计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160195
53(8):1740-1752,2016

挖掘专利知识实现关键词自动抽取

陈忆群[1,2] 周如旗[2] 朱蔚恒[3] 李梦婷[4] 印 鉴[1]
[1](中山大学计算机科学系 广州 510275)
[2](广东第二师范学院计算机科学系 广州 510303)
[3](暨南大学信息科学技术学院 广州 510632)
[4](珠海魅族科技有限公司 广东珠海 519085)
(chenyiqun@gdei.edu.cn)

Mining Patent Knowledge for Automatic Keyword Extraction

Chen Yiqun[1,2], Zhou Ruqi[2], Zhu Weiheng[3], Li Mengting[4], and Yin Jian[1]
[1](*Department of Computer Science, Sun Yat-sen University, Guangzhou* 510275)
[2](*Department of Computer Science, Guangdong University of Education, Guangzhou* 510303)
[3](*College of Information Science Technology, Jinan University, Guangzhou* 510632)
[4](*Meizu Telecom Equipment Co. Ltd, Zhuhai, Guangdong* 519085)

Abstract Keywords are important clues that can help a user quickly decide whether to skip, to scan, or to read the article. Keyword extraction plays an increasingly crucial role in information retrieval, natural language processing and other several text related researches. This paper addresses the problem of automatic keyword extraction and designs a novel automatic keyword extraction approach making use of patent knowledge. This approach can help computer to learn and understand the document as human being according to its background knowledge, finally pick out keywords automatically. The patent data set is chosen as external knowledge repository because of its huge amount of data, rich content, accurate expression and professional authority. This paper uses patent data set as the external knowledge repository serves for keyword extraction. An algorithm is designed to construct the background knowledge repository based on patent data set, also a method for automatic keyword extraction with novel word features is provided. This paper discusses the characters of patent data, mines the relation between different patent files to construct background knowledge repository for target document, and finally achieves keyword extraction. The related patent files of target document are used to construct background knowledge repository. The information of patent inventors, assignees, citations and classification are used to mining the hidden knowledge and relationship between different patent files. And the related knowledge is imported to extend the background knowledge repository. Novel word features are derived according to the different background knowledge supplied by patent data. The word features reflecting the document's background knowledge offer valuable indications on individual words' importance in the target document. The keyword extraction problem can then be regarded as a classification problem and the

收稿日期:2016-03-21;**修回日期**:2016-05-30
基金项目:**国家自然科学基金项目**(61472453,U1401256,U1501252);**广东省科技计划基金项目**(2012A010701013)
This work was supported by the National Natural Science Foundation of China (61472453, U1401256, U1501252) and the Research Foundation of Science and Technology Plan Project in Guangdong Province (2012A010701013).
通信作者:**印鉴**(issjyin@mail.sysu.edu.cn)

support vector machine (SVM) is used to extract the keywords. Experiments have been done using patent data set and open data set. Experimental results have proved that using these novel word features, the novel approach can achieve superior performance in keyword extraction to other state-of-the-art approaches.

Key words background knowledge; keyword extraction; patent data; support vector machine (SVM); information retrieval

摘 要 关键词是人们快速判断是否要详细阅读文件内容的重要线索,关键词自动抽取在信息检索、自然语言处理等研究领域均有重要应用.设计了一种新的关键词自动抽取方法,使计算机能够像人类专家一样,利用知识库对目标文本进行学习和理解,最终自动抽取出关键词.专利数据因其数据量庞大、内容丰富、表达准确、专业权威而被选中作为知识库来源.详细讨论了专利数据的特性,挖掘不同专利间的知识关联,针对某一知识领域构造背景知识库,在此基础上进行目标文本的关键词自动抽取.与目标文本相关的专利文集中每个专利的专利发明人、权利人、专利引用和分类信息都被用于在不同的专利文档之间发现关联性,利用关联信息扩充背景知识库,获得目标文档在各个相关知识领域的背景知识库.基于背景知识库设计了词知识特征值,以反映词在目标文本背景知识中的重要程度.最后,把关键词抽取问题转化为分类问题,利用支持向量机(support vector machine, SVM)抽取出目标文本的关键词.在专利数据集和开放数据集的实验结果证明明显优于现有算法.

关键词 背景知识;关键词抽取;专利数据;支持向量机;信息检索

中图法分类号 TP391

关键词是快速获取文档主题的重要方式,广泛应用于新闻报道、科技论文等领域,以方便人们高效地管理和检索文档.在大数据时代,每时每刻都有大量信息产生,采用传统的人工方式标注关键词已不再可行.社会迫切需要自动为文档标注关键词的技术,因此,关键词自动抽取已成为自然语言处理和信息检索的研究热点和重点.

传统的关键词人工标注方法依靠人类专家所具备的背景知识,理解目标文本,最终标注出目标文本的关键词.现有的关键词抽取算法多关注于目标文本的数据特征,依靠统计信息进行分析,从中抽取出关键词.但目标文本的信息量有限,其数据特征信息(如词频、结构等)尚不能完整表征目标文本的语义内涵和外延.本文针对关键词抽取问题,为目标文本建立相应的背景知识库,使计算机能够像人类专家一样,根据背景知识去"理解"目标文本,自动抽取出关键词.

人类专家的知识来源于学习与经验积累,其背景知识包括常识、专业知识及相关领域知识等.在接触新的知识范畴时,人类专家能通过查阅相关文献资料,结合自身的背景知识进行有针对性地学习,以便理解并掌握新知识内容.为了使计算机能够像人类一样针对目标文本进行"学习"和"理解",需要为计算机提供相应的背景知识库,提供相关文献资料以便计算机查阅、学习.首先,相关文献资料内容必须正确且具有权威性,保证计算机学到的知识是准确、可靠的;其次,计算机需要从海量信息中获取知识,因此相关文献资料必须数量庞大、内容丰富.随着科学技术的快速发展,世界各国的专利文献数量呈不断上升的趋势.专利文献作为科学技术进步与创新的主要载体,不仅内容丰富、用语标准、阐述清晰、覆盖全面,而且具有相当的权威性,可以很好地充当背景知识的来源,辅助计算机"学习"和"理解"目标文本.美国专利商标局所出版的专利数据集(US patent)[1]以格式良好的 XML 文件方式存储着专利的大量信息,包括专利标题、专利号、专利发明人、专利权利人、专利摘要、专利阐述、专利声明、专利引用、专利分类信息等.海量的专利数据集是一个巨大的知识宝藏.因此,本文选用专利数据集作为相关文献资料充当背景知识来源.

目前国内外尚未有其他研究者利用专利数据集作为外部知识来源构造背景知识库服务于关键词抽取.本文分析、利用专利数据集的数据特征,挖掘专利数据集中隐含的知识关联和链接关系,建立专利数据集构成的知识库内容.针对目标文本在专利数据中发起关联查询,以构造目标文本相应的背景知

识库,包括相关知识库、工作知识库、相关领域知识库、先验知识库和同类知识库,并针对不同的背景知识库定义了新的关键词知识特征值,设计了基于专利知识库的特征值计算公式.新的知识特征值能够有效反映词语在目标文本背景知识中的重要程度,从而反映其在目标文本中的重要程度,最终利用特征值将关键词自动抽取转化为分类问题,有效地解决关键词抽取问题.

1 相关工作

1.1 关键词自动抽取算法

根据关键词自动抽取算法的基本思想可将关键词自动抽取算法分为:基于统计特征的关键词自动抽取、基于主题模型的关键词自动抽取和基于词图模型的关键词自动抽取算法.

