

基于分类号和引文的专利相似度 测量方法研究*

王鑫, 赵蕴华, 高芳

(中国科学技术信息研究所, 北京 100038)

摘要: 面向专利数据领域, 根据专利引文数据、IPC分类号、德温特分类号构建一种专利相似度的测量方法, 以LTE移动通信领域专利为样本进行实证分析, 发现该方法对专利相似程度区分性较好。

关键词: 专利相似度; 专利分类号; 专利引文; 专利文本挖掘

中图分类号: G305

DOI: 10.3772/j.issn.1673—2286.2015.01.011

专利作为知识产权的核心代表, 是一个企业、行业乃至国家竞争力的集中体现。充分利用专利信息可以分析出企业战略布局、研发进度、技术导向、未来市场瞭望等众多难以获得的企业情报。有效运用专利情报, 平均可缩短研发时间60%, 节省研发费用40%^[1], 助力企业知己知彼、紧跟行业步伐, 引领市场潮流。相似专利的检索不仅在专利布局、专利预警、专利价值衡量、新技术识别、专利规避中占据重要地位, 对于新专利的申请、侵权检测、专利诉讼等也意义重大。提高专利相似度测量的准确性和有效性对于专利分析意义深刻。

1 专利相似度测量现状

专利具有标题、摘要等文本信息及分类号、引用关系等结构信息, 还包括权利要求项、申请日、优先权等独特字段。由专利审查员授予一个或多个分类号, 对内容范畴的界定统筹也更加准确。专利引文除专利申请人引用外也还包括审查员引用。由此便产生了两种相似度

测量方法, 基于文本的专利相似度测量和基于引用的专利相似度测量。二者分别从语法和技术演化关系的角度反映专利间的相似度。

1.1 基于文本的专利相似度测量

随着文本挖掘技术的不断发展, 基于专利标题、摘要、权利要求项的文本相似度测量迅速成为专利相似度研究的重点领域, Magerman等测试了这种方法的可能性和准确性^[2]。专利文本相似度测量多采用基于共词的向量空间模型, 依据词语匹配程度构建相似矩阵, 这种方法对语词切分词典有较大的依赖性, 语词的识别与切分度直接制约着相似度测量的准确度^[3]。语词本身的多义性、歧义性等都严重制约着该方法的应用。王晋引入领域本体的词间关系来计算词汇语义相似度^[1]。Bergmann等提取出文档的主谓宾结构, 通过文档的结构相似性来测量专利相似度^[4]。Huang等在分析专利文档结构的基础上提出了结合专利结构和内容的相似度计算方法^[5]。胡侠等度量了专利摘要文本和IPC分类号

* 本研究得到中国科学技术信息研究所预研基金项目“中国TD-LTE产业发展中的标准扩散与政策研究”(编号: YY-201404)资助。

的相似性并据此进行了专利聚类^[6]。刘志辉等分别运用关键词与分类号描述专利主体内容,实证发现分类号方法所揭示的内容粒度要更细^[7]。

1.2 基于引用的专利相似度测量

创新大多不是独立事件,某项创新所含的技术常常能引发或促成一系列相关的创新集合。创新的这种聚集性反映在专利上即为专利引用,它在一定程度上能够体现创新的集群、技术的承接。相似专利间所描述的相似技术常基于共同的基础技术演化而来,专利审查员在新颖性判别时会对照相似专利并标注为引用,故引文能在一定程度上刻画专利间的相似性,而且这种刻画粒度比较细,能够精确到具体的每一篇专利。专利引用从语义层面反映出技术间的演进和借鉴关系,由此产生了基于引用从技术演化的语义层面测量专利相似度的方法。虽然在引用动机上专利与论文有较大差异^[8],但文献引用的一般特征在专利中也不例外,专利引文中蕴含的知识活动能部分反映专利间相似度^[9]。Lai等依据专利间的引用关系来测量专利相似度并据此构建了专利分类系统^[10]。McGill采用企业间的专利互引来测量专利相似度^[11]。李睿等对比了同被引与引用耦合两种方法,发现引用耦合在专利间相似度测量上更胜一筹^[12]。

