特定主题概念关联的挖掘及其表示式的实现

丁泽亚[1,2]　缪建明[2]　张全[2]

1. 中国科学院研究生院，北京，100049

2. 中国科学院声学研究所，北京，100190

Email: zeya.ding@gmail.com

摘要：本文提出了一种特定主题概念关联知识挖掘的方法。在实际特定主题的语料基础上，本文使用了信息增益的方法选取出主题关键词，并对关键词所对应的关联概念进行统计，结合对应关键词的重要程度值，赋予这些关联概念以权重值，从而根据权重值得到特定主题下的概念关联知识及其表达式。通过实验证明，该方法是有效的。

关键词：概念关联　主题关键词　信息增益

Mining Concept Related Knowledge and the Realization of its Expressions

Ding Zeya[1,2]　Miao Jianming[2]　Zhang Quan[2]

1. Graduate University of Chinese Academy of Sciences, Beijing, 100049

2. Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190

Email: zeya.ding@gmail.com

Abstract: This paper proposes a method of mining concept related knowledge in specific topics. Based on the actual corpus of a specific topic, we use the information gain to select keywords of the topic. Then we get the statistics of the related concepts of keywords, and give weight values to the related concepts combined with importance degree of keywords. According to the weight of the concepts, we can obtain concept related knowledge and expressions of a specific topic. Proved by the experiment, this method is effective.

Keywords: Related concept, Topic keyword, Information gain

1. 概述

伴随着社会经济的发展，一些方面的信息或新闻会成为人们关注的焦点，同时由于信息的发布和获取变得越来越容易，形成了一些信息的聚集，如果对这些信息加以分析和利用，对于舆论方面的研究有非常大的用处。

我们把人们关注的一些社会焦点称为“主题”，比如上海世博会、地震、拆迁等。在对这些主题的语料进行信息挖掘或是舆论研究时，常常需要进行文本分类。目前比较常用的文本分类方法主要是基于统计的方法，而在HNC（概念层次网络）理论的基础上，我们希望利用领域知识来对文本进行分类，尤其是特定主题的文本，从而通过语义信息来提高分类效果，而不是仅仅用统计的方法。利用领域知识进行文本分类时，非常重要的部分是每个类别的领域框架，而设计类别领域框架的关键之一就是概念关联知识的挖掘。概念关联知识用来描述两个或多个概念之间有什么样的关联性（强关联或关联），它是领域句类表示式的一部分。用经验知识对概念关联进行挖掘时，常常不够全面或者不够准确，所以本文在统计方法的基础上，结合实际语料，对特定主题的概念关联进行了挖掘，并得出了概念关联的表示式。

2. 相关知识

2.1 关于 HNC（概念层次网络）理论

HNC 理论，即概念层次网络理论，是由中科院声学所黄曾阳研究员创立的、不同于传统语言研究的自然语言处理的理论体系。这种理论体系建立了以作用效应链为核心的多层级概念基元符号体系，并在概念基元空间的基础上，通过语义块和句类、领域句类这些知识的符号表示，分别构造出了词语（含短语）、语句、句群的表示式，从而对自然语言词语、语句和段落、篇章的进行研究[1]。

HNC 理论立足于语言概念空间，把语言概念空间划分为四个层级，即 “基层一第一介层一第二介层一上层”。基层由语言概念基元符号体系构成，大体对应着语言空间的词语；第一介层由句类符号体系构成，大体对应着语言空间的语句；第二介层由语境单元符号体系构成，大体对应着语言空间的句群；上层由语境符号体系构成，大体上对应着语言空间的段落及篇章。这四个层级都有对应的数学物理表示式。

HNC 理论中的领域划分建立在概念基元符号体系的基础上，是对人类活动所属范畴的划分。HNC 理论中的领域划分为 10 类[2]，涵盖了整个人类活动以及其他生命体的本能活动、自然界灾祸状态的内容，如表 1 所示[3]。

表 1 HNC 领域分类表

编号	描述内容	对应的扩展基元符号	基元符号类型
1	心理活动及精神状态	71,72	第一类扩展基元
2	人类思维活动	8	
3	专业及追求活动（第二类劳动）	a, b	
4	理念活动	d	
5	第一类劳动	q6	第二类扩展基元
6	业余活动	q7	
7	信仰活动	q8	
8	本能活动	6m (m=0-5)	
9	灾祸	3228α(α=8-b)	
10	状态	503，50α(α=8-b)	