基于统计特征的关键词自动抽取算法关注目标文本的词汇基本特征(词频、词的位置、词性和词语长度等),是一种简单易行的常用方法.如扩展的KP-Miner[2]算法首先基于词频和位置标记出候选关键词,使用基于带权重的TF-IDF(term frequency-inverse document frequency)算法计算候选关键词的权重,并设计了1个增强因子用于组合关键词组,最后进一步提炼使最后的关键词列表综合考虑长关键词和短关键词的分布:某个关键词如果出现在另一个组合关键词中,则此关键词的权重被降低.此类方法容易忽略重要的低频词语和文档内部的主题分布语义特征.

基于主题模型的关键词自动抽取算法以基于LDA(latent Dirichlet allocation)的关键词自动抽取算法应用最为广泛[3-5].LDA是一种无监督机器学习技术[6],通过大量已知的"词语-文档"矩阵和一系列训练,推理出隐藏在内部的"文档-主题"分布和"主题-词语"分布,出现在目标文本主要主题中的词语更有可能被识别为关键词.主题模型通过对数据进行训练而得到,关键词抽取的效果与训练数据的主题分布关系密切,因此抽取结果对训练数据集的依赖较大.

基于词图模型的关键词自动抽取算法以TextRank[7]为代表,通过把文本分割成若干组成单元并建立图模型,将目标文本的每一个句子/词视为1个节点,句子/词之间的相似度作为节点之间的边值,利用投票机制对目标文本中的重要节点进行排序.算法认为1个词的重要性由链向它的其他词的重要性来决定.此类算法仅利用单篇文档本身的信息即可实现关键词自动抽取,因其简洁有效而得到了广泛应用.但是,此类算法只利用了文本内部的信息进行关键词自动抽取,没有考虑到文本的背景知识.因此,研究者引入各种知识帮助关键词排序,如考虑文档近邻[8-9]、与文档摘要互相增强补充[10-11]、考虑文档标题的作用[12]等.

此外,近年来涌现大量借助外部数据进行关键词自动抽取的研究工作.研究者提出的引入文档以外的外部数据以辅助关键词自动抽取的算法主要分为2类:1)利用标签(tags)数据.Web2.0网站向用户提供了为感兴趣的对象自由标注标签的功能,这些标签便于用户分享、管理、收藏和检索对象,具有表征意义,因此标签作为一种外部知识可引入到关键词自动抽取中,如Tag-TextRank[13]算法.该方法在TextRank基础上,通过将目标文档的每个标签引入相关文档来估计词项图的边权重并计算得到词项的重要度,最后将不同标签下的词项权重计算结果进行融合.2)引入外部知识的算法主要是利用维基百科(Wikipedia)丰富的百科词条.将每个维基百科词条看作是一个独立的概念(concept),1个词的语义信息可以用维基百科概念上的分布来表示.其中,在某个概念上的权重可以用这个词在该概念词条中的TF-IDF值来表示,这样就可以通过比较2个词的概念向量来度量他们的相似度.Wikify[14]将目标文本的重点概念指向维基百科中的相关页面,通过链接结构得到了关键词的新特征.由此也发展了一些扩展算法,如利用维基百科的结构来构建一个词语之间的语义图以抽取关键词[15].文献[16]不仅考虑了词汇的本身特性,同时引入词汇间的3种关系值,这3种关系值分别来源于文档层、语料库层(与目标文本类似的文档)以及知识层(维基百科)的词汇间相似度计算结果.Maui[17]系统在KEA[18]的基础上扩展了3个基于维基百科的特征值,包括:1)维基百科词条.1个词作为维基百科词条的可能性.2)语义关联.从维基百科计算的词汇语义值.3)链接值.链向维基百科页面的链接数.文献[19]利用维基百科内部链接、外部链接以及目录信息等计算目标文本的特征值,以实现关键词自动抽取的算法.文献[20]提出利用维基百科的文档标题以及目录图来为给定某领域的短文档集找出关键词(这些词不一定出现在文档集中)的方法.研究者对如何利用维基百科做了充分的讨论,并利用维基百科提高关键词抽取质量.

从内容上看,标签数据依赖于网站维护,用户人工标注的方式具有一定的参考价值,但在表达准确性、用词的专业性和权威性方面不能得到有效保证.维基百科数据量大、内容丰富,每天都有来自世界各地的许多参与者进行数百万次的编辑,其数据也对社会产生较大影响,许多研究生的论文和某些媒体甚至会引述维基百科的内容,维基百科的内容得到了广泛认可.但其内容基本上都是由普通用户所撰写,其大部分页面都可以由任何人修改,虽然内容经过管理者审查,但显然其科学权威性和准确性无法与专利文件相媲美.专利文献的每一份材料都是经过科学考察检验的精确描述.用专利数据集作为外部知识来源具有数据庞大、用语科学准确和权威性强等优势,能够为计算机提供内容精准有保障的正确知识.海量的专利数据集是一个巨大的尚未被开发的知识宝藏,因此,本文选用专利数据集作为背景知识来源.

目前国内外尚未有其他研究者利用专利数据集作为外部知识来源构造背景知识库服务于关键词抽取.从内容组织及结构特性方面考虑,标签数据和维基百科数据内容与专利数据内容组织各不相同,具有较大的差异性.基于标签数据和基于维基百科数据的利用外部数据进行关键词抽取的方法对利用专利数据作为背景知识来源可以起到一定的启发作用,但不能沿用作为专利数据的利用方法.因此,本文设计了利用专利数据作为背景知识来源的方法,通过分析专利数据的数据特征,挖掘专利数据中隐含的知识关系,建立了带索引的专利数据集,利用专利数据集为目标文本构造背景知识库.针对关键词自动抽取问题,定义了候选词的5个新知识特征值,对目标文本中的每个词计算其知识特征值和统计特征值,应用于关键词分类器的训练和关键词抽取中.

1.2 专利数据研究

格式良好、内容丰富、数量庞大的专利数据集已在信息检索及自然语言处理方面引起研究兴趣.目前围绕专利数据集开展的工作主要是针对专利文献的翻译、检索和自动分类研究.如CLEF-IP会议[21]提出针对文本的专利在先搜索(prior art candidate search)、专利分类以及基于图形的分类和检索等工作任务;PatentMT[22]会议专注于专利文献自动翻译等工作.

专利数据集包含丰富的信息,如专利文献中的发明人、权利人、专利标题、专利摘要、详细描述、专利引用、专利类别信息等,这些丰富的资料可以作为背景知识库帮助理解目标文本.图1展示US patent数据集中专利号为S-08621694-B2的专利文献信息(因篇幅所限,只展示部分信息,整个专利文档共651行).

Fig. 1 Example of patent file from US patent.

图1 US patent中的专利文献文档示例

专利文件以XML格式存储,专利标题存储在〈invention-title〉标签中,〈abstract〉标签存储专利摘要,〈description〉标签内容存储专利描述,包括专利申明(claim)等.

专利发明人信息存储在〈applicants〉标签中,1个专利可能不止1个发明人,因此通过sequence属性表明发明人排序情况,如第1位为001.发明人信息包括发明人姓名、通信地址等详细内容.

专利权利人信息存储在〈assignees〉标签中,当权利人为公司时,公司名存储在〈orgname〉中,当权利人为独立个人时,存储在〈first-name〉〈last-name〉标签中.

每份专利文件在撰写过程中可能引用现有的专利文件;在提交专利申请时,专利文件内容经过审查,包括检查其内容与之前已发布的专利是否存在重复.〈references-cited〉包括了在该专利审查过程中引用到的所有专利文件,从〈country〉标签和〈doc-number〉标签内容可以组合出引用的专利文件号.

