《智能信息处理》课程作业

基于概念格的语义相关度分析

王明

作业	分数 [20]
得分	

2020 年 11 月 13 日

基于概念格的语义相关度计算

王 明

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要 概念格，也称为 Cralois 格，又叫做形式概念分析，是基于属于某一概念的所有对象与其共有属性之间的二元关系而建立的一种概念层次结构，是一种知识分析与处理的有效工具。概念格结构模型来源于形式概念分析（FCA）理论，是 FCA 中的核心数据分析工具，它本质上描述了对象（样本）与属性（特征）之间的关联。概念格由德国教授 Wille R 于 1982 年首先提出，它提供了一种支持数据分析的有效工具，作为一种优良的数学工具，概念格已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。为了提高检索信息的相关度和检索效率，本文基于形式概念分析的理论，提出了一种利用概念来计算词汇之间语义相关度的计算方法，并将其应用到信息检索当中。实验结果表明，基于概念格的语义相关度计算方法是有效的，且该方法在信息检索系统的语义理解方面能起到很好的支持作用。

关键词 形式概念分析，概念格，语义相似度

中图法分类号 TP391 **文献标识码** A

Semantic relevancy computing based on Concept Lattice

Wang Ming

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract Concept lattice, also known as formal concept analysis, is a concept hierarchy based on the binary relationship between all objects belonging to a concept and their common attributes. It is an effective tool for knowledge analysis and processing. Concept lattice structure model comes from formal concept analysis (FCA) theory. It is the core data analysis tool in FCA. It essentially describes the relationship between objects (samples) and attributes (features). Concept lattice was first proposed by German professor wille r in 1982. It provides an effective tool to support data analysis. As an excellent mathematical tool, concept lattice has been widely used in many fields, such as knowledge representation, data mining, information retrieval and so on. In order to improve the relevance and retrieval efficiency of retrieval information, based on the theory of formal concept analysis, this paper proposes a method to calculate the semantic relevancy between words by using concepts, and applies it to information retrieval. The experimental results show that the semantic relevance calculation method based on concept lattice is effective, and this method can play a good supporting role in the semantic understanding of information retrieval system.

Keywords Formal Concept Analysis; Concept Lattice ; Semantic Relevancy

0 引言

概念格也称 Galois 格，是由 R.Wille 根据二元关系提出的一种概念层次结构，是数据分析和知识发现的一种有效工具[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。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[2]。

传统的信息检索系统大都是基于关键字匹配的检索技术，使得用户在检索时经常得到大量与查询无关的结果。为了提高用户对检索系统的满意度，可以通过扩展检索系统对语义查询和动态查询的支持来现[3]。将语义相关度计算引入信息检索技术的研究中，正是为了提高检索系统对用户查询信息的语义处理能力，从而提高检索效率，使系统更具智能性。

在信息检索中，语义相关度要反映的是文本或者用户查询在意义上的符合程度。目前语义相关度计算的研究都是建立在对语义相似度

收稿日期：2021-11-30

作者简介：王明，男，1999 年生，硕士生在读，主要研究领域为自然语言处理，Email:1173037855@qq.com

研究的基础上的。常见的相似度计算方法有 2 种[4]，一种是根据世界知识或者分类体系计算；一种是利用大规模的语料库进行统计。本文提出一种利用概念格的特性，来计算词汇之间语义相关度的方法并对其在语义，检索中的应用做研究。该方法属于基于语料库的计算方法。由于概念格具有知识聚类的特点，使得该方法在一定程度上降低了数据疏松和数据噪声的影响。

1 概念格里的理论基础

概念格的理论基础来源于形式概念分析 FCA（Fomal Concept Analysis）。FCA 的产生源于对“概念 ”这一哲学思想的理解：概念包括外延和内涵．概念的外延表示属于这个概念的所有对象的集合，而内涵则是这些对象具有的所有属性的集合．一个领域中的所有对象和所有属性之间的二元关系构成一个形式背景．概念格就是根据数学中的格理论由形式背景得到的一种具有概念层次结构的格，其中概念格中的 1 个结点表示领域中的 1 个概念．概念格的构造过程实质上就是一个知识或者概念的聚类过程。

