《智能信息处理》课程作业

**基于形式概念分析的词汇相似度计算**

杜伟

|  |  |
| --- | --- |
| 作业 | 分数[20] |
| 得分 |  |

2020年11月12日

第 43 卷 第 12 期 计 算 机 科 学 技 术 Vol.43 No.12

2020 年 11 月 Computer Science December 2020

基于形式概念分析的词汇相似度计算

杜伟

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

**摘要** 形式概念分析是由德国的 Wille 教授在 20 世纪 80 年代初所提出的，其核心数据结构概念格，也 被称为 Galois 格，准确而简洁地描述了概念之间的层次关系，已成为一种极其重要的知识表示方法。作为一种优良的数学工具，概念格己经被广泛的应用于知识表示、数据挖掘、信息检索等众多领域。本文介绍了形式概念分析，讨论了概念格的基本原理，本文基于形式概念分析理论，提出一种基于主题相似度定量刻画词汇相似度的计算方法，有效利用了词汇对应的主题语义关系，能更好地反映词语之间的关联性。

**关键词** 形式概念分析，概念格，词汇相似度

**中图法分类号** TP311 **文献标识码** A

Lexical similarity calculation based on Formal Concept Analysis

DuWei

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China) **Abstract** .Formal concept analysis was proposed by Professor wille of Germany in the early 1980s. Its core data structure concept lattice, also known as Galois lattice, accurately and concisely describes the hierarchical relationship between concepts, and has become an extremely important knowledge representation method. As an excellent mathematical tool, concept lattice has been widely used in knowledge representation, data mining, information retrieval and many other fields. This paper discusses the basic principle of concept lattice, introduces the related construction algorithms of concept lattice, and discusses various algorithms of building lattice, then analyzes the application in data mining, and finally puts forward the related research direction of concept lattice in the future.It can better reflect the relevance between words.

**Keywords** Formal Concept Analysis,Concept Lattice ,Lexical Similarity

1 引言

形式概念分析（Formal Concept Analysis，FCA）是Wille提出的从形式上下文中进行数据分析和规则提取的强大工具[1]。形式概念分析是以数学为基础的，它在形式上代表了主题的概念、属性和关系，是以概念为基础的。作为一个单元构造。它清楚地展示了本体的结构，这样的对象构造过程是半自动化的。在设计阶段，需要相关方面的专家的参与，识别领域中的对象属性，建立它们之间的关系。在生成该概念之后，它构造语句，使用概念网格的生成算法CLCA，并自动生成本体。形式概念分析，以人类认知为中心，并提供传统的统计数据分析和知识代表了完全不同的方法。它已成为人工智能的重要研究对象，广泛应用于机器学习、数据挖掘、信息检索等领域。

而概念格则是 FCA 的核心数据结构。概念格理论最早由 Wille R 等提出，是应用数学的分支，它来源于哲学相关领域内对概念的理解。随着研究的深入，很多学者逐渐认识到概念格自身结构的巨大优势，研究从开始的单纯理论扩展发

收稿日期：2020-11-12

作者简介：杜伟（1997-）男，硕士生在读。

展到理论与实际应用相结合，并且融合交叉多个相关 理论，成为许多专家学者关注的热点。作为数据 分析和知识处理的形式化研究方法，概念格在知 识发现、信息检索等方面均得到了广泛的应用[2]。 概念格理论的研究不仅能用于解决知识发现领域中所涉及的关联规则、蕴含规则[3]、分类规则的提 取，还能够实现对信息的有机组织[4]，减少冗余度， 简化信息表，所以对于概念格理论及其构造方法 的研究具有十分重要的意义[5]。

本文首先介绍了形式概念分析的基本概念和其形式背景，又介绍了概念格的基本概念以及构 建形式背景和概念格，进行简单的概念分析，并 从分析中获得关联规则，再通过对构造概念格的算法进行论述，使人们更加了解构造概念格的算法，最后介绍了概念格在词汇相似度方面的应用。

