一种专利智能推荐算法设计与软件实现①

辛 阳[1], 文益民[1], 曾德森[1], 彭文乐[1], 申孟杰[1], 刘文华[2]

[1](桂林电子科技大学 计算机科学与工程学院, 桂林 541004)
[2](湖南工学院 计算机与信息科学系, 衡阳 421002)

摘 要: 针对科技研发人员从事创新活动而需要频繁检索专利的需求, 以及当今专利检索智能程度不高的现状, 提出一种专利智能推荐算法并开发了相应的软件. 算法的输入是用户输入的检索词, 输出结果中不仅包括检索系统输出的专利还包括一批推荐的专利. 本算法首先实现专利间的关联, 进而计算专利关联度, 并根据关联度对推荐专利进行排序, 构成一个有序的推荐专利集合. 实验表明推荐的专利与检索词之间的确存在关联.

关键词: 专利推荐; 关联度; 专利检索; 信息过载; 信息检索

Design and Implementation of an Intelligent Recommendation Algorithm for Patents

XIN Yang[1], WEN Yi-Min[1], ZENG De-Sen[1], PENG Wen-Le[1], SHEN Meng-Jie[1], LIU Wen-Hua[2]

[1](School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)
[2](Computer and Information Science Department, Hunan Institute of Technology, Hengyang 421002, China)

Abstract: For the demand of scientific and technological researchers engaged in innovative activities which need frequently retrieving patents from database, and the intelligent degree of the current patent databases is not enough high for retrieving, this paper proposed an intelligent recommendation algorithm and introduced the development of the corresponding software. Inputted with the retrieval words, the algorithm outputs the results including not only patents which the retrieval system provides but also a group of recommended patents. The algorithm first explores the correlation among the patents, and then calculates the correlation degree among them, finally sorts all the recommended patents according to the correlation degree in order to form an orderly recommended patents set. Experimental results illustrates that there indeed exits correlation between the recommended patents and the retrieval word.

Key words: patent recommendation; correlation degree; patents search; information overload; information retrieval

随着科技的迅速发展和经济的全球化, 专利的作用越来越得到人们的重视. 如今, 从某种程度上说, 国家之间的竞争等同于科学技术之间的竞争. 进入 21 世纪以来, 专利信息增长尤为迅速. 我国每年公布的专利说明书也呈快速增长趋势. 根据国家统计局的数据——2010 年受理国内外专利申请 122.2 万件, 2011 年受理国内外专利申请 163.3 万件. 专利信息的如此快速增长带来了信息超载, 即科技研发人员从海量的专利信息里寻找感兴趣的专利将成为一件不轻松的工作. 专利推荐算法作为一种信息过滤的重要手段, 是解决专利信息超载的一种重要的、有潜力的方法.

目前主流的推荐算法主要包含以下几大类[1]: 基于内容的推荐, 协同过滤的推荐, 基于知识的推荐和组合推荐. 仲伟炜通过跟踪和记录用户的访问操作行为, 分析专利查询者经常查阅的专利文献, 利用关联规则来分析专利文献的相关性, 以实现专利文献的个性化推荐[2]. 该算法本质上属于协同过滤推荐, 需要跟踪大量用户的专利检索行为, 所推荐专利是一群专利用户的共同兴趣. 而对于科技研发人员来所, 经常需要检索与本身研究目的相关的专利. 通过专利检索, 了解当前研究现状, 同时拓展研究思路. 在这种情况下文献[2]中提出的算法将变得不再适应, 而本文提出

① 基金项目:国家大学生创新性实验立项项目(101059513);湖南省科技计划(2010GK3047,2011GK3150)
收稿时间:2012-06-07;收到修改稿时间:2012-07-23

算法则适合该问题的解决.

在我国, 较有权威和影响力的专利检索网络平台包括: 中国国家知识产权局网站、中国知识产权网、中国专利网、中国专利信息网、Soopat 专利搜索和 Patentics[3]以及中国期刊网. 这七大专利检索平台采用的检索形式与传统信息检索类似, 采用字段检索, 输入检索词或按照“*” (与)、“+”(或)、“一”(非)等组成字段内或字段间逻辑关系式. 根据以上专利检索网络平台的这个特点, 本文提出的算法向用户推荐专利标题以及摘要中不包含检索词, 但其在内容上又和检索词存在一定语义关联的专利. 本文提出的算法属于基于内容的推荐算法[4,5]. 算法基于向量空间模型, 建立了专利间语义关联模型, 通过计算专利关联度, 构成有序的专利推荐集合.