每份专利文件都列出了本专利的分类信息.其专利分类有国际分类号、美国分类号等,本文从

〈classification-ipcr〉标签获取每个专利的国际分类号作为分类信息依据. 1 份专利文件可能同属多个分类号.

庞大的专利数据中隐藏着隐含的知识关联,每一份专利文件不应该是一个信息孤岛,而应该是庞大的专利数据网中的一个节点. 如何为专利文件建立链接,将众多专利文件组成一个背景知识网,使得从 1 个知识点就可以找到相关联的其他知识点,这是本文重点考虑的问题. 通过分析发现,在专利数据中有 4 个隐含的链接关系. 1)专利发明人链接. 在不同的专利文件,如果发明人姓名、通信地址信息完全一致,则认可是同一位发明人. 对专利数据集中具有相同发明人的专利文件建立关联链接以便发现关联信息. 2)专利权利人链接. 在不同的专利文件中,如果专利权利人/组织的姓名/组织名和通信地址完全一致,则认可是同一权利人/组织,本文方法为具有相同权利人/组织的专利文件建立了关联链接. 3)引用链接. 一份专利的撰写和审核过程中,需要引用已有的专利文件,这种引用关系体现了知识的发展和关联,因此,对专利文件间的引用关系建立了关联链接. 4)类别链接. 每份专利都有自己的分类号,可能同时属于多种类别,这种类别属性也体现了知识领域跨度信息. 因此,为同一个分类号下所有专利建立链接关系. 显然,这些链接关系将能成为专利数据集的知识索引,通过建立这 4 种链接关系,可以将孤立的专利文件组织成为一个庞大的知识网,为相关语义信息检索提供强有力的支持.

2 挖掘专利知识库实现关键词自动抽取

为表述清楚,本文提出如下定义:

1) 目标文本. 待抽取关键词的文本信息.

2) 背景数据集. 用于产生背景知识库的数据集,本文采用 US patent 数据集作为背景数据集.

3) 背景知识库. 针对目标文本,在背景数据集中搜索相关内容,构造背景知识库. 每个目标文本都有自己不同层次的背景知识库.

4) 查询词. 从目标文本中获得的代表目标文本主要内容及方向的词.

本文设计的基于专利知识的关键词自动抽取算法如算法 1 所示.

算法 1. 基于专利知识的关键词自动抽取算法.

输入:目标文本;

输出:目标文本的关键词.

步骤 1. 产生查询词:对目标文本使用改进的 TextRank 算法,获取目标文本的查询词.

步骤 2. 构造背景知识库:利用查询词在专利数据集中检索,将相关度较高的专利文件集构成相关文集. 对相关文集合中的每一份专利文件,抽取其专利发明人、权利人、专利引用和分类信息,建立与相关文件的关联关系,抽取专利文件的标题和摘要构造不同的背景知识库,包括相关知识库 PAI、工作知识库 IF、相关领域知识库 AS、先验知识库 CI、同类知识库 CL(详细算法见算法 2).

步骤 3. 利用背景知识库,计算词的知识特征值及其他文本特征值.

步骤 4. 把关键词自动抽取转化为词的分类问题,利用词特征值和训练数据集训练分类器,使用分类器对目标文本词汇进行分类,将属于关键词类别的词作为算法结果输出.

步骤 5. 对判断为关键词的词列表进行检查,将在原文中相邻的词组合成复合词作为关键词.

本文方法首先对专利数据集进行数据预处理. 专利数据集中每个专利文件以 XML 文件格式独立存放,不具备链接关系. 1 名专利发明人可能具有多个发明专利,这几个发明专利因共同的发明人而具有了链接关系. 1 名专利权利人可能拥有多项专利权利,这些专利也因共同的权利人而产生链接关系. 1 个专利对其他专利的引用也产生了链接关系. 同一个分类属下的不同专利之间也有链接关系.

因此,对专利数据中的专利发明人、专利权利人、引用及类别关系,系统为其分别构建链接关系,建立索引,形成带索引的背景数据集,以便于提高后续工作中的相关知识库的构造效率.

对每一份目标文本,系统先为其找出查询词,利用查询词在背景数据集中进行检索,找出关联度较高的相关专利文件构成背景数据集. 进一步利用背景数据集中的专利发明人链接、专利权利人链接、引用链接及类别链接,为目标文本构造背景知识库包括相关知识库、工作知识库、相关领域知识库、先验知识库、同类知识库. 利用背景知识库为目标文本的每个词计算词知识特征值,再结合其他文本特征值得到目标文本中每个词的 10 个特征值,这 10 个特征值综合目标文本的背景知识库内容,反映了该词汇的语义特征和在目标文本中的重要程度. 最后,将关键词抽取转化为分类问题,10 个词特征值服务于支持向量机的训练和关键词抽取工作. 基于专利知识的关键词自动抽取过程如图 2 所示:

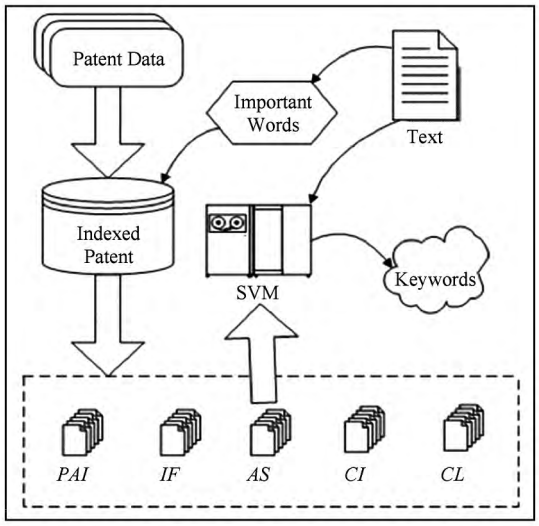


Fig. 2 The framework of our approach.

图 2 系统流程

2.1 产生查询词

为了从专利数据集中获取相关的背景知识,本文将从专利数据中检索出与目标文本相关的专利文件作为背景知识来源,用于构造背景知识库.为检索出相关文件,需确定查询词.用查询词检索出的相关文件如果和目标文本相关度不高,则背景知识库与目标文本的相关度不高,对理解目标文本帮助不大,甚至容易产生误导.因此,查询词具有重要意义,需代表目标文本的主要内容和方向.本文使用改进的基于词图模型的关键词自动抽取算法来构造查询词.

本文构造查询词的算法与经典的词图模型 TextRank 算法主要不同之处在于相似度衡量方法.在 TextRank 中,句子间的相似度由 2 个句子中单词的重叠度(overlap)衡量.本文定义了句子间相似度新的衡量方法.

对 2 个句子 C 和 D,将句子 C 中包含的词(记为 C_i)与句子 D 中包含的词(记为 D_j)进行两两对比,计算语义相似度来得到句子 C 和 D 之间的相似度,记为 $sem\_sim(C,D)$.

$$sem\_sim(C,D)=\left(\sum_{i=1}^{m}\frac{c_i}{m}+\sum_{j=1}^{n}\frac{d_j}{n}\right)\Big/2, \quad (1)$$

其中,c_i 为词 C_i 与句子 D 中每个词之间相似度的最大值(见式(2)).同理,d_j 为词 D_j 与句子 C 中每个词之间相似度的最大值(见式(3)).

$$c_i=\max(s(C_i,D_1),s(C_i,D_2),\cdots,s(C_i,D_n)), \quad (2)$$

$$d_i=\max(s(D_j,C_1),s(D_j,C_2),\cdots,s(D_j,C_n)), \quad (3)$$

其中,$s(C_i,D_j)$ 表示词 C_i 和词 D_j 间的语义相似度.通过先利用 QTAG[23] 判断词 C_i 和词 D_j 的词性,再利用 wordNet[24] 提供的语义工具 Similarity,可以获取词 C_i 和词 D_j 在指定词性下的语义相似度,从而得到 $s(C_i,D_j)$ 的值.