2 基于形式概念分析的主题定义及表示模型

2.1 形式背景

形式概念是现实世界中各种概念的抽象,通过概念外延与内涵之间的关系形式化地刻画抽象概念。在形式概念分析中,数据是用形式背景表示的。形式概念分析是Wille提出的一种从形式背景进行数据分析和规则提取的强有力工具，形式概念分析建立在数学基础之上，对组成本体的概 念、属性以及关系等用形式化的语境表述出来，然后根据语境，构造出概念格(concept lat-tice)，即本体，从而清楚地表达出本体的结构。这种本体构建的过程是半自动化的，在概念的形成阶段，需要领域专家的参与，识别出领域内的对象、属性，构建其间的关系，在概念生成之后，可以构造语境，然后利用概念格的生成算法，自动产生本体。形式概念分析强调以人的认知为中心，提供了一种与传统的、统计的数据分析和知识表示完全不同的方法，成为了人工智能学科的重要研究对象，在机器学习、数据挖掘、信息检索等领域得到了广泛的应用。

2.2 概念格

概念格理论，也称形式概念分析理论，首先由德国的数学家于 1982 年提出。形式概念分析理论是一种基于概念和概念层次的数学化的表达的应用数学的一个分支。因此，在形式概念分析理论时，需要用数学的思维方式进行概念数据分析和知识的处理。形式概念分析中的“形式”一词表示我们正在处理数学领域的工作，通过与这些工作相联系的结构化的概念的联系，发现可理解的、有意义的知识。概念格作为形式概念分析中核心的数据结析的两个基本柱石。概念格理论的主要思想是在形式背景中寻找所有的概念并构造出格结构以此刻画出数据集中对象与属性之间的关系。构造概念格是概念格应用的前提，但构造概念格已被证明是NP 问题；因此，人们在构造概念格之前希望在保持格结构不变的情况下，尽可能的简化数据。目前，概念格约简的研究包括对象约简、属性约简、纵横向维护和内涵约简等。本文引入形式概念分析的原理 和方法，将文献视为对象，关键词视为属性，将概念 视为由文献集合（外延）和关键词集合（内涵）所组成 的知识单元。它们之间的关联关系构成形式背景，而对形式背景进行分析和挖掘可以揭示文献集合中 隐含的主题及其层次关系，并通过概念格展示出来。 下面给出文献集合隐含主题的相关定义：

定义1：文献隐含主题的形式背景 M 是一个三元组(D，K，I)，其中D是文献（对象）的集合，K是关键词（属性）的集合 ，I是D和K的二元关系，即I⊆D×K。dIk或者(d，k)∈I则表示文献d拥有词汇k。表1是一个由5篇文献和5个关键词所构成的形式背景。其中“×”表示文献di包含关键词kj，空格表示文献di未包含关键词kj。

定义2：文献集合的主题及形式概念：设A是文档集合D的一个子集，定义f（A）={ k ∈ K|∀d∈ A，dIk}表示A中文档共享的关键词集合，即主题；同理设T是词汇集合K的一个子集，代表一组词 汇构成的主题，定义g（T）={d∈D∀k∈T，dIk}，表示共享主题T的文档集合。若f（A）=T且g（T）=A，则称二元组(A，T)为形式背景M的一个形式概。假设A1={d1，d2，d4，d5}，T1={ k2，k5}，满足f(A1)=T1且g(T1)=A1，则(A1，T1)是M上的一个形式概念。

定义3：概念格：设(A1，T1)和(A2，T2)是形式背景M上的两个形式概念。如果A1⊆A2（或T2⊆T1），那么(A1，T1)被称为(A2，T2)的子概念，(A2，T2)被称为(A1，T1)的超概念，并且记为(A1，T1)≤(A2，T2)。关系“≤”为形式概念之间的序。形式背景M上所有形式概念用这种序组成的集合被表示为 β ( D，K，I)，并且被称为形式背景 M 的概念格。

