《智能信息处理》课程作业

基于形式概念分析处理文本数据

李安宁

作业	分数[20]
得分	

2021 年 11 月 27 日

基于形式概念分析处理文本数据

李安宁
(大连海事大学 电子信息，大连 116026)

摘要 随着不断向“万事互联网”时代推进，因特网作为最大的数据来源，数据的种类愈加多元化，不再仅仅是单一的非结构化数据，而现在大量的相关工作和研究，包括人工智能领域等都会涉及到非结构化的文本数据的处理，例如从自然语言文本中抽取事实信息、知识图谱补全采取结构信息和文本信息进行联合建模、利用数量庞大的文本和网页等非结构化资源构建本体，因此文本数据处理对下游任务至关重要。

但传统基于空间向量模型来表示文本的处理方法使得文本中的特征词的维数过高，导致复杂的计算量，而概念格能够自动的从形势背景中提取概念，并组织成概念的层次结构，为海量的、非结构化文本集建立基于文档的形式背景，因此能够为文本中所有的概念建立起一个清晰的概念层次结构。综上所述，本文引入形式概念分析对文本进行数据分析和处理，以达到更优的性能。

关键词：形式概念分析；概念格；文本处理中图法分类号：TP18

文献标识码：A

Processing Text Data Based on Formal Concept Analysis

Shi Xueying
(Computer science and technology, Dalian maritime university, Dalian 116026, China)

Abstract With the continuous advancement of the "everything Internet" era, the Internet as the largest source of data, the variety of data is more diverse, not just a single unstructured data, but now a lot of related work and research, including artificial intelligence, etc. It involves the processing of unstructured text data, such as extracting fact information from natural language texts, complementing knowledge maps with structural information and text information for joint modeling, and constructing unstructured resources such as large amounts of text and web pages. Ontology, so text data processing is critical to downstream tasks.

However, the traditional method based on space vector model to represent text makes the dimension of the feature words in the text too high, resulting in complex calculations, and the concept lattice can automatically extract concepts from the situation background and organize them into conceptual hierarchies unstructured text set, thus establishing a clear conceptual hierarchy for all concepts in the text. In summary, the paper introduces the formal concept analysis to analyze and process the text, which has achieved better performance.

Keywords: formal concept analysis; concept lattice; text processing

1 引言

自因特网诞生以来，其高速的计算能力、便捷的网络服务等极大改变了人类的交互方式。如今随着网络以及信息技术的快速发展，互联网正逐步转型为万事互联网（IoE），它以一种更加智能和紧密的形式将人、过程、数据和网络聚集在一起，将信息转化巨大的网络资源，以追求“万物互联”终极目标。IoE 将人、事、物等实体和关系融合和集成到网络中，必然会带来前所未有数据量，导致网络信息存储呈爆炸式增加。

如今互联网作为最大的数据存储库，是最主要的信息源，数据的种类的形式愈加多样化，已经从普通的静态文本、图像的存储媒介，转向交互式的网络信息平台发展。面对如此大规模的文本数据信息,需进行文本信息处理，实现信息有效地收集、检索和选择关，在海量的信息中发现新的概念,并自动分析它们之间的关系[1]。在这样的需求驱动下,高效的文本数据处理方法极为重要，文本数据处理是从文本中抽取有效、可理解的、有价值知识，并且利用这些知识更好地组织信息，应用于其他处理过程，如信息抽取、本体构建和知识图谱补全等任务。但是，现有的文本处理方法存在一些问题：很多方法大多是用空间向量模型来表示文本，造成文本特征词的维数过高，从而使计算复杂度增加。而用形式概念分析中的概念来表示文本能很好的解决上述问题。

2 相关介绍

形式概念分析，即 FCA（FormalConceptAnalysis），又称作概念格，由 WilleR 于 1982 年首先提出。形式概念分析能够自动的从形式背景中进行数据分析，提取概念与概念间的规则[2]，识别那些具有共同特征和属性的一组)对象，从而获得概念与概念之间的相关关系，并且基于该关系构建格结构，形成概念间的层次结构，并通过 Hasse 图实现对数据的可视化，能够轻松地构建和展现出数据之间的依赖或因果关系模型，所以概念格被作为进行数据分析的有力工具。以下是对概念格的相关定义[3]：