1 专利智能推荐算法

1.1 概念定义

为描述算法方便起见, 先定义以下三个概念:

1) 用来描述专利集合 C 中全部专利内容的 K 维向量称为目标专利特征向量, 记为 TF_C. TF_C 的构造方法是: 首先根据检索词获得一个专利集合, 称为, 提取 C 中各篇专利的摘要; 然后对各摘要实施分词; 过滤掉量词和副词等词语后, 留下名词和动词两类词语; 然后统计各个词语出现的总频率; 按词频从高到低排序, 取前 K 个词语对应的词频, 构成 TF_C, 这 K 个词语构成的集合定义为词表 V. 在本文中取 K=10.

2) 用于描述第 i 篇专利的内容的 K 维向量被称为专利特征向量, 记为 F_i^V. F_i^V 的构造方法是: 提取第 i 篇专利的摘要; 对其进行中文分词, 过滤掉量词和副词等词语, 留下名词和动词两类词语; 然后统计各个词语的词频, 根据词表中词语的顺序, 定义一个 K 维向量. 若词表中的某词不在分词结果中, 则填入 0, 否则填入该词的词频.

3) 第 i 篇专利在内容上与专利集合 C 的目标专利特征向量 TF_C 相近的程度被称为第 i 篇专利与专利集合 C 的关联度, 记为 $S_{(F_i^V \cdot TF_C)}$. 关联度的计算公式如下:

$$S_{(F_i^V \cdot TF_C)} = \frac{F_i^V \cdot TF_C}{|F_i^V| \times |TF_C|} \quad (1)$$

1.2 专利智能推荐算法描述

输入: 专利在线检索引擎平台的网址、检索词、向量空间维数 K;

输出: 推荐专利的有序集合 C_r.

①定义推荐专利集合 $C_r = \Phi$, Φ 表示空集;

②根据用户输入的检索词通过某个专利在线检索引擎平台检索得到专利集合 C;

③将专利集合 C 中各专利的标题进行中文分词, 过滤掉量词、副词等语义表达能力不强的词语, 留下动词和名词构成检索词集合 W;

④利用检索词集合 W 中的各个词, 通过专利在线检索引擎平台再进行检索, 得到专利集合 $\tilde{C}$. 定义备选推荐专利集合 $\overline{C} = \tilde{C} - C$;

⑤针对专利集合 C, 提取词表 V 和目标专利特征向量 TF_C;

⑥针对专利集合 $\overline{C}$, 提取 $\overline{C}$ 中第 i 篇专利的专利特征向量 F_i^V, $1 \le i \le Num(\overline{C})$, $Num(\overline{C})$ 表示取集合 $\overline{C}$ 的元素个数. 根据(1)式计算第篇专利与专利集合 C 的关联度 $S_{(F_i^V \cdot TF_C)}$, 并从集合中删除第 i 篇专利, 将该专利加入到推荐专利集合 C_r 中, 重复步骤 6 直至集合 $\overline{C}$ 为空止;

⑦根据关联度大小将 C_r 中各专利排序后输出. 关联度最大的专利排在最前面.

2 软件实现

本论文中使用 VC6.0 对提出的算法进行了实现. 整个软件包括三个主要模块: 专利信息获取模块、中文分词模块、专利推荐模块.

2.1 专利信息获取模块

在本文中数据源是 SooPAT 专利网站(www.soopat.com). 专利信息获取模块执行的流程图如图 1 所示. 该模块需要调用如下三个函数:

CString CPatentsRecommDlg:: FormatSearchWord (CString csAnsiString);

bool CPatentsRecommDlg:: GetHtmlByUrl(CString csServerUrl, CString &csContent, CString &csDescript);

CString CPatentsRecommDlg:: Search(CString csSearchWord, CString &csSearchResult, int ResultNum, CString csExistNames, bool bRecomm).

函数 FormatSearchWord 主要是把人工输入的检索词转换成专利在线检索引擎平台能够识别的格式. 参数 csAnsiString 表示输入的检索词.

