# 基于闭频繁项集挖掘的技术演化研究方法*

陈亮 张志强 尚玮姣

[摘 要] 本文以专利中的技术术语作为事务、以术语中的词汇作为项,通过闭频繁项集挖掘方法,对专利文献中的技术术语的结构变化情况进行时序分析,以从新的角度来研究技术演化趋势,之后以硬盘驱动器磁头技术为例进行实证分析,实证结果表明,该方法能够对技术演化过程中所产生的技术变化进行有效识别。

[关键词] 闭频繁项集 技术演化 技术术语 文本挖掘 硬盘驱动器
[分类号] G353.1
DOI:10.7536/j.issn.0252-3116.2013.19.017

## 1 引 言

技术演化(technological evolution)的产生源自人们观察到技术发展和生物进化的相似性,期望采用演化这一隐喻方式来描述技术发展变化。目前尚无关于技术演化的统一定义,一般而言,它指在某一特定技术领域,从技术出现到当前阶段,技术领域内部的技术活动、子技术或技术主题随着时间推移的发展、继承和变化的过程[1]。关于技术演化的本质,W. B. Arthur[2]认为它是一个组合过程,现有技术构建起新技术,而新技术又为未来技术出现提供了可能的构建材料;L. Fleming[3]也认为现有部件和新产生部件的重新组合实现了技术发展,F. Kodama[4]从市场层面观察,发现技术间的互补能够赋予发明新的特性,从而开启新的市场。他进一步指出基于技术互补的技术融合正在取代技术突破成为发明的新模式,越来越多的突破性技术来自技术融合。

鉴于技术演化在国家、企业战略管理中的重要地位,世界主要国家投入大量人力物力对其展开研究,并形成较大的技术演化研究方法家族,具体见表1[1]。

当前技术演化研究方法家族中的定量方法以专利信息作为主要数据来源,以专利中的引文关系、分类编号或文本内容作为承载技术的信息载体,对技术演化过程中一系列行为特征(如技术扩散、技术转移、技术融合以及技术之间关联程度等)展开分析,然而由于这些信息载体难以准确刻画技术演化的发生形式——技术的拆分重组,从而使研究者难以在更加具体、细化的层次上对技术演化的特征规律展开研究,虽然当前也

表1 常见主要技术演化研究方法

| 方法分类 | | 主要内容 |
|---|---|---|
| 文献计量学方法 | 专利引文法 | 1949年提出专利引文分析的概念,1981年后逐步被证实,20世纪90年代起开始发展完善,具体方法有专利同被引分析法、专利耦合分析法、专利引文时序分析等。 |
| | 专利分类法 | 按照专利分类标准对不同技术领域专利申请或授权量进行统计,以了解该领域的技术构成和技术焦点。 |
| | 生命周期法 | 包括数学模型法和指标分析法,代表性的数学模型法有 Logistic 和 Gompertz 模型,指标分析法基于特定目的度量,采用指标来分析技术生命周期,可以发掘出技术发展影响因素的实证信息。 |
| TRIZ 法 | | 该方法在大规模专利分析基础上,将产生新工作过程的原理具体化,并提出一系列规则、算法与发明创造原理,形成一套比较完整的创新设计理论。 |
| 文本挖掘方法 | | 该方法包括较基本的词频分析法和旨在寻找能够区分数据重要程度的潜在变量(也即技术挖掘结构)的高级方法。 |
| 形态分析法 | | 将研究对象视为一个系统,通过系统分析方法将其分解为相对独立的子系统,子系统所实现的功能成为基本元素,实现各子系统功能的技术手段成为基本形态,通过排列与组织方法可以得到多种可行解,经过删选可从中确定系统的最佳方案。 |
| 技术路线图法 | | 通过对目标技术领域相关专利信息进行搜集、处理和分析,并将分析全部结果可视化地表达出来,使复杂多样的专利情报得到方便有效地理解。 |
| 专家调查法 | | 以专家作为索取信息的对象,依靠专家的知识和经验,对技术演化趋势做出判断、评估和预测。 |

