

一种基于潜在引用网络的专利价值评估方法

冯 岭^{1,2} 彭智勇^{1,2} 刘 斌^{1,2} 车敦仁³

¹(软件工程国家重点实验室(武汉大学) 武汉 430072)
²(武汉大学计算机学院 武汉 430072)
³(南伊利诺伊斯大学计算机科学系 美国伊利诺伊州卡本戴尔 IL62901)
(peng@whu.edu.cn)

A Latent-Citation-Network Based Patent Value Evaluation Method

Feng Ling^{1,2}, Peng Zhiyong^{1,2}, Liu Bin^{1,2}, and Che Dunren³

¹(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072)
²(School of Computer, Wuhan University, Wuhan 430072)
³(Department of Computer Science, Southern Illinois University, Carbondale, Illinois, USA IL62901)

Abstract Patent value refers to the exchange value of a patent in the purchase and transaction. It can provide precious information for the decision-makings of patent owners and buyers. Existing patent value evaluation methods are mainly based on training or citation analysis. However, the methods based on training depend too much on the experimental parameters, which results in weak credibility. On the other hand, the methods based on citation analysis only consider direct citations during the evaluation leaving indirect citations and novelty of patent neglected. For these reasons, this paper presents a latent-citation-network based patent value evaluation method to evaluate the value of each patent in which direct citations, indirect citations and novelty are all considered. First, the latent citation association is discovered utilizing similarity between patents and a latent citation network is established. Then, a basic algorithm is implemented to effectively evaluate the value of each patent on the network. Further, an improved algorithm is proposed to solve the problem of inefficiency of the basic algorithm. Finally, to handle the value variations caused by the arrival of a new patent, a dynamic patent value evaluation algorithm is designed to efficiently update the value of the original patents. As shown in the experiments, the proposed methods in this paper are effective.

Key words patent value; patent association; latent citation network; value evaluation; dynamic updating

摘 要 专利价值指的是在专利购买交易时的交换价值,它可以为专利的拥有者和购买者提供大量关于专利的珍贵信息,为专利的交易提供决策支持.目前的专利价值评估主要是基于训练或引用的方法.前者过于依赖于人工选择的参数,导致专利价值评估结果的可信度较低;后者往往只考虑了直接引用关联,而忽略了间接引用关联和新颖度对专利价值的影响.针对这些问题,提出一种基于潜在引用网络的专利价值评估方法,并设计了相应的算法来评估各个专利的价值.针对基本算法效率较低的问题,提出了专利价值评估改进算法,极大地提高了专利价值评估的效率.最后,由于新专利加入专利集合时各个

专利的价值会发生变化,提出了专利价值评估动态更新算法,用于快速地更新各个专利的价值.实验表明,提出的方法是有效的.

关键词 专利价值;专利关联;潜在引用网络;价值评估;动态更新

中图法分类号 TP391.10

专利是科技信息及知识的载体、是世界上最大的知识产权信息源.根据世界知识产权组织的统计,专利数据包含了世界上 95% 的研发成果.若能有效地利用专利信息,不仅可以提高科研项目研究起点和水平,而且可以节约 60% 的研究时间和 40% 的研究经费^[1].由于专利的重要性,目前已有大量的文献对其进行了研究,并在专利检索^[2-4]、专利自动分类^[5-7]、竞争对手发现^[8]、专利趋势分析^[9-11] 和合作推荐^[12] 等方面取得了丰硕的成果.

专利价值一般指专利购买和交易过程中专利的交换价值.在专利购买和交易过程中往往需要对专利的价值进行评估,专利价值评估是专利购买过程中至关重要的一环,直接关系到企业在进行技术引进后的未来利润,因此显得格外重要.2011 年,中国技术交易所提出了“专利价值分析指标体系”,通过法律价值度、技术价值度和经济价值度 3 方面来衡量专利的价值,但其中的指标完全依赖于人工,需要耗费大量的人力资源和时间资源^[13].

目前的专利价值评估方法一般可以分为 2 种:一种是基于训练的方法;另一种是基于引用的方法.

基于训练的专利价值评估方法首先利用经验找出影响专利价值的因素.之后,在专家已判定好专利价值的训练数据集中,通过训练确定各个因素在专利价值中所占的权重,得到价值评估公式.最后,利用该公式计算测试集中各个专利的价值.通过这种基于训练的专利价值评估方法,已有许多文献提出利用多因素来评估专利的价值.文献^[14]基于元特征、新颖度特征、写作质量特征、趋势特征、发明人特征构建了一个专利预测模型,用该预测模型来判断是否应对特定专利进行续费.文献^[15]基于声明的数量、同族专利数量、专利授权时间、专利申请人的权威度等特征,利用因子分析法来构建因子模型,并利用该因子模型来度量各个专利的价值.文献^[16-17]则通过专利授权的速度、同族专利数量、专利被续费的时间、专利是否被其他专利权人反驳等多个特征的组合来构建不同的分类器,并验证各个分类器在对专利价值进行评估有效性,以得到准确的专利价值评估模型.然而,由于影响专利价值的因

素通过经验得到,具有很强的主观性,且价值评估公式过度依赖于训练数据集,导致基于训练方法的专利价值评估结果的可信度较低.

专利价值也可以通过专利之间的引用关联进行评估.与训练的方法相比,基于引用关联的评估方法更为客观,且不依赖于训练集的优劣,具有更高的可信度.一般来说,一个专利被越多的专利所引用,则其专利价值越高.在美国知识产权咨询公司 CHI 提出的指标评价体系中,引证指数(专利被引用的次数)被用来评估目标专利的影响力和价值,专利被引用的次数越多,代表该专利所包含的技术属于基础性或领先技术,处于核心技术或位于技术交叉点.引证指数已经成为判断专利价值的主要指标之一,许多文献都采用该方法来度量专利的价值^[18-19].但这种方法只考虑了专利之间的直接引用关联,而忽略了专利之间的间接引用关联,且没有考虑专利的新颖度对专利价值的影响,因此不能很好地评估专利的价值.此外,中文专利中引用关联的缺失也给这种方法的应用带来了困难.

针对已有专利价值评估方法的缺陷,本文提出了一种基于潜在引用网络的专利价值评估方法.首先利用专利间的相似度建立专利潜在引用关联,在此基础上同时考虑直接引用关联和间接引用关联及专利的新颖度对专利价值的影响,提出了专利价值评估基本算法,该算法可以有效地评估专利的价值.针对基本算法时间花费较高的问题,本文提出了专利价值评估改进算法,提高了算法的效率.在此基础上,当新的专利加入专利集合时,本文提出了专利价值评估动态更新算法,可以快速地更新各个专利的价值.

本文的主要贡献是:1)提出了专利之间的潜在引用关联;2)提出了基于潜在引用网络的专利价值评估基本算法,并改进了算法的效率;3)提出了专利价值评估动态更新算法;4)通过一系列实验验证了本文提出的方法的有效性和高效性.

1 相关知识

给定专利文档的集合 $D=\{d_1,d_2,\cdots,d_n\}$,专利

价值评估需要计算任意一个专利 d_i 的价值. 在 CHI 公司提出的指标评价体系中, 引用关联被用来评估专利的价值. 根据专利之间的引用关联, 专利集合可以被映射到一个有向图中.

例 1. 图 1 中表示了一个专利集合 $D=\{a,b,c,d,e,f,g\}$ 对应的专利引用网络 $G=(V,E)$, 在该网络中, 节点表示所对应的专利文档, 节点之间的边表示专利文档之间的引用关联, 节点的横轴坐标表示专利发表的时间. 例如节点 a 表示对应的专利文档 a , 其发表时间为 2001 年, 节点 b 指向 a 的边表示专利 b 引用了专利 a .

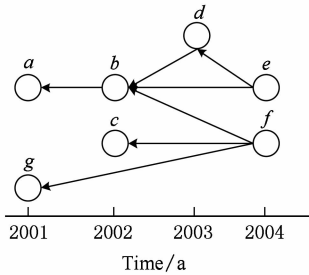


Fig. 1 An example of patent citation network.

