《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

王悦

作业	分数[20]
得分	

2020 年 11 月 13 日

基于形式概念分析的词汇相似度计算

王悦

(大连海事大学 信息科学技术学院 大连 中国 116026)

摘要:在文献层和词汇层之间加入主题层，研究一种新的词汇相似度计算方法。阐述基于形式概念分析（FCA）的主题定义和表示模型，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。该方法依赖文献关键特征词抽取的质量。基于形式概念分析的词汇相似度计算方法有 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: 词汇相似度 形式概念分析 概念格 主题

中图法分类号 TP18

Calculating Word Similarities Based on Formal Concept Analysis

WANG Yue

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

Abstract: This paper tries to add a topic layer between document and word layers, aiming to calculate word similarities effectively.First, we proposed a topic defintion and representation model based on the theory of formal concept analysis. Then, we mapped words to the topic layer. Finally, we developed an algorithm to calculate word similarities with the help of topic-to-topic relationship. The proposed method relies on the quality of extracted feature words of documents.The proposed method utilizes the semantic relations among associated topics, and effectively calculate word similarities.

Keywords: Words Similarity Formal Concept Analysis Concept Lattices Topic

1 引 言

词汇是人类语言和思维的基本单元，而词汇相 似度计算是对词汇间复杂关系的定量度量。词汇 相似度计算作为一项基础技术在文本类、主题提取、信息检索、机器翻译以及知识问答等多个领域有 着重要应用。由于词汇相似度涉及到词法、句法、语义、语用等多个层面，学者们从不同角度有不同的理解和定义。但一致认为现实中很少存在能够在文章中进行互换而不影响原来 句子表达的相似词汇，因此词汇相似度计算更多是从语义距离的角度考量，本文仍采用“词汇相似度”，但所指宽泛的相似，既包含词汇语义相似关系也包含语义相关关系。

已有不少学者针对词汇相似度进行研究与测试，所提方法大致分为两类：基于知识库和基于统计的方法。前者利用语义词典和语义网络中规范的知识体系结构计算词汇之间的相似度，后者基于词汇在大规模语料库（包括传统语料库和 Web 语料库）中的共现关系测量词汇相似度。基于知识库的方法具有坚实的语义学基础，但是收录的词汇有限。而基于统计的方法虽然词汇覆盖面广，却存在数据噪音和语义误差问题。这是因为以往的统计方法大多基于词汇在文献中的上下文信息，只考虑了词汇和文献之间的对应关系（见图 1 中A 部分），而忽略了起桥梁作用的主题信息。本文认为文献虽然由词汇组成，但词汇是通过主题与文献 建立联系的。一篇文献的语义实际上是通过主题标注的（B 部分），而主题又是由不同

的词汇表达的（C 部分）。词汇和文献之间实际上是间接关系。为了提高词汇相似度计算的精准度，有必要在文献层和词汇层之间加入主题层。

本文基于形式概念分析理论（Formal Concept Analysis，FCA），挖掘文献集合中的隐含研究主题，将词汇项映射到主题层级，并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，以期能更精准地测量出词汇的语义相似度。

具体来说，本文着重研究三个问题：

（1） 基于 FCA 的主题定义和表示模型；

（2） 基于 FCA 的词汇相似度计算方法；

（3） 以信息检索领域为例，对基于 FCA 的词汇相似度计算进行实证研究。

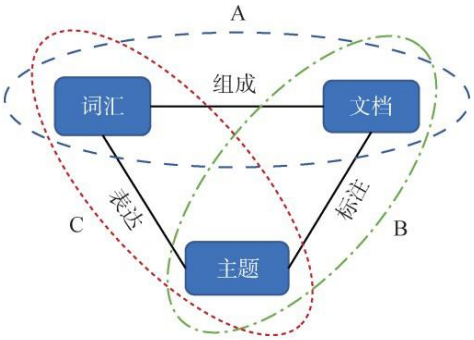

图 1　文献、主题、词汇三者之间的关系

基于形式概念分析的主题定义及表示 模型任何一个文献集合都是形式与内容的 统一。形式是外在物理层次的文献集合，内容是文献内在主题的集合。文献之间存在着隐含且固有的语义关联，对应于文献隐含主题的关联关系，这种关联通常体现在共同关键词上，称之为关键词耦合。Morris 等认为， 通过共同的词语联系在一起的文献可能表 示一个共同的研究主题。而文献中共享的关键词集合则作为共同研究主题的表征。本文利用形式概念分析探测文献中隐含的研究 主题及其关联关系。

2. 形式概念分析

形式概念分析是一种建立在数学基础之上，用于对数据集中概念结构的识别、排序和显示的数学分析理论。以人的认知为中心，将概念诠释为由外延（对象）和内涵（属性）两个部分组成的思想单元。概念的外延被理解为属于这个概念的所有对象集合，而内涵则被认为是所有这些对象所共有的特征集。概念和概念间的泛化和例化关系可以构成一个概念格，因此形式概念分析通过概念格对概念及其层次关系进行形象化描述。形式概念分析中对概念的定义巧妙地解读了文献与关键词的关系。由一组共同关键词联系在一起的特定文献集合代表一个概念 的外延，而这些文献共享的关键词集合代表该概念的内涵（即该组文献包含的主题）。因此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 $I \subseteq D \times K$。dIk 或者$(d, k) \in I$ 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。