前期工作实验证实,从语义角度衡量句子间的相似度比从词的拼写重合度来衡量句子相似度更加准确,从而能更好的选出目标文本中的查询词[25].

2.2 构造背景知识库

查询词代表目标文本的主要内容和主题方向.本文利用这些查询词,在专利数据集中搜索与目标文本相关度较高的专利文件.这些专利文件内容与目标文本紧密相关,将有利于帮助理解目标文本的语义信息,作为目标文本的相关知识,称为相关知识库.

从人类知识的角度,一位计算机科学工作人员拥有的专业背景知识显然与文学工作人员不同,很多在计算机科学领域视为常识的内容显然文学工作者并不具备,而 2 位工作人员日常相关工作涉及的内容也各不相同.因此,为了让计算机像人类一样,能够获得专业领域的专业知识、工作上常接触的知识、相近领域的相关知识、先验知识、同类领域中的知识等内容,本文挖掘专利数据集中隐含的知识关联.专利发明人、权利人一般会专注密切相关的业务领域.通常,专利发明人所发明的不同专利不会跨越多个不同领域,本文将发明人的所有发明专利中的知识归纳为工作知识;专利权利人所处理的专利领域通常跨度也不会太大,一般具有辅助或合作关系,因此,将专利权利人所代理的专利知识归类为相关领域知识;而专利的引用信息表示该专利说明中引用的其他专利,是该专利的先验知识.与该专利具有同一分类号的专利文件则讨论了该专利文件的同类领域知识,与该专利属于同类知识.因此,以上内容包含了目标文本的相关背景知识,有利于掌握目标文本对应的领域知识,帮助理解目标文本.本文将使用以上知识关联,抽取相关专利文集信息构造背景知识库.本文设计的背景知识库构造方法如算法 2 所述.

算法 2. 背景知识库构造算法.

输入:目标文本的查询词;

输出:目标文本的背景知识库.

步骤 1.用查询词在专利数据集中进行关键词检索,将相关度较高专利文件组成相关文集,记为 PAI,称为相关知识库. PAI 中的第 r 份文件记为 p_r,其查询相关度记为 $z(p_r)$.

步骤 2. 对每个专利文件 p_r 抽取专利发明人信息,并通过背景数据集搜索得到每位发明人的发明专利文件集合,将每位发明人的相关文件集合并作去重处理,得到文件 p_r 的发明人相关文集,记为 $IF(p_r)$,称 IF 为工作知识库.

步骤 3. 对每个专利文件 p_r 提取专利权利人信息,并通过背景数据集得到权利人的其他专利文件集合,最后将该专利的每位权利人的相关文集合并作去重处理,得到文件 p_r 的专利权利人相关文集,记为 $AS(p_r)$,称 AS 为相关领域知识库.

步骤 4. 对每个专利文件 p_r 提取专利引用信息. 将每个引用的专利号在背景数据集中进行搜索,最终得到文件 p_r 的引用专利文集,记为 $CI(p_r)$,称 CI 为先验知识库.

步骤 5. 对每个专利文件 p_r 提取专利国际分类号信息,检索与此专利具有同一分类号的专利文件,组成同类文集,记为 $CL(p_r)$,称 CL 为同类知识库.

步骤 6. 对以上 5 个知识库,抽取库中全部专利文件的标题和摘要去掉停用词后的单词来构建知识库内容分别为 $LT_{PAI}(p_r)$, $LT_{IF}(p_r)$, $LT_{AS}(p_r)$, $LT_{CI}(p_r)$, $LT_{CL}(p_r)$构成了整个背景知识库.

注意在算法 2 中,通过检索出与目标文本紧密相关的专利文件构成相关知识库,进一步抽取每个相关专利的发明人、权利人、引用专利和专利分类号信息,检索出属于该专利发明人的其他专利文件集构成工作知识库,属于该专利的每位专利权利人的其他专利文件构成相关领域知识库,将该专利引用的其他前沿专利文件构成先验知识库,将具有该专利相同分类号的其他专利文件构成同类知识库,这些知识库内容都是围绕与目标文本内容紧密相关的专利而构成的,都有利于帮助理解目标文本,因此可用来构成目标文本的背景知识库. 由于每一份专利文件的标题和摘要都尽量简明扼要地描述专利内容,标题和摘要中含有的词汇具有重要表征意义. 因此,将专利文件标题和摘要抽取出来作为背景知识库的内容,使得背景知识库包括了帮助理解目标文本内容的关键词汇.

2.3 特征值的计算

本文对目标文本中的每个词,计算其在背景知识库中的重要程度,称为知识特征值. 在背景知识库的构造过程中,本文利用针对目标文本产生的查询词在专利数据集中检索出与目标文本密切相关的专利文件,从中提取专利的发明人、权利人、专利引用和同类专利相关信息,抽取专利的标题和摘要构成了目标文本的各种背景知识库. 相关知识库 PAI 与通过挖掘 PAI 中专利文件隐含的信息而得到的工作知识库 IF、相关领域知识库 AS、先验知识库 CI、同类知识库 CL 的相关性各不同,重要程度也各不相同,不能等同视之. 因此,针对不同的知识库,本文分别设计了知识特征值计算公式,计算目标文件中每个词汇在不同知识库中词汇的相似度. 目标文本中的词与知识库的词语义相似度越高,则其特征取值越高,说明它在此知识库中越重要.

本文对目标文本中每个词 x_i 计算 10 个特征值. 其中,前 5 种知识特征值是利用各种背景知识库进行计算的结果,其他文本特征值在目标文本内部进行计算,结合各个特征值进行分类器的训练和使用,辅助关键词的判断.

1) 相关知识特征值

针对相关知识库 PAI,取出 PAI 中每份专利文件 p_r,计算统计出目标文本中每个词 x_i 的相关知识特征值 f_{PAI}:

$$f_{PAI}(x_i)=\frac{\sum\limits_{p_r\in PAI}\left[z(p_r)\sum\limits_{k\in LT_{PAI}(p_r)}\sigma_1(x_i,k)\right]}{\sum\limits_{p_r\in PAI}z(p_r)\,|LT_{PAI}(p_r)|},\quad(4)$$

其中,PAI 是用查询词从专利数据中检索得到的相关文件集,p_r 是 PAI 中的每个专利文件,$LT_{PAI}(p_r)$ 是目标文本的相关知识库内容,即 p_r 的标题和摘要去除停用词后的词汇总;$|LT_{PAI}(p_r)|$ 是 $LT_{PAI}(p_r)$ 包含的词语总数;k 是 $LT_{PAI}(p_r)$ 中的每个词;$\sigma_1(x_i,k)$是词 x_i 和 k 的语义相似度,在确定词性后使用 WordNet 计算(计算方法见 2.1 节介绍). $z(p_r)$是每个相关文件 p_r 的查询相关度,作为权重,调节语义相似度值. 在本文实验中利用开源搜索引擎使用查询词,检索专利数据集中的相关文件. 搜索引擎对返回的搜索结果文件列表中的每个文件 p_r,都会给出该文件与查询词的相关度值,作为 $z(p_r)$ 的值.