图1该形式背景中隐含的12个概念及其层次关系。由于每个概念的内涵对应一个主题，因此概念格也 同时展示了能反映主题泛化和特化关系的层次结构。可以看到每个下层主题都继承了上层主题中的所有词汇属性。越往概念格的下层走，主题越具体，对应的外延（即文献）也越少。例如：最上层的概念C1的内涵（主题）只有一个词汇 k5，所有 5 篇文档都 共享这个主题；第二层概念C4的内涵（主题）包含两个词汇{k1，k5}，共享这个主题的文献有4篇文档，分别是d1、d3、d4d5；第三层概念 C6 的内涵（主题）包含三个词汇T1= {k1，k2，k5}，共享这个主题的文档有3篇，分别是d1、d4和d5；第四层概念C9的内涵（主题）包含4个词汇{ k1，k2，k3，k5}，共享这个主题的文档有2篇，分别是d4和 d5；第五层概念C12的内涵（主题）包含5个词汇{ k1，k2，k3，k4，k5}，共享这个主题的文献有1篇，即d5

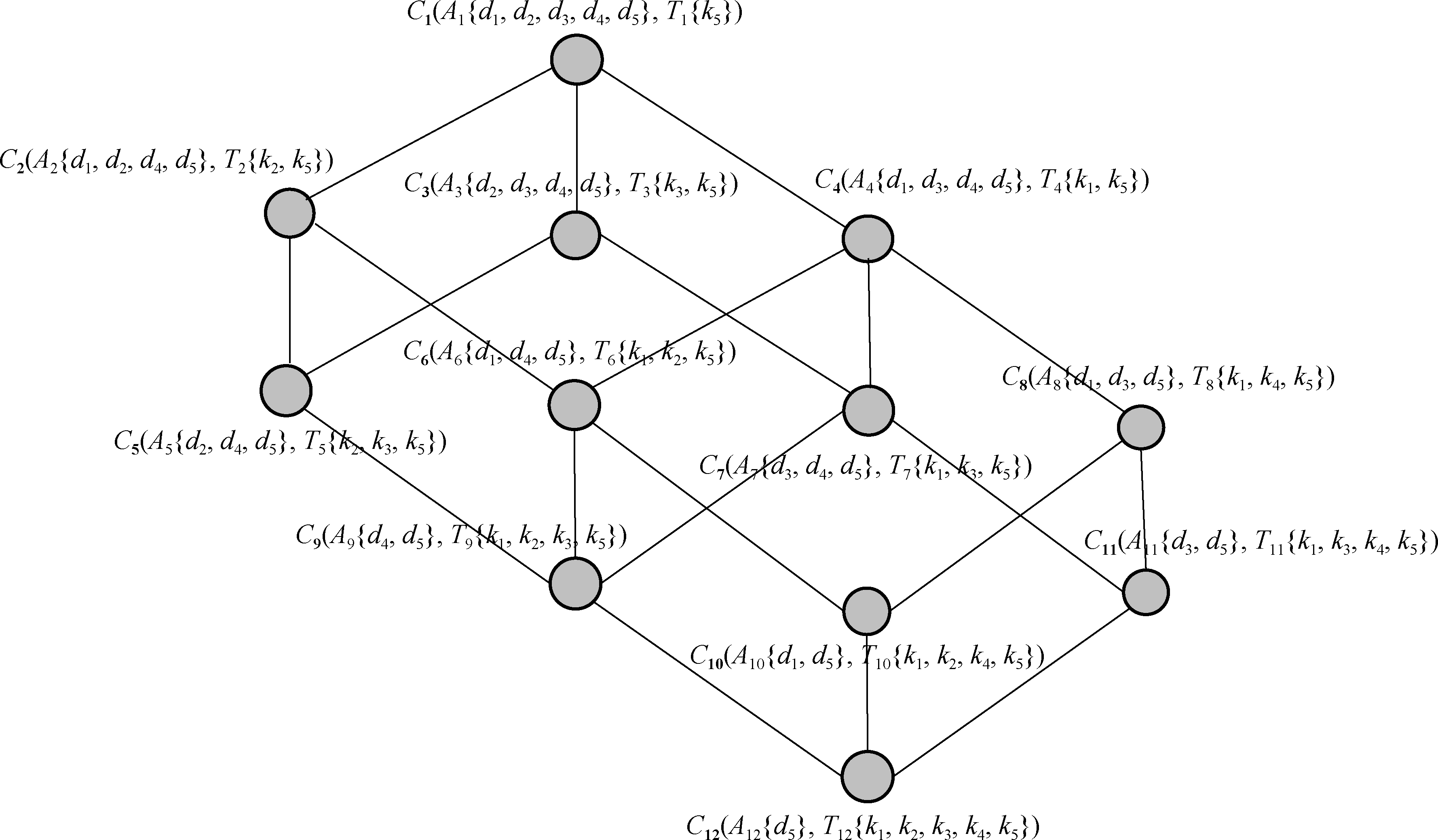


图 1 基于形式背景生成的概念格

3 基于FCA的词汇相似度计算

3.1基于主题的词汇表征

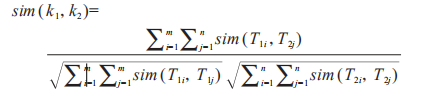
上文描述了通过形式概念分析的方法在文档和词汇之间引入一个潜在的主题层，将文档和词汇都 映射到具有层次结构的主题空间。揭示主题的过程是探究文档主题凝聚词语的过程。不同的文档共享 不同粒度的主题，当主题粒度宽泛时，对应的词汇少、文献多；当主题粒度具体时，对应的词汇多、文献少。词汇间的语义关系可以从主题的角度进行度 量。当两个词汇共同出现在一个主题中，表示这两 个词汇具有一定的语义关系。当两个词汇共同出现在 多个主题中，表示这两个词汇具有较强的语义关系。设主题 T 与词汇集 K 的关联矩阵为 L = { lij}，当 Ti 包含关键词 kj 时，lij = 1，否则 lij = 0，对 L 进行转置 可以得到词汇 K 与主题 T 的关联矩阵 LT。基于图 1 概念格得到 T - K 关联矩阵 L，如表 1 所示。

表 1 基于概念格的 *T* - *K* 的关联矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *k*1 | *k*2 | *k*3 | *k*4 | *k*5 |