领域句类表示式是以句类表示式为表述基础，融合了领域概念所蕴含的领域知识的知识表述框架。而领域句类表示式和概念关联知识是领域句类知识包含的两个方面的内容[4,5]。

2.2 概念关联知识

概念关联知识在句类表示式和领域句类表示式中都存在着，但是两者是不同的。在句类表示式中，概念关联知识表现为语义块之间的关联知识[4]。而在领域句类表示式中，除了已有的语义块间的关联知识，还包含了语义块构成知识及领域概念之间的关联知识。语义块构成知识描述了语义块在该领域下应该由何种概念构成，具有领域针对性。领域概念关联知识描述了该领域概念和其他概念有什么样的关联。如下面例子：

泛组织的概念节点为 a03，其概念延伸表示式如下：

a03:(t=b)

a03t　泛组织的 3 种基本形式

a039　　行业

a03a　　流派

a03b　超组织

其领域句类表示式为：

SCD(a03)=Cn-1PrXY0-a*21J//XY0-b*21J

+~XY10-1S0-b*322J

从上面可以看到，共有两组表示式：第一组描述某一个具体泛组织机构（GBK1）正在进行或者已经完成了（EK）某一个具体的专业活动的内容（GBK2）；第二组描述完成以后，这一活动（GBK1）对人、社会或者具体机构（GBK2）产生了（EK）什么样的效应状态（GBK3）。

领域知识的重点体现在专业活动的完成及其产生的效应上。其中的领域句类的语义块定义如下：

Cn-1:=j10(时间);Pr:=l16(原因);

A:=pea03b\k(泛组织机构);YC:=ra00β (具体专业活动);

B:=pea01(组织机构);SC:=r50a1//r50a2(好//坏情况)

在对泛组织（a03）的领域概念本身节点分析后我们发现，它的三项延伸结构之间天然就具有相关特性，而具体的泛组织类型又必然与其专业活动领域具有强相关特性。HNC 符号映射化后得到如下概念关联式：

a039=a03t　(行业与其它泛组织相关联)

a03a=a03t　(流派与其它泛组织相关联)

a03b=a03t　(超组织与其它泛组织相关联)

a03b\1≡a1　(政治超组织与政治强相关联)

a03b\2≡a2　(经济超组织与经济强相关联)

a03b\3≡a3　(文化超组织与文化强相关联)

a03b\4=a　(其他非政府组织与专业活动领域相关联)

3. 基于统计方法的特定主题概念关联的挖掘

在之前的介绍中，概念关联知识及其表示式是通过分析概念节点本身及领域句类表示式来获得的。对于某一特定主题而言，这样主题一般都具有一定的新颖性，由于先验知识存在一定的缺失，如果使用上面的分析方法对其进行概念关联挖掘，则主观性会比较强，并且不能保证概念关联知识的准确性和完备性。所以我们在实际特定主题的语料基础上，使用了统计的方法，首先通过信息增益的方法判定主题关键词；然后对关键词所对应的关联概念进行统计，并结合这些关联概念所对应关键词的重要程度值，赋予这些关联概念以权重值，从而来判定概念之间的关联强度。

3.1 主题关键词获取

文本分类中经常使用信息增益来进行文本特征项的选择。一个特征项的信息增益表明了这个

特征项能为分类提供多少信息量，从而来衡量此特征项的重要程度[6]。在概念关联的挖掘中，我们借鉴了文本分类中信息增益的定义。这里的信息增益是为了描述某个词为区分出此特定主题所提供的信息量，也就是表明了此词对于此特定主题的重要程度或关键性。