图 1 一个专利引用网络的例子

在 CHI 的指标体系中, 专利被引用的次数被用来评估专利的价值, 在网络中即节点的入度. 在图 1 中, 专利 a 只被一个专利所引用, 因此其专利价值 $pvalue(a)=1$. 类似地, $pvalue(b)=3$, $pvalue(c)=pvalue(d)=pvalue(g)=1$, $pvalue(e)=pvalue(f)=0$.

CHI 指标虽然在一定程度上反映了专利的价值, 但也存在一定的缺陷, 具体分析有如下 4 点:

1) 不仅专利的直接引用会对专利价值产生影响, 间接引用也会对专利价值产生影响. 而 CHI 指数只考虑了专利间的直接引用关联, 忽略了专利间的间接引用关联. 例如, 在图 1 中计算专利 a 的价值时, CHI 指标只考虑专利 b 对专利 a 的直接引用. 然而, 对于专利 d, e 和 f , 尽管它们并没有直接引用 a , 但都通过专利 b 对专利 a 进行了间接引用, 因此专利 d, e 和 f 也应该对专利 a 的价值产生影响.

2) 专利的新颖度是影响专利价值的一个重要因素. 一般来说, 由于专利中所述技术会随时间的变化而逐渐变得落后, 导致其专利价值也会相应地衰减. 而在 CHI 指标中忽略了专利的新颖度对专利价值的影响. 例如, 在图 1 中, 专利 c 和专利 g 都被专利 f 所引用, 但由于专利 g 发表的时间比专利 c 早, 因此其价值应该比专利 c 低, 而在 CHI 指标中, 专

利 c 和专利 g 具有相同的专利价值.

3) 尽管有些专利并未被其他专利所引用, 但也具有一定的专利价值. 而在 CHI 指标中, 未被引用的专利的价值为零, 如图 1 中的专利 e 和 f .

4) 中文专利中不存在专利间的引用信息, 导致基于引用的专利价值评估方法不能被应用.

2 基于潜在引用网络的专利价值评估方法

针对 CHI 指标专利价值评估方法的缺陷, 本文提出了一种基于潜在引用网络的专利价值评估方法. 在该方法中, 首先利用专利间的相似度建立了专利潜在引用关联, 在此基础上同时考虑直接引用关联、间接引用关联及专利的新颖度对专利价值的影响, 提出了专利价值评估基本算法. 针对基本算法时间花费较高的问题, 提出了专利价值评估改进算法, 提高了算法的效率.

2.1 潜在引用关联

由于中文专利中不存在引用信息, 导致基于引用的方法很难应用到中文专利的价值评估中. 此外, 跟相似度关联相比, 引用关联很少考虑专利之间的潜在语义关系, 并且由于专利申请者的知识面有限, 可能会存在一些误引用或缺少引用的情况, 导致基于引用关联的分析方法往往得不到良好的效果^[20]. 基于以上原因, 本文根据专利间的相似度及发表时间提出了专利间的潜在引用关联, 即如果 2 个专利文档的相似度大于给定的相似度阈值, 那么后申请的专利潜在引用先申请的专利. 跟引用关联相比, 基于相似度关联的潜在引用关联不仅可以很好地模拟专利引用关联, 而且考虑了专利之间的潜在语义关系.

定义 1. 潜在引用关联(latent citation). 专利集合 $D=\{d_1, d_2, \dots, d_n\}$, 对应的专利发表时间的集合 $T=\{T_1, T_2, \dots, T_n\}$, 对于任意 2 个专利文档 d_i 和 d_j , 如果 d_i 和 d_j 之间的相似度 $sim(d_i, d_j)$ 大于等于给定的相似度阈值 α , 且专利 i 的发表时间大于专利 j 的发表时间, 则存在专利 i 到专利 j 的潜在引用关联; 否则不存在专利 i 到专利 j 的潜在引用关联.

$$lc\langle d_i, d_j \rangle = \begin{cases} 1, & sim(d_i, d_j) \geq \alpha \text{ and } T_i > T_j, \\ 0, & \text{else,} \end{cases} \quad (1)$$

其中, $lc\langle d_i, d_j \rangle = 1$ 表示专利 i 潜在引用专利 j , $lc\langle d_i, d_j \rangle = 0$ 表示不存在专利 i 到专利 j 的潜在引用关联.

例 2. 专利文档集合 $D=\{a,b,c\}$, 对应的专利

发表年份分别为 2005, 2007 和 2001. 给定相似度阈值为 0.5, 专利文档之间的相似度分别为 $sim(a, b) = 0.6$, $sim(a, c) = 0.7$, $sim(b, c) = 0.3$. 根据式(1), 因为 $sim(a, b) > 0.5$, 且专利 b 的发表时间 2007 大于专利 a 的发表时间 2005, 则专利 b 到专利 a 的潜在引用关联 $lc\langle b, a \rangle = 1$. 类似地, $lc\langle a, c \rangle = 1$, $lc\langle b, c \rangle = 0$.

根据定义 1, 要计算专利之间的潜在引用关系, 首先需要计算专利文档之间的相似度. 常用的文档之间的相似度计算方法包括 Jaccard 系数、余弦相似度等等, 本文中采用余弦相似度来度量专利文档之间的相似度. 具体步骤如下:

1) 用一组关键词的空间向量来表示专利集合 D 中的每一篇专利文档.

采用现有的中文分词软件, 如中国科学院分词软件 ICTCLAS^[21] 对所有的专利文档进行分词; 根据公用的停用词词库去除文档中的停用词, 其中停用词为没有实际含义的功能词, 如“和”、“一个”、“的”等等; 对于剩余的词项, 根据 TF-IDF^[2] 公式计算每个词项在各个专利文档中的权重:

$$w_{ik} = \frac{tf(t_k, d_i) \times \text{lb}(N/n_{t_k})}{\sqrt{\sum_{t_k \in d_i} [tf(t_k, d_i) \times \text{lb}(N/n_{t_k})]^2}}, \quad (2)$$

其中, w_{ik} 为词项 t_k 在专利文档 d_i 中的权重, $tf(t_k, d_i)$ 为词项 t_k 在专利 d_i 中的词频, N 为专利集合 D 的大小, n_{t_k} 为专利集合 D 中出现词项 t_k 的专利文档数.

得到各个词项的 TF-IDF 权重后, 采用空间向量表示每一篇专利文档 d_i , 记作 $\mathbf{d}_i = \langle w_{i1}, w_{i2}, \dots, w_{in} \rangle$, 其中 n 为关键词的数目.

2) 采用余弦相似度计算公式计算两两专利文档之间的相似度. 其计算公式为

$$sim(d_i, d_j) = \frac{\sum_{k=1}^n w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^n w_{ik}^2) \times (\sum_{k=1}^n w_{jk}^2)}}, \quad (3)$$

其中, w_{ik} 为第 k 个词项在专利文档 d_i 中的权重, w_{jk} 为第 k 个词项在专利文档 d_j 中的权重.

2.2 专利价值评估基本算法

获得专利间的潜在引用关联之后, 可以建立专利集合的潜在引用网络. 对于专利集合 $D = \{d_1, d_2, \dots, d_n\}$, 建立潜在引用网络 $G = (V, E)$, 在该网络中, 任意节点 v_i 对应于 D 中的专利文档 d_i , 节点之间的边表示专利文档之间的潜在引用关联, 边上的权重表示专利之间的相似度. 即对于任意 2 个

专利文档 d_i 和 d_j , 如果 d_i 到 d_j 的潜在引用关联 $lc\langle d_i, d_j \rangle = 1$, 则在 G 中建立 v_i 到 v_j 的有向边, 边上的权重为 d_i 与 d_j 的相似度, 即:

$$w\langle v_i, v_j \rangle = \begin{cases} sim(d_i, d_j), & lc\langle d_i, d_j \rangle = 1, \\ 0, & \text{else}, \end{cases} \quad (4)$$

其中, $w\langle v_i, v_j \rangle = sim(d_i, d_j)$ 表示存在 v_i 指向 v_j 的边, 且边上的权重为 d_i 与 d_j 之间的相似度; $w\langle v_i, v_j \rangle = 0$ 表示不存在 v_i 到 v_j 的边. 可知, 边上的权重代表专利之间的潜在引用关联的强度, 即 2 个专利文档的相似度越高, 它们之间的潜在引用关联越强.