知识特征值 f_{PAI} 计算词 x_i 与相关知识库 $LT_{PAI}(p_r)$ 的知识关联度,词 x_i 与相关知识库中的词语义相似度越高,则其特征取值越高. LT_{PAI} 是与目标文本紧密相关的专利文集构成的相关知识库,基于统计学理论:在同一个训练集中频繁共同出现的词语会在同一个领域的其他文档中共同出现. 因此,目标文本的关键词、关键词的同义词、近义词在相关知识库中必然有着与其他非关键词不同的出现频率和特征,关键词的 f_{PAI} 值将与非关键词有不同的特征,这反映了一种特征,可用于辅助关键词判断.

2) 工作知识特征值

$$f_{IF}(x_i)=\frac{\sum_{p_r\in PAI}\left[z(p_r)\sum_{k\in LT_{IF}(p_r)}\sigma_1(x_i,k)\right]}{\sum_{p_r\in PAI}z(p_r)\,|LT_{IF}(p_r)|},\quad(5)$$

式(5)计算目标文本中每个词 x_i 与工作知识库 $LT_{IF}(p_r)$的知识关联度 f_{IF}. 其中,$LT_{IF}(p_r)$是目标文本的工作知识库内容,即 p_r 的所有发明人的其他发明专利的标题和摘要去除停用词后的词汇总;k 是 $LT_{IF}(p_r)$中的每个词;$\sigma_1(x_i,k)$是词语 x_i 和 k 的语义相似度. $|LT_{IF}(p_r)|$是工作知识库 $LT_{IF}(p_r)$中的词个数. 注意一个专利文件 p_r 有 1 个或多个发明人,每个发明人至少有 1 个发明专利 p_r 或多个发明专利,工作知识库 $LT_{IF}(p_r)$中记录的是每个发明人的所有发明文件合并去重后的结果. 从而工作知识库 $LT_{IF}(p_r)$的内容会比 PAI 更多,提供了对 PAI 的补充信息. 通过式(5),可以得到目标文本中每个词 x_i 的工作知识特征值,代表其在工作知识库中的语义关联度,作为一种特征值辅助进行关键词判断.

3) 相关领域知识特征值

$$f_{AS}(x_i)=\frac{\sum_{p_r\in PAI}\left[z(p_r)\sum_{k\in LT_{AS}(p_r)}\sigma_1(x_i,k)\right]}{\sum_{p_r\in PAI}z(p_r)\,|LT_{AS}(p_r)|},\quad(6)$$

同理,式(6)中的各个变量与式(5)计算方法相同. 其中,$LT_{IF}(p_r)$是目标文本的相关领域知识库内容,$|LT_{AS}(p_r)|$是相关领域知识库 $LT_{AS}(p_r)$中的词个数. 计算目标文本中每个词 x_i 与相关领域知识库 $LT_{AS}(p_r)$的知识关联度 f_{AS}. 对相关文集中的每份专利文件 p_r,通过找出此专利的专利权利人的其他专利文件,取得文件标题和摘要来获取其相关领域知识,结合专利 p_r 的相关度,计算 x_i 与相关领域知识库中每个词的语义相关度,统计得到其知识关联度. 与相关领域知识库中的词语义近似度越高的词其知识关联度 f_{AS}值越高.

4) 先验知识特征值

$$f_{CI}(x_i)=\frac{\sum_{p_r\in PAI}\left[z(p_r)\sum_{k\in LT_{CI}(p_r)}\sigma_1(x_i,k)\right]}{\sum_{p_r\in PAI}z(p_r)\,|LT_{CI}(p_r)|},\quad(7)$$

f_{CI}计算目标文本中每个词 x_i 与先验知识库 $LT_{CI}(p_r)$的知识关联度,见式(7). 同理,$|LT_{CI}(p_r)|$是先验知识库 $LT_{CI}(p_r)$中的词个数. 对相关文集中的每份专利文件 p_r,通过找出此专利的引用专利文件,取其文件标题和摘要的词来组成先验知识库,结合专利 p_r 的相关度,计算 x_i 与先验知识库中每个词的语义相关度,统计得到其知识关联度. 与先验知识库中的词语义近似度越高的词其知识关联度 f_{CI}值越高.

5) 同类知识特征值

$$f_{CL}(x_i)=\frac{\sum_{p_r\in PAI}\left[z(p_r)\sum_{k\in LT_{CL}(p_r)}\sigma_1(x_i,k)\right]}{\sum_{p_r\in PAI}z(p_r)\,|LT_{CL}(p_r)|},\quad(8)$$

同理,可以使用同类知识库 $LT_{CL}(p_r)$计算目标文本中每一个词 x_i 在同类知识库 CL 的知识关联度 f_{CL}. 根据统计学理论,关键词在同类知识库中的知识关联度将与非关键词有所不同,因此,这些知识特征值可以协助进行关键词分类判断.

6) TF-IDF

TF-IDF 是信息检索领域中的一种统计方法,用以评估一个词语对于文本集合中某个特定文本的重要程度,其值为 $TFIDF$. 计算公式[26]如下:

$$tf=\frac{\text{该文档中该词出现次数}}{\text{该文档的词总数}},\quad(9)$$

$$idf=-\lg\frac{\text{训练集中出现该词的文档数}}{\text{训练集中的文档总数}},\quad(10)$$

$$TFIDF=tf\times idf.\quad(11)$$

7) 词的平均位置 WAP (word average position)[26]

对于特定位置上的一个词语,计算词语的位置:

$$pos=\frac{\text{在该词之前出现的词总数}}{\text{该文档的词总数}}.\quad(12)$$

每个词语在文本中可能出现不止一次,对该词语每次所出现的位置求均值,以获得词语的平均位置,见式(13).

$$WAP=\frac{\text{在该词之前出现的词总数}}{\text{该文档的词总数}}.\quad(13)$$

8) 特殊名字(specific name, SN)

记录某词语是否指代了特殊的人名或地名. 若有指代,则 SN 为 1,否则为 0.

9) 单词长度(word length, WL).

10) 单词是否出现在总结性句子中(con-clusion sentence, CS)

若词语所在的句子包含了总结性的单词或短语(如"in summary","in conclusion","finally"等),其总结特征 CS 则为 l,否则为 0.

2.4 样本训练及关键词抽取

在计算词特征值之后,本文应用机器学习方法来抽取关键词,分成 2 个步骤:样本训练及抽取.

本文将关键词的抽取视为分类问题,假设待处理的目标文本中共有 n 个词,这些词分成 2 类:关键词和非关键词.对每个词计算以上 10 个维度的特征值,当前每个词的类标号尚未确定(若为关键词,则类标号为 1;否则类标号为 0),见表 1 所示.从词语多个维度上的特征属性映射到关键词类别(是关键词/非关键词),利用分类模型对全部词语进行分类,即完成对关键词的抽取.