*本文系中国科学院知识创新工程重要方向性项目"影响经济社会重大体系的战略性科技问题分析"(项目编号:Y110071)研究成果之一。
[作者简介] 陈亮,中国科学院国家科学图书馆兰州分馆,中国科学院大学博士研究生,E-mail:chenliang@mail.las.ac.cn;张志强,中国科学院国家科学图书馆副馆长、兰州分馆馆长,研究员,博士生导师;尚玮姣,中国林业科学研究院图书馆助理馆员,硕士。
收稿日期:2013-06-14 修回日期:2013-09-15 本文起止页码:107-111 本文责任编辑:徐健

就通过技术重组进行未来技术分析的研究方法如形态分析法,它基于专利关键词构建技术形态结构,通过对不同技术排列组合结果的评估来筛选和优化未来可能的创新方案[5],但其过于依赖专家的缺陷导致该方法较适合在具体创新实践中提供支持,而无法应用于涵盖内容更加广泛的技术演化分析。

较之前述技术信息载体,闭频繁项集的主要优点如下:第一,它依据词汇的共现频度进行文本抽取,因而可以筛选出重要的技术术语以及相互关联密切的技术术语组合,并可以据此分析技术演化过程中不同技术组合所形成的技术替换和技术互补现象;第二,这种信息载体可以按照层次结构进行组织,研究者不仅可以藉此标识技术及其子技术的上下位关系,还可以通过父技术与子技术之间置信度的变化来判断技术演化过程中技术结构是否发生改变;第三,闭频繁项集可以通过计算机从大量专利文本中抽取出来,减少了对专家的依赖,从而使研究工作更加高效、准确、客观。

本文将在下面详细介绍基于闭频繁项集的技术演化方法的分析框架和具体流程,并以硬盘驱动器磁头技术为例进行实证分析、总结,指出目前存在的缺陷及后续的研究内容。

## 2　基于闭频繁项集的技术演化研究方法

### 2.1　相关概念定义及方法框架

本文从频繁模式挖掘角度出发,提出一种基于闭频繁项集划分的技术演化研究方法,首先,对本文中涉及的概念进行必要的解释。

关联规则:关联规则是形如 X - > Y 的蕴含表达式,其中 X 和 Y 是不相交的项集,即 $X \cap Y = \emptyset$[6];

支持度:确定规则可以用于给定数据集的频繁程度[6]。

频繁项集:项集支持度大于用户指定最小阈值,这样的项集叫做频繁项集[6];

闭项集:项集 X 是闭的,如果它的直接超集都不具有和它相同的支持度计数[6];

闭频繁项集:一个项集是闭频繁项集,如果它是闭的,并且它的支持度大于或者等于最小支持度阈值[6]。

基于闭频繁项集的技术演化研究方法基于以下观察,新技术术语有三种构成方式:①之前从未出现过的全新技术术语;②现有技术术语组合形成新技术术语;③现有技术术语和全新技术术语组合形成新技术术语。

其中,第一种技术术语代表新技术,第二种技术术语代表现有技术间的应用融合(application convergence),第三种技术术语代表现存技术与已有技术的横向融合(lateral convergence)[7]。鉴于重要程度较高的技术术语所包含的词汇组合在技术术语集合中普遍共现频次较高,本文以专利中的技术术语作为事务(transa ction),技术术语中的词汇作为项(item),通过技术术语集合中术语词汇所组成的闭频繁项集,反推技术术语,术语的重要程度与该术语的支持度有关,术语之间的关系强度与这些术语之间的置信度有关,从而基于闭频繁项集之间的包含关系建立技术术语层次结构,通过技术演化过程中技术术语层次结构以及它所包含闭频繁项集的支持度和置信度变化,在更加具体、细化的层次上对技术演化的特征规律展开研究。

本文提出的分析框架如图 1 所示:


图 1　总体分析框架

### 2.2　具体分析流程

2.2.1　专利文本预处理　首先,确定待分析的技术领域,并借助领域专家从相关专利中选择若干关键词从专利数据库中检索专利,然后抽取术语并进行异常数据删除、术语去重、词干提取和同义词替换等操作,最终得到清洗后的技术术语集合。

2.2.2　抽取闭频繁项集　当前闭频繁项集挖掘算法较多,如 A - close、CLOSET、CLOSET +、FPCLOSE、CHARM 等[6],本文采用 A - close 算法[8],输出结果是由词汇所组成的闭频繁项集及其支持度。