例 3. 图 2 表示了一个潜在引用网络 $G = (V, E)$, 每一个节点代表相应的专利文档, 边表示专利间的潜在引用关联, 边上的权重为专利文档之间的相似度, 其中相似度阈值 $\alpha = 0.3$. 例如 v_1 对应于专利文档 d_1 , v_2 指向 v_1 的边表示专利文档 d_2 潜在引用 d_1 , 且专利文档 d_1 与 d_2 的相似度为 0.5.

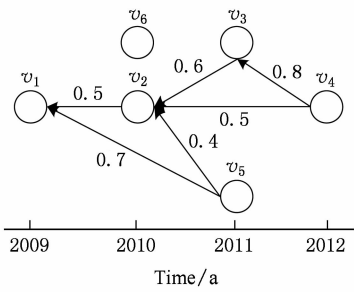


Fig. 2 An example of patent latent-citation-network.
图 2 专利潜在引用网络示例

对于建立好的潜在引用网络, 本文考虑从直接和间接潜在引用关联、新颖度、潜在引用关联的强度 4 个方面衡量每一个专利的价值. 本文中提出的专利价值评估方法基于 4 个经验结论:

- 1) 专利被直接和间接引用的越多, 该专利的价值越高.
- 2) 专利的发表时间距离当前时间的间隔越长, 其专利价值会相应地降低.
- 3) 专利之间的相似度越高, 则潜在引用关联越强.
- 4) 所有专利的专利价值都不为零.

基于以上结论, 本文首先定义了专利的初始价值, 以保证当专利不被其他专利直接或间接引用时, 其专利价值仍不为 0. 之后, 对于每一个对其进行直接或间接潜在引用的专利, 计算它们对该专利的价值的贡献度. 最后, 综合考虑专利的初始价值及其他专利对该专利价值的贡献度, 计算各个专利的价值. 其具体方法如下:

假设专利价值随时间指数衰减,首先初始化各个专利的初始价值,其计算如下:

$$Ipv(v_i) = 1 \times e^{-\beta t}, \quad (5)$$

其中, $Ipv(v_i)$ 为 v_i 在当前时间的初始价值,“1”表示发表时间相同的各个专利的初始价值相同, β 为给定的价值衰减因子, t 为专利 i 从发表时间到当前时间的间隔,则:

$$t = T_0 - T_i, \quad (6)$$

其中, T_i 为专利 i 的发表时间, T_0 为当前时间.

例 3 中,假设当前的时间为 2012 年,给定价值衰减因子 $\beta=0.5$,则初始化 v_4 的专利价值 $Ipv(v_4) = 1 \times e^{-0.5 \times (2012-2012)} = 1$. 类似地, $Ipv(v_3) = e^{-0.5}$, $Ipv(v_5) = e^{-0.5}$, $Ipv(v_2) = e^{-1}$, $Ipv(v_6) = e^{-1}$, $Ipv(v_1) = e^{-1.5}$.

从图 2 中可以看出,任意 2 个节点 v_j 和 v_i 之间的直接和间接潜在引用关联在网络中反映为一条从 v_j 到 v_i 的路径,且可能存在多条 v_j 到 v_i 的路径. 例如 v_4 到 v_2 存在 2 条路径,其中路径 $\langle v_4, v_2 \rangle$ 为 v_4 到 v_2 的直接引用关联,路径 $\langle v_4, v_3, v_2 \rangle$ 为 v_4 到 v_2 的间接引用关联. 根据专利间的关联,本文定义了专利贡献度.

定义 2. 专利贡献度(patent contribution). 对于潜在引用网络中的任意节点 v_i 和 v_j ,如果存在 v_j 到 v_i 的路径,则 v_j 对 v_i 的专利贡献度为 v_j 到 v_i 的所有路径对 v_i 的专利价值的影响之和,记作 $pc(v_j, v_i)$,则:

$$pc(v_j, v_i) = \sum_{p_m \in P(v_j, v_i)} pvc(v_j, v_i | p_m), \quad (7)$$

其中, $P(v_j, v_i)$ 为 v_j 到 v_i 的所有路径的集合, p_m 表示 $P(v_j, v_i)$ 中的任意一条路径. $pvc(v_j, v_i | p_m)$ 为在路径 p_m 上 v_j 对 v_i 的路径价值贡献度,其定义如下:

定义 3. 路径价值贡献度(path value contribution). 对于潜在引用网络中的任意节点 v_i, v_j ,如果存在一条路径 $p_m = \langle v_j, v_{j+1}, v_{j+2}, \dots, v_{i-1}, v_i \rangle$ (假设 $j < i$),那么 v_j 在路径 p_m 上对 v_i 的专利价值造成的影响称为 v_j 在路径 p_m 上对 v_i 的路径价值贡献度,记作 $pvc(v_j, v_i | p_m)$,则:

$$pvc(v_j, v_i | p_m) = Ipv(v_j) \times e^{-\beta(T_j - T_i)} \times \prod_{k=j}^{i-1} \frac{\text{sim}(d_k, d_{k+1})}{\text{deg}(v_k)}, \quad (8)$$

其中, $\text{deg}(v_k)$ 为节点 v_k 的出度, $(T_j - T_i)$ 为专利 v_i 和 v_j 的专利发表时间的间隔.

路径价值贡献度的定义采用随机游走模型的原理, $1/\text{deg}(v_k)$ 表示当前节点具有相等的概率将其影

响度转移到与它相邻的节点. 此外, $e^{-\beta(T_j - T_i)}$ 反映了专利 v_j 与 v_i 的发表时间的间隔对路径价值贡献度的影响,时间间隔越大则表示 v_i 对 v_j 的影响传播速度越慢,路径价值贡献度越小; $\text{sim}(d_k, d_{k+1})$ 反映了专利间的语义关联对路径价值贡献度的影响,专利间的语义相似度越高,则路径价值贡献度越大.

例 3 中, v_4 在路径 $p_1 = \langle v_4, v_3, v_2 \rangle$ 上对专利节点 v_2 的路径价值贡献度 $pvc(v_4, v_2 | p_1) = 1 \times e^{-0.5(2012-2010)} \times 0.4 \times 0.6 = 0.24e^{-1}$. 类似地, v_4 在路径 $p_2 = \langle v_4, v_2 \rangle$ 上对 v_2 的路径价值贡献度 $pvc(v_4, v_2 | p_2) = 0.25e^{-1}$. 根据定义 2, v_4 对 v_2 的专利贡献度 $pc(v_4, v_2) = 0.24e^{-1} + 0.25e^{-1} = 0.49e^{-1}$. 同理, v_3 对 v_2 的专利贡献度 $pc(v_3, v_2) = 0.6e^{-1}$, v_5 对 v_2 的专利贡献度 $pc(v_5, v_2) = 0.2e^{-1}$.

定义 4. 专利价值(patent value). 对于潜在引用网络中的任意节点 v_i ,其专利价值为 v_i 初始专利价值与所有与 v_i 存在直接和间接潜在引用关联的节点对 v_i 的专利贡献度之和,记作 $pvalue(v_i)$. 则:

$$pvalue(v_i) = Ipv(v_i) + \sum_{\{v_j \parallel P(v_j, v_i) \mid \neq 0\}} pc(v_j, v_i), \quad (9)$$

其中, $P(v_j, v_i)$ 为 v_j 到 v_i 的所有路径的集合.

根据专利价值的定义,任意专利 v_i 的价值由 v_i 的初始价值及其他专利对 v_i 的专利价值贡献度决定,其中 v_i 的初始价值保证了在不存在直接和间接潜在引用 v_i 的专利时, v_i 的专利价值不为 0.

