●徐文海，温有奎（西安电子科技大学 经济管理学院，陕西 西安 710071）

一种基于TFIDF方法的中文关键词抽取算法

摘 要：本文在海量智能分词基础之上，提出了一种基于向量空间模型和TFIDF方法的中文关键词抽取算法。该算法在对文本进行自动分词后，用TFIDF方法对文献空间中的每个词进行权重计算，然后根据计算结果抽取出科技文献的关键词。通过自编软件进行的实验测试表明该算法对中文科技文献的关键词自动抽取成效显著。

关键词：关键词抽取；向量空间模型；算法

Abstract: On the basis of Massive Intelligent Segmentation, this paper proposes a Chinese keyword extracting algorithm based on Vector Space Model and TFIDF method. After automatic segmentation of text, this algorithm calculates the weight of every word in document space with TFIDF method and extracts the keywords of scientific and technical documents according to the calculation result. The experimental test with self-compiled software indicates the algorithm improves the efficiency of automatic keyword extraction of Chinese scientific and technical documents obviously.

Keywords: keyword extraction; VSM; algorithm

“新摩尔定律”指出：因特网上的信息正以每6个月翻一番的速度爆炸般地产生，它使任何上网寻求信息的人都难以选择。面对这如海潮般涌来的五光十色、瞬息万变的信息，如果没有一个强有力的工具来帮助寻找、发掘有用的信息，人们就会被湮没在信息的海洋中，迷失方向[1]。

关键词是为了文献标引工作，从报告、论文中选取出来用以表示全文主题内容信息的单词或术语。关键词自动抽取是依靠计算机从文档中选择出反映主题内容的词，也称作关键词自动标引，在文献检索、自动文摘、文本聚类/分类等方面有着重要的应用[2]。关键词可以为文档提供一个简短的概括，使读者能够在短时间内了解文档的大概内容。关键词还是信息检索系统中对文档进行索引、聚类等操作的基础[3]。

关键词一要反映论文的主题内容，二要具有专指性。在学术期刊中，关键词主要是名词和术语。目前学术刊物中的论文一般都有作者自行确定的关键词，其他的各类文章还很少提供关键词，通常需要在编辑整理时手工抽取。手工抽取关键词不仅费时费力，而且主观性强，抽取不当往往会对下一步的应用造成消极影响。因此关键词的自动抽取具有一定的研究价值。

如今信息资源已经成为人们竞争的重点，如何快速获得有价值的信息已经成为一种新的财富，而借助于某些工具，自动抽取出能准确代表文本中的关键词，是一种解决信息危机的有效方案。

1 国内外的研究现状

国外对于关键词自动抽取的研究较早，已经建立了一些实验系统。Turney设计的GenEx[4]系统将遗传算法和C4.5决策树机器学习方法用于关键短语的抽取；Witten[5]采用朴素贝叶斯技术对短语离散的特征值进行训练，获取模型的权值，然后从文档中抽取关键短语。Hulth[6]提出了一种在学术论文的摘要中自动提取关键词的方法，她采用了一种叫做Rule Induction的学习算法，利用词频和词性等特征对样本进行学习，结果使得正确率达到了29.7%。

中文文本没有显式的词边界，使得关键词的自动抽取增加了一定难度，Yang Wenfeng[7]提出了基于PAT树结构获取新词，并采用互信息等统计方法进行关键词抽取，但建立获取候选词的PAT树需要大量的存储空间，实现起来比较复杂。李素建等[8]提出了利用最大熵模型进行关键词自动标引的方法，由于特征的选择以及估计特征参数时不够准确，最大熵模型在关键词标引中的应用并不理想。王军[9]提出了一个用于自动标引的文献主题关键词抽取方法，它限于从已标引的结构化语料库中元数据的标题中抽取关键词。笔者提出了一种建立在海量智能分词基础之上，结合向量空间模型（VSM）和TFIDF特征项权重计算方法的中文关键词自动抽取新方法。

2 基于海量智能分词的中文自动分词新算法

2.1 海量智能分词技术

由海量智能计算技术研究中心开发的海量智能分词技术在国内处于领先水平，其分词准确率达到99.6%，分词效率为2 000万字/min。海量智能分词技术很好地解决了分词领域中的两大技术难题，即歧义切分和新词的识别。海量智能分词能对绝大多数的组合歧义进行正确地切分，在新词的识别上，针对不同类型采用了不同识别算法，其中包括对人名、音译词、机构团体名称、数量词等新词的识别，准确率同样达到了同行业的领先水平。图1是海量智能分词软件在对《向知识标引进军》一文进行分词的结果。