定义 1 **形式背景** 通常形式背景(formal context)采用高度结构化的形式来表示，即一个形式背景就是一个三元组 K=(G, M, I) , 其中 G 是对象的集合, M 是属性的集合, I 是 G 和 M 之间的二元关系。对于∀ g∈G, m∈M, 若（g, m)∈I, 就说 g 具有属性 m, 记做 gⅠm。

定义 2 **形式概念** 给定一个形式背景 K=（G,M,R）。序偶（E,I）为形式背景 K 上的一个形式概念，其中对象集 E⊆G,属性集 I⊆M。则称 E 和 I 分别称为概念的外延和内涵。

定义 3 **概念格** 给定两个形式概念 G1=（A1,B1）和 G2=（A2,B2），若满足 A1 ∈ A2，

称(A1, B1) 为子概念) ,(A2, B2) 为超概念, 且满足（A1,B1）≤（A2,B2）的偏序关系。这种由形式背景中所有形式概念，通过此偏序关系构建出来的完全格即为概念格。

知识图谱是谷歌于 2012 年提出，旨在让机器理解用户搜索词所代表的的具体含义，提高收索引擎质量，为用户返回更加精准和有效的信息，是大数据背景下提出的一种知识表示和管理方式。通常，知识图谱以高度结构化的形式表示，其本质上是一个关系的网络，提供了从“关系”的角度看待问题的方式。然而现有的知识图谱所包含的实体和关系的数量是有限的，因此需要不断进行补全使知识图谱达到更全的覆盖面，如今的大部分现有的表示模型只能从知识图谱内已有的实体和关系进行推理发现隐藏的新事实，而无法自动的处理知识图谱外的新实体和关系从而扩大知识图谱的规模[4]。如何从文本中发现显得实体和关系，以及利用实体丰富的文本信息进行补全是一大关键，这就需要文本处理的支持。

3 文本数据的应用

如今，互联网以及各种衍生的网络服务应用中的数据不再仅仅是单一的结构化数据，数据的种类的形式愈加多样化，包括音频、语音、图像、文本等，其中 web 网页的非结构化的文本数据成为信息资源的主要来源。而现在大量的相关工作和研究，包括人工智能领域等都会利用文本信息进行，例如从自然语言文本中抽取事实信息、知识图谱补全利用结构信息和文本信息进行联合建模、利用数量庞大的文本和网页等非结构化资源构建本体，因此文本数据处理对下游任务至关重要。

3.1 信息抽取

信息抽取(informationextraction，IE)实现的文本数据结构化，是一种直接从自然语言文本中抽取事实信息，并以结构化的形式描述信息的过程[4]。其中，应用于 Web 页面检索并组织信息的 Web 信息抽取(Web information extraction，WebIE)是将 Web 作为信息源的一类信息抽取。经过处理的信息能很好支持一些决策系统、搜索任务的进行，如今信息处理技术有了很大进展，国内外已经开发出多种工具系统用于信息抽取。因为各种工具采取的原理和技术不同，Web 信息抽取大致可以划分为五种方式，包括基于页面抽取语言的信息抽取、基于 HTML 结构的信息抽取、基于自然语言处理 的信息抽取、包装器归纳方式的信息抽取和基于模式的信息抽取。

3.2 知识图谱补全

知识图谱是谷歌于 2012 年提出，旨在让机器理解用户搜索词所代表的的具体含义，提高收索引擎质量，为用户返回更加精准和有效的信息，是大数据背景下提出的一种知识表示和管理方式。通常，知识图谱以高度结构化的形式表示，其本质上是一个关系的网络，提供了从“关系”的角度看待问题的方式。然而现有的知识图谱所包含的实体和关系的数是有限的，因此需要不断进行补全使知识图谱达到更全的覆盖面，如今的大部分现有的表示模型只能从知识图谱内已有的实体和关系进行推理发现隐藏的新事实，而无法自动的处理知识图谱外的新实体和关系从而扩大知识图谱的规模[5]。如何从文本中发现显得实体和关系，以及利用实体丰富的文本信息进行补全是一大关键，这就需要文本处理的支持。