1.3 基于分类号的专利相似度分析

综上,文本相似度测量主要从语法层面基于共词假设的语词匹配进行专利相似度刻画;引用层面上主要从引文反应的技术演化关系以及审查员检测角度反应专利相似度。这两种专利相似度测量方法各有其优势与特点。除此之外,专利分类号所体现的专利间的类属关系也能从一定程度上对相似专利进行刻画,这种刻画角度是上述两种方法所无法描述的。基于相同的基础技术可以演化出不同甚至差异较大的技术,对此引用无法有效描述。相似的技术采用不同的实现方式,在措辞上可能会有很大差异,对此基于语词匹配的文本相似度测量方法也无法进行有效刻画。而这正是专利分类号的用武之处,归属于同一类别的专利,无论其是否基于相同的技术引用,是否采用接近的措辞描述,分类号都可对其相似度进行有效反映。同时专利分类号,类别多样,都由权威组织定期更新维护,在体系结

构上也较为细致全面。IPC国际专利分类号侧重功能分类,德温特手工代码侧重应用分类^[13],不同分类体系从不同的角度对专利相似度进行描述。文章将探索结合分类号和专利引用的相似度测量方法。

2 基于分类号和引文的专利相似度测量方法

专利包含名称、摘要、分类号、优先权、引文等众多信息,综合利用各种专利信息,尤其是分类号、引文在语义甚至语用空间的相似度分布描述对于提高专利相似度测量具有重要意义。本文结合分类号的功能应用描述和引文的技术继承演化描述,提出了一种基于引用和多种分类号(IPC专利分类号、德温特手工分类号(以下简称手工分类号)、德温特分类代码(以下简称分类代码))的相似度测量方法。

对于n件专利集合P,对专利 P_i 提取其专利号m及引文c、IPC专利分类号f、手工分类号a、分类代码b等五维特征值构成特征向量,即: $p_i = (m, c, f, a, b)$ 。

生成集合p对应的5维专利特征矩阵。

对于引文c、IPC分类号f、手工代码a、分类号b这四维数据分别用向量空间模型(VSM)表示。

例如,对于专利 P_i 的引文数据可表示为:

$$c_i = (w_1 c_i^1, w_2 c_i^2 \dots w_j c_i^j \dots)$$

其中, c_i^j 表示专利 P_i 的第j篇引文, w_j 表示引文 c_i^j 对应的权重。可对专利的发明人引用和审查员引用因其动机不同可赋予不同权值。

对于专利 P_i 的IPC分类号可表示为:

$$f_i = (w_1 f_i^1, w_2 f_i^2 \dots w_j f_i^j \dots)$$

其中, f_i^j 表示专利 P_i 的第j个IPC分类号, w_j 表示IPC分类号 f_i^j 对应的权重。一个专利可能拥有多个专利分类号,排名越靠前的专利号通常越能反映专利内容,对此可赋予不同权重。手工代码a和分类号b的向量表示同IPC分类号类似。

由此可得到专利对 (P_i, P_j) 各自的引文数 (c_i, c_j) , IPC类号数 (f_i, f_j) , 手工分类号数 (a_i, a_j) 分类代码数 (b_i, b_j) , 通过集合相交运算得到两专利共有的引文数 c_{ij} 、IPC类号数 f_{ij} 、手工分类号数 a_{ij} 、分类代码数 b_{ij} 。可通过相对比例来衡量四个维度下专利对间的相似度分量,即: $(p_i, p_j) = (p_{ij}^c, p_{ij}^f, p_{ij}^a, p_{ij}^b)$ 。

其中 $p_{ij}^c, p_{ij}^f, p_{ij}^a, p_{ij}^b$ 分别代表专利 P_i, P_j 的相对共有引文比率,相对共有IPC分类号比率,相对共有手

工分类号比率, 以及相对共有分类代码比率。各项相对共有比例为两专利的共有绝对数值比上平均数值, 以消除数量规模的影响, 即: $p_{ij}^c = \frac{c_{ij}}{c_i + c_j}$, $p_{ij}^f = \frac{f_{ij}}{f_i + f_j}$, $p_{ij}^a = \frac{a_{ij}}{a_i + a_j}$, $p_{ij}^b = \frac{b_{ij}}{b_i + b_j}$ 。 $p_{ij}^c, p_{ij}^f, p_{ij}^a, p_{ij}^b$ 取值范围[0,1], 值越大相似度越高。基于引文、IPC分类号、手工分类号、分类代码四维度的相对共有比率, 加权得到专利间的相似度 $sim(p_i, p_j) = \sum w_s \times s$, 其中 $s = \{c, f, a, b\}$, w_s 表示s的权重, $\sum w_s = 1$ 。 $sim(p_i, p_j)$ 值越大, 专利 $p_i \cdot p_j$ 越相似, 其取值范围为[0,1]。