2.2.3　建立技术术语层次结构网络　本文以查找全部闭频繁项集的最大闭频繁子集的方式,计算关联规则的置信度并通过关联规则建立技术术语层次结构网络,所谓最大闭频繁子集即假设有闭频繁项集 A,B 且 $A \subset B$,如果 $\forall C$ 使得 $A \subset C$ 且 $C \subset B$,那么 A 即为 B 的最大闭频繁子集,计算关联规则的置信度的算法伪代码如下:

(1) 筛选出项数大于 1 的闭频繁项集,按照项数从大到小存入队列 Queue 中

(2) while queue 不为空

(3)　　closet←从 queue 队头取出元素

(4)　　max_subclosets_queue←查找 closet 的全部直接闭频繁

子集

（5） while max_subclosets_queue 不为空

（6） max_subcloset←从 max_subclosets_queue 取出队头元素

（7） confidence(closet→max_subcloset) = $\frac{support(max\_subcloset)}{support(closet)}$

（8） 将规则 closet→max_subcloset 及置信度放入结果队列 rules

（9） 将 max_subclosets_queue 清空

（10）返回 rules

查找闭频繁项集 closet 的全部直接闭频繁子集的算法伪代码如下：

（1）将全部闭频繁项集存入队列 $queue_1$

（2）while $queue_1$ 不为空

（3） closet_temp←取出 $queue_1$ 队头元素

（4） If closet 是 closet_temp 的真子集

（5） If 结果存储队列 max_subclosets_queue 为空

（6） Max_subclosets_queue←closet

（7） else

（8） if Max_subclosets_queue 中不存在 closet 的超集

（9） Max_subclosets_queue←closet

（10） If Max_subclosets_queue 中存在 closet 的真子集

（11） 将该真子集从 Max_subclosets_queue 中移除

（12）返回结果队列 Max_subclosets_queue

2.2.4 基于技术术语层次结构网络进行技术演化分析 将在不同时间段中持续出现的闭频繁项集筛选出来，其中项数较少、支持度较高、技术含义明确的闭频繁项集代表着创新行为最活跃的技术领域，以这些闭频繁项集为基础，通过查找它们的闭频繁超集对每个技术领域进行时序分析。

## 3 实证分析

本文以美国专利商标局所提供的专利作为数据源，选择硬盘驱动器的磁头技术为例进行实证分析。美国是硬盘驱动器技术的发源地和主要技术国家，美国专利商标局所涵盖的专利信息具有代表性，选择硬盘驱动器的磁头技术进行技术演化分析基于两点考虑：其一是相比传统工业，硬盘驱动器技术的发展和革新较快，更适合较短时间内的技术演化分析[9]，磁头作为硬盘驱动器的关键技术，能够集中体现硬盘驱动器的技术水平；其二是硬盘驱动器领域历来重视专利申请，因而专利信息能够充分反映该领域的技术发展过程。自 1956 年 IBM 制造出第一块硬盘 RAMAC 开始，硬盘驱动器的磁头经历了铁氧体磁头、薄膜磁头和磁阻磁头三代技术[10]。本文对 1976 - 1999 年间硬盘驱动器技术中薄膜磁头相关专利所包含的术语进行技术演化分析。

本文检索策略采用关键词和专利引文相结合方式，并辅以人工判读，具体检索过程如下所示：

第一步：以 ABST/“thin film head” AND APD/1/1/1976 - >31/12/2003 作为检索式，得到专利 125 个。

第二步：搜索 1976 - 1999 年间与这 125 个专利存在引用或者被引关系的专利，得到专利 1 416 个，最终得到总专利 1 543 个。

第三步：分析 1 543 个专利所形成的专利引文网络，得到 12 个独立元组，其中巨元组包含专利 1 369 个，经判读知其反映了薄膜磁头中主要技术的演化过程，本文将最终得到包含 1 369 个专利的数据集。

第四步：以 4 年为单位将全部时间划分出 6 个时间段，对每个时间段的专利使用 c-value 方法从专利文本中抽取术语，对术语预处理之后进行闭频繁项集的抽取，并据此建立闭频繁项集层次结构网络（以下“简称结构网络”）。本方法在各个时间段中的统计数据如表 2 所示：