3.3 本体构建

“本体（Ontology）”一词源于哲学领域，用来描述事物的本质，引入计算机等领域用来描述概念和概念之间的关系。知识图谱描述的现实世界中实体以及实体之间的关系，而本体是一个概念网络模型，是“对共享概念模型明确的形式化规范说明”[5]。

构建本体所用到的资源主要有两种：一种是结构化资源，主要包含主题词表、关系数据库等；另一种是非结构化资源，主要是随互联网发展而大量产生的数量庞大的文本、网页、电子文档等作为研究对象。两种方式各有其自身特点：使用结构化资源构造领域本体效率极高，结构化资源的结构层次关系，为领域本体概念提取、关系识别提供参考。但这类领域本体存在不易修改和更新的缺点，主要是因为以固定层次结构的资源为基础。仅依靠单一结构的资源来构建的本体，已难以满足领域发展的需要，利用数量庞大的文本、网页构建覆盖范围广、语义丰富、能够快速更新等的领域本体，已成为本体发展的必然需求。这一类方法的缺点是不同形式的非结构化文本资源无法使用同一结构来表示，增加了构建本体的工作量。

因此需要需求更加有效地文本处理，以便于构建概念及语义关系更为丰富的领域本体。

4 基于概念格处理文本数据

4.1 问题概述

基于上述文本数据的应用，需要在文本处理过程中从数据中抽取有用的信息和知识，将非结构化数据以结构化的形式进行存储。现有的大多数的文本表示方法都是基于空间向量模型，将文本映射到向量空间中，将其表示为 $t\in R^K$ 维的向量空间。对于给定的文本，表示为 $t\in R^K$ 维的向量空间。对于给定的文本，将其表述为 D={d1, d2, d3,…, dn}。其中 di=<(t1, w1) ,(t2, w2) , …,(tn, wn) >，ti 为特征词，wi 为此特征词 ti 在文档 di 中的权重。但是，将将其表述为 D={d1, d2, d3,…, dn}。其中 di=<(t1, w1) ,(t2, w2) , …,(tn, wn) >，ti 为特

征词，wi 为此特征词 ti 在文档 di 中的权重。但是，将文本表示为空间向量会导致特征词的维数过多，带来复杂的计算量，不利于文本处理。由于概念中包含多个属性, 与空间向量模型相比维数有所降低, 从而降低了计算的复杂度，因此利用概念来表示文档能很好的解决这个问题。

本文提出了一种基于概念格的文本概念处理方法, 首先把文本集表示成形式背景, 即文本中的特征词为形式背景中的属性, 文本作为形式背景中的对象, 这样通过为文本建立一个概念集合，为文本中的概念建立了清晰地来看清概念的层次结构 , 从而在文本和概念之间处理和挖掘它们的关系，而概念格及其 Hasse 图体现了概念内涵和外延的统一[6], 反映了对象和属性间的联系以及概念间的泛化和特化关系。基于这种情况, 文中提出了利用概念格来处理概念间的潜在关系，实现了基于概念格的文本处理。

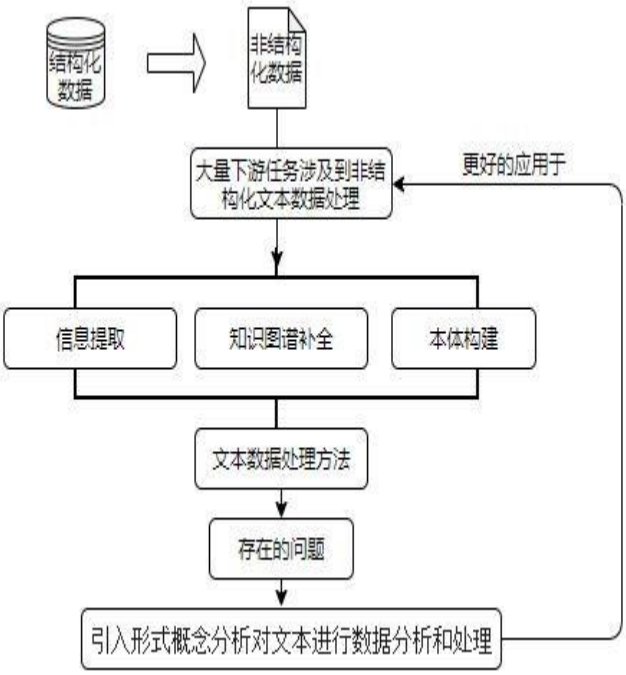

图 1 研究问题背景及目标

4.2 文本处理中概念格表示模型

因为概念格处理文本数据 ,基本思想是将本文设定为概念格中的对象集, 文档中的属性和特征构成属性集，从而建立起一个基于文档的形式背景，定义如下:

定义 4 :**基于文档的形式背景** 一个基于文档的形式背景是一个三元组 C = (G , M , I), 其中 G 是文档(对象)集, M 是特征(属性)集，即集合 G ={g1 , g2 , … , gm}为一个包含 m 个文本的集合;集合 M =m1 ∪ m2 …∪ mj , 其中 mi ={mi1 , mi2 , …, min }, i =1 , 2 , …, j。mi 为第 i 个 Web 文本的 n 个特征组成的特征集合 ;I 为 G 和 M 间的二元关系。

形式背景可以用一个二维表来表示, 二维表的行表示文档, 列表示特征属性。如表 1 所示的一个形式背景, 文档集 G ={d1 , d2 , d3,d4}, 关键词集 T ={t1 , t2 , t3 , t 4 , t 5}, I =D ×T ,表中的“1” 表明文档中有该关键词 ,“0” 表示文档中没有该关键词。

表 1　形式背景

I	m1	m2	m3	m4	m5	m6
d1	1	0	1	0	1	1
d2	1	0	1	0	0	1
d3	0	1	0	1	0	0
d4	1	0	0	1	0	1

为了构建一个概念层次结构, 必须找到形式背景的所有概念，生成在新的形式背景下的概念格，就可以清楚地看出文档集和属性集之间的内在关系。

4.3 基于概念格的文本处理

根据建立的形式背景 K =(G, M , I), 可以通过批处理算法和渐进式算法构造概念格。设 B ={m|m∈ M}(m’ 表示具有属性 m 的对象的集合 ,称 m’ 为 m 的属性外延), 形式背景 K=(G,M,I)的基 B(Basis)是 K 的所有属性外延组成的集合，用 FB 表示集合, 它是由基 B 的交集形式生成, 即 F B ={m’|I 2B(表示 B 的幂集)}, 对任意 F ∈ FB, 用 γ (F)表示满足条件的 M 的子集 .对于 M 中的每一个元素 m , F 是 m′ 的子集 , 即 γ (F)={m ∈ M| F∈m′ }。在算法中:B 为添加新数据前形式背景的基 ; FB 为由基 B 生成的所有概念的对象集合; γ (FB)为概念集合 FB 的内涵的集合 γ (FB)={γ (m′)m′ ∈ B};B′ 为添加新数据后形式背景的基 ; FB’ 为由基 B’ 生成的所有概念的对象集合。算法如下 :

```
Input : 基 B′, 原始的格 L
Output:修改的格
begin
   初始化 FB’
   Let FB’={G}, γ(G)为空集
For each m′∈ B′do
if m′=G then γ(G)=γ(G)∪{m}
for each m′∈ B′ do
   for each F ∈ FB ’ do
   begin
   F0' =F∩ m′
   If    F0’⊄ FB’ then
   FB ’= FB ’ ∪ F0'
```

γ($F_0^{'}$)=γ($F_0^{'}$) ∪ {m}
// 修改已经存在格的外延
if exits F′∈ FB and γ F0')=γ(F′)
then F′=F′∪ F0'
else
// 添加一个新的概念到已经存在的格
Insert concept (F0' , γ(F′0)) to the existed lattice L
end if
end
end

基于上述算法，在给定原始形式背景 K =(G, M , I)所对应的原始概念格通过修改格的外延和格节点的更新来生成在新的形式背景下的概念格[7]。由图 1 中的概念格就可以清楚地看出文档集和属性集之间的内在关系。

