

基于权利要求结构信息的中文专利无效检索模型^{*}

刘玉琴, 汪雪锋, 吕 琳

(北京理工大学 管理与经济学院, 北京 100081)

摘 要: 中文专利独立权利要求分为前序部分和特征部分。文中构建的专利无效检索模型, 充分考虑了这一信息, 从专利数据库中统计出 40 个分割词对独立权利要求进行分割处理。具体检索中采用两步检索: 第一步进行布尔检索以提高召回率; 第二步对申请专利与第一步返回专利独立权利要求的前序部分和特征部分分别进行相似度计算, 适当组合后作为整体的相似度。实验中对分割前后以及分割后不同的词语权重选择方法对检索效果的影响作了比较, 结果显示该模型是非常有效的。

关键词: 中文专利; 专利检索; 无效检索; 权利要求; 相似性

中图分类号: TP301 文献标志码: A 文章编号: 1001-3695(2008)07-2068-03

Invalidity search model for Chinese patent based on claim structure information

LIU Yu-qin, WANG Xue-feng, LV Lin

(School of Management & Economics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Chinese patent independent claim contains a preamble portion and a characterizing portion. Invalidity search model for Chinese patent proposed in the paper draws on the structure information. Forty split words were extracted from patent database artificially; these words could divide independent claims into preamble portion and characterizing portion effectively and automatically. For it was impossible to compute similarity on the whole database two-step search method was used in practice: at 1-step Boolean query was applied to improve recall, at 2-step vector space model was used to compute similarities of preamble portion and characterizing portion between applying patent (query) and previous patents (documents) obtained at 1-step respectively, and then combined them properly to sort the search results in order to improve precision. Experiment data set comes from SIPO; search results with split claims are contrasted with that without them; different methods of term-weighting are compared. Evaluation results show that the model works well.

Key words: Chinese patent; patent search; invalidity search; claim; similarity

伴随着经济全球化竞争愈演愈烈, 商业模式已经有所改变。许多公司尤其是国际型大公司采用获取、利用、管理知识产权的战略。专利信息作为知识产权战略中的关键因素, 已经得到越来越多国家、企业的重视, 他们努力保护自己的知识产权, 争夺每一项可能给他们带来利润的专利技术。

在这种环境下, 专利的重要性日趋明显。当进行产品开发前、专利申请前或判断一个已经申请的专利的有效性时, 一项非常重要的工作就是搜索专利数据库寻找相关专利。这些相关专利很有可能使某项专利发明无效。通常称这种目的的专利检索为专利无效检索。然而对于专利发明者、申请者、专利审查员来说, 在庞大的数据库中找出相关专利并不容易。网络上有许多免费专利检索系统, 这些系统大多使用布尔型检索进行简单的匹配, 即没有采用有效的检索算法, 又没有考虑到专利文本的结构特征, 检索效果低下。因此, 有必要设计一个准确、高效的检索模型, 本文的研究致力于解决相关的问题。

1 专利文本特点与专利检索研究现状

1.1 专利文本特点

专利是一种具有法律效力的半结构化文本, 含有大量的结构化与非结构化信息。中国专利申请包含如下信息: 申请时

间、申请号、授权时间、授权号、专利类型、国际分类号、申请人、发明人、题目、摘要、权利要求等三十余项字段。摘要和权利要求是典型的非结构化信息, 使用了许多技术术语和新名词, 甚至有些还含有图片。另外, 专利权人为扩大权利范围, 经常使用一些概念模糊的词语, 使得专利更加难以理解。一般情况下, 进行专利检索也是依据这两个字段, 同时配以结构化信息对检索范围加以限定。

1.2 专利检索研究现状

商业专利检索系统已经存在很长时间了, 但是相关研究直到近些年才得到信息检索和自然语言理解研究人员的重视, 相继召开了有关国际会议(SIGIR 2000^[1]、ACL 2003^[2]、NTCIR-3 2002^[3]、NTCIR-4 2004^[4]、NTCIR-5 2005^[5])。这些研究主要针对英文、日文、韩文专利进行的, 部分涉及到中文的也只是将外国专利翻译成中文, 并不是真正意义上的中文专利。检索的任务多样化包含有专利分类、技术调查、无效检索、跨语言专利检索、专利地图绘制等。本文的研究是针对中文专利无效检索进行的。