例 3 中,专利节点 v_2 的专利价值为 v_2 的初始价值与 v_3, v_4 和 v_5 对 v_2 的专利贡献度之和,即 $pvalue(v_2) = e^{-1} + 0.6e^{-1} + 0.49e^{-1} + 0.2e^{-1} = 2.29e^{-1}$. 类似, $pvalue(v_1) = 2.495e^{-1.5}$, $pvalue(v_3) = 1.4e^{-0.5}$, $pvalue(v_4) = 1$, $pvalue(v_5) = e^{-0.5}$, $pvalue(v_6) = e^{-1}$.

根据定义 2 和定义 4,任意节点 v_i 的专利价值等于 $Ipv(v_i)$ 与所有和 v_i 存在关联的节点到 v_i 的路径对 v_i 的路径价值贡献度之和,即:

$$pvalue(v_i) = Ipv(v_i) + \sum_{\{v_j \parallel P(v_j, v_i) \mid \neq 0\}} \sum_{p_m \in P(v_j, v_i)} pvc(v_j, v_i | p_m).$$

由于和 v_i 没有关联的节点不存在到 v_i 的路径,所以和 v_i 存在关联的节点到 v_i 的所有路径即潜在网络中所有可以抵达 v_i 路径的集合,因此 v_i 的专利价值也可以表示为 v_i 初始专利价值与所有可以抵达 v_i 的路径对 v_i 的路径价值贡献度之和,即:

$$pvalue(v_i) = Ipv(v_i) + \sum_{p_m \in P(v_i)} pvc(v_i | p_m), \quad (10)$$

其中, $P(v_i)$ 表示可以抵达 v_i 的所有路径的集合, p_m 为 $P(v_i)$ 中的任意一条路径, $pvc(v_i | p_m)$ 为可以抵达 v_i 的任意路径对 v_i 的路径价值贡献度.

根据式(10), 例 3 中节点 v_2 的专利价值为 v_2 的初始价值与路径 $\langle v_4, v_3, v_2 \rangle, \langle v_4, v_2 \rangle, \langle v_3, v_2 \rangle, \langle v_5, v_2 \rangle$ 对 v_2 的路径价值贡献度之和, 即

$$pvalue(v_2) = e^{-1} + 0.24e^{-1} + 0.25e^{-1} + 0.6e^{-1} + 0.2e^{-1} = 2.29e^{-1}.$$

根据定义 2~4, 基于潜在引用网络的专利价值评估基本算法首先根据专利发表时间计算专利的初始价值, 然后在潜在引用网络中找到其他专利节点到该专利节点的所有路径, 并计算每条路径对该专利的路径价值贡献度. 最后累加专利的初始价值和所有路径价值贡献度来得到专利的价值. 其具体算法如算法 1 所示:

算法 1. 专利价值评估基本算法 (basic patent value evaluation, BPVE).

输入: 潜在引用网络 $G=(V, E)$, 专利发表的时间集合 $T=\{T_1, T_2, \dots, T_n\}$;

输出: 各个专利的价值.

- ① 根据式(5)(6)初始化各个专利的价值;
- ② for 每一个专利 v_i
- ③ 找出所有可以抵达 v_i 的路径的集合 $P(v_i)$;
- ④ for $P(v_i)$ 中的每一条路径 p_m
- ⑤ 根据式(8)计算 p_m 对 v_i 的路径价值贡献度;
- ⑥ endfor
- ⑦ 根据式(10)计算专利 v_i 的专利价值.
- ⑧ endfor

2.3 专利价值评估改进算法

尽管 BPVE 算法可以有效地评估算法的价值, 但该算法首先要在潜在网络中找出可以抵达目标专利节点的所有路径, 因此需要耗费大量的时间. 实际上, 可以通过专利价值的特性对算法进行改进.

从图 2 中可以看出, 对于任意专利节点 v_i , 路径 p_m 对该专利的路径价值贡献度通过 v_i 的邻接节点传递给 v_i , 因此本文定义了传递价值度.

定义 5. 传递价值度 (transmitted patent value). 对于潜在引用网络中的任意 2 个专利节点 v_i 和 v_j , 所有可以抵达 v_i 且以 v_j 为倒数第 2 跳节点的路径的集合为 P_j , 那么 v_j 对 v_i 的传递价值度 $tpv(v_j, v_i)$ 定义为 P_j 中的所有路径对 v_i 的路径价值贡献度之和, 即:

$$tpv(v_j, v_i) = \sum_{k=1}^{|P_j|} pvc(v_i | p_{jk}), \quad (11)$$

其中, p_{jk} 为 P_j 中的第 k 条路径, $|P_j|$ 表示集合 P_j 中的路径数目.

根据传递价值度的定义, 我们证明了下面 2 个定理.

定理 1. 对于潜在引用网络 $G=(V, E)$, v_i 是 G 中的任意一个专利节点, 假设专利节点集合 C 为指向 v_i 的所有邻接节点的集合, 那么专利节点 v_i 的专利价值等于 v_i 的初始专利价值与 C 中所有节点对 v_i 的传递价值度之和.

证明. 根据式(10), 专利节点 v_i 的专利价值等于 v_i 的初始专利价值与所有可以抵达 v_i 的路径对 v_i 的路径价值贡献度之和. 假设 $P=\{p_1, p_2, \dots, p_n\}$ 为所有可以抵达 v_i 的路径的集合, $C=\{v_1, \dots, v_k\}$ 是指向 v_i 的所有邻接节点的集合, 那么 P 中所有路径的倒数第 2 跳节点一定属于 C . 按路径的倒数第 2 跳节点划分, P 可以划分 k 个不相交的集合 $P_1, P_2, \dots, P_k$, 其中 P_j 表示以 v_j 为倒数第 2 个节点的路径的集合. 根据定义 5, P_j 中所有路径对 v_i 的路径价值贡献度之和即为 v_j 对 v_i 的传递价值度. 那么 C 中所有节点对 v_i 的传递价值度之和等于 $P_1, P_2, \dots, P_k$ 中所有路径对 v_i 的路径价值贡献度之和. 因此, 节点 v_i 的专利价值等于 v_i 的初始专利价值与 C 中所有节点对 v_i 的传递价值度之和. 证毕.

根据定理 1, 对于潜在网络中的任意专利节点 v_i , 其专利价值的计算为

$$pvalue(v_i) = Ipv(v_i) + \sum_{j=1}^{|C|} tpv(v_j, v_i), \quad (12)$$

其中, C 为指向 v_i 的所有邻接节点的集合, $|C|$ 表示集合 C 的大小.

根据式(12), 例 3 中专利节点 v_1 的专利价值也可以表示为

$$pvalue(v_1) = Ipv(v_1) + tpv(v_2, v_1) + tpv(v_5, v_1).$$

定理 2. 对于潜在引用网络 $G=(V, E)$ 中的任意专利节点 v_i , v_{i-1} 是指向 v_i 的任意邻接节点, 那么 v_{i-1} 对 v_i 的传递价值度 $tpv(v_{i-1}, v_i)$ 等于节点 v_{i-1} 的专利价值 $pvalue(v_{i-1})$ 与 $\frac{sim(d_{i-1}, d_i) \times e^{-\beta(T_{i-1}-T_i)}}{deg(v_{i-1})}$

的乘积; 其中 $sim(d_{i-1}, d_i)$ 为专利节点 v_{i-1} 与 v_i 的文本相似度, $deg(v_{i-1})$ 为节点 v_{i-1} 的出度, T_i 和 T_{i-1} 分别为专利 v_i 和 v_{i-1} 的发表时间.

