面向语料的领域主题词表构建算法

安亚巍[1] 操晓春[2] 罗 顺[1]
(上海通用识别技术研究所 上海 201112)[1] (中国科学院信息工程研究所 北京 100093)[2]

摘 要 针对大规模领域主题词表提取的问题,提出根据给定语料中词共现特征构建词共现特征矩阵的方法。在此基础上进行词簇划分,进而计算出每个词簇的中心词,并以中心词为核心重新组织每个词簇,最终实现面向语料的主题词表的自动构建。实验结果表明,该算法具有较高的准确率和召回率。

关键词 主题词表,词共现特征,词簇划分,语料挖掘
中图法分类号 TP391 **文献标识码** A

Construction Method of Domain Subject Thesaurus Based on Corpus

AN Ya-wei[1] CAO Xiao-chun[2] LUO Shun[1]
(Shanghai General Recognition Technology Institute, Shanghai 201112, China)[1]
(Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)[2]

Abstract To achieve a massive domain corpus oriented subject thesaurus, a method based on feature matrix which is set up by computing words co-occurrence was proposed. By operating on this feature matrix, words are divided into clusters, and central word for each words cluster is calculated. Lexical bundles are finally gained by re-organizing words clusters using central word as a core. The experiment indicates that the proposed method can achieve good precision rate and recall rate.

Keywords Subject thesaurus, Words co-occurrence feature, Words cluster dividing, Corpus mining

1 引言

主题词表是用于表达特定领域中事物和概念等内容的词汇集合。它的本质是对自然语言中的词汇进行选择、规范,以揭示其间的相关关系,由此形成具有领域信息组织和描述功能的受控词汇列表。

主题词表形成于 20 世纪 50 年末,是在吸取元词法、标题法及分面组配分类法等知识组织方法优点的基础上发展起来的。从 1959 年美国杜邦公司编制第一部主题词表至今,主题词表已得到了很大发展。著名的主题词表有美国国会标题表 LCSH、医学标题表 Mesh、工程和科学词汇叙词表 TEST、艺术和建筑叙词表 AAT[1],以及我国国务院办公厅秘书局于 1997 年 12 月修订的国务院公文主题词表等。另一方面,现有主题词表大都是针对通用领域或综合性领域分类的,缺乏面向特定专业或细分领域的主题词表。通常,为优化自然语言处理模型的性能和效果,需要根据训练语料的特征来构建与之相适应的领域主题词表。

在主题词聚类和提取研究方面,肖健等提出了基于语义模板与基于统计工具相结合来提取多词表达的方法[2],陈炯等提出了一种通过挖掘词汇间语义关联来构造特定领域的概念词典的方法[3],葛宁等提出了一种从语料库中自动提取领域知识和标引知识的方法[4]。

本文提出一种面向给定语料的领域主题词表提取算法,通过对训练语料的统计学习,构建语言处理模型,实现领域词汇的自动获取,并通过模型的不断迭代和语料的不断丰富,实现自适应、自增长的词表自动更新。

2 语料准备

选取一定数量的文本,并对其中属于目标领域的文本进行标注,以文本全集作为从中抽取目标领域主题词汇表的样本语料。相关语料的准备主要包含以下 3 部分工作:中文分词、词汇统计和词汇过滤。

1)中文分词:分词采用中科院的 ICTCLAS 分词系统。ICTCLAS 的分词精度高于 98.45%[5],是当前较好的汉语分词工具。

2)词汇统计:初步的词汇统计主要包括词汇频数、包含某一词汇的文本数、样本语料文本总数等统计特征。

3)词汇过滤:包括词性过滤和停用词过滤。其中,词性过滤采用名词作为候选词汇;停用词过滤是指根据停用词表剔除候选词中的停用词,将一些具有一般意义或功能的名词如“今天”“先生”“我们”等剔除。

本文受国家自然科学基金项目(61422213,U1636214)资助。

安亚巍(1978-),男,硕士,工程师,主要研究方向为数据分析处理、知识工程、信息安全,E-mail:ywan20@163.com;**操晓春**(1980-),男,博士,研究员,主要研究方向为多媒体内容安全、计算机视觉;**罗 顺**(1982-),男,硕士,工程师,主要研究方向为数据分析处理、知识工程、信息安全。