定义 1 一个形式背景 K:=(G,M,I)由集合 G，M 以及它们之间的关系 I 组成，G 的元素称为对象，M 的元素称为属性描述一个对象 g 和一个属性 m 之间的二元关系 I，可以写成 gIm 或(g, m) ∈I 表示对象 g 有属性 m。

一个形式背景实际上表示的是对象集和属性集 之间的二元关系，所以可以用交叉表来表示。表中各 行用对象名标识，各列用属性名标识。g 行与 m 列的 交叉点表示对象 g 具有属性 m。

定义 2 形式背景(G，M，I)中的一个形式概念，是一个对(A, B)，其中 A⊆G，B⊆M，满足 A′＝B 且 B′＝A 的条件。A，B 分别称为形式概念 (A, B)的 外延和内涵。

对于 K 上 2 个形式概念(A_1, B_1)和(A_2, B_2)如果 $A_1 \subseteq A_2$，则称(A_1, B_1)是(A_2, B_2)的子概念，称(A_2, B_2)是(A_1, B_1)的超概念，记为 (A_1, B_1)≤(A_2, B_2)，一个概念格就是由形式背景中的形式概念按照超概念－子概念这样的编序关系得到的格结构。概念格能够清晰地表明概念间的泛化例化关系，也具有明显的概念层次结构可视化表示为与其相对应的 Hasse 图。

2 基于概念格的相关度

2.1 相关度的基本思想

词汇语义相关度反映的是词汇之间语义关联的程度，由区间[0，1]上的一个实数表示。无论哪种相关度的计算方法，都要先对词汇进行一些语义上的描述，然后根据这些描述计算词汇间的语义相关度[5]。例如，传统的大规模语料库进行统计的方法是事先选择一组特征词，然后计算这一组特征词与每一个词的相关性，于是对于每一个词都可以得到一个相 关性的特征词向量，然后利用这些向量之间的相关度（一般用向量的夹角余弦来计算）作为这 2 个词 的相关度。

观察概念格结构发现，自顶向下形式概念中包含的对象数依次减少，属性数依次增加，形式概念中的对象之间的相关性也依次变大．所以可以认为，形式概念越靠近 Hasse 图的底端，包含在其中的对象之间的相关度越大；而处在同一层次的形式概念（可按属性数或者对象数的多少分层，属性数或对象数相同的形式概念在图中的同一层），包含在其 中的对象之间的相关度是相同的。由此，引出形式概念相关度和对象之间的相关度的定义如下。

定义 3 概念格 L 中的形式概念 C 的相关度

$$Rel(C)=m/n$$

其中，n 为 Hasse 图的层数；m为形式概念 C 所在的层号。

定义 4 对于概念格 L，对象 g_1 和 g_2 之间的相关度

$$Rel(g_1, g_2):=Max(Rel(C))$$

其中，C⊆L；｛g_1,g_2｝⊆Ext(C)。

Rel(C)实际上体现的是形式概念 C 中对象之间的相关度,显然，Rel(C)是一个在 0 和 1 之间的实数。Ext(C)是表示 C 中的外延。计算概念格 L 中 2 个对象 g_1 和 g_2 的相关度，Rel(g_1, g_2)首先需要计算包含着 2 个对象的形式概念的相关度 Rel(C)，然后取数值最大的 Rel(C)作为这 2 个对象的相关度。

2.2 相关度计算方法的实现

结合概念格的结构特点以及相关度计算的基本思想，本文提出的相关度计算方法如下：根据领域知识生成形式背景，再由形式背景构造概念格，最后根据概念格的层次结构计算对象（关键词）之间的语义相关度。

2.2.1 形式背景的生成

对于某个确定的领域知识，可以相应地得

到一个形式背景，即用对象和属性这样的二元关系来描述领域知识。形式背景生成的关键是确定其对象集和属性集以及对象和属性间的二元关系[6]。实际确定三元组K中的3个集合的方法为：统计的方法得到对象集即对该领域知识的训练文档集作统计,从而选出能够表示该领域知识的关键词,把选出来的关键词作为对象集；选出最能够反映对象集特征的特征词构成属性集，借鉴本体构建中基于字典构建方法的思想，在一个关键词语义词典和领域专家的共同作用下，确定其属性集；以1和0来表示对象和属性的二元关系I，如果对象（关键词）包含属性（特征词）的语义，用1来表示I，否则用0表示。