3 实证研究

本文以德温特创新索引 (DII, Derwent Innovation Index) 为数据源, 选取4G移动通信标准领域的PCT授权专利为样本对象进行实证研究。根据4G移动通信标准确定关键词通过PCT限定于2014年7月采集到6900件专利, 扣除部分无引文专利得到4936条样本数据。这些数据中共涉及22469篇去重被引用专利, 1114个不同的IPC分类号, 1072个不同的手工分类号, 78个不同的分类代码。平均每篇专利包含6篇专利引文, 5个IPC分类号, 6个手工分类号, 2个分类代码。

3.1 实证分析

基于前述方法分别计算引文与分类号等四维度

的分量相似度。对于IPC分类号, 手工分类号, 分类代码这三种分类号维度基于数据量的考虑剔除部分共有项数少的专利数据对, 分别得到129648条分类号值大于2的专利, 364437条分类号值大于3的专利, 1496条分类号值大于3的专利。对于引用, IPC分类号, 手工分类号, 分类代码的相似权重分别赋予为0.2, 0.3, 0.3, 0.2, 权重取值源于多次尝试后的相对较优结果, 此处引用权重取值较小主要原因是专利间的引用矩阵比较稀疏。

汇总得到486799条专利数据对各维度相似度及加权综合相似度。所有样本专利的平均相似度为0.307。

最相似的是专利WO2014003506-A1和专利WO2014003508-A1, 相似度0.89, 均为LG旗下关于测量并传递信道信息方法的专利, 二者拥有相同的发明人、IPC分类号、德温特手工分类号、德温特分类代码以及两篇共同的引用文献, 通过阅读专利摘要及全文证实二者相似度的确很高。相似度最低的是专利US2009086704-A1和US2008132230-A1, 相似度0.097, US2009086704-A1是关于无线链路控制协议的专利, US2008132230-A1是用于无线网络控制器不连续传输和接收的控制方法的专利, 二者在内容上相似性不强。相似度分布参见表1, 其中相似度r分布区间包含最小值不含最大值。前10对最相似的专利对参见表2, 图中数值k/(i,j)中, k标示二者共有的数目, i、j分别表示两件专利各自的数目。

表1 基于引用和分类号的相似度分布

相似度	[0.8,1]	[0.7,0.8]	[0.6,0.7]	[0.5,0.6]	[0.4,0.5]	[0.3,0.4]	[0.2,0.3]	[0.1,0.2]	[0,0.1]
频次	2	68	1205	9495	47765	175766	234092	19102	1
比重	0.000	0.001	0.003	0.019	0.098	0.361	0.481	0.039	0.000

3.2 该方法与基于共词的相似度测量的对比分析

为验证试验结果的有效性, 对比该方法与一般专利文本相似计算方法, 选取相似度最高的前20对专利和相似度最低的前20对专利, 选取这40对专利为样本, 以标题, 摘要字段进行基于共词的专利相似度测量分析。样本中共有836个单词。经过去停用词、部分高频无

实义词、词干化后得到715个原形单词, 统计各篇专利的语词及其词频分布, 每对专利的共有词汇及其频率, 并据此计算出每对专利的累积共有词频 (累积共有词频= $\sum$ 每个共有词汇的频率)。为消除不同专利间文本 (尤其是摘要) 长度引起的词频差异, 文章中用专利对的累积共词频率与平均文本 (标题及摘要) 长度的相对比例表示专利相似度 (相似度=累积共有词频/标题和摘要的平均长度)。基于文本的相似度并不是本文的重点内容故