无效检索的目的是找出与某一专利权利要求相关的专利, 通过这些专利使该权利要求无效, 甚至使整个专利无效。它是一种专利对专利的检索方式, 通常由知识产权部门专利审查员

收稿日期: 2007-07-02; 修回日期: 2007-09-18 基金项目: 国家自然科学基金重点资助项目(70031010); 国家“985”工程经费资助项目
作者简介: 刘玉琴(1979-), 男, 博士研究生, 主要研究方向为数据挖掘、科技管理、专利检索(liuyuqin2004@126.com); 汪雪锋(1976-), 男, 博士研究生, 主要研究方向为数据挖掘、科技管理; 吕琳(1976-), 女, 博士后, 主要研究方向为数据挖掘、科技管理。

进行。在检索中采用时间进行检索限制可以达到技术新颖性和专利侵权性检索。因此, 随着企业对专利的重视, 企业内部也逐渐开始进行相关的检索, 以达到实施自己专利战略的目的。L. S. Larkey 等人在 USPTO 支助下采用分布式检索方法设计一个专利检索系统^[6]。该系统同时能够对专利进行分类处理。A. Shinmori、J. H. Kim 等人介绍了在日文专利深加工的基础上进行相关性检索的方法^[7, 8]。实验证明该方法非常有效。

2 中文专利权利要求的结构特征^[9]

我国专利法规定一项发明或实用新型应当只有一项独立权利要求, 并且写在同一发明或实用新型的从属权利要求之前。独立权利要求从整体上反映发明或实用新型的技术方案, 记载解决技术问题的必要技术特征。从属权利要求: 如果一项权利要求包含了另一项同类型权利要求中的所有技术特征, 且对该另一项权利要求的技术方案作了进一步的限定, 则该权利要求为从属权利要求。

- 独立权利要求撰写时应当包括前序部分和特征部分。
- 1) 前序部分 写明要求保护的发明或实用新型技术方案的主题名称和发明或实用新型主题与最接近的现有技术共有的必要技术特征。
- 2) 特征部分 使用“其特征是……”或类似的用语, 写明发明或实用新型区别于最接近的现有技术的技术特征。这些特征和前序部分写明的特征合在一起, 限定发明或实用新型要求保护的范围。

本文设计的检索模型充分地利用了权利要求的结构特征, 因此为了使叙述更加清晰, 定义如下:

定义 1 权利要求分割词。中文发明或实用新型专利权利要求中能够将前序部分和特征部分分割开的惯用词语, 如“其特征是”“其特点”“方法为”等。

独立权利要求分两部分撰写的目的, 在于使公众更清楚地看出独立权利要求的全部技术特征中哪些是发明或实用新型与最接近的现有技术所共有的技术特征, 哪些是发明或实用新型区别于最接近的现有技术的特征。当然, 某些情况下独立权利要求也可以不分前序部分和特征部分。例如开拓性发明、由几个状态等同的已知技术整体组合而成的发明、已知方法的改进发明等。

3 中文专利无效检索模型

3.1 确定权利要求分割词

现有的统计方法与机器学习理论可以用来为专利检索服务。但专利文本是一种特殊的半结构化文本, 有其自身的特点, 专利检索中应该考虑这些特征。在专利无效检索中要求较高的准确率, 如果仅仅是用权利要求或摘要进行布尔匹配或使用简单的向量空间模型进行相似性计算, 对结果进行排序, 均难以获得满意的结果。因此, 结合中文专利权利要求的特点, 设计下面的检索模型。基本思想就是: 布尔检索与向量空间模型相结合提高召回率和准确率; 前序部分与特征部分分别处理进一步提高准确率。

对 2000 年申请的发明专利中具有独立权利要求的 55 866 件专利进行人工观察, 统计出常用的 40 个分割词, 将它们分为三类, 即特征类、组成类和过程类, 如表 1 所示。这些分割词

