

（1）支持本体多种表示语言和存储形式，具有本体导航功能。

（2）支持本体的基本操作（如本体学习、本体映射、本体合并等）。

（3）提供本体版本管理功能，支持本体的可扩展性和一致性。

1.1.3 语义网络与本体的联系

作为知识表示的工具，语义网络与本体非常相似，是知识表示的不同形式，均可以通过带标记的有向图进行知识表示，适合于逻辑推理。但是，两者之间也存在一定的区别，主要是描述的对象或范围不同，有以下几点：

（1）语义网络可以被看作一种带有标记的有向图，最初用于表示命题信息，现在被广泛应用于专家系统中表示知识。其中，节点用于表示物理实体、概念或状态，边用于表示关系。由于对节点和边都没有特殊规定，所以语义网络能够描述的范围比本体更广。而本体是面向特定领域的概念模型，是对共享概念模型的规范说明，其概念在某个特定领域是公认的。

（2）在知识表示的深度上，语义网络对建模没有特殊要求。然而，构建本体需要五要素，分别为概念、关系、函数、公理和实例，本体依赖这五要素来严格、正确地刻画所描述的对象。

（3）在建模的条件上，语义网络不要求有相关领域的专业知识，因而不必有专家参与，相对容易构建；而本体的建立必须有专家参与，相对更加严格和困难。

语义计算通过理解、解释自然语言中各个组成单元的具体含义，构建语言单元的语义表示，它在自然语言处理、信息检索等领域发挥着十

分重要的作用，与机器学习、数据挖掘技术紧密相关，被广泛应用于语义消歧、文本分类、舆情分析等任务中。由于自然语言存在丰富的特性（如指代、转折、多义等），现在已经成为计算机理解并解释人类写作与说话方式的障碍。尤其在文本处理过程中，计算机面临大量语义消歧的问题，导致语义计算需要大量知识（包括与句法、词法、语义有关的语言学知识和与语言规则无关的世界知识）。

1.2 语义相似度

当前，作为信息传播的主要载体，文本已成为网络大数据的重要组成部分，如网页、报表、电子邮件、XML 文档等。人们希望借助语义计算技术从海量、多源的文本数据中获取有价值的信息，以此应对重复数据、垃圾数据和歧义数据给文本理解与分析带来的挑战。其中，语义相似度计算作为关键技术之一，在提高计算机的文本理解能力方面起着重要作用。

语义相似度指两个对象在语义内容（含义）上相似的程度，是一个从定量的角度表示对象之间相似性的指标。心理学认为，相似性是人们受到两个对象之间关系的刺激所产生的心理感知以及对对象进行定性比较的心理过程，是人类思想和语言中最基本的元素[2]。例如，人们面对一对父子的外貌信息会产生“很相似”的心理反应。为了量化对象之间的相似性，研究者们提出了相似度的概念。计算机科学侧重于利用人工智能模拟人类对于相似性的判断行为，以关于相似性的假设为基础，从特定的知识表述中计算出对象在语义层面上的相似度。

依据不同粒度的表现形式，文本的基本组成单元包括词、句、段落（篇章），它们的语义内涵具有抽象层次递进的关系。词、句、段的表示学习方法存在较大的差异，往往针对不同类型的任务，考虑文本理解的粒度。词是文本的最小组成单元，因此词的语义表示通常是文本语义计算的基础，不仅可以用于句子、段落等长文本的表示学习，也被用于具体的任务中（如智能问答等）。

正因为如此，如何衡量词汇之间的语义相似度对于自然语言处理具体任务的性能提升起着关键性作用。然而，在传统的信息检索技术中，基于关键词的文本匹配方法没有考虑检索词的语义，只停留在词汇的表层，导致检索结果包含大量无关信息，无法理解和满足用户的真实需求。因此，进行词汇的语义计算、从语义上理解词汇的内涵以及量化词汇之间的语义关系，已成为提升文本挖掘、机器翻译等应用的人工智能水平的关键技术之一。基于词形匹配的相似度计算方法难以深入挖掘词汇的语义，尤其是词汇语义的异构性和歧义性使其在文本分类、文本主题抽取等任务中的适用性较低。此外，语义标注的文本在实际中通常难以获得。因此，基于语义的词汇相似度计算对于提升文本处理任务的性能显得尤为重要。