表 2 由表 1 获得的所有形式概念

步骤	关键词	外延	内涵
1		{d1, d2, d3, d4}	{}
2	t1, t6	{d1, d2, d4}	{t1, t6}
3	t2	{d3} {}	{t2, t4} {t1, t2, t3, t4, t5, t6}
4	t3, t5	{d1, d2}	{t1, t3, t5, t6}
5	t4	{d3, d4} {d4}	{t4} {t1, t4, t6}

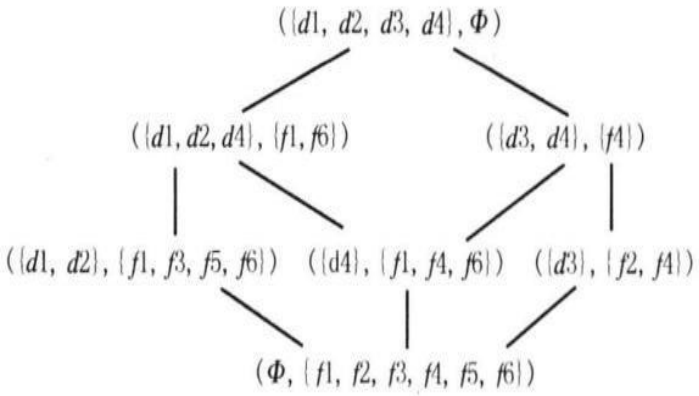

图 2 表 1 中形式背景的 Hasse 图

基于概念格对文本进行处理，这样通过为文本中的概念建立了清晰的层次结构 ，从而在文本和概念之间处理和挖掘它们的关系，并用概念格及其 Hasse 图体现了概念内涵和外延的统一, 反映了对象和属性(特征)间的联系以及概念间的泛化和特化关系，实现了利用概念格来处理文本间的潜在关系。

5 结语

本文开篇通过论述互联网时代快速激增的非结构化数据以及文本处理对下游任务的支撑作用，分析了传统的文本处理方法所存在一些问题，即基于空间向量模型来表示文本，使得文本中的特征词的维数过高，导致复杂的计算量，不利于实现。

基于上述问题，由于概念格能够自动的从形式背景中提取概念，并组织成概念的层次结构，因此本文提出采用概念格模型来表示文本数据[8]，通过为海量的、非结构化文本集建立形式背景，将文档作为形式背景中的对象，而包含在文档中的特征词作为概念格的属性,从而得到一个基于文档集的形式背景，获得所有的概念,为文本中所有的概念建立起一个清晰的概念层次结构。由概念格模型处理的文本数据可以看出 ,利用构造的概念格能够有效地抽取隐含在文本中潜在的、有价值的知识,极大消除了文本非结构化及异构问题，提高文本质量，更好地用于下游任务。

参考文献

[1] 黄埔. 文本信息抽取优化关键技术研究与系统实现[D]. 北京邮电大学, 2019.

[2]Wille R.Conceptual graphs and formal concept analysis [C] //International Conference on Conceptual Structures. Springer Berlin Heidelberg,1997: 290-303.

[3]Ganter B, Wille R. Applied lattice theory:Formal concept analysis[C] //In GeneralLattice Theory,G.Grätzer editor, Birkhäuser,1997.

[4]ZHONG H,ZHANG J,WANG Z,et al.Aligning knowledge and text embeddingsbyentity descriptions[C]//the 2015 Conference on Empirical Methods in Natural Language Processing,2015:267-272

[5] 陈晓美, 毕强. 面向文本的领域本体学习方法与应用研究综述[J]. 图书情报工作, 2011,55(23):27-31.

[6] Godin R,Missaoui R, Alaui H .Increment al concep t f ormation algorithm based on galois (concept) lattice[J] .Compu tati onal Itelli-gence,1995, 11(2):246-26.

[7]Wille R.Restructuring lattice theory: an approach based on hierar- chies of concepts [M] // RivalI.OrderedSets.Dordrecht: Reidel, 1982: 445- 470.

[8]Sun XB,Li BX,211ang S,Tao CQ.Using lattice of class and method dependence for change impact analysis of object oriented programs l AJ.Proceedings of the Symposium on Applied Computing[C].Tai Chung,Taiwan:ACM,2011.1439-1444 .