能有效地将权利要求前序部分与特征部分分割开, 分割率约为 94. 6%。

表 1 中文专利权利要求分割词

特征类	组成类	过程类
特点在于; 其 特征; 其特点; 其特性; 特征 是; 特征为; 特 征	其复合组分; 括; 包含; 含有; 成是;; 组成;; 组成; 其组分; 下组分; 成分为;; 成分:	主要由; 是由; 包有; 包 过程为;; 工艺步骤为;; 工艺为;; 步骤;; 如下方 法; 如下配方; 生产方 法; 方法是;; 方法为;; 方法;; 方法;; 制备工艺

3.2 无效检索模型

两步检索在信息检索领域广泛应用, 因此依据已经确定的分割词, 设计如下两步无效检索模型: a) 进行布尔初检, 尽可能扩大检索的范围提高检索召回率; b) 对输入独立权利要求和 a) 返回的专利独立权利要求进行分割处理, 分别计算前序部分相似度和特征部分相似度, 根据两部分的相关性最终获得返回结果的相似度, 从而完成结果的排序工作。这样既节省了计算时间, 又提高了检索的准确性, 因为在整个专利数据库中进行排序计算是不现实的。相应的流程如图 1 所示。

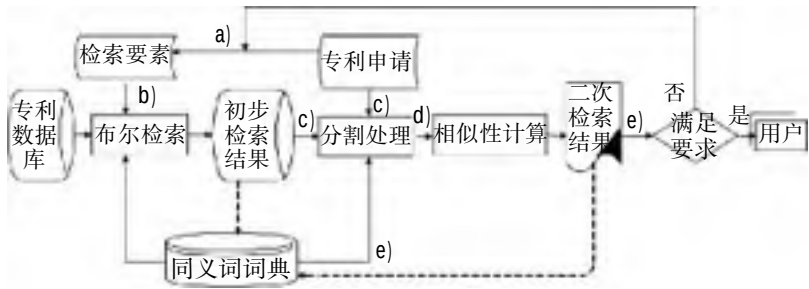


图 1 中文专利无效检索模型

a) 确定检索要素进行布尔检索。对新申请的专利进行分析, 根据独立权利要求确定检索要素后, 利用同义词词典进行扩展, 确定检索关键词。基本检索要素可以根据技术领域、技术问题、技术手段、技术效果等方面进行确定。

b) 利用关键词在专利数据库中进行布尔检索, 获得初步检索结果并临时存储。

c) 分割处理。对申请专利的独立权利要求和 a) 返回结果的权利要求进行分割处理。

d) 相似性计算。对初步检索结果采用向量空间模型进行相似性计算。计算方法如下:

$$\text{sim}(D_i, Q_i) = \alpha \times \text{sim}(D_i^p, Q_i^p) + (1 - \alpha) \text{sim}(D_i^c, Q_i^c); 0 \leq \alpha \leq 1 \quad (1)$$

其中: $D_i^p = (td_{i1}^p, td_{i2}^p, \dots, td_{in}^p)$, $D_i^c = (td_{i1}^c, td_{i2}^c, \dots, td_{in}^c)$, $D_i = (td_{i1}, td_{i2}, \dots, td_{in})$, 表示返回结果专利独立权利要求的前序部分、特征部分和整个独立权利要求, td_{ij} 、 td_{ij}^p 、 td_{ij}^c 表示返回结果中相应部分关键词权重。

同样地, $Q_i^p = (tq_{i1}^p, tq_{i2}^p, \dots, tq_{in}^p)$, $Q_i^c = (tq_{i1}^c, tq_{i2}^c, \dots, tq_{in}^c)$, $Q_i = (tq_{i1}, tq_{i2}, \dots, tq_{in})$, 表示申请专利独立权利要求前序部分、特征部分和整个独立权利要求, tq_{ij} 、 tq_{ij}^p 、 tq_{ij}^c 表示相应部分关键词的权重。 $\text{sim}(D_i^p, Q_i^p)$ 、 $\text{sim}(D_i^c, Q_i^c)$ 采用夹角余弦:

$$\text{sim}(D_i^p, Q_i^p) = \left(\sum_{i=1}^n td_{ij}^p \times tq_{ij}^p \right) / \sqrt{\sum_{i=1}^n td_{ij}^{p2} \sum_{i=1}^n tq_{ik}^{p2}} \quad (2)$$

$$\text{sim}(D_i^c, Q_i^c) = \left(\sum_{i=1}^n td_{ij}^c \times tq_{ij}^c \right) / \sqrt{\sum_{i=1}^n td_{ij}^{c2} \sum_{i=1}^n tq_{ik}^{c2}} \quad (3)$$