证明. 假设可以抵达 v_{i-1} 的路径的集合为 P' , 可以抵达 v_i 且以 v_{i-1} 为倒数第 2 跳节点的所有路

径的集合为 P_{i-1} , 那么如果 P' 中存在一条路径 $p'_l = \langle v_1, v_2, \dots, v_{i-1} \rangle$, 则在 P_{i-1} 中必定存在相应的路径 $p_{(i-1,l)} = \langle v_1, v_2, \dots, v_{i-1}, v_i \rangle$. 但由于路径 $\langle v_{i-1}, v_i \rangle$ 属于 P_{i-1} , 且在 P' 中不存在对应的路径, 因此 $|P_{i-1}| = |P'| + 1$. 除路径 $\langle v_{i-1}, v_i \rangle$ 外, P_{i-1} 中任意一条路径 $p_{(i-1,l)}$ 对 v_i 的路径价值贡献度为

$$\begin{aligned} pvc(v_i | p_{(i-1,l)}) &= Ipv(v_1) \times \\ &e^{-\beta(T_1-T_i)} \times \prod_{k=1}^{i-1} \frac{sim(d_k, d_{k+1})}{deg(v_k)} = \\ &\frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times e^{-\beta(T_{i-1}-T_i)} \times Ipv(v_1) \times \\ &e^{-\beta(T_1-T_{i-1})} \times \prod_{k=1}^{i-2} \frac{sim(d_k, d_{k+1})}{deg(v_k)} = \\ &\frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times e^{-\beta(T_{i-1}-T_i)} \times pvc(v_{i-1} | p'_l). \end{aligned}$$

路径 $\langle v_{i-1}, v_i \rangle$ 对 v_i 的路径价值贡献度为

$$\begin{aligned} pvc(v_i | \langle v_{i-1}, v_i \rangle) &= Ipv(v_{i-1}) \times \\ &e^{-\beta(T_{i-1}-T_i)} \times \frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})}. \end{aligned}$$

那么 v_{i-1} 对 v_i 的传递价值度为

$$\begin{aligned} tpv(v_{i-1}, v_i) &= \sum_{l=1}^{|P_{i-1}|} pvc(v_i | p_{(i-1,l)}) = \\ &Ipv(v_{i-1}) \times \frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times e^{-\beta(T_{i-1}-T_i)} + \\ &\sum_{l=1}^{|P'|} \frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times e^{-\beta(T_{i-1}-T_i)} \times pvc(v_{i-1} | p'_l) = \\ &\frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times e^{-\beta(T_{i-1}-T_i)} \times (Ipv(v_{i-1}) + \\ &\sum_{l=1}^{|P'|} pvc(v_{i-1} | p'_l)). \end{aligned}$$

可知:

$$Ipv(v_{i-1}) + \sum_{l=1}^{|P'|} pvc(v_{i-1} | p'_l) = pvalue(v_{i-1}),$$

则:

$$\begin{aligned} tpv(v_{i-1}, v_i) &= \frac{sim(d_{i-1}, d_i)}{deg(v_{i-1})} \times \\ &e^{-\beta(T_{i-1}-T_i)} \times pvalue(v_{i-1}). \quad \text{证毕.} \end{aligned}$$

根据定理 1 和定理 2, 对于潜在网络中的任意专利节点 v_i , 假设 $C = \{v_1, \dots, v_m\}$ 是指向 v_i 的所有邻接节点的集合, 那么专利 v_i 的专利价值计算公式可以表示为

$$\begin{aligned} pvalue(v_i) &= Ipv(v_i) + \sum_{k=1}^m \frac{sim(d_i, d_k)}{deg(v_k)} \times \\ &e^{-\beta(T_k-T_i)} \times pvalue(v_k). \end{aligned} \quad (13)$$

根据潜在引用网络的构建方法, 对于网络中的任意 2 个节点 v_i 和 v_j , 如果 v_i 的发表时间 T_i 小于

v_j 的发表时间 T_j , 那么一定不存在 v_i 到 v_j 的路径, 即 v_i 不会影响 v_j 的专利价值. 因此, 可以首先对专利节点按发表时间排序, 然后根据式 (13) 依次计算各个专利的价值. 其算法如下:

算法 2. 专利价值评估改进算法(improved patent value evaluation, IPVE).

输入: 潜在引用网络 $G = (V, E)$, 发表时间集合 $T = \{T_1, T_2, \dots, T_n\}$;

输出: 各个专利的价值.

- ① 根据式 (5) (6) 初始化各个专利的价值;
- ② 将所有专利节点按专利的发表时间从大到小排序;
- ③ for 对排序后的每一个节点 v_i
- ④ 根据式 (13) 依次计算 v_i 的专利价值.
- ⑤ endfor

如算法 2, 对于图 2 中的潜在引用网络, 首先根据式 (5) (6) 初始化 $v_1, v_2, v_3, v_4, v_5, v_6$ 的价值分别为 $e^{-1.5}, e^{-1}, e^{-0.5}, 1, e^{-0.5}, e^{-1}$. 之后, 将专利节点按专利的发表时间从大到小进行排序, 得到的次序为 $v_4, v_3, v_5, v_2, v_6, v_1$. 最后, 根据式 (13) 依次计算各个专利的价值, 得到 $pvalue(v_4) = 1, pvalue(v_3) = 1.4e^{-0.5}, pvalue(v_5) = e^{-0.5}, pvalue(v_2) = 2.29e^{-1}, pvalue(v_6) = e^{-1}, pvalue(v_1) = 2.495e^{-1.5}$.

从算法 2 中可以看出, IPVE 算法只需要首先对专利文档按发表时间进行排序, 然后依次计算各个专利节点的价值. 假设潜在网络中的专利节点数目为 V , 边的数目为 E , m 为专利节点的平均入度, 则 $E = mV$. 使用快速排序的方法对所有专利文档进行排序的时间复杂度为 $O(V \log V)$, 计算所有专利节点价值的复杂度为 $O(mV) = O(E)$, 因此 IPVE 算法的时间复杂度为 $O(V \log V + E)$.

3 专利价值评估动态更新算法

IPVE 算法可以评估在当前时间 t 时专利集合中各个专利的价值. 然而, 当有新的专利在时间 $t+1$ 加入专利集合时, 各个专利的价值将发生改变. 尽管可以在时间 $t+1$ 将新的专利加入潜在引用网络, 并使用 IPVE 算法重新计算时间 $t+1$ 各个专利的价值, 但这种方法存在大量重复的计算. 因此, 需要提出一种专利价值评估的动态更新算法, 快速地更新时间 $t+1$ 下各个专利的价值.

由 2.2 节的分析可知, 专利的价值与该专利的潜在引用关联以及发表时间与当前时间的间隔有

关. 因此, 当新的专利在时间 $t+1$ 加入时, 各个专利的价值会受到 2 方面的影响:

1) 专利的价值随其发表时间与当前时间间隔的增加而衰减.

2) 新专利对其的直接和间接引用造成其专利价值的变化.

基于以上分析, 可以在时间 $t+1$ 不重新建立潜在引用网络, 而是采用动态更新的方法计算时间 $t+1$ 每一个专利的价值.

当专利 v_j 在时间 $t+1$ 加入时, 首先依据式(5)(6)初始化 v_j 的专利价值为 $Ipv^{(t+1)}(v_j) = 1 \times e^{-\beta(T_j - T_0)}$, 其中 T_0 为时间 $t+1$ 的当前时间, T_j 为专利 v_j 的发表时间. 因为 $T_0 = T_j$, 所以 v_j 的初始价值为 1. 之后, 将 v_j 加入到潜在引用网络中, 其步骤为: 计算专利 v_j 与时间 t 的专利集合中所有专利的相似度, 如果两者的相似度大于等于给定的相似度阈值, 则在潜在引用网络中建立 v_j 到该专利节点的边.

将 v_j 加入到潜在引用网络后, 首先找出与 v_j 存在路径的节点集合 V_j . 根据上述经验结论, 对不属于 V_j 的节点, 其专利价值只受到价值随时间衰减的影响, 它在时间 $t+1$ 的专利价值为时间 t 的专利价值与 $e^{-\beta}$ 的乘积, 即:

$$pvalue^{(t+1)}(v_i) = pvalue^{(t)}(v_i) \times e^{-\beta}. \quad (14)$$

对于 V_j 中的节点, 新加入的专利 v_j 会对其造成影响. 因此, 同 IPVE 算法, 根据式(5)(6)初始化各个节点在时间 $t+1$ 的初始价值, 并将 V_j 中所有专利按发表时间从大到小排序, 然后根据式(13)按序依次计算 V_j 中各个专利的价值.