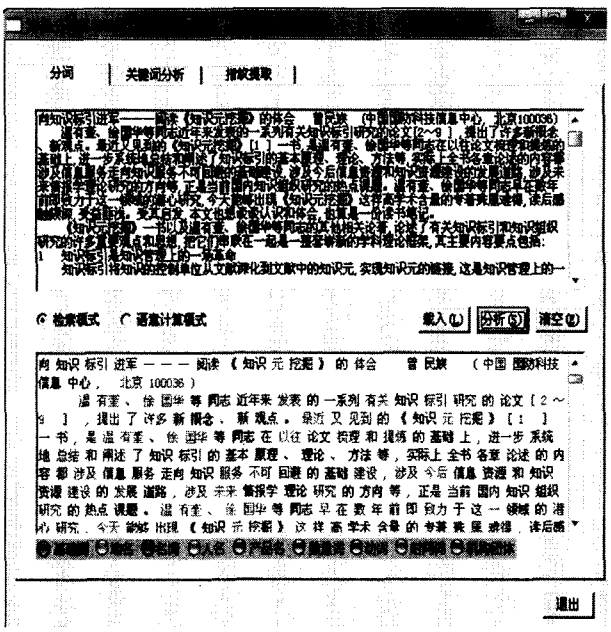

图1 海量智能分词软件分词结果

从图1中可以看出，海量智能分词尽管能把大部分词语分开，但对于包含创新点的科技文章来说，其中的创新点词和未登录词尚不能准确地识别，分词结果还不尽如人意。如文中的“知识标引”和“知识元”没有被看成一个词，而这些词比较好地反映了该文的主旨，所以该方法还有待改进。

2.2 基于海量智能分词的中文自动分词新算法

海量智能分词对大部分词都作出了切分，但不能得到能反映中文科技文献的关键词和创新点。为了能得到如实反映文章语义的更精确的分词结果，笔者提出了一种基于海量智能分词初步结果的新算法。

1）文本过滤。由于表征文章主旨的词主要是实词，我们首先建立了一个虚词表，如果海量智能分词结果中词的词性为虚词表中的任一种或标点符号，则将这个虚词或标点符号用空格代替。经过处理后的文本只包含了全部实词和空格，这里用ArrayList存储这个经过过滤的文本。

2）统计词数。通过编写的全文检索函数对文章中的每个词进行词数统计，并用一个HashTable分别存储该词的词名、词性以及在文中出现的次数。这里只是针对海量智能分词结果进行统计，下一步将对该结果进行改进。

3）通过全文遍历，修正词频统计结果。文本过滤后，从文章开始处，如果开始不是空格并且连续两个词不为空格，则统计这两个词出现的次数，如果该次数大于给定的阈值，就将这两个词连接起来并修改词性为new，并代替ArrayList中原来的两个词；再将这个词看成一个新词，如果它的下一个词为非空，并且这个新词与下一个词的出现词数大于给定阈值，则将这个新词和下一个词又组成新词，并把词性改成new；如果连续两个词的词数不大于给定阈值，或遇到空格或空值，则跳过，寻找下一个非空实词。经过遍历，ArrayList中的分词结果有了反映科技文章主旨的能力。同样，用全文检索函数对新的ArrayList中的词进行了次数统计。该算法流程如图2所示。