表 2 基于专利术语的闭频繁项集挖掘统计数据

| 时间段 | 时间段 1 | 时间段 2 | 时间段 3 | 时间段 4 | 时间段 5 | 时间段 6 |
|---|---|---|---|---|---|---|
| 专利个数 | 42 | 74 | 104 | 181 | 248 | 351 |
| 闭频繁项集个数 | 47 | 119 | 152 | 253 | 434 | 561 |
| 结构网络中非孤立节点数 | 25 | 69 | 89 | 162 | 281 | 331 |

由于篇幅所限，本文仅列举出第一时间段中除去孤立节点后的结构网络（见图 2），其中节点大小代表该闭频繁项集的支持度，连线粗细代表关键规则的置信度，不同连通分量采用不同颜色节点表示。从图中可以看出，有些闭频繁项集本身就是组件如 head 和 layer，或能表达明确的技术如 MR，有些闭频繁项集则需要和其他闭频繁项集配合来表示组件或者技术，如 thin、film 和 head 组成 thin film head 也即金属薄膜磁头，在不引起混淆的情况下，本文将前者称为第一类闭频繁项集，将后者称为第二类闭频繁项集，在第一类闭频繁项集与第二类闭频繁项集通过组合形成第一类闭频繁项集的过程中，技术演化得以进行。

将在不同时间段中持续出现的第一类闭频繁项集筛选出来，并将其命名为集合 S，S 中项数为 1 且支持度最高的两个闭频繁项集{head}和{layer}，这标志着它们是与硬盘驱动器的磁头技术关联最密切且创新最