由以上步骤可知, 任意节点 v_i 在时间 $t+1$ 专利价值的计算公式可表示为

$$pvalue^{(t+1)}(v_i) = \begin{cases} pvalue^{(t)}(v_i) \times e^{-\beta}, & v_i \notin V_j, \\ Ipv^{(t+1)}(v_i) + \sum_{k=1}^m \frac{sim(d_i, d_k) \times e^{-\beta(T_k - T_j)}}{deg^{(t+1)}(v_k)} \times (15) \\ pvalue^{(t+1)}(v_k), & \text{otherwise,} \end{cases}$$

其中, $deg^{(t+1)}(v_k)$ 为时间 $t+1$ 专利节点 v_k 在潜在引用网络中的出度, $Ipv^{(t+1)}(v_i)$ 为时间 $t+1$ 时 v_i 的专利初始价值.

基于以上步骤, 专利价值评估动态更新算法的描述如算法 3:

算法 3. 专利价值评估动态更新算法(dynamic patent value evaluation, DPVE).

输入: 专利集合 D 、时间 t 各个专利的价值的集

合 $Value = \{pvalue^{(t)}(v_1), \dots, pvalue^{(t)}(v_n)\}$ 、潜在引用网络 $G = (V, E)$ 、发表时间集合 $T = \{T_1, T_2, \dots, T_n\}$ 、时间 $t+1$ 新加入的专利 v_j ;

输出: $t+1$ 时间下各个专利的价值.

- ① 初始化 v_j 的专利价值为 1;
- ② for 专利集合 D 中的每一个专利 v_i
- ③ 计算专利节点 v_i 与 v_j 的文本相似度 $sim(d_i, d_j)$;
- ④ if $sim(d_i, d_j) \geq$ 给定的相似度阈值
- ⑤ 建立 v_j 到 v_i 的边;
- ⑥ endif
- ⑦ endfor
- ⑧ 找出专利节点集合中与 v_j 存在路径的节点的集合 V_j ;
- ⑨ for 不属于 V_j 的每一个节点 v_i
- ⑩ 根据式(14)计算 v_i 的价值;
- ⑪ endfor
- ⑫ 对 V_j 中的专利节点按发表时间从大到小排序;
- ⑬ 根据式(5)(6)初始化 V_j 中各个专利 v_k 在时间 $t+1$ 的初始价值;
- ⑭ for V_j 中的每一个节点 v_k
- ⑮ 根据式(13)依次计算 v_k 的专利价值.
- ⑯ endfor

定理 3. 对于专利集合中的任意专利 v_i , 采用 DPVE 算法得到的专利价值与在时间 $t+1$ 采用 BPVE 算法计算得到的专利价值相同.

证明. 对于潜在引用网络 $G = (V, E)$, 假设时间 $t+1$ 加入的专利为 v_j , 其初始专利价值为 1. 根据 DPVE 算法, 首先找出与 v_j 存在直接或间接引用关联的节点集合 V_j .

对于不属于 V_j 的任意节点 v_i , 若采用 BPVE 算法分别计算时间 t 和时间 $t+1$ 时 v_i 的专利价值, 由于时间 $t+1$ 加入的专利 v_j 与 v_i 不存在潜在直接或间接引用关联, 因此时间 t 可以抵达 v_i 的路径与时间 $t+1$ 可以抵达 v_i 的路径相同. 但由于时间 $t+1$ 专利的初始专利价值等于时间 t 的初始专利价值与 $e^{-\beta}$ 的乘积, 即 $Ipv^{(t+1)}(v_i) = Ipv^{(t)}(v_i) \times e^{-\beta}$, 因此根据式(8), 任意一条路径 p_m 对 v_i 在时间 $t+1$ 的路径价值贡献度等于时间 t 路径价值贡献度与 $e^{-\beta}$ 的乘积, 即 $pvc^{(t+1)}(v_i | p_m) = pvc^{(t)}(v_i | p_m) \times e^{-\beta}$. 之后, 根据式(10)可得, v_i 在时间 $t+1$ 的专利价值等于时间 t 专利价值与 $e^{-\beta}$ 的乘积, 即 $pvalue^{(t+1)}(v_i) = pvalue^{(t)}(v_i) \times e^{-\beta}$. 根据式(14)可知, 该结果与

在时间 $t+1$ 采用 DPVE 算法得到的 v_i 的专利价值相同. 因此, 对于不属于 V_j 的节点, 定理 3 成立.

对于 V_j 中的任意节点 v_i , 直接采用 IPVE 算法重新计算时间 $t+1$ 时 v_i 的专利价值. 因此, 对于 V_j 中的任意节点 v_i , 定理 3 成立. 证毕.

如图 3 是图 2 在时间 $t+1$ 的潜在引用网络, v_7 是在时间 $t+1$ 新加入的专利. 由于 v_7 在时间 $t+1$ 不会被直接或间接引用, 根据式 (13), $pvalue^{(t+1)}(v_7) = Ip v^{(t+1)}(v_7) + 0 = 1$. 根据算法 3, 将 v_7 加入潜在网络, 其中虚线为时间 $t+1$ 建立的边. 之后, 找出 v_7 直接或间接引用的节点集合 $V_7 = \{v_1, v_2, v_5\}$.

对于不属于 V_7 的节点 v_3, v_4 和 v_6 , 按式 (13) 计算其专利价值得 $pvalue^{(t+1)}(v_3) = pvalue^{(t)}(v_3) \times e^{-0.5} = 1.4e^{-1}$, $pvalue^{(t+1)}(v_4) = e^{-0.5}$, $pvalue^{(t+1)}(v_6) = e^{-1.5}$. 对 V_7 中的节点, 首先按专利的发表时间从大到小排序, 得 $V_7 = \{v_5, v_2, v_1\}$. 之后, 根据式 (5) (6) 分别初始化 v_5, v_2, v_1 的初始专利价值为 $e^{-1}, e^{-1.5}, e^{-2}$, 并按式 (13) 依次计算 v_5, v_2, v_1 的专利价值, 得 $pvalue^{(t+1)}(v_5) = Ip v^{(t+1)}(v_5) + pvalue^{(t+1)}(v_7) \times e^{-1} \times 0.5 = 1.5e^{-1}$, $pvalue^{(t+1)}(v_2) = 2.39e^{-1.5}$, $pvalue^{(t+1)}(v_1) = 2.72e^{-2}$.

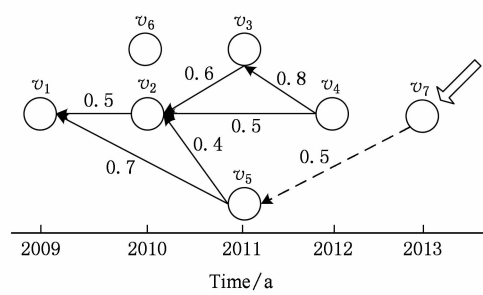


Fig. 3 A latent-citation-network at time $t+1$.
图 3 时间 $t+1$ 的潜在引用网络

由图 3 中的例子可知, DPVE 算法不需要在新专利加入时根据 IPVE 算法重新计算每一个专利的价值, 而是只重新计算那些与新加入专利存在直接或间接潜在引用关联的专利的价值, 对不存在关联的专利进行简单的基于时间衰减的价值更新. 因此与在时间 $t+1$ 重新计算所有节点专利价值的 IPVE 算法相比, DPVE 算法降低了专利价值评估的时间复杂度.

4 实验结果与分析

为了验证本文方法的有效性和效率, 本文在真实数据集上分别实现了 BPVE, IPVE 和 DPVE 算

法, 对算法的效果进行了评估, 并测试了算法的运行效率. 本文使用的数据为从中国知识产权局下载的 2003—2012 年期间专利分类号 G06T 下的 14 582 条专利数据. 所有算法均采用 C++ 编程, 运行在 2.5 GHz 的双核处理器上, 内存为 4 GB.