表2 基于引用和分类号的相似度TOP10

专利对 (专利号标示)	引用	IPC	手工 分类号	分类 代码	相似度
WO2014003506-A1; WO2014003508-A1	2/(5,4)	1/(1,1)	1/(1,1)	1/(1,1)	0.89
US2013040640-A1; US2013039268-A1	2/(2,2)	1/(2,3)	1/(1,1)	1/(1,1)	0.8
CN102355680-A; WO2011005373-A1	0/(13,5)	1/(1,1)	4/(4,4)	1/(1,1)	0.8
WO2013127424-A1; CN102355680-A	0/(1,13)	1/(1,1)	4/(4,4)	1/(1,1)	0.8
WO2013127424-A1; WO2011005373-A1	0/(1,5)	1/(1,1)	4/(4,4)	1/(1,1)	0.8
WO2013180477-A1; WO2013180487-A1	1/(3,3)	3/(3,3)	4/(4,4)	1/(2,2)	0.7667
WO2010082877-A1; WO2010019902-A1	0/(5,3)	4/(5,4)	1/(1,1)	1/(1,1)	0.7667
WO2013109176-A1; WO2010140907-A1	0/(3,9)	1/(1,1)	4/(5,4)	1/(1,1)	0.7667
WO2010149100-A1; WO2010019902-A1	0/(4,3)	3/(3,4)	1/(1,1)	1/(1,1)	0.7571
WO2013035981-A1; WO2013036090-A1	4/(4,4)	1/(1,2)	1/(2,2)	1/(1,1)	0.75

采用了简化的计算方式,得出大致结果趋势以便和文章中相似度结果进行对比。结果参见表3。

介于原始累积共有词频与平均文本长度比值过小,区分度不明显,为便于比较专利间的相对强弱此处将累积共有词频整体放大7倍,未出现超出平均文本长度者,比率范围仍为[0,1],放大倍数7来源于多次尝试以保障对比效果更佳。此处计算共词的相似度只是为了验证其结果与基于分类号和引用的相似度测量结果的一致性,只做两种方式下的一致性趋势比较,不做横向绝对值大小比较,故结果整体放大不会影响内部趋势。对比结果参见图1。

图1中横坐标专利对1-20对应表3中从上到下的专利对,“前20”对应表3中相似专利对前20,“后20”对应表3中相似专利对后20。“前20”、“后20”分别对应表3中的相似专利对前20和相似专利对后20。“分类号-引用”表示基于分类号和引用的相似度测量方法下的相似度(以下简称分类号-引用相似度),“原始共词”表示基于共词的相似度测量方法下的相似度(以下简称共词相似度),“7倍共词”表示放大7倍后的“原始共词”相似度(以下简称7倍共词相似度)。图中红色虚线表示相似度为0.5的水平线。

表3 基于两种方法测度的专利相似度

相似专利对前20	共词的相似度	分类号和引 用的相似度	相似专利对后20	共词的相似度	分类号和引 用的相似度
WO2014003506-A1; WO2014003508-A1	0.1044	0.8889	WO2009002656-A2; US2008132230-A1	0.0746	0.1136
US2013040640-A1; US2013039268-A1	0.0725	0.8	WO2012131400-A1; EP2093983-A1	0.0541	0.1132
CN102355680-A; WO2011005373-A1	0.0795	0.8	US2008132230-A1; EP1915017-A2	0.0699	0.1129
WO2013127424-A1; CN102355680-A	0.0573	0.8	US2010239031-A1; US2008132230-A1	0.0485	0.1126
WO2013127424-A1; WO2011005373-A1	0.0765	0.8	JP5044017-B; US2008132230-A1	0.0706	0.1121
WO2013180477-A1; WO2013180487-A1	0.1424	0.7667	US2009109942-A1; US2008080399-A1	0.0695	0.1119
WO2010082877-A1; WO2010019902-A1	0.0663	0.7667	US2008132230-A1; US2008107161-A1	0.0341	0.1115
WO2013109176-A1; WO2010140907-A1	0.0559	0.7667	US2008132230-A1; WO2007079579-A1	0.0501	0.1111
WO2010149100-A1; WO2010019902-A1	0.0628	0.7571	WO2009000790-A1; US2008132230-A1	0.05	0.1102