在已有的语义相似性度量的研究中，语义相似性被认为是语义相关性（Semantic Relatedness）的一种特例。一些研究指出，“语义相似”不等同于“语义相关”。“语义相关度”衡量的是语义上的关联程度，比“语义相似度”的概念更广、更通用[3,4]。

下面以如图 1 – 2 所示的台式电脑、平板电脑和鼠标的关系为例，说明“语义相似”与“语义相关”的区别。

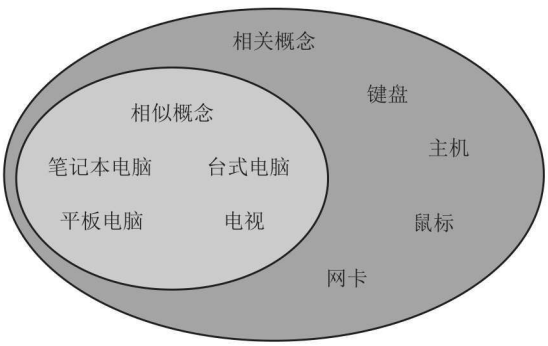

图 1 – 2　举例说明“语义相似”与“语义相关”的区别

“台式电脑”与“平板电脑”具有许多共同的特性，可以上网、播放视频等，因此两者具有语义相似性；而“台式电脑”与“鼠标”虽然没有相同的特性，并不相似，但“台式电脑”依赖于“鼠标”输入

数据信息，两者存在一定的语义相关性。

再例如，词“银行”和“利息”虽然具有不同的含义，但能频繁地同时出现在经济类文章中，因此具有语义相关性。这一点符合人类的直观感受。

针对大数据的检索和处理不仅依赖于云计算等技术提高运行效率[5]，对数据挖掘算法和语义计算等相关技术提出了更高的要求。其中，文本之间的语义相似度常被用于对千万量级的数据进行合并和去重，减少数据冗余。语义相似度计算在文本聚类[6]、服务发现[7]、问答系统、机器翻译[8]、推荐系统[9]、舆情分析[10]等领域都具有广泛的应用。在商用搜索引擎中，谷歌、百度、YouTube 等公司均已实现了基于语义的信息检索，从语义层面理解和处理检索请求，借助挖掘语义关联、消除词汇歧义[11]，以达到增强信息检索的智能性和灵活性的目的。

在语义搜索引擎中，词汇之间的同现关系、语义关系被用于扩展查询词。例如，当用户搜索“苹果”时，语义搜索引擎不仅能够给用户返回“苹果—水果”，还可能将“苹果公司”“苹果手机”“小米手机”等作为查询结果或推荐结果反馈给用户；当用户搜索“捷豹”时，语义搜索引擎能够结合用户的搜索历史，为用户展示其感兴趣的某汽车品牌下的所有车型图片，而不是一张大型猫科动物的图片。

此外，词汇语义相似性计算的准确性也已成为提高智能问答系统性能的关键因素之一[12,13]。典型的问答系统（如微软的小冰、苹果的 Siri、谷歌的 Now 语音助手、百度语音助手等）均采用了语义计算和推理技术。与此同时，通信 4G 时代和智能移动终端推动了移动互联的巨大发展，文本数据出现了一个明显的特征变化，即短小、频繁。中国互联网络信息中心（CNNIC）在 2018 年 1 月发布的第 41 次《中国互联网络发展状况统计报告》显示，台式电脑、笔记本电脑的使用率均出现下降，使用智能手机作为互联网接入终端的比率持续增长，人们利用碎片化的时间，通过微博、微信、短评等即时通信媒介传播信息。在这种情况下，文本语义相似度计算的重要性不言而喻，能够在很大程度上影响