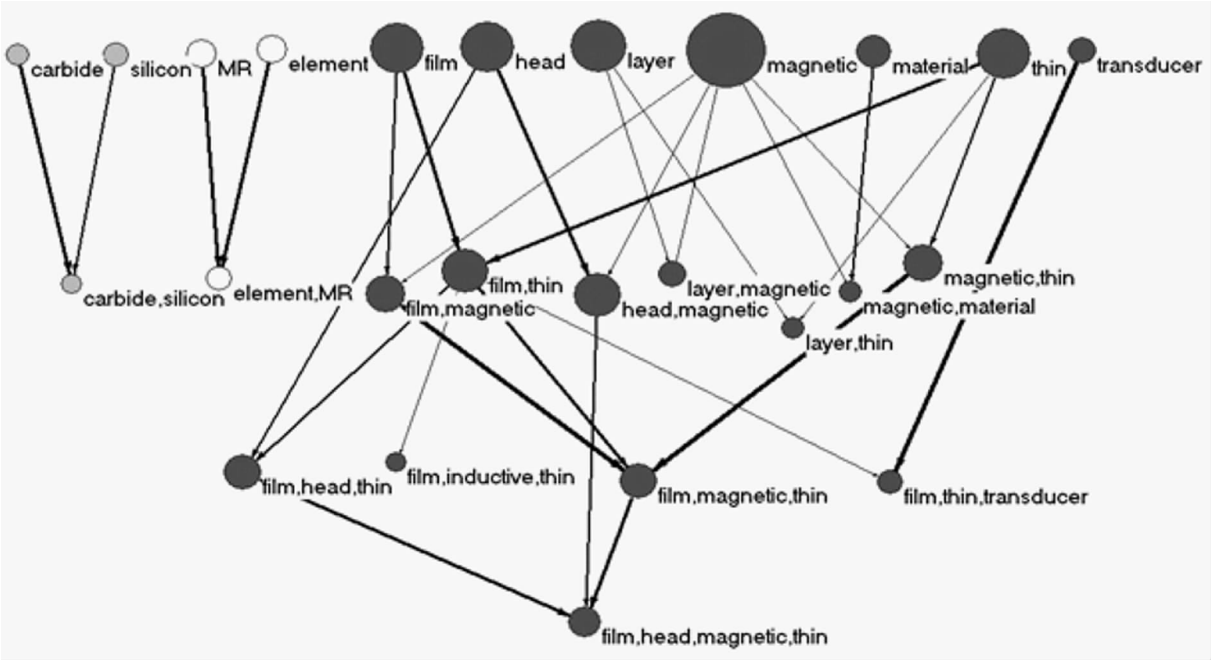


图 2　1976 – 1979 年间的技术结构网络

活跃的技术领域，将 S 中分别与这两个闭频繁项集关联密切（也即闭频繁项集之间关联规则的置信度较高）的第一类闭频繁项集进行汇总，如图 3、图 4 所示：

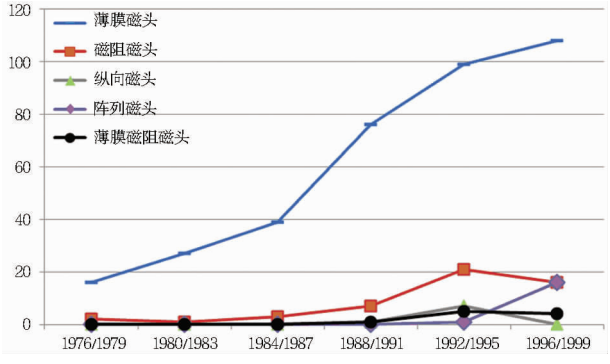


图 3　磁头技术演化时序

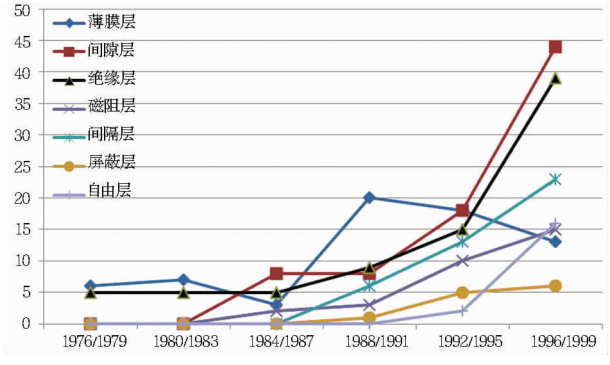


图 4　涂层技术演化时序

由图 3 可知，1976 – 1983 年期间磁头技术以薄膜磁头为主，虽然这一时期出现了磁阻磁头技术但发展缓慢，1984 – 1995 年间薄膜磁头技术发展明显加快，磁阻磁头与薄膜磁头之间关联愈发密切，并出现了磁阻与薄膜技术相结合的复合型磁头，之后该类型磁头一直呈平稳态势发展。此外，如图 4 所示，1976 – 1983 年间磁头涂层成分以薄膜层和绝缘层为主，基本构成保持稳定，自 1984 年起，间隙层、间隔层、屏蔽层、自由层、磁阻层相继进入硬盘磁头技术领域并获得了迅速发展，以至 1996 – 1999 年期间，各涂层所在闭频繁项集的支持度比例完全迥异于 1976 – 1979 年期间，这表明了伴随着磁头技术的迅速发展，磁头技术的内部结构发生了重大变化。

事实上，这一时期内磁头技术领域出现了巨磁阻技术和阵列磁头技术两项重大进展，其中巨磁阻技术通过提升磁头灵敏度来促使磁盘存储容量快速增长，巨磁阻磁头的薄膜材料包括扎钉层、间隔层、自由层、交换层和屏蔽层，其中间隔层是绝缘层，用于将自由层和钉扎层的磁化分开[11]；阵列磁头技术出现于 1994 年，在 1996 – 1999 年期间增长迅速，实际上随着硬盘驱动器由单磁片发展为多磁片结构，由多磁头组成的阵列磁头成为可对多磁片并行读写的必要技术[12]；纵向磁头技术在 1992 – 1995 年间曾经短暂出现，但由于纵向磁记录中较高的自退磁现象为存储密度的提升带来极大困难[13]，该技术随后被淘汰。

从整个时间段来看，一方面薄膜磁头技术的发展呈 S 形，根据技术生命周期理论，1999 年后其发展势头将继续放缓，但巨磁阻技术会继续其迅猛势头，成为磁头技术的重要创新点，另一方面，阵列磁头属于结构创新，与薄膜磁头处于不同的技术发展维度，它们之间