函数 GetHtmlByUrl 实现向专利在线检索引擎平

台发送获取信息请求和接受获取的数据. 参数 csSearchWord 表示专利在线检索引擎平台的 URL 地址; 参数 csContent 表示网页内容; 参数 csDescript 表示状态描述字符串.

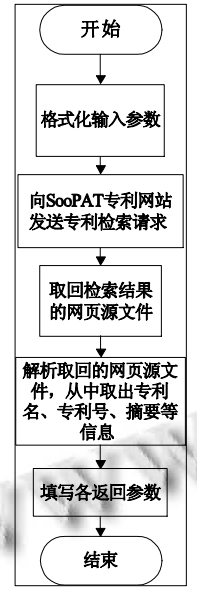

图 1　专利信息获取模块流程图

函数 Search 用于从专利在线检索引擎平台获取专利网页文件, 并网页源文件中提取专利信息. 各参数的含义如下: csSearchWord 为用户输入的检索词; csSearchResult 用于表示完整的检索结果, 包含检索到的所有专利的名称、专利号、摘要和对应的网页链接; ResultNum 参数用于控制从专利在线检索引擎平台检索的专利的数量, 以便控制算法的反应时间; 由于在推荐专利时也用到这个函数, 用 csExistNames 这个参数来输入先前已检索出的专利名, 以避免重复推荐; bRecomm 用来区分这个函数是在直接根据检索词检索时调用, 还是在推荐专利时调用. 函数返回一个包含检索结果中所有专利名的字符串.

2.2 中文分词模块

中文分词模块使用的是中国科学院计算技术研究所汉语词法分析系统 ICTCLAS, 其官方网站(http://ictclas.org/)有提供可使用的链接库下载. 中文分词模块的执行流程图如图 2 所示. 该模块需要调用如下五个函数:

ICTCLAS_Init();

ICTCLAS_GetParagraphProcessAWordCount();

ICTCLAS_ParagraphProcessAW();

ICTCLAS_KeyWord();

ICTCLAS_Exit();

第一个函数完成中文分词系统初始化工作; 第二个函数计算输入字符串分词后的词语的数量; 第三个函数实施分词; 第四个函数将分词结果存放到特定的结构体数组中以便使用; 第五个函数终止中文分词系统, 释放资源.

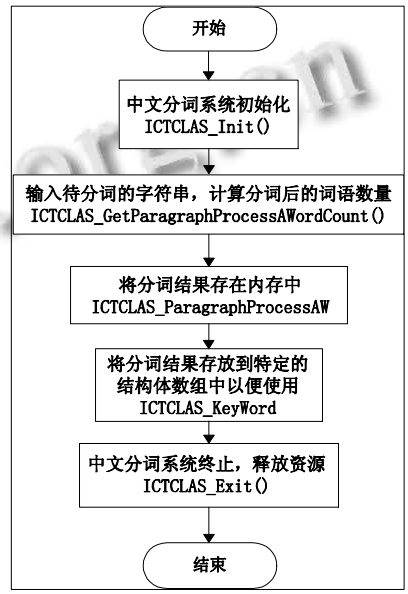

图 2　中文分词模块流程图

2.3 专利推荐模块

专利推荐模块是本软件中最重要的一个模块, 执行的流程图如图 3 所示, 其主要调用了如下函数:

CString CPatentsRecommDlg:: Search(CString csSearchWord, CString &csSearchResult, int ResultNum, CString csExistNames, bool bRecomm);

CString CPatentsRecommDlg:: GetAbstract(char* szAbstDir, CString& csAbstName, CString& csAbstNum);

double CPatentsRecommDlg:: ChooseRelated (double iaVector[], double &iResult);

函数 Search 被调用多次, 用于从专利在线检索引擎平台获取专利网页文件, 并网页源文件中提取专利信息, 其参数已经在前文描述.

函数 GetAbstract 用于获取某个专利的摘要文本. 参数 szAbstDir 表示专利信息的存储地址; csAbstName 和 csAbstNum 用于返回专利名和专利号; 函数返回值是一个字符串, 就是摘要文本;

函数 ChooseRelated 计算专利特征向量与目标专利特征向量之间的关联度. 参数 iaVector 表示目标专利特征向量; 参数 iResult 表示专利特征向量.