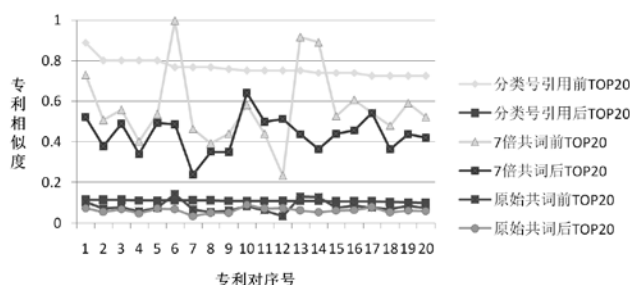


图1 两种测量方法下专利对相似度趋势对比图

如图所示, 基于分类号和引用的专利相似度测量方法下, 前20专利相似度明显高于后20专利相似度, 前20专利对可判定为相似, 后20专利对可判定为不相似, 参见图中分类号-引用前20曲线和分类号-引用后20曲线。基于文本的相似度测量方法下, 图中原始共词前20曲线也整体高于原始共词后20曲线, 即前20专利相似度高于后20专利相似度, 前20专利对可判定为相似, 后20专利对可判定为不相似。两种方法大致趋势一致, 共词相似度测量方法对相似专利集合与不相似专利集合的区分度较低, 分类号-引用相似度测量方法区分度较高。对于前20、后20专利集合内部共词相似度测量方法区分度相对较好, 但由于结果太小不明显, 故采用7倍共词相似度进行内部分析。

对于7倍共词相似度, 见图中7倍共词前20曲线和7倍共词后20曲线。其中若以0.5为是否相似的判定阈值, 则对于“分类号-引用”中判定为相似的前20专利对, 结果大致上也判定为相似, 对于分类号-引用中判定为不相似的后20专利对, 结果大致上也判定为不相似。后20专利对的相似度整体低于前20专利对的相似度, 整体趋势大致相同。根据7倍共词曲线可见对于相邻或相近的节点区分度较大, 对曲线两端节点以及前20节点、后20节点差异不大, 即对于序号接近的专利区分度较大, 对于序号相距较远的专利区分度不足。共词相似度测量对于特别相似和特别不相似的专利区分度不足对于相似度一般的专利区分度较好。

综上基于分类号引用的相似度测量方法对特别相似和特别不相似的专利对之间的区分度较大, 对相似度一般的专利间区分度不大, 即图中“分类-引用”曲线两端的区分度大, 中间的区分度低。说明: 文章中的特别相似专利和特别不相似专利同属于同一个大技术领域下专利, 并非跨领域的专利, 跨技术领域的专利间通常可比性较小。而基于共词的专利相似度测量方法则与之互补, 对于相似度一般的区分度较大, 对相似

度的极值区分度较小。这也符合一般规律, 即特别相似与特别不相似的专利在专利分类号和引用上会表现明显, 但由于共属于一个大领域在专利措辞上同样会存在较多的共有词汇。对于相似度一般的专利之间常共同归属于一个具体的细分领域有着相近的引文, 故分类号和引文的区分度不强, 共词却能更好的体现二者之间的差异。

3.3 异常值判定

对表3, 图1中极端异常值, 前20对、后20对专利集合中分类号-引用相似度和7倍共词相似度相差最远的专利对。前20集合中专利对12: WO2014003508-A1; WO2011020269-A1 (7倍共词相似度0.231, 分类号-引用相似度0.75, 共词测量判定为不相似, 分类号-引用测量判定为相似)。后20集合中专利对10: WO2008100488-A1; US2008132230-A1 (7倍共词相似度0.6398, 分类号-引用相似度0.1107, 共词测量判定为相似, 分类号-引用测量判定为不相似)。对这四篇专利进行标题及摘要的人工阅读判定。WO2014003508-A1描述了在LTE-A等无线通信系统中通过CSI-RS (信道状态信息测量导频) 和RRM (无线资源监测) 来测量信道状态及参数信息的方法。WO2011020269-A1描述了一种LTE系统访问方法, 该方法可根据用户终端带宽来对上传/下载链路的聚合载波进行分组, 并生成聚合载波的物理随机访问信道参数。WO2008100488-A1描述了一种通过演进型节点B (eNB, evolved Node-B) 组建重定位命令、隧道、切换命令的无线通信网络间切换方法。US2008132230-A1描述了一种不连续发送/接受控制方法, 该方法用Node-B表现偏移量并通过高速共享控制通道 (HS-SCCH) 传输。相比之下, 前20专利对12更为相似, 后20专利对10: WO2008100488-A1; US2008132230-A1较为不相似, 实际结果更偏向分类号-引用相似度测量结论。但实际相似度又不如分类号-引用相似度那么极端, 更加符合两种方法的中间态, 用二者的加权平均值或许更能反映实际相似度。其他专利对的随机抽查阅读也基本符合两种方法的差异规律。