表 1　形式背景示例

Table 1　An Example of Formal Context

	k_1	k_2	k_3	k_4	k_5
d_1	×	×		×	×
d_2		×	×		×
d_3	×		×	×	×
d_4	×	×	×		×
d_5	×	×	×	×	×

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 $f(A) = \{ k \in K \mid \forall d \in A, dIk \}$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 $g(T)=\{d \in D \mid \forall k \in T, dIk\}$，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 的一个形式概念。以表 1 为例，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 $A1 \subseteq A2$（或 $T2 \subseteq T1$），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（$A1$，$T1$）$\leqslant$（$A2$，$T2$）。关系“$\leqslant$”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 $\beta(D, K, I)$，并且被称为形式背景 M 的概念格。

图 2 展示了由表 1 形式背景所构建的概念格，揭示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 $d1$、$d3$、$d4$ 和 $d5$；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 $d5$。

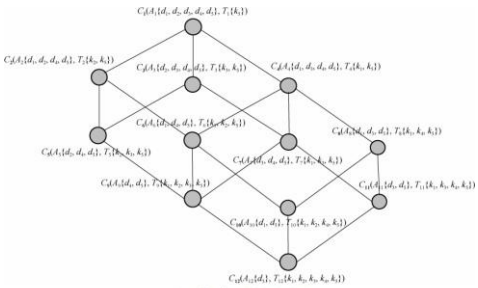

图 2　基于表 1 生成的概念格

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 2 概念格得到 $T-K$ 关联矩阵 L，如表 2 所示。$K-T$ 关联矩阵 LT 如表 3 所示。

Table 2　The Association Matrix of $T-K$ Based on the Concept Lattice in Fig.2

	k_1	k_2	k_3	k_4	k_5
T_1	0	0	0	0	1
T_2	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_9	1	1	1	0	1
T_{10}	1	1	0	1	1
T_{11}	1	0	1	1	1
T_{12}	1	1	1	1	1

表3 基于图2概念格的$K-T$关联矩阵

Table 3 The Association Matrix of $K-T$ Based on the Concept Lattice in Fig.2

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
k_1	0	0	0	1	0	1	1	1	1	1	1	1
k_2	0	1	0	0	1	1	0	0	1	1	0	1
k_3	0	0	1	0	1	0	1	0	1	0	1	1
k_4	0	0	0	0	0	0	0	1	0	1	1	1
k_5	1	1	1	1	1	1	1	1	1	1	1	1

基于词汇集 K 与主题 T 的关联矩阵 LT 可以得到基于主题的词汇表征，如公式（1）所示。

$ki=[wi1, wi2, \cdots, wij, \cdots, wip]$ (1)

其中，wij 的值反映词汇 ki 在主题 Tj 中的所属情况，当 Tj 包含 ki 时，$wij=1$，否则 $wij=0$。该表征可以简化为公式（2）。

$ki=\{Tj \mid ki \in Tj\}$ (2)

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 $k2$，$k1=\{T11, T12, \cdots, T1m\}$，$k2=\{T21, T22, \cdots, T2n\}$ 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。词汇相似度的计算如公式（3）所示。

$$sim(k_1,k_2)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}sim(T_{1i},T_{2j})}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}sim(T_{1i},T_{1j})}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}sim(T_{2i},T_{2j})}} \quad (3)$$

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。主题之间的相似度可以通过其在概念格中的结构位置来计算，主要考察两个主题节点的最近公共父节点深度和最短路径长度。具体来说任意两个主题 Ti 和 Tj 的相似度计算如公式（4）所示。

$$sim(T_i,T_j)=\frac{2\times depth(LCS)}{2\times depth(LCS)+distance(T_i,T_j)} \quad (4)$$

其中，Ti，Tj 分别表示概念格结构中的两个主题节点；LCS 代表 Ti 和 Tj 的最近公共父节点；$depth$（LCS）表示最近公共父节点所处的深度（设根节点的深度为1），$distance(Ti, Tj)$ 代表两个节点之间的最短路径长度。例如选取图2概念格中主题 $T6$ 和主题 $T7$，它们的最近公共父节点是 $T4$，$T4$ 的深度为2。主题 $T6$ 和主题 $T7$ 之间的最短路径为2，因此相似度计算如下所示。

$$sim(T_6,T_7)=\frac{2\times depth(T_4)}{2\times depth(T_4)+distance(T_6,T_7)}=\frac{2\times 2}{2\times 2+2}=\frac{2}{3}$$

4 结语

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。

参考文献

[1] 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

[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, v.4;No.41(05):70-78