| *T* | 0 | 0 | 0 | 0 | 1 |
| *T* | 0 | 1 | 0 | 0 | 1 |
| *T*3 | 0 | 0 | 1 | 0 | 1 |
| *T*4 | 1 | 0 | 0 | 0 | 1 |
| *T*5 | 0 | 1 | 1 | 0 | 1 |
| *T*6 | 1 | 1 | 0 | 0 | 1 |
| *T*7 | 1 | 0 | 1 | 0 | 1 |
| *T*8 | 1 | 0 | 0 | 1 | 1 |
| *T* | 1 | 1 | 1 | 0 | 1 |
| *T* | 1 | 1 | 0 | 1 | 1 |
| *T*11 | 1 | 0 | 1 | 1 | 1 |
| *T*12 | 1 | 1 | 1 | 1 | 1 |

3.2基于概念格结构的词汇相似度计算

给定两个词汇k1和k2，k1={T11，T12，...，T1m}，k2= {T21，T22，...，T2n}分别表示 k1 对应 m 个主题 ，k2 对应 n 个主题。那么词汇 k1 和 k2 之间的相似度不仅 取决于这两个词汇对应的相同维度的主题数量，还取决于两个主题集合中不同的主题对之间的语义关 联度。若两个集合中包含的相同主题越多，或两个 集合中不同的主题对关联度越高，则两个词汇的相似度越大。如公式所示：



其中，T1i为k1对应的第i个主题，T2j为k2对应的第 j 个主题，sim (T1i，T2j) 表示主题对 (T1i，T2j) 的相似度。

4 实验

4.1实验数据的获取与预处理

本文以信息检索领域为例，选取 2006 年-2016 年国际会议 SIGIR 收录的论文为实验数据，共计 693 篇。下载全文后对文章进行预处理，一般英语的分 词是以空格拆分单词的，但考虑到有些词组不应该 被拆分成独立的词，所以首先从每篇文章作者给出 的关键词中抽取词组，构建基础词组库，然后以此为 基础对全文进行分词。去除停用词后，再利用 TF- IDF 方法从每篇文章的全文中抽取权重排名靠前的 词汇以扩充文章关键词，将每篇文章的关键词拓展 为 8 个（文章的关键词一般不超过 8 个）。