Table 1 Keyword with 10 Features Value

表 1 带有 10 个维度特征的关键词

Word	Feature 1	…	Feature 10	Keyword
w_1	v_{11}	…	v_{110}	0/1
w_2	v_{21}	…	v_{210}	0/1
⋮	⋮		⋮	⋮
w_n	v_{n1}	…	v_{n10}	0/1

建立在统计学理论和结构风险最小化原则基础上提出的支持向量机(support vector machine, SVM)能够根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力(或称泛化能力).显然,支持向量机与其他学习机相比,具有良好的推广能力,在处理非线性识别和小样本学习方面具有良好的特性,可以很好地应用基于多特征的分类,适用于本文工作.本文应用 LibSVM[27] 开源模式识别软件包进行实验实现.采用支持向量机分类器对文本进行关键词抽取时,分类器的核函数和参数设置对分类效果有很大影响.其中对分类器性能影响最大的参数为支持向量机核函数参数 γ 及支持向量机分类惩罚因子 C.在支持向量机的实际应用中,最为常用的参数调优方法为"网格法",即对每个参数设置多个值,对每个参数对(γ, C)均进行 1 次分类,进而选取分类效果最好的参数对作为实际分类中的参数值.我们在前期工作(见文献[25])详细讨论了支持向量机在关键词自动抽取工作中的使用情况,经过论证实验,设置分类惩罚因子 $C=256$,采用了 LibSVM 提供的径向基核函数并设置 $\gamma=16$ 能获得较好效果:

$$\kappa(x_i,x_j)=\ell^{-\gamma\|x_i-x_j\|^2},\gamma>0. \tag{14}$$

样本训练过程如下:样本数据包括文本集和每个文本对应的关键词列表.每份文本作为 1 份目标文本,其中包含的非停用词及其 10 个特征值,以及该词是否为关键词将成为 1 份训练数据.对目标文本的每个非停用词 w_i,按照上文所述算法 1 和算法 2,计算在专利数据背景知识库中的知识特征值及其他文本特征值共 10 个特征值(见表 1 所示),结合 10 个特征值和是否为关键词标记(0 表示非关键词,1 表示关键词)来训练分类器.

得到分类器后,利用分类器可实现对目标文本的关键词抽取.首先对目标文本中的每个非停用词按上文所述方法计算其 10 个特征值,应用分类器对这些词进行分类.将归到关键词类的词进行进一步的检查工作:当 2 个被标记为关键词的词在目标文本中相邻时,则这 2 个候选关键词被合成为 1 个关键词.例如,"grid"和"comput"这 2 个词都被归为关键词类,在目标文本中这 2 个词相邻出现过,因此,"grid comput"被合成作为 1 个关键词.注意在处理过程中,所有的词都被抽取为词干以简化计算.

3 实 验

3.1 数据集

专利数据集作为背景知识数据集:本文下载了 US patent 数据集 10 年(2003 至 2013 年)的数据作为背景数据集.在实验中,本文算法 2 步骤 1 采用开源搜索引擎 Lucene[28] 对专利文集进行检索.Lucene 提供基于关键词匹配查询功能,可以按相关度排序返回检索到的相关文件列表.在本文实验中使用查询词,检索专利数据集中的相关文件,对返回的文件列表中的每个文件 p_r,Lucene 都会返回该文件与查询词的相关度值,记为查询相关度 $z(p_r)$.

本文实验提取前 50 篇专利文件构造相关知识库,并保存每篇专利文件的相关度 $z(p_r)$.

专利文集:我们从专利数据集文件中随机选取了 3 000 份独立专利文献(属于同一分类)进行关键词抽取.对于每份专利文献,待抽取关键词的目标文本为专利文献中的摘要(abstract).专利文献标题(headline)简明扼要的体现专利文件内容,标题中的词比较可能成为关键词.但是,标题能容纳的词量有限,而且某些情况下标题为了吸引眼球,其中的词不一定代表文件最主要内容.因此,我们邀请了 30 名硕士生,对这 3 000 份文本进行人工标注关键词.每份文件由 4 名学生标注,要求标注 5~10 个关键词.另外,将标题去除停用词后剩下的词作为第 5 份答案.如果 1 个词在这 5 份答案中获得 3 份以上的认同,则将此词标记为关键词答案.通过这个方法整理

出来的3 000份文本的关键词平均个数为每份文本6.7个.在实验中对比了不同训练样本数据对测试结果的影响.

SemEval[29]数据集:SemEval数据集是关键词抽取领域的开放数据集,含有244份科技文件,每份有6～8页,内容涵盖计算机网络通信(computer-communication networks)、信息储存与检索(information storage and retrieval)、人工智能(artificial intelligence)和计算机应用(computer applications)4个ACM分类的研究领域.数据集提供了3份关键词答案,1份是文件作者自己提供的关键词,1份是人工标注的关键词,1份是前面2个结果的融合,本文采取了作者和人工标注的关键词融合作为标准答案.平均75%的关键词由人工标注得到,25%是文件原作者提供.在本文实验中,此数据集被分出144篇(含2 265关键词)用于训练而剩下的100篇(含1 443关键词)用于测试.

3.2 评价方法

目前国内外尚未有其他研究者利用专利数据集作为外部知识来源构造背景知识库服务于关键词抽取的工作发表.文献[17]使用SemEval数据集实验评估了当前最新的关键词抽取算法和商用系统,结果证明目前最新关键词抽取算法和系统中,性能最佳的3个方法为Alchemy keyword(Alch_key)[30],KP-Miner[2],Maui[17].其中Alch_key是商用系统AlchemyAPI提供的文本分析服务接口,如实体提取、语义分析和文本分类等.KP-Miner是基于统计特征的关键词自动抽取,Maui是使用维基百科作为背景知识的有效代表性方法,其工作原理在相关工作中已进行了讨论.这3个系统返回的关键词列表带有相关度权值,当以权值排序选取前15个为最终结果时得到的实验结果最佳.为了评估本文算法的有效性,将本文算法与这3个系统在SemEval数据集上进行实验对比,见实验2.

本文采用分类问题中较为流行的准确率(precision, P)(式(15))、召回率(recall, R)(式(16))及综合评价指标(F1-score, F)(式(17))对关键词的抽取性能进行评估.

$$P=\frac{\text{实验得到的准确结果数量}}{\text{实验得到的结果数量}}\times 100\%,\quad(15)$$

$$R=\frac{\text{实验得到的准确结果数量}}{\text{全部准确结果数量}}\times 100\%,\quad(16)$$

$$F=\frac{2PR}{P+R}.\quad(17)$$

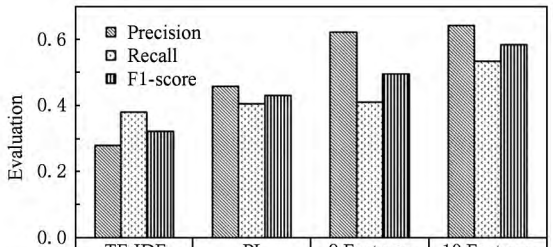
3.3 实验分析

实验1.验证背景知识库的有效性:使用不同背景知识库在专利数据集上进行的对比实验.

本文算法利用专利数据集作为背景知识库来源,构造了背景知识库,利用查询词在专利数据集中首先获得了相关知识库PAI,从PAI中的每一份专利文件,根据专利文件间隐含的知识关联和链接关系,进一步得到其他知识库(工作知识库IF、相关领域知识库AS、先验知识库CI和同类知识库CL)生成了新的知识特征值,包括相关知识特征值、工作知识特征值、同类知识特征值、相关领域知识特征值及先验知识特征值,并结合词频、词长、词的位置等多种特征进行词的分类以抽取关键词.

背景知识库的引入增加了关键词抽取的计算量.特别是同类专利文集的引入产生了同类知识库,大大增加了背景知识库的内容,同类知识特征值较大地增加了算法的计算量.为验证背景知识库的必要性,本文实验比较了只使用传统统计数据进行关键词抽取的结果(简称为TF-IDF);只使用相关知识库PAI进行实验(简称PI);只使用相关知识库PAI和从相关知识库PAI中衍生得到的工作知识库IF、相关领域知识库AS、先验知识库CI加上统计特征值计算的9种特征值的关键词抽取结果(简称为9 Features);使用全部10种特征值进行关键词抽取效果(简称为10 Features).