3　主题词表的构建

3.1　候选词选取

根据语料准备结果，得到领域词汇集 $W=\{w_i \mid i=1,2,\cdots,n\}$ 和领域文本集 $D=\{d_j \mid j=1,2,\cdots,m\}$，主题词表候选词的选取采用 TFIDF 算法。

Salton 等[6]于 1973 年提出了 TFIDF 算法，其主要思想是：如果某个词汇在某一类型文章中出现的频率高，且在其他类型文章中较少出现，则认为该词汇具有较好的类别区分能力。在领域主题词表提取中，可以利用 TFIDF 方法来量化计算词汇的这种领域性区分能力。TFIDF 算法的具体计算如下。

1）TF(Term Frequency)，即词汇在文本内的频率，计算公式如下：

$$TF_{ij}=\frac{f_{ij}}{\max_k f_{kj}}$$

其中，f_{ij} 表示词汇 w_i 在文本 d_j 中的频数，$\max_k f_{kj}$ 表示文本 d_j 中所有词汇频数的最大值。

2）DF(Document Frequency)，即词汇的文本频率，计算公式如下：

$$DF_i=\frac{n_i}{N}$$

其中，n_i 表示文本中包含词汇 w_i 的文本频数，N 表示样本语料中的文本总数。

3）IDF(Inverse Document Frequency)，即词汇的反文本频率，计算公式如下：

$$IDF_i=-\log_2 DF_i$$

从而，词汇 w_i 的 $TFIDF$ 值，即词汇 w_i 区分文本 d_j 和样本语料中其他文本的能力为：

$$TFIDF_{ij}=TF_{ij}\times IDF_i=\frac{f_{ij}\times\log_2\frac{N}{n_i}}{\max_k f_{kj}}$$

事实上，为计算词汇 w_i 区分文本集 D 和样本语料中其他文本的能力，并进行归一化表示，本文采用如下改进的 $TFIDF$ 的计算公式：

$$TFIDF(w_i)=\sum_{d_j\in D}TF_{ij}\times IDF_i$$

$$=\sum_{j=1}^{m}\frac{\frac{f_{ij}}{\max_k f_{kj}}\times\log_2(\frac{N}{n_i}+0.01)}{\sqrt{\sum_{k=1}^{N}(\frac{f_{ij}}{\max_k f_{kj}}\times\log_2(\frac{N}{n_i}+0.01))^2}}$$

分别计算领域词汇集中每一个词汇的 $TFIDF$ 值，过滤出其中 $TFIDF$ 值大于阈值 h 的词汇作为主题词表的候选词汇集：

$$\{(w_i,TFIDF(w_i)) \mid i=1,2,\cdots,r,r\leqslant n\}$$

3.2　词共现特征阵的构建

词共现模型的主要思想是：在某领域语料中，当一些词汇经常在同一自然语言窗口单元中出现时，则认为这些词汇在该语料环境下存在语义相关性[7]。将该语料领域中所有具有语义相关的词汇聚集在一起，以词汇间相关关系或相关强度为权重，即可得到该领域中主题候选词的词共现特征阵。

词共现度的计算模型如下：

$$C(w_i,w_j)=\frac{1}{2}(\frac{f(w_i,w_j)}{f(w_j)}+\frac{f(w_j,w_i)}{f(w_i)})$$

其中，$f(w_i)$ 为词汇 w_i 在领域文本集 D 中出现的次数，$f(w_i,w_j)$ 为词汇 w_i 和 w_j 在领域文本集 D 中同一自然语言窗口单元中共同出现的次数。本文将一个段落作为一个自然语言窗口单元，$C(w_i,w_j)$ 是该自然语言窗口单元中的词共现度量。

显然，词共现度量模型是对称的，即有：

$$C(w_i,w_j)=C(w_j,w_i)$$

从而得到如下词共现特征阵：

$$P=\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_r \end{array}\begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_r \end{array} \\ \begin{pmatrix} 1 & c_{12} & \cdots & c_{1r} \\ c_{21} & 1 & \ddots & \vdots \\ \vdots & \cdots & \ddots & c_{(r-1)r} \\ c_{r1} & c_{r2} & \cdots & 1 \end{pmatrix} \end{array}$$

其中，$c_{ij}=C(w_i,w_j)$。由于 $C(w_i,w_j)=C(w_j,w_i)$，因此 $c_{ij}=c_{ji}$，即词共现特征阵 P 是 r 阶实对称的。