例如，对于对象集 G＝｛电脑，计算机，内存，存储器｝和属性集 M＝｛输入，输出，运算，存储｝可以得到表1所示的形式背景。

表1 形式背景示例

I	输入	输出	运算	存储
电脑	1	1	1	1
计算机	1	1	1	1
内存	0	0	0	1
存储器	0	0	0	1

2.2.2 概念格的构造

目前概念格的构造算法有批处理算法和渐进式构造算法。本文采用的是一种对渐进式Godin 算法的改进算法[7]。由于该算法的思想比较成熟而且已经多次应用到概念格构造的应用中，在这里就不作更多的叙述。

2.2.3 语义相关度的计算

根据定义3和定义4的论述，可以计算一个概念格上对象之间的相关度，不过计算时需要考虑概念格如何分层。分层的问题实际上是Hasse 图的构图问题,但它对相关度计算的结果有重要影响。经多次试验发现：按照属性数相同分层可能导致一些相关度较高的概念由于其包含的属性数偏少而得到较低的值；按照对象数相同分层，可能导致一些相关度较高的概念由于包含的对象数偏多而得到较低的值。解决办法是：分别按属性数相同的方式和对象数相同的方式分层，计算出 Relatt(C)（属性数分层的结果）和 Relobj(C)（对象数分层的结果）把 2 个值的加权平均数作为形式概念C的相关度 Rel(C)按属性数分层的权值和按对象数分层的权值分别设为 α 和 β 其中 α+β=1。实际计算中可以适当调节二值的大小以得到更精确的结果，目前 α，β 都设定为 0.5。综上，计算对象之间的相关度算法描述如下。

算法 计算 Rel(g_1, g_2)

Procedure RelCaculate(L, g_1, g_2)

输入：概念格 L，中的对象 g_1，g_2

输出：Rel(g_1, g_2)

```
BEGIN
 FOR L 中的每个形式概念 C DO
  计算 Relatt(C)
  计算 Relobj(C)
  Rel(C)←α* Relatt(C)+ β* Relobj(C)
 END FOR
  FOR L 中的每个形式概念 C DO
    IF g1, g2同时包含在 C 的对象集中
     THEN
       IF Rel(g1, g2)<Rel(C) THEN
        Rel(g1, g2)←Rel( C)
       END IF
    END IF
   END FOR
END
```

3 在信息检索方面的应用

在用户并不十分清楚自己查询的情况时，系统如果能够给用户提供一些与查询相关的内容反馈FF0C那么就可以使用户查询变得更加明确，从而提高检索的效率[8]。可以通过给出用户输入关键词的相关词的方式来实现系统的语义理解。

在确定了领域知识的前提下，很容易就能找到与输入关键词的语义相关的词。具体方法是：对该领域设定一个阈值 λ。一个关键词与用户输入关键词的语义相关度大于 λ 时，返回该关键词。

考虑到一个查询关键词可能出现在多个领域（形式背景）的情况，必须进行跨领域的操作，即在不同的概念格中查找其语义相关词[9]。具体方法是：由于关键词在不同的领域中出现的频度不同，所以要对每个关键词确定其对每个领域的隶属度，然后通过隶属度加权的方法计算词汇之间的相关度，最后根据相关度大小确定关键词的语义相关词。

4 实验及分析

以计算机硬件为领域构建形式背景，其中对象集 G＝｛金手指，显卡，声卡，显示器，

内存，存储器 Cache，主存，外存，南桥，北桥，硬盘，CPU，主板，网卡，键盘，鼠标，电脑，计算机｝，属性集M＝｛存取，指令，数据，运算，控制，芯片，通信，系统，设备，存放，输入，计算，插槽，信号，中断，多媒体，处理，接口，输出，适配器，图像，显示，传输，碟片，驱动器，板卡，硬件｝。由此可以得到一个具有 43 个形式概念结点、属性数分层 12 层、对象数分层 13 层的概念格。根据本文提出的相关度计算算法，可以算出关键词之间的相关度。把本文算法计算的相关度同向量空间算法计算的相关度做比较，其部分结果如表 2 所示。