4.2论文主题形式背景及概念格的构建

基于形式概念分析理论，以 693 篇文档作为对 象集合 D = { d1，d2，…，d693}，以 180 个高频关键词作为属性集合 K = { k1，k2，…，k180}，通过文章关键词关 联关系矩阵构建形式背景，如表 5 所示，并在此基础 上构建概念格。概念格的构造实际上是形式背景中 隐含概念及其层次关系识别的过程。本文采用批处 理算法将形式背景中的对象及其属性转换成概念格中具有偏序层次的概念节点，共挖掘出 1308 个概念节点。由于每个概念的内涵对应一个主题，因此也 得到 1308 个主题及其层次结构。

4.3对比分析结果

通过神经网络模型生成词向量是近年词汇相似度计算统计方法中研究较多的方法，出现了许多词 向量模型，也表现出较好的性能，如 FastText。因 此，为验证本文方法的有效性，选择 FacebookAI 研 究院（FAIR）基于维基百科和普通爬网数据集训练 的 FastText 词向量模型进行词汇相似度计算的对 比实验。进行对比的词汇类型有两种：单词和词组。 以往词汇相似度实证研究中通常只测试单个词汇的 相似度，忽略了词组之间的相似度。本文认为词组 之间的相似度计算在实际应用中也尤为重要，因此 增加了词组相似度的计算，该 FastText 模型考虑了 词 汇 的 形 态 构 成 而 且 自 带 未 登 录 词（Out-Of- Vocabulary，OOV）功能，可以解决原本词典中不存 在的二元或多元词组问题。表8分别列出利用本文方法和FastText 方法得出的词汇相似度计算结果（词汇对后括号内的数字代表人工判断的结果，“1” 表示相似，“0”表示不相似）。在每一类型的词汇对中，分别挑选三组词代表 相似词汇和不相似词汇进行比较。从表 8 中可以发 现通过本文方法与 FastText 方法计算的词汇相似度 结果有些相差比较大，已在表中加粗标记。

5 结束语

词汇相似度计算是自然语言处理的基础研究问题，且具有广泛的应用领域。现有的词汇相似度计算无论是知识库的方法还是基于统计的方法，都有着自身难以逾越的瓶颈。本文从一个新的视角，即从主题层面研究关键词的关系。在文档和词汇之间加入主题层，通过形式概念分析挖掘隐含主题及层次关系，将词汇项映射到主题层级，提出一种基于主题相似度定量刻画词汇相似度的计算方法。实验结果证明基于 FCA 的词汇相似度计算性能在精确率与召回率上都明显优于 FastText 词向量方法，展现了该方法的有效性和潜在优势。但如果要更严格地区分词汇的相似关系和相关关系，仅使用本文的方法还是有局限的。下一步笔者将结合带有词汇语义相似关系的词典优化结果，并探索该方法在文本分类中的具体应用。

参考文献

[1] 张硕望,欧阳纯萍,阳小华,等.融合《知网》和搜索引擎的词汇语义相似度计算[J]. 计算机应用, 2017, 37(4):1056-1060.

[2] 朱新华,马润聪,孙柳,等.基于知网与词林的词语语义相似度计算 [J]. 中 文 信 息 学 报 , 2016, 30(4): 29-36.

[3] 池哲洁,张全.基于概念基元的词语相似度计算研究[J].电子与信息学报,2017,39(1):150-158.

[4] 彭丽针,吴扬扬 .基于维基百科社区挖掘的词语语义相似度计算 [J].计算机科学,2016,43(4):45-49.

[5] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information[J] .Transactions of the Association for Computational Linguistics, 2019, 5:135-146.