关键词权重的计算, 对排序结果有影响, 因此有必要对不同的权重计算方法进行比较, 择优而定。

e) 人工观察检索结果是否达到检索要求, 如果没达到, 返回 a) 重新确定检索要素; 否则, 结束检索。该步同时进行同义词词典的扩充: 阅读检索结果, 找出明显的与检索要素同义的词语加到同义词典中。

4 实验与结果评价

4.1 实验数据与实验设计

为了验证该模型的有效性,对比不同词语权重方法的效果进行两组实验:第一组对比权利要求分割前后检索效果的不同,用“分割前”表示;第二组是在分割处理前提下对比词语权重的选择对检索效果的影响,权重选择方法一为 $f_{q,t} \times f_{d,t} \times idf_t$,用“分割后 A”表示,二为 $b_{q,t} \times b_{d,t}$,用“分割后 B”表示。其中: $f_{x,t}$ 表示 x 中词 t 出现的频度; idf_t 表示词 t 的倒排文档频度; $b_{x,t}$ 表示 x 中词 t 存在(1)与否(0)。

为此,随机选取五件专业性不强的发明专利作为申请专利:95118 × ×1, 02121 × ×7, 200410096 × ×7, 200510000 × ×7, 200610049 × ×4, 分别用 NO. 1 ~5 表示,确定其检索要素并进行同义词扩展。被检索专利为申请时间在 1985. 01. 01 ~ 2006.2. 28 内的所有发明专利。初步检索返回结果分别为 2087、628、67、123、140, 进行分割处理时不能用表 1 中分割词进行分割的独立权利要求,采用人工分割。但这样的数据很少,分别占初步检索结果的 4. 3%、5. 4%、6. 8%、1. 7%、2. 3%。

4.2 结果评价

对于检索返回的结果,采用人工方式评价其与输入申请专利的相关性。检索的目的是快而准地找到与申请专利密切相关的专利,所以只评价排序在前 20 位的返回结果。为综合比较两组实验结果,将结果数据绘制到同一个图中。图 2 是返回结果中排序前 20 位的专利与相应的申请专利之间具有相关性的专利数量。

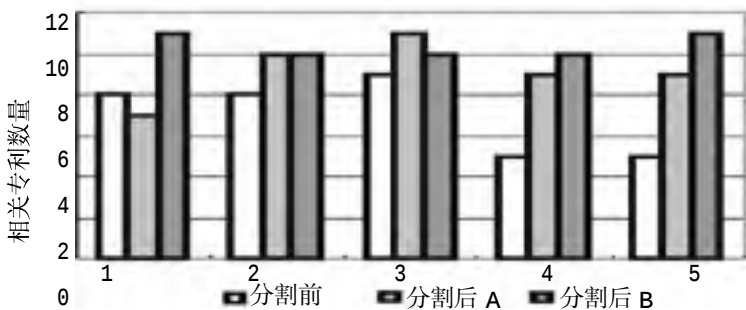


图 2 检索结果对比图

从图中可以看出: a) 权利要求分割处理后(分割后 A), 除 No. 1 外其余检索效果均有所改善, 尤其是 No. 5 和 6 的改善效果更加明显; b) 当采用 $b_{q,t} \times b_{d,t}$ 词语权重时(即分割后 B), 除 No. 3 外其余检索效果比采用 $f_{q,t} \times f_{d,t} \times idf_t$ 词语权重(即分割后 A)要好, 这与一般的信息检索有所不同; c) 分割处理后, 并采用 $b_{q,t} \times b_{d,t}$ 词语权重方法(即分割后 B), 五件发明