在实验1中,本文使用200份已标注好答案的专利文件作为训练数据,以其他100份文件作为测试数据.对每一份文件,首先构造查询词,然后使用查询词在专利数据文集中搜索相关专利文集,按照算法2方法构造背景知识库,并计算目标文件中非停用词的在专利数据集上的特征值和常规统计特征值,使用支持向量机利用专利文集中已经具有人工标注答案的200份训练文件及其词特征值和关键词标记进行训练,得到分类器.最终利用分类器对目标文本所包含的词结合其特征值进行分类,从而找出目标文件的关键词.实验结果如图3所示,本文提出的知识特征值能够较好的反映词在目标文本背景知识中的重要程度,与单纯使用传统统计数据的方法(TF-IDF)相比,当加入相关知识库PAI时,其分类效果有非常突出的改进,得到较佳的关键词抽取结果.从PAI衍生的知识库能提供有效的知识补充,提升分类效果;同类知识库得到的同类知识特征值能有效帮助关键词抽取工作,其关键词抽取结果明显优于前面几个方法.



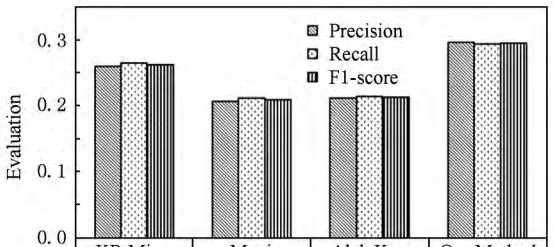
	TF-IDF	PI	9 Features	10 Features
Precision	0.280	0.458	0.621	0.643
Recall	0.381	0.404	0.411	0.534
F1-score	0.323	0.429	0.495	0.583

Fig. 3 Experiments on 9 features and 10 features.

图 3 9 个特征值和 10 个特征值的实验

实验 2. 验证关键词抽取算法的有效性:在 SemEval 数据集上将本文算法与其他现有先进系统进行对比试验.

在 SemEval 数据集上,本文算法的执行过程如下:对 SemEval 数据集上每一份目标文件,首先构造查询词.然后使用查询词在专利数据文集中搜索相关专利文集,按照算法 2 方法构造目标文件的背景知识库,并按照 3.3 节的方法计算获得目标文件中非停用词在专利数据集上的 5 个知识特征值和 5 个常规统计特征值,最后使用支持向量机利用 SemEval 数据集 144 份测试文件的关键词答案和文件中各非停用词 10 个特征值进行分类器训练,得到分类器.利用分类器对 SemEval 数据集的 100 份测试文件的词结合词特征值进行分类,从而找出关键词.图 4 展示了本文算法与其他 3 种当前最优算法 Alch_key[30], KP-Miner[2], Maui[17] 在 SemEval 数据集上实验的结果对比.可以看到,本文基于背景知识库的关键词抽取算法能够有效提高准确率、召回率及综合评价指标.

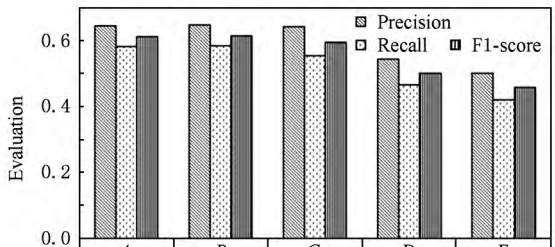


	KP-Miner	Maui	Alch Key	Our Method
Precision	0.258 7	0.206 7	0.210 8	0.295 7
Recall	0.264 7	0.211 5	0.213 5	0.293 4
F1-score	0.261 6	0.209 0	0.212 1	0.290 0

Fig. 4 Comparison with the state-of-the-art systems.

图 4 与其他现有先进系统的对比实验

实验 3. 验证关键词抽取算法在大规模数据中的有效性:为验证算法在大规模数据中的有效性,方法 A 在专利数据集中采用 2 000 份专利文件及关键词标记,计算专利文件中每个非停用词的 10 个特征值为样本数据,训练分类器,之后再利用分类器将剩下的 1 000 份专利文件作为目标文件进行关键词抽取. 方法 B 调整训练样本大小为 1 500 份专利文件进行分类器训练,同样对前面方法 A 的 1 000 份目标专利文件做关键词抽取. 方法 C 调整训练样本大小为 1 200 份专利文件进行分类器训练,同样对前面方法 A 的 15 000 份目标专利文件做关键词抽取. 方法 D 调整训练样本大小为 1 000 份专利文件进行分类器训练,同样对前面方法 A 的 1 000 份目标专利文件做关键词抽取. 方法 E 调整训练样本大小为 500 份专利文件进行分类器训练,同样对前面方法 A 的 1 000 份目标专利文件做关键词抽取. 实验结果如图 5 所示,可以看到,本文算法在大规模数据中同样能得到较好的结果.而且,训练样本数据在达到一定 1 500 份时已经能取得较好的效果,不需要再增加.

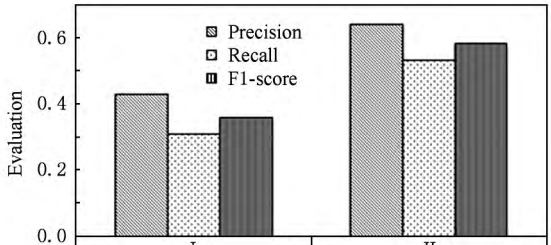


	A	B	C	D	E
Precision	0.645	0.647	0.643	0.543	0.501
Recall	0.581	0.583	0.553	0.465	0.421
F1-score	0.611	0.613	0.595	0.501	0.458

Fig. 5 Experiments on mass data and different training data.

图 5 大数据集及样本数量实验

实验 4. 自动化训练实验:现有科学文献关键词的标准答案都以人类专家标注的方法获得,例如前文使用的关键词抽取领域的公开测试集 SemEval 数据集,关键词答案由文章作者和人类专家标注组成.为提高自动化,本文尝试用专利文件标题中含有的非停用词作为关键词答案,自动构成训练样本.方法 I 随机抽取了 200 份专利文件,以其标题所含非停用词为关键词答案构成训练数据,对另外 100 份专利数据作为测试数据.按照算法 1 的方法获得分类器后,对测试数据文本进行关键词抽取.测试文本仍以人工标注答案为正确答案.实验结果与实验 1 中 10 个特征值(10 Features)的实验结果进行对比,如图 6 所示.



	I	II
Precision	0.430	0.643
Recall	0.310	0.534
F1-score	0.360	0.583

Fig. 6 Experiments on automatic training.

图 6 自动化训练实验

正如前文所述,文本标题可以简明扼要地体现文件内容,标题中的词比较可能成为关键词.但是,标题能容纳的词量有限,而且某些情况下标题为了吸引读者,其中的词不一定代表文件最主要内容和方向.因此,以标题内容为标准答案构造训练数据的做法得到的结果较差.

4 结束语

随着信息时代的飞速发展,在海量数据资料面前,人们需要更快地明确接触到的信息是否符合自己的兴趣范围.关键词是人们进行判断的重要线索.在大数据时代,传统的人工标注关键词的方法已不再可行,因此自动准确地从文本中抽取关键词成为一个非常重要的任务.关键词自动抽取在信息检索、文本挖掘和自然语言处理等各领域具有重要应用.