3.3　主题词簇图

由 3.2 节从样本语料中得到目标领域主题候选词的词共现特征阵 P。如何从词共现度中刻画词汇的语义相关性，并对词汇进行簇聚类，进而刻画词汇簇与目标领域的语义相关性，就成为了主题词表生成的关键。

首先，设定一个阈值 λ，如果词共现特征阵 P 中的元素 $c_{ij}>\lambda$，则认为词汇 w_i 与 w_j 是语义相关的。在图模型中，若将每一个词汇都看成一个节点，则认为当 $c_{ij}>\lambda$ 时，节点 w_i 和 w_j 是连通的且被选中用于构建目标领域主题词表。

假设共现度大于阈值 λ 的词汇共有 t 个，即图模型中共有 t 个节点，最少可用 $t-1$ 条边就能把这些节点连成一个连通图。构建如下语义关联权重模型：

$$Q(w_i,w_j)=\frac{c_{ij}(TFIDF(w_i)+TFIDF(w_j))}{2}$$

计算所有可连通词汇在样本语料中的语义关联权重，并选取其中权重最大的 $t-1$ 条边，这样词共现图就可分成若干连通子图，即簇集，如图 1、图 2 所示。

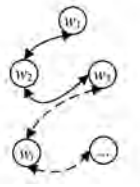

图 1　单连通词簇

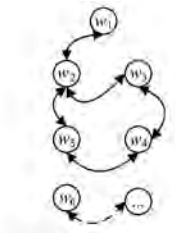

图 2　多连通词簇

若词共现图是一个单连通图，如图 1 所示，则表示目标领域主题候选词的主题是单一的；若词共现图不是单连通的，如图 2 所示，则表示目标领域主题候选词被分割成多个单连通词簇，每个词簇对应着一个子主题。

3.4　阈值 λ 的选取

阈值 λ 的选取，会影响到词共现连通图的规模和结构。若 λ 选取过小，会导致词共现连通图规模过大，结构过于复杂，进而使得词表的主题领域性被弱化；若 λ 选取过大，则会导致词共现连通图规模过小，结构过于简单，进而影响词表的领域词覆盖质量。

考虑到对于语料中规模巨大的共现词对，无法采用人工一一判定的方法进行选取，本文采用概率检测方法，即初步设

（下转第 410 页）

big data [J]. Knowledge-Based Systems, 2017, 117: 3-15.
[13] ZHAI J H, WANG X Z, PANG X H. Voting-based instance selection from large data sets with mapreduce and random weight networks[J]. Information Sciences, 2016, 367: 1066-1077.
[14] SONG G, ROCHAS J, BEZE L E, et al. K nearest neighbour joins for big data on mapreduce: a theoretical and experimentalanalysis[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(9): 2376-2392.
[15] 刘军，林文辉，方澄. Spark 大数据处理-原理、算法与实例[M]. 北京：清华大学出版社，2016.
[16] 翟俊海，郝璞，王婷婷，张明阳. MapReduce 并行化压缩近邻算法[J]. 小型微型计算机系统，2017(12)：2678-2682.

(上接第 397 页)

定一个较低数值的 λ(如 0.1)，然后随机抽取几次(如 10 次)一定规模的词对样本(如 100 个)交由专家判定。若专家判定样本中的词对确实具有语义关联性，则此时的 λ 即可作为阈值输出；否则，对 λ 递增一定步长，再重复随机抽取和专家判定的过程。

事实上，随着 λ 的递增，所抽取样本中专家判定具有语义关联性的词对比率会同步递增。因此，初期可将 λ 递增的步长设置为一个较大的数值(如 0.01)，随着 λ 的递增再将步长逐步减小(如 0.005)。

3.5 词簇中心的选取

假设 $W_1=\{w_i \mid i=1,2,\cdots,s, s\leqslant t\}$ 是一个连通词簇，即目标领域主题中的一个子主题。定义词簇中某一个词汇 w_i 与词簇 W_1 的语义关联度 $O(w_i)$ 如下：

$$O(w_i)=TFIDF(w_i)+\sum_{\substack{j=1\\ j\neq i}}^{s}Q(w_i,w_j)$$

通过归一化，可得到词汇 w_i 对词簇 W_1 的语义贡献度为：$O(w_i \mid W_1)=\dfrac{O(w_i)}{\sum\limits_{w_i\in W_1}O(w_i)}$。