专利检索效果较分割前均有所改善。

5 结束语

本文构建的中文专利无效检索模型,充分考虑了中文专利的独立权利要求结构特征信息,借助分割词进行分割处理。在具体检索中采用两步检索方法,使得检索的准确率和召回率均得到提高。同时,对词语权重选择方法 $b_{q,t} \times b_{d,t}$ 和 $f_{q,t} \times f_{d,t} \times idf_t$ 的检索效果进行比较。实验结果显示,该模型能有效地对检索结果进行排序,使审查员能够迅速地寻找到相关专利,从而完成专利的审查工作。

参考文献:

[1] ANDO N, LEONG M K. Workshop on patent retrieval SIGIR 2000 workshop report[J]. ACM SIGIR Forum Archives, 2000, 34(1): 28-30.

[2] ACL 2003. Proceedings of ACL 2003 workshop on patent corpus processing [EB/ OL]. (2003) [2007- 03- 05]. <http://www.slis.tsukuba.ac.jp/~fujii/acl2003ws.html>.

[3] IWAYAMA M, FUJII A, KANDO N. Overview of patent retrieval task at NTCIR-3[C] // Proc of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering. Tokyo: [s. n.], 2003: 21-24.

[4] FUJII A, IWAYAMA M, KANDO N. Overview of patent retrieval task at NTCIR-4[C] // Proc of the 4th NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization. Tokyo: [s. n.], 2004: 225-232.

[5] FUJII A, IWAYAMA M, KANDO N. Overview of patent retrieval task at NTCIR-5[C] // Proc of the 5th NTCIR Workshop on Evaluation of Information Access Technologies, Information Retrieval, Question Answering and Cross-lingual Information Access. Tokyo: [s. n.], 2005: 269-277.

[6] LARKEY L S. A patent search and classification system[C] // Proc of the 4th ACM Conference on Digital Libraries. 1999: 79-87.

[7] SHINMORI A, OKUMURA M. Patent claim processing for readability-structure analysis and term explanation[C] // Proc of ACL Workshop on Patent Corpus Processing. Tokyo: [s. n.], 2003: 56-65.

[8] KIM J, HUANG Jin-xia, HAYONGJ Ha-yong. Patent document retrieval and classification at KAIST[C] // Proc of the 5th NTCIR Workshop on Evaluation of Information Access Technologies, Information Retrieval, Question Answering and Cross-lingual Information Access. Tokyo: [s. n.], 2005.

[9] 中华人民共和国国家知识产权局. 审查指南[K]. 北京: 知识产权出版社, 2006: 218-242.

(上接第 2063 页)

[2] 高刚毅. 分布式地理信息系统[D]. 杭州: 浙江大学, 2004.

[3] 马修军, 刘晨, 谢昆青, 等. P2P 环境中的全局空间数据目录研究[J]. 地理与地理信息科学, 2006, 22(3): 23-25.

[4] CRESPO A, GARCIA-MOLINA H. Routing indices for peer-to-peer systems[C] // Proc of the 22nd Int Conf on Distributed Computing Systems. Washington DC: IEEE Computer Society, 2002.

[5] 杨峰, 李凤霞, 余宏亮, 等. 一种基于分布式哈希表的混合对等发现算法[J]. 软件学报, 2007, 18(3): 715-721.

[6] 邱彤庆, 陈贵海. 一种令 P2P 覆盖网络拓扑相关的通用方法[J]. 软件学报, 2007, 18(2): 381-389.

[7] 钱卫宁. 对等计算系统中的数据管理[D]. 上海: 复旦大学,

2003.

[8] 吴信才, 吴亮. 面向服务的分布式空间信息支撑平台[J]. 地球科学, 2006, 31(5): 585-589.

[9] MECELLA M, ANGELACCIO M, KREK A, et al. Workpad: an adaptive peer-to-peer software infrastructure for supporting collaborative work of human operators in emergency/disaster scenarios[C] // Proc of IEEE International Symposium on Collaborative Technologies and Systems. 2006: 173-180.

[10] MacEACHREN A M, BREWER I, CAI G, et al. Visually-enabled geocollaboration to support data exploration and decision-making[C] // Proc of the 21st International Cartographic Conference. Durban, South Africa: [s. n.], 2003.