4.1 专利价值评估的有效性对比

本节首先比较了本文提出的 BPVE 算法、文献 [18-19] 中使用的 CHI 引证指数及经典的基于链接分析来度量网页重要度的 PageRank 算法 3 种方法的专利价值评估效果. 由于中文专利中缺乏引文信息, 无法直接应用引证指数和 PageRank 算法 [22] 计算专利的价值, 因此本文首先构建了专利之间的潜在引用关联, 之后采用引证指数和 PageRank 算法计算各个专利的价值.

一般来说, 曾被购买或转让的专利具有较高的专利价值. 因此, 本文通过比较引证指数、PageRank 算法与 BPVE 算法的实验结果中曾被转让专利所占的比例来评估 3 种方法的有效性. 本文分别采用 3 种方法计算了各个专利的价值, 并对专利按照其价值从大到小排序, 通过排序列表前 k 个价值最高的专利中曾被转让的专利占有所有被转让专利的比例来评估 3 种方法的专利价值评估效果, 前 k 个价值最高的专利中曾被转让专利的比例越高, 则价值评估的效果越好.

在比较 3 种方法的价值评估效果时, 本文首先将数据集分为测试集和训练集, 其中训练集为 2003—2007 年之间发表的专利, 用来训练参数的值, 2008—2012 年之间发表的专利为测试集, 用来比较 3 种算法的价值评估效果.

图 4 是反映了在测试集上引证指数、PageRank 算法和 BPVE 算法的运行结果中价值最高的前 k

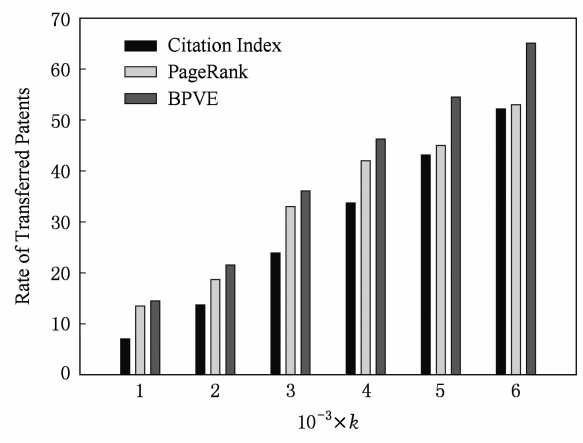


Fig. 4 Comparisons on effectiveness.
图 4 算法有效性上的对比

个专利中被转让专利所占的比例,可以看到,在 $k=1\,000, 2\,000, 3\,000, 4\,000, 5\,000, 6\,000$ 时, BPVE 算法中被转让专利所占的比例都高于引证指数和 PageRank 算法,即 BPVE 算法具有比引证指数和 PageRank 算法更好的价值评估效果.

4.2 时间效率上的对比

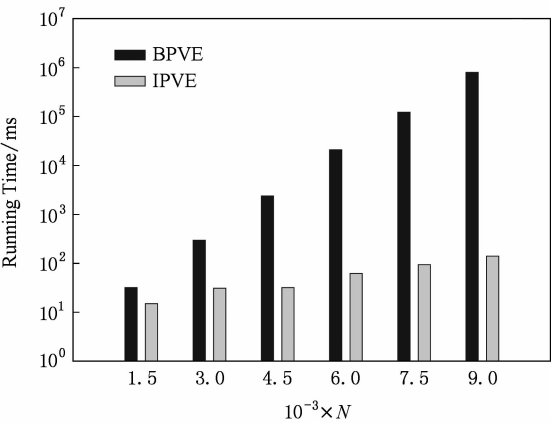
本文比较了 BPVE 算法和 IPVE 算法的运行效率以及 DPVE 算法与 IPVE 算法在新专利加入时进行价值更新的运行效率. 从 2 个算法的步骤可知, β 并不会对算法的运行时间造成影响. 因此,在以下的实验对比中,只比较了不同的数据量和相似度阈值下算法的时间效率.

图 5 首先比较了 BPVE 算法和 IPVE 算法的运行效率. 图 5(a)是在相似度阈值 $\alpha=0.3$ 时 BPVE 算法和 IPVE 算法的时间效率对比. 可以看到, BPVE 算法和 IPVE 算法的运行时间都随专利集合的增大而单调递增,但 IPVE 算法的运行时间的增长较为缓慢,而 BPVE 算法的运行时间急剧增长,且 IPVE 算法总比 BPVE 算法具有更小的时间花

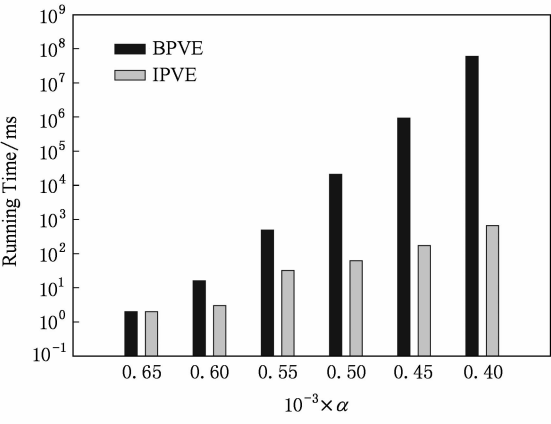
费. 图 5(b)比较了在专利数据集大小 $N=6\,000$ 时 BPVE 算法和 IPVE 算法的时间效率. 可以看到,当取不同的相似度阈值 α 时, BPVE 算法和 IPVE 算法的运行时间都随相似度阈值的减小而单调递增,且 IPVE 算法的运行时间总小于 BPVE 的运行时间. 通过对图 5 中各图的观察,可以看到 IPVE 算法具有比 BPVE 算法更小的时间花费.

由于当前不存在专利价值的动态更新算法,因此文中比较了专利价值评估的动态更新算法 DPVE 与 IPVE 算法在新专利加入专利集合时进行专利价值更新的运行效率. 为了验证 DPVE 算法的高效性,本文将 2003—2012 年专利分类号 G06T 下的所有专利分为 2003—2011 年发表的专利集合与 2012 年发表的专利集合 2 部分,并预先使用 IPVE 算法计算了 2003—2011 年的专利集合中各个专利在 2011 年的专利价值. 之后,将 2012 年发表的专利作为时间 $t+1$ 新加入的专利,并加入到潜在引用网络中. 然后分别采用 DPVE 算法和 IPVE 算法对各个专利在 2012 年的专利价值进行了更新计算.

图 6 是在时间 $t+1$ 使用 IPVE 算法和 DPVE 算法在时间效率的比较. 图 6(a)比较了在相同的相似度阈值下 IPVE 算法和 DPVE 算法的时间效率,其中相似度阈值为 $\alpha=0.3$. 可以看到,当有较少的专利在时间 $t+1$ 加入时, DPVE 算法的运行时间随新加入专利数量 n 的减少而单调递减,而 IPVE 算法的运行时间基本保持不变. 图 6(b)比较了时间 $t+1$ 新加入专利的数量相同的情况下 IPVE 算法和 DPVE 算法的运行效率,其中 $n=60$. 当给定的相似度阈值增大时, IPVE 算法和 DPVE 算法的运行时间都单调递减,尽管在某些相似度阈值上 DPVE 算法的运行时间略高于 IPVE 算法,但随着相似度阈值的增大, DPVE 算法具有更快的运行效率. 从图 6 可知,时间 $t+1$ 新加入的专利数量与潜在引用网络的稀疏程度都会对 DPVE 算法的运行效率产生影响. 当潜在引用网络过于稠密时,新加入的专利会与已有专利集合中较多的专利存在潜在引用关联,需要找出这些专利并重新计算它们的价值,此时 DPVE 运行效率相对较低;而当潜在网络较为稀疏时,与新加入的专利存在引用关联的专利数目较少,即需要重新计算价值的专利数量较少,此时 DPVE 算法比 IPVE 算法具有更高的时间效率. 另外,时间 $t+1$ 新加入专利的数量也会对 DPVE 算法产生影响. 当新加入的专利数量较多时, DPVE 算法进行价值更新的效率优势并不明显,而当新加入的专利数量较少