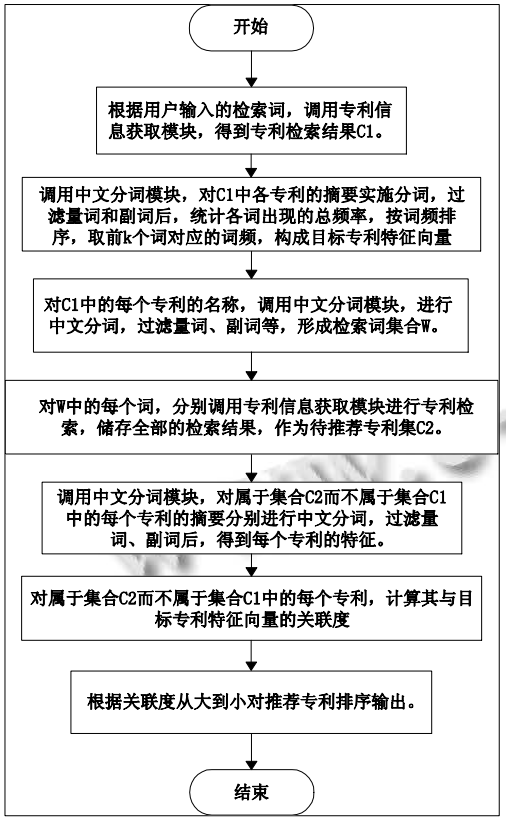

图 3 专利推荐模块流程图

3 实验结果验证与分析

为了检验本论文提出的算法的合理性和设计的软件的可用性, 本论文设计了如下实验——实验进行 3 次, 选择的检索词分别为: 海水发电、地沟油检测、高空逃生. 在实验中, 检索词的选择并没有特别的要求, 只是考虑了使用该检索词能检索到一定数量的专利. 使用上述 3 个检索词于 2011 年 11 月 8 日 21: :30 通过 Soopat 专利搜索平台检索到的专利数量分别为: 28、26、36.

实验结果提供了软件推荐的前五个关联度最高的专利, 软件推荐的结果如表 1 所示. 由于篇幅的限制, 本实验没有提供使用这 3 个检索词通过 Soopat 专利搜索平台检索到的专利的详细情况. 但是, 通过比较软件推荐的专利的名字和对应的检索词可以看出: 软件推荐的专利与检索词之间的确存在关联; 推荐的这些专利的确能拓展科技开发人员实施专利检索的思路.

由于到目前为止还没有发现同类型的专利推荐算法, 因此没有就本文提出的专利推荐算法与其他专利推荐算法进行比较实验.

表 1 使用各检索词时的前四个专利推荐结果

检索词	推荐专利名称及发明(申请)号	推荐专利与查询的关联度
海水发电	发电机 200810010273.9	0.663
	浮力引擎 200780051333.5	0.596
	发电机 200810210448.0	0.592
	发电机组 200480031294.9	0.590
地沟油检测	品质原因量度显示 01116490.5	0.683
	方法 00809815.8	0.598
	地沟盖板 201020144436.5	0.126
	地沟转窑 03218630.4	0.126
高空逃生	高层救生器 200410046374.3	0.425
	高层救生伞 200910131455.6	0.400
	城际轻轨飞列 200910117128.5	0.181
	消防车 95213391.1	0.130

4 结语

本文提出了一种专利智能推荐算法. 算法首先实现专利间的关联, 进而计算专利关联度, 并根据关联度对推荐专利进行排序, 构成一个有序的推荐专利集合. 实验表明该算法在专利推荐上具有一定效果. 由于本算法在提取专利文本特征向量上只考虑了词频这一特性, 所以该算法还存在着诸多改进之处. 专利推荐的精确度在很大程度上取决于描述检索目标的目标专利特征向量和描述专利的专利特征向量的精确度, 这方面的改进将是以后的研究方向.

参考文献

1 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究.软件学报,2009,20(2):350−362.
2 仲伟炜.专利文献分类及关联推荐技术应用研究[硕士学位论文].南京:南京航空航天大学,2009.
3 胡晓,魏雪梅.我国网络专利检索平台分析和评价.科技管理研究,2010,(14):75−81.
4 吴良杰,刘红祥,张立垄,况振东.个性化服务中网页推荐模型的研究.计算机应用研究,2005,(6):83−84.
5 李容.基于 K 均值聚类算法的图书商品推荐仿真系统.计算机仿真,2010,27(6):346−349.