从而，选取对词簇 W_1 语义贡献度最大的词汇作为词簇中心 w_1^*，即：

$$w_1^*=\max_{w_i\in W_1}O(w_i \mid W_1)$$

词簇中其他词汇即作为词簇中心词汇的连接词汇。

3.6 主题词表的生成

由 3.1 节得到目标领域主题候选词，由 3.3 节得到基于词共现的主题词簇图，由 3.4 节得到每个词簇的中心词汇，从而给出如下主题词表生成算法。

Step1 在主题词簇集 W 中选择任意初始词汇节点，设为 w_0；

Step2 遍历词汇 w_0 所在的词簇，设为 $W_0\subset W$，并计算词簇 W_0 的词簇中心，设为 w_0^*；

Step3 以词汇 w_0^* 为中心词汇，以词簇 W_0 中其他词汇为连接词汇，构建目标领域主题词表 $(w_0^*, O(w_0^*)):\{(w_i, O(w_i)) \mid w_i\in W_0\}$；

Step4 若主题词簇集 $W\backslash W_0=\varnothing$，则退出；否则选择词汇节点，设为 $w_1\neq w_0$，重复 Step2 和 Step3。

4 算例

为验证本文所提方法对目标领域主题词表抽取的效果，选用某时政类报刊上 715 篇新闻报道进行实验。每篇报道不短于 1000 字，并标注其中 223 篇是关于主题“美国海外军事活动”的。

以这 715 篇新闻报道作为样本语料，以标注具有主题属性的 223 篇新闻报道作为领域文本集。抽取出目标领域主题词表共 377 个词汇，29 个词簇，表 1 列出了部分抽取结果。

表 1 算例计算结果

核心词汇	链接词汇
局势(0.18)	危机(0.11) 原则(0.09) 政治(0.07) 事态(0.06) 立场(0.06) …
外交部(0.23)	发言人(0.13) 外长(0.10) 公约(0.07) 应对措施(0.04) 大使(0.02) …
航空母舰(0.3)	护卫舰(0.21) 战机(0.17) 海军基地(0.05) 战斗群(0.03) 武器(0.01) …
水域(0.27)	军事行动(0.15) 主权(0.14) 安全(0.12) 岛链(0.03) 射程(0.03) …
演习(0.4)	联合军事演习(0.33) 作战单位(0.05) 演练项目(0.04) 编队(0.04) 救援(0.03) …
…	…

通过与该领域专家人工抽取结果的比较表明，该词表具有较好的准确率和召回率。

结束语 对特定领域主题词表的提取，一方面能为进一步优化自然语言的处理性能和效果提供基础，另一方面能为本体构建提供概念的构成要素和关系结构[8]。

本文提出一种面向给定语料的领域主题词表提取算法，构建了一种从语料中自动获取主题词表的语言处理模型，实现了对特定专业或细分领域主题词表的统计学习。该方法是一种基于计算的经验主义方法，不受人为主观因素的影响，计算结果具有客观性。

参 考 文 献

[1] 常春，卢文林. 叙词表编制历史、现状与发展[J]. 农业图书情报学刊，2002(5)：25-28.
[2] 肖健，徐建，徐晓兰，等. 英中可比语料库中多词表达自动提取与对齐[J]. 计算机工程与应用，2010，46(31)：130-134.
[3] 陈炯，张永奎. 一种基于词聚类的文本特征描述方法[J]. 计算机系统应用，2011，20(2)：211-215.
[4] 葛宁，王军. 领域 Ontology 的自动丰富——基于 ADL 地名表的实例研究[J]. 计算机科学，2007，34(9)：156-162.
[5] 奉国和，郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作，2011，55(2)：41-45.
[6] SALTON G, CLEMENT T Y. On the construction of effective vocabularies for information retrieval[C]// Proc. of 1973 Meeting on Programming Languages and Information Retrieval. New York, USA: ACM Press, 1973.
[7] 丁国栋，白硕，王斌. 一种基于局部共现的查询扩展方法[J]. 中文信息学报，2006，20(3)：84-91.
[8] 李勇，李苹. 主题词表到领域本体的转化研究[J]. 现代计算机，2013(5)：12-15.