概念关联挖掘时，除需要特定主题语料外，同时还需要一般语料，以便于进行相关的概率统计计算，这和文本分类中的特征项信息增益计算是类似的。其中，信息量的多少是由熵来衡量的，因此，词的信息增益即不考虑任何词时文档的熵和考虑该词后文档的熵的差值。下面给出词的信息增益的计算公式：

$$\begin{aligned}\mathrm{Gain}(w_i) &= \mathrm{Entropy}(S) - \mathrm{Expected\ Entropy}(S_{w_i}) \\ &= \{-[P(T)\times \log P(T) + P(\overline{T})\times \log P(\overline{T})]\} \\ &\quad - \{P(w_i)\times [-(P(T|w_i)\times \log P(T|w_i) + P(\overline{T}|w_i)\times \log P(\overline{T}|w_i))] + P(\overline{w_i}) \\ &\quad \times [-(P(T|\overline{w_i})\times \log P(T|\overline{w_i}) + P(\overline{T}|\overline{w_i})\times \log P(\overline{T}|\overline{w_i}))]\}\end{aligned}$$

其中，$P(T)$表示此特定主题T的文档在整个语料中出现的概率，$P(\overline{T})$表示一般语料即非此主题的文档在整个语料中出现的概率，$P(w_i)$表示整个语料中包含此词w_i的文档的概率，$P(\overline{w_i})$表示整个语料中不包含此词w_i的文档的概率，$P(T|w_i)$表示文档包含词w_i时是属于此主题T的条件概率，$P(\overline{T}|w_i)$表示文档包含词w_i时不属于此主题T的条件概率，$P(T|\overline{w_i})$表示文档不包含词w_i时是属于此主题T的条件概率，$P(\overline{T}|\overline{w_i})$表示文档不包含词$w_i$时不属于此主题T的条件概率。

计算得到整个语料中每个词的信息增益后，按信息增益的大小从高到低进行排序，选择信息增益较高的前 N 个词作为主题关键词，表示这些词对此主题的确定做出了较大的贡献或是为从语料中区分出此主题语料提供了很大的信息量，是这个特定主题的关键词。

我们以 1 为阶度来量化这 N 个关键词的重要程度，即关键词中信息增益最小的关键词的重要度为 1，其次为 2，……，信息增益最大的关键词重要度为 N。

设 N 个关键词按信息增益从大到小排序为$\{kw_n, kw_{n-1}, \ldots kw_i, kw_{i-1}, \ldots kw_1\}$，其中$n = N$且$0 \le i \le N$，那么，定义关键词的**重要度**$\mathrm{Value}(kw_i) = i$。

3.2 关联概念统计及权重计算

在 HNC 知识库中，每个词条都有关联概念节点，列出了词条所关联的概念，例如：“地基”这个词的关联概念节点有三个，分别是 54、65、w，表示其与“结构”、“人类特有的本能活动”和“物”这三个概念是相关联。

因此，我们所要做的就是首先从 HNC 知识库中找出每一个主题关键词的关联概念节点（一个或多个），然后对所有的这些关联概念进行统计，即查找出每一个涉及到的概念分别关联了哪些关键词，每一个概念可能会关联 k 个关键词（$0 \le k \le N$），则给这个概念的**关联频次**记为 k。给每个关联概念（concept）赋予一个权重，公式如下：

$$W(\mathrm{concept}) = \log_2\left(\frac{k + \mathrm{Value}_{avg}(kw_{related})}{\mathrm{Count}(\mathrm{concept})}\right)$$

其中，$\mathrm{Value}_{avg}(kw_{related})$表示与此概念相关联的关键词的重要度的算术平均值，$\mathrm{Count}(\mathrm{concept})$表示统计出来的所有关联概念（节点）的个数。

得到每个关联概念的权重值以后，设定一个权重阈值，选出所有权重大于这个阈值的概念，那么，我们认为在这个特定的主题下这些概念之间是相关联的，权重越接近的概念之间的关联强度越大。

4. 特定主题的概念关联表示式的实现

我们所用到的语料来源于网络新闻和博客，实验所设定的主题是“拆迁”，即进行“拆迁”主题中关联概念的挖掘。整个语料包括2200篇拆迁类语料，3000篇其他混合类语料（不包含拆迁相关内容）。

首先运用信息增益的方法进行主题关键词选取，我们选取了信息增益较高的1000个关键词，同时得到它们的重要度Value(kw)，结果如下表2所示：

表2 主题关键词

主题关键词	重要度Value(kw)	关联概念节点
拆除	1000	03a
补偿	999	321b\1
房屋	998	a219\11*9\1
土地	997	jw539
房子	996	a219\11*9\1
强制	995	003a
协议	994	a009aa
开发	993	a21aa
钉子	992	54-00 w
改造	991	3519
房地产	990	a219\i
法律	899	a5
……	……	……