为了使计算机能够像人类一样针对目标文本,利用相关文献资料查阅出相关知识,对目标文本的内容进行"学习"和"理解",最终标记出关键词.本文首次提出利用专利数据集获取目标文本背景知识的关键词抽取方法.专利数据集因数据庞大、内容准确、科学权威而被选作外部知识来源.本文分析了专利数据集的语义和结构特征,提出了利用专利数据集为目标文本构造背景知识库的方法,为目标文本构建了包括相关知识库、工作知识库、同类知识库、相关领域知识库及先验知识库等背景知识库,定义了表征词汇特征的知识特征值及计算公式,定义了基于背景知识库的关键词抽取算法.海量、准确、严谨、权威的专利信息为目标文本提供权威可靠的背景知识库,新的知识特征值能反映词语在目标文本的背景知识库中的重要程度,很好地补充了基于目标文本计算的传统词特征值,使计算机能够像人类专家一样,根据背景知识,对目标文本进行"理解",从而标注出关键词.

本文提出专利数据集作为背景知识来源的利用方法不仅可用于关键词抽取工作,在信息检索及自然语言处理的其他相关工作中也可进一步使用.另外,在专利文献数据集中,除了本文使用到的专利标题、专利摘要、专利发明人、权利人、分类号、专利引用信息及专利分类信息外,还有大量信息,如专利发表时间、专利权利声明、专利图片等尚未加以利用.这些专利描述信息也具有非常大的价值,在今后的工作中,若利用得当,相信可以更好地提高关键词抽取效率.因此,在未来的工作中将进一步考虑以上 2 方面的工作内容.

参 考 文 献

[1] The United States: Patent and Trademark Office. Patent Grant Full Text [DB/OL]. [2012-02-03]. http://www.google.com/googlebooks/uspto-patents-grants-text.html

[2] El-Beltagy S R, Rafea A. Kp-miner: A keyphrase extraction system for English and ? Arabic documents [J]. Information Systems, 2009, 34(1): 132-144

[3] Claude P. Task5: Single document keyphrase extraction using sentence clustering and latent Dirichlet alocation [C] // Proc of the 5th Int Workshop on Semantic Evaluation. Stroudsburg, CA: Association for Computational Linguistics, 2010: 154-157

[4] Shi Jing, Li Wanlong. Topic words extraction method based on LDA model [J]. Computer Enginering, 2010, 36(19): 81-83 (in Chinese)
(石晶, 李万龙. 基于 LDA 模型的主题词抽取方法[J]. 计算机工程, 2010, 36(19): 81-83)

[5] Liu Jun, Zou Dongsheng, Xing Xinlai, et al. Keyphrase extraction based on topic feature [J]. Application Research of Computers, 2012, 29(11): 4224-4227 (in Chinese)
(刘俊, 邹东升, 邢欣来, 等. 基于主题特征的关键词抽取[J]. 计算机应用研究, 2012, 29(11): 4224-4227)

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet alocation [J]. Journal of Machine Learning Research, 2003, 1(3): 993-1022

[7] Mihalcea R, Tarau P. Textrank: Bringing order into texts [C] //Proc of Conf on Empirical Methods in Natural Language. Piscataway, NJ: IEEE, 2004: 404-411

[8] Wan Xiaojun, Xiao Jianguo. CollabRank: Towards a collaborative approach to single-document keyphrase extraction [C] //Proc of IEEE COLING'08. Piscataway, NJ: IEEE, 2008: 969-976

[9] Wan Xiaojun, Xiao Jianguo. Single document keyphrase extraction using neighborhood knowledge [C] //Proc of IEEE AAAI'08. Piscataway, NJ: IEEE, 2008: 855-860

[10] Zha H. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering [C] //Proc of IEEE SIGIR'02. Piscataway, NJ: IEEE, 2002: 113-120

[11] Wan Xiaojun, Yang Jianwu, Xiao Jianguo. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction [C] //Proc of IEEE ACL'07. Piscataway, NJ: IEEE, 2007: 552-559

[12] Li Decong, Li Sujian, Li Wenjie. A semi-supervised key phrase extraction approach: Learning from title phrases through a document semantic network [C] //Proc of IEEE ACL'10. Piscataway, NJ: IEEE, 2010: 296-300

[13] Li Peng, Wang Bin, Shi Zhiwei, et al. Tag-TextRank: A webpage keyword extraction method based on tags [J]. Journal of Computer Research and Development, 2012, 49 (11): 2344-2351 (in Chinese)
(李鹏, 王斌, 石志伟, 等. Tag-TextRank: 一种基于Tag的网页关键词抽取方法[J]. 计算机研究与发展, 2012, 49 (11): 2344-2351)

[14] Mihalcea R, Csomai A. Wikify!: Linking documents to encyclopedic knowledge [C] //Proc of ACM IKM'07. New York: ACM, 2007: 233-242

[15] Grineva M, Lizorkin D. Extracting key terms from noisy and multitheme documents [C] //Proc of ACM WWW'09. New York: ACM, 2009: 661-670

[16] Zhang Wei, Feng Wei, Wang Jianyong. Integrating semantic relatedness and words intrinsic features for keyword extraction [C] //Proc of IEEE IJCAI'13. Piscataway, NJ: IEEE, 2013: 2225-2231

[17] Louis J. An assessment of online semantic annotators for the keyword extraction task [G] //LNAI 8862: Proc of ICAI. Berlin: Springer, 2014: 548-560

[18] Medelyan O, Frank E, Witten I H. Human-competitive tagging using automatic keyphrase extraction [C] //Proc of IEEE EMNLP'09. Piscataway, NJ: IEEE, 2009: 1318-1327

[19] Xu Songhua, Yang Shaohui, Lau F C M. Keyword extraction and headline generation using novel word features [C] //Proc of IEEE AAAI. Piscataway, NJ: IEEE, 2010: 1461-1466

[20] Qureshi M, O'Riordan C, Pasi G. Short-text domain specific key terms/phrases extraction using an *n*-gram model with Wikipedia [C] //Proc of ACM IKM'12. New York: ACM, 2012: 2515-2518

[21] IFS. CLEF-IP [EB/OL]. [2010-03-01]. http://www.ifs.tuwien.ac.at/~clef-ip

[22] NTCIR. PatentMT [EB/OL]. [2010-03-01]. http://ntcir.nii.ac.jp/PatentMT/

[23] University of Birmingham. QTag [CP/OL]. [2010-03-01]. http://web.bham.ac.uk/O.Mason/software/tagger/

[24] Miller G A. Wordnet: A lexical database for english [J]. Communications of the ACM, 1995, 38(11): 39-41

[25] Chen Yiqun, Yin Jian, Zhu Weiheng. Novel word features for keyword extraction [G] //LNCS 9098: Proc of the 16th Int Conf on Web-Age Information Management. Berlin: Springer, 2015: 148-160

[26] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2010

[27] Chang Chih-Chung, Lin Chih-Jen. Libsvm—A library for support vector machines [J]. ACM Trans on Intelligent Systems & Technology, 2011, 2(3): 389-396

[28] Apache. Lucene [CP/OL]. [2010-03-01]. http://lucene.apache.org/

[29] Kim S N, Medelyan O, Kan M Y, et al. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles [C] //Proc of the 5th Int Workshop on Semantic Evaluation. Stroudsburg, CA: Association for Computational Linguistics, 2010: 21-26

[30] Alchemy. Alchemyapi [CP/OL]. [2010-03-01]. http://www.alchemyapi.com/api/keyword-extraction

Chen Yiqun, born in 1979. PhD candidate, associate professor. Her main research interests include information retrieval, data mining and artificial intelligent.

Zhou Ruqi, born in 1971. PhD candidate, associate professor. His main research interests include machine learning and artificial intelligent.

Zhu Weiheng, born in 1976. PhD, lecturer. His main research interests include data mining and information retrieval.

Li Mengting, born in 1988. PhD. Her main research interests include data mining and artificial intelligent.

Yin Jian, born in 1968. PhD, professor and PhD supervisor. His main research interests include data mining and artificial intelligent.