表 2 **相关度计算结果**

关键词 1	关键词 2	本文算法	向量空间算法
计算机	电脑	0.881	1.000
计算机	CPU	0.718	0.540
计算机	键盘	0.513	0.354
键盘	鼠标	0.513	1.000
内存	网卡	0.080	0.125

从表 2 可以看出，本文方法的计算结果符合人们对该领域的认知：“计算机”与“电脑”的相关度要高于“计算机”与“CPU”的相关度，而“计算机”与“CPU”的相关度要高于“计算机”与“键盘”的相关度。本文的方法是出于对整个领域的词汇之间相关性的考虑，例如在实验的领域里“计算机”和其他关 键词之间的相关度都是比较大的。而用向量空间的=计算方法由于只计算特征向量之间的差异，在理解=整体相关性上存在着偏差，所以该算法在实验中出 现了比较明显的错误：“键盘”和“鼠标”的相关度为 1，这是由于它们包含的特征向量完全相同（数据疏松，即缺失某些能够区分 2 个关键词的特征向量）造成的。而对于数据噪声可能出现某些特征向量冗余或者某些特征向量选取不恰当的情况，由于概念格本身的结构特点，使得这些特征向量对概念格结构不会有太大影响，从而使其并不能过多影响计算结果。例如，大部分关键词都包含“设备”这个特征向量，这就使大部分形式概念包含“设备”，不过它对计算结果没有太大影响。

如果设定此概念格的阈值 λ 为 0．7，那么对于用户查询的关键词“计算机”，可以返回“电脑”和“CPU”；对于用户查询的关键词“显示器”，可以返回“显卡”和“计算机”。由此可以为用户检索提供一些语义相关的帮助。

5 结语

本文提出了一种基于概念格的相关度计算方法，并将其应用到对信息检索的语义支持中．通过对算法的实现和实验说明，该方法在语义相关度计算方面是有效的，并降低了数据噪声和数据疏松对计算结果的影响，且该方法容易实现在信息检索中对 语义检索的支持。目前的研究是在单值形式背景的情况下进行的，考虑到特征词可以有不同的权重，则需要在多值形式背景下计算语义相关度，这是下一 步的工作重点。

参考文献

[1] 降惠. 概念格理论研究进展与发展综述[J]. 办公自动化, 2019,24(09):18-21+28.

[2] 刘萍, 彭小芳. 基于形式概念分析的词汇相似度计算[J]. 数据分析与知识发现, 2020, 4(05):66-74.

[3] 李晓光,王大玲,于戈. 基于统计语言模型的信息检索[J]. 计算机科学, 2005, 32(08):124－127.

[4]许云, 樊孝忠, 张锋. 基于知网的语义相关度计算. 北京理工大学学报, 2005, 25(05):411－414.

[5] 裴培, 丁雪晶. 基于本体的语义相似度计算综述[J]. 合肥学院学报(综合版), 2020, 37(05):68-74.

[6]M Priya, Aswani Kumar Ch. A novel method for merging academic social network ontologies using formal concept analysis and hybrid semantic similarity measure[J]. Library Hi Tech,2019,38(2):

[7] 沈夏炯,韩道军,刘宗田,等. 概念格构造算法的改进[J]. 计算机工程与应用, 2004, (24):100－103.

[8]蒋运承, 李璞, Akram AFTAB. 一种面向形式概念分析的语义相似度计算框架（英文）[J]. 华南师范大学学报(自然科学版), 2016, 48(03):44-52+2.

[9]Shivani Jain, Seeja.K.R,Rajni Jindal. A New Method for Semantic Similarity Assessment using Fuzzy Formal Concept Analysis & Fuzzy Set Similarity Measure[J]. International Journal of Recent Technology and Engineering (IJRTE),2018,7(4):