接下来，对这些关键词进行关联概念的统计，去除一些噪音关联概念，如p、w（人、物）等，最后得到的关联概念共有497个，结果如表3所示：

表3 关联概念统计结果

关联概念节点	关联频次	所关联的关键词
65	13	住户 楼房 应急 新房 抵触 公房 矿工 出面 老伴 矿井 出行 拖延 地基
a219\11*9\1	6	房屋 房子 住宅 平房 别墅 屋子
a219\11	6	建筑 房产 业主 商品房 装修 建筑物
451	6	使用权 用途 适用 用户 滥用 动用
a219\1	6	房地产 不动产 工地 施工 修建 扩建
54-	5	新房 公房 窗户 厂房 矿井
a02	5	进行 实施 办理 事宜 举办
……	……	……

计算每个关联概念的权重值，我们设定权重经验阈值为0.65，所有权重值大于0.65的关联概念按照权重大小排序，结果如表4所示：

表 4 关联概念挖掘结果

关联概念节点	权重值W(concept)	概念含义	关键词范例
03a	1.0101	免除约束（约束免除）	拆除
a219\11*9\1	0.8905	住宅建筑	房屋
35	0.8893	立与破	拆毁
65	0.8566	人类特有的本能活动	抵触、拖延
a009aae21	0.8032	权益	权利
003a	0.7911	强制性一轮运作	强制
a21aa	0.7481	开发	开发
a219\1	0.7350	基建（不动产）	房地产
a009aa	0.6976	契约	协议
j86e26	0.6701	消极	粗暴
a5	0.6698	法律	法律

从表 3 中我们可以看到，这些概念都是相关联的，其中一些相邻的概念是强相关联的，那么，“拆迁”主题的概念关联表示式可以表示为如下：

03a ≡ a219\11 * 9\1　（免除约束（约束免除）与住宅建筑是强相关的）
a219\11 * 9\1 ≡ 35　（住宅建筑与立与破是强相关的）
a009aae21 ≡ 003a　（权益与强制性一轮运作是强相关的）
a21aa ≡ a219\1　（开发与基建（不动产）是强相关的）
j86e26 ≡ a5　（消极与法律是强相关的）
35 = 65　（立与破与人类特有的本能活动是相关的）
65 = a009aae21　（人类特有的本能活动与权益是相关的）
……

其他的概念关联表示式在这里就不再一一列出。从上面的概念关联表示式中，我们可以很明显地看到，通过这种方法挖掘的概念关联知识与人对于“拆迁”这一主题的认识相比，具有高度的一致性，而且完备性也比较好。

5. 结束语

本文提出了一种特定主题概念关联知识挖掘的方法。由于单纯从概念节点的领域句类分析来进行关联知识挖掘的方法主观性较强，并且不够全面，所以我们在实际特定主题的语料基础上，使用了信息增益的方法选取出主题关键词，然后对关键词所对应的关联概念进行统计，并结合这些关联概念所对应关键词的重要程度值，赋予这些关联概念以权重值，根据权重值的大小及差异程度来得到在特定主题下的概念关联知识。实验中我们主要关注了“拆迁”主题，从最后得到的概念关联表示式中，可以看出实验结果与人的认知具有较高的一致性，并且比较全面。本文将来的工作是在所得到的概念关联知识的基础上，在基于领域知识的文本分类中进行领域框架的设计。

参考文献

[1] 缪建明.专业活动领域句类的设计与知识表示[D].北京: 中国科学院声学研究所博士学位论文,2007.

[2] 黄曾阳.HNC（概念层次网络）理论[M].北京：清华大学出版社，1998.

[3] 黄曾阳.语言概念空间的基本定理和数学物理表示式[M].北京:海洋出版社,2004.

[4] 贾宁.基于概念知识关联的中文人名和机构名称识别[D].北京: 中国科学院声学研究所博士学位论文,2008.

[5] 缪建明，张全等.基于文章标题信息的汉语自动文本分类[J].计算机工程.2008,34(20):13-17.

[6] 宗成庆.统计自然语言处理[M].北京：清华大学出版社，2008.