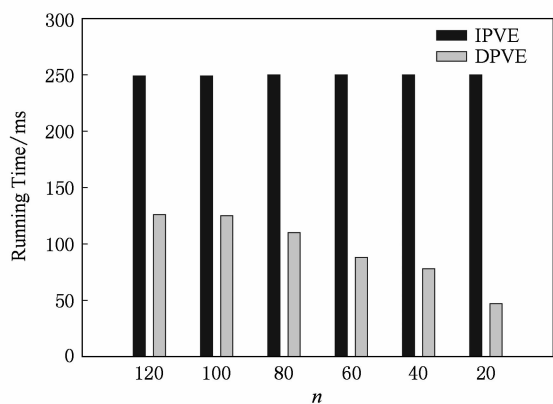
(a) Running time with different dataset size ($\alpha=0.3$)



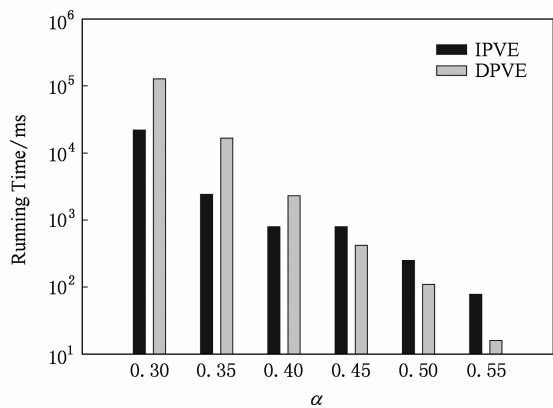
(b) Running time with different similarity threshold ($N=6\,000$)

Fig. 5 Comparisons of BPVE and IPVE on running time.
图 5 BPVE 算法和 IPVE 算法在时间效率上的对比

时,DPVE 算法的运行效率远高于 IPVE 算法. 由以上分析可以看出,在时间 $t+1$ 新加入的专利数量较少或潜在引用网络较为稀疏时,DPVE 算法比 IPVE 算法具有更小的时间花费. 一般来说,相对于时间 t 已有的专利数据集合的大小,时间 $t+1$ 新加入的专利数量较少,且新加入专利与已有专利数据集中较少的专利存在潜在引用关联. 因此,在一般情况下,DPVE 算法比 IPVE 算法具有更少的时间花费.



(a) Running time with different size of new arrival patents ($\alpha=0.3$)



(b) Running time with different similarity threshold ($n=60$)

Fig. 6 Comparisons of IPVE and DPVE on running time.

图 6 IPVE 算法和 DPVE 算法在时间效率上的对比

5 总 结

本文针对已有专利价值评估方法中存在的缺陷,提出了一种基于潜在引用网络的专利价值评估方法,并设计了相应的算法对专利的价值进行评估. 新的方法考虑了专利之间潜在的直接和间接引用以及专利的新颖度和潜在引用强度对专利价值的影响,可以有效地评估各个专利的价值. 针对基本算法效率较低的问题,本文提出了改进算法,极大提高了算法的效率. 最后,针对新专利加入专利集合时各个

专利的价值会发生变化的情况,提出了专利价值评估的动态更新算法,可以快速地更新专利的价值.

参 考 文 献

- [1] Liu S J, Shyu J. Strategic planning for technology development with patent analysis [J]. International Journal of Technology Management, 1997, 13(5/6): 661-680
- [2] Chen Y L, Chiu Y T. An IPC-based vector space model for patent retrieval [J]. Information Processing and Management, 2011, 47(3): 309-322
- [3] Hasan M A, Spangler W S, Griffin T, et al. COA: Finding novel patents through text analysis [C] //Proc of the 15th ACM SIGKDD Conf. New York: ACM, 2009: 1175-1184
- [4] Lupu M. Patent information retrieval: An instance of domain-specific search [C] //Proc of the 35th ACM SIGIR Conf. New York: ACM, 2012: 1189-1190
- [5] Chen Y L, Chang Y C. A three-phase method for patent classification [J]. Information Processing and Management, 2012, 48(6): 1017-1030
- [6] Kim J H, Choi K S. Patent document categorization based on semantic structural information [J]. Information Processing and Management, 2007, 43(5): 1200-1215
- [7] 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, 41(2): 313-330
- [8] Tang Jie, Wang Bo, Yang Yang, et al. PatentMiner: Topic-driven patent analysis and mining [C] //Proc of the 18th ACM SIGKDD Conf. New York: ACM, 2012: 1366-1374
- [9] Shih M J, Liu D R, Hsu M L. Mining changes in patent trends for competitive intelligence [G] //LNCS 5012: Proc of the 12th PAKDD Conf. Berlin, Springer, 2008: 999-1005
- [10] Chang P L, Wu C C, Leu H J. Using patent analyses to monitor the technological trends in an emerging field of technology: A case of carbon nanotube field emission display [J]. Scientometrics, 2010, 82(1): 5-19
- [11] Dereli T, Durmusoglu A. A trend-based alert system for technology watch [J]. Journal of Scientific & Industrial Research, 2009, 68(8): 674-679
- [12] Wu S, Sun J, Tang J. Patent partner recommendation in enterprise social networks [C] //Proc of the 6th ACM WSDM Conf. New York: ACM, 2013: 43-52
- [13] China Technology Exchange. Patent value analysis [OL]. [2014-05-02]. <http://www.ctex.cn/article/zlpt/zscqpg/> (in Chinese)
- [14] Jin X, Spangler S, Chen Y, et al. Patent maintenance recommendation with patent information network model [C] //Proc of the 11th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2011: 280-289

[15] Baron J, Delcamp H. Patent quality and value in discrete and cumulative innovation [J]. Scientometrics, 2012, 90 (2): 581-606

[16] Zeebroeck N V. The puzzle of patent value indicators [J]. Economics of Innovation and New Technology, 2011, 20(1): 33-62

[17] Reitzig M. What determines patent value? Insights from the semiconductor industry [J]. Research Policy, 2003, 32(1): 13-26

[18] Dang Qianna. Patent analysis methods and main indications [OL]. (2005-11-21) [2014-05-02]. <http://www.istis.sh.cn/list/list.asp?id=2402>(in Chinese)
(党倩娜. 专利分析方法和主要指标[OL]. (2005-11-21) [2014-05-02]. <http://www.istis.sh.cn/list/list.asp?id=2402>)

[19] Wang Xuefeng, Liu Xiaoxuan, Zhu Donghua. Patent value evaluation research [J]. Scientific Management Research, 2008, 26(6): 115-117 (in Chinese)
(汪雪锋, 刘晓轩, 朱东华. 专利价值评价指标研究[J]. 科学管理研究, 2008, 26(6): 115-117)

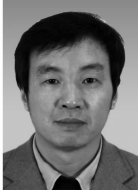
[20] Yoon B, Park Y. A text-mining-based patent network: Analytical tool for high-technology trend [J]. Journal of High Technology Management Research, 2004, 15(1): 37-50

[21] Galaxy. ICTCLAS Chinese segmentation system [CP/OL]. (2012-07-03) [2014-05-02]. http://ictclas.cn/ictclas_download.aspx (in Chinese)
(北京中科天玑科技有限公司. ICTCLAS 汉语分词系统 [CP/OL]. (2012-07-03) [2014-05-02]. http://ictclas.cn/ictclas_download.aspx)

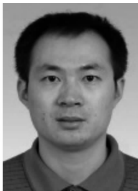
[22] Lawrence P, Sergey B, Rajeev M, et al. The PageRank citation ranking: Bringing order to the Web [R/OL]. Stanford, CA: Stanford InfoLab, 1998 [2013-11-01]. <http://ilpubs.stanford.edu:8090/422/>



Feng Ling, born in 1986. PhD candidate. Student member of China Computer Federation. His research interests include patent analysis and mining (fengling@whu.edu.cn).



Peng Zhiyong, born in 1963. PhD, professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include complex data, trusted data and Web data management.



Liu Bin, born in 1975. PhD, lecturer. Member of China Computer Federation. His main research interests include complex data management and data mining (binliu@whu.edu.cn).

Che Dunren, born in 1964. PhD, professor. IEEE senior member. His main research interests include database, data mining and cloud computing(dche@cs.siu.edu).