4 结语

基于分类号和引用的相似度测量从外部从属类别、技术演化关系的角度反映专利相似度。基于共词的

相似度测量方法从专利内容的字面匹配角度刻画专利相似度。二者各有所长,互为补充。结合分类号和引用的测量方法对相似度较高和较低的专利区分度更好,共词下的相似度测量更擅长于区分相似度一般的专利,二者的结合度量方法效果更佳。对这两种测量方法可进行有效的加权合并以便更加综合全面的测量专利相似度。文章通过专利对之间的相对共有比例反映引文和专利号各维度下的专利相似度,也可通过向量空间模型SVM中的向量夹角余弦反映各维度下的专利相似度,此处采用相对共有比例主要在于缓解向量特征值数量大且分散带来的庞大计算量负荷。在效用上基于特征值共现计算具体维度的专利间相似度,语法层面的共现频次才是相似度的重要体现,因而可以适当舍弃以简化运算提高效率。文章中为减轻计算负荷简化了相关运算,虽然结果的精度不高,但也大致反映出了两种方法的优劣。可通过特征向量的余弦计算相似度,各项指标权值的进一步探究等使得结果更加精准,对两种方法的有效合并也还需进一步探究。

参考文献

- [1] 王晋.基于领域本体的专利地图研究[D].苏州: 苏州大学,2011.
- [2] Magerman T, Van Looy B, Song X. Exploring the Feasibility and accuracy of Latent Semantic Analysis Based Text Mining Techniques to Detect Similarity Between Patent Documents and Scientific Publications[J]. Scientometrics, 2010(2):289-306.
- [3] 尹远明.专利技术相似度评价以及系统应用研究[D].上海: 华东理工大学,2010.
- [4] Bergmann I, Butzke D, Walter L, et al. Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: the Case of DNA Chips[J]. R&D Management, 2008 (5):550-562.
- [5] Huang SH, Ke HR, Yang WP. Structure Clustering for Chinese Patent Documents [J]. Expert Systems with Applications, 2008(4):2290-2297.
- [6] 胡侠, 林晔, 汪亮, 盛夏, 王灿. 基于树距离规范化的专利聚类方法[J]. 情报学报, 2012(11):1187-1193.
- [7] 刘志辉, 赵筱媛. 基于专利形态相似性的竞争态势分析方法研究[J]. 情报理论与实践, 2012(6):67-69.
- [8] 谢黎, 邓勇, 张苏闽. 论文引用与专利引用比较研究[J]. 情报杂志, 2012(4):18-21.
- [9] 杨中楷, 梁永霞, 刘倩楠. 专利引用过程中的知识活动探析[J]. 科研管理, 2010(3): 171-176.
- [10] Lai K K, Wu S J. Using the patent Co-citation approach to establish A New Patent Classification System [J]. Information Processing and Management, 2005(2):313-330.
- [11] McGill J P. Technological Knowledge and Governance in Alliances among Competitors [J]. International Journal of Technology Management, 2007(2):69-89.
- [12] 李睿, 张玲玲, 郭世月. 专利同被引聚类与专利引用耦合聚类的对比分析[J]. 图书情报工作, 2012(8):91-95.
- [13] 顾震宇. 德温特手工代码与IPC国际专利分类号的分类比较: 以燃料电池为例[EB/OL][2015-01-03]. www.istis.sh.cn/list/list.aspx?id=5229.

作者简介

王鑫, 男, 1990年生, 硕士生, 研究方向: 专利与政策分析。
赵蕴华, 女, 1967年生, 副研究馆员, 研究方向: 科技政策、科技领域分析。
高芳, 女, 1980年生, 博士, 讲师, 研究方向: 重点科技领域信息分析。

A Method for Assessing Patent Similarity Based on International Patent Classification and Patent Citation

WANG Xin, ZHAO YunHua, GAO Fang

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Oriented to patent data field, we build a patent similarity measurement method according to the patent citation, International patent classification, Derwent classification number. Patents of LTE mobile communication field were taken as an example for empirical analysis. Experimental comparison with previous work indicates that the proposed method is fit for patent data field.

Keywords: Patent similarity; International patent classification; Patent citation; Patent mining

(收稿日期: 2015-01-06)

编辑: 刘伟