并无替代关系，也即虽然薄膜磁头专利数量的增速变缓，但并不影响阵列磁头在薄膜磁头相关专利中的扩散速度，阵列磁头专利数量将继续保持原有增长幅度。

## 4 总 结

基于闭频繁项集的技术演化研究方法是一种基于专利数据的文本挖掘方法，它通过技术术语的构成变化情况对技术演化趋势进行分析。相比其他技术演化分析方法，它可以在更加具体、细化的层次上对技术演化展开研究。从实证结果来看，本方法有效排除了充斥于非结构化文本中的噪音数据，减少了分析过程中对专家的依赖程度，从而使研究工作更加高效、准确和客观。

然而，通过术语变化来反映技术演化趋势存在其固有局限，很多技术需要多个术语的搭配来表达，比如巨磁阻磁头（giant magneto resistive head），它很少直接出现在专利文献中，大多数情况下是或者通过 giant magneto resistive 直接表达，或者使用副词将其与 element、sensor 或 transducer 等连接起来表达。此外，该方法无法挖掘不具备包含关系的术语之间的相互关系，比如巨磁阻磁头与自由层之间的技术关联强弱，在本文中只是借助专家知识对这些关系进行定性判断，并没有定量度量。然而，本方法为下面的研究工作奠定了基础，接下来的工作可以沿用本文的研究框架，以专利文献作为事务、以频繁术语作为项进行技术挖掘，以期更加完整和准确地进行技术演化趋势分析。

**参考文献：**

[1] 陈亮，张志强. 技术演化研究方法进展分析[J]. 图书情报工作，2012，56(17)：59－66.

[2] Arthur W B. The nature of technology [M]. New York：Free Press，2011.

[3] Fleming L. Recombinant uncertainty in technological search[J]. Management Science，2001，47(1)：117－132.

[4] Fumio K，Emerging patterns of innovation：Sources of Japan's technological edge [M]. Boston：Harvard Business School Press，1995.

[5] 黄鲁成，李欣，吴菲菲. 技术未来分析理论方法与应用[M]. 北京：科学出版社，2010.

[6] 陈封能，斯坦巴赫，库玛尔. 数据挖掘导论[M]. 范明，范宏建，等译. 北京：人民邮电出版社，2011.

[7] Hacklin F. Howincremental innovation becomes disruptive：The case of technology convergence[EB/OL]. [2013－08－05]. http://ieeexplore. ieee. org/stamp/stamp. jsp? arnumber=01407070.

[8] Pasquier N，Bastide Y，Taouil R，et al. Discovering frequent closed Itemsets for association rules [EB/OL]. [2013－08－05]. http://hal. archives－ouvertes. fr/docs/00/46/77/47/PDF/Discovering_frequent_closed_itemsets_for_association_rules_Pasquier_et_al._ICDT_1999. pdf.

[9] Christensen C. 创新者的窘境[M]. 胡建桥，译. 北京：中信出版社，2010.

[10] Edward G，Roger F. Future trend in hard disk drives[J]. IEEE Transaction on Magnetic，1996，32(3)：1850－1854.

[11] GMR－Head Technology[EB/OL]. [2013－08－05]. http://www. personal. psu. edu/sbk142/gmr. htm

[12] 唐朔飞. 计算机组成原理（第二版）[M]. 北京：高等教育出版社，2008.

[13] 杜明辉. 垂直磁记录介质薄膜的制备及其晶体结构和磁性研究[D]. 北京：华北电力大学，2011.

## A Research Method of Technological Evolution Based on Frequent Closed Itemset Mining

Chen Liang[1,2] Zhang Zhiqiang[1] Shang Weijiao[3]

[1]Lanzhou Branch of National Science Library，CAS，Lanzhou 730000

[2]University of Chinese Academy of Sciences，Beijing 100190

[3]Library of Chinese Academy of Forestry，Beijing 100091

[Abstract] This paper makes a time series analysis to the dynamics of technological terminologies´construction via close frequent set mining by taking technological terminologies from patent documents as transitions and words in technological terminologies as item，on purpose of researching technological evolution on a new perspective. After that we employ head techniques of hard disk drive to execute our empirical analysis. The result shows that our method can recognize changes during technological evolution process efficiently.

[Keywords] frequent closed itemset technological evolution technological terminology text mining hard disk driver