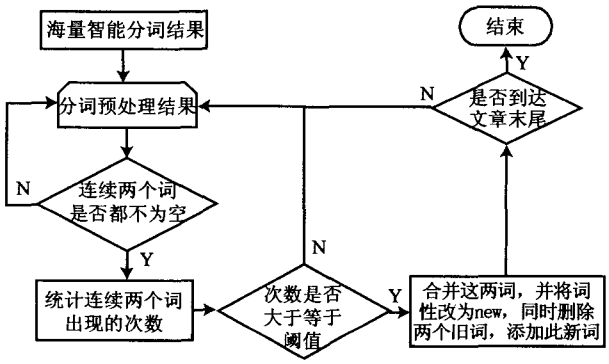

图2 改进后的中文科技文献分词算法流程图

笔者用Java语言编写了一个测试软件，图3为用本算法得到的分词结果。

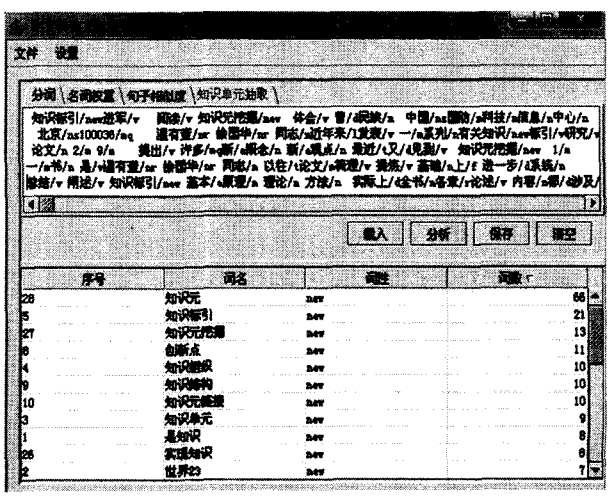

图3 改进后的新算法所得的分词结果

从利用新算法所得的分词结果可以看出，“知识元”、“知识元标引”、“知识元挖掘”和“知识元链接”等词能被识别出来，并分别用“new”加以标记。但也出现了一些如“是知识”，“世界23”等词，这主要是因为这种连续两个词组合的出现次数比较大，新算法也将它们作为一个词提取了出来。在以后的工作中，我们将对新词进行过滤，以达到更好的分词效果，下一节将就所得的结果进行特征项权重分析。

3 基于向量空间模型的文本表示及特征项权重计算

3.1 基于向量空间模型的文本表示

向量空间模型是由Salton等人于20世纪60年代末提出并成功地应用于著名的SMART（System for the Manipulation and Retrieval of Text）系统之后，该模型及其相关的技术，包括项的选择、加权策略，以及采用相关反馈进行查词优化等技术，在文本分类、自动索引、信息检索等许多领域得到了广泛的应用。VSM已成为最简便高效的文本表示模型之一。由于VSM的这些特点，在文本过滤领域，VSM也是广泛采用的文本表示模型。VSM的基本概念如下：

1）文档（Document）。泛指一般的文本或文本中的片断（段落、句群或句子），一般指一篇文章。尽管文档可以是多媒体对象，但在本文的讨论中我们只认为是文本对象，并且对文本与文档不加以区别。

2）项（Term）。文档的内容特征常常用它所含有的基本语言单位（字、词、词组或短语等）来表示，这些基本的语言单位统称为项，即文档可以用项集（TermList）表示为 $D(t_1, t_2, \cdots, t_N)$，其中 t_k 是项，$1 \leqslant k \leqslant N$。

3）项的权重（TermWeight）。对于含有N个项的文档 $D(t_1, t_2, \cdots, t_N)$，项 t_k 常常被赋予一定的权重 W_k，表示它们在文档 D 中的重要程度，即：$D = D(t_1, w_1; t_2, w_2; \cdots; t_N, w_N)$，简记为 $D = D(w_1, w_2, \cdots, w_N)$。这时我们说项 t_k 的权重为 w_k，$1 \leqslant k \leqslant N$。

4）向量空间模型（VSM）。给定一自然语言文档 $D = D(t_1, t_2, \cdots, t_N)$，由于 t_k 在文档中既可以重复出现又应该有先后次序的关系，分析起来仍有一定的难度。为了简化分析，可以暂不考虑 t_k 在文档中的先后顺序并要求 t_k 互异（即没有重复）。这时可以把 $t_1, t_2, \cdots, t_N$ 看成一个 N 维的坐标系，而 $w_1, w_2, \cdots, w_N$ 为相应的坐标值，因而 $D(w_1, w_2, \cdots, w_N)$ 被看成是 N 维空间中的一个向量（如图4中的 D_1，D_2）。我们称 $D(w_1, w_2, \cdots, w_N)$ 为文档 D 的向量表示或向量空间模型。

5）相似度（Similarity）。两个文档 D_1 和 D_2 之间的（内容）相关程度（Degree of Relevance）常常用它们之间的相似度 $\mathrm{sim}(D_1, D_2)$ 来度量。当文档被表示为VSM，我们可以借助于向量之间的某种距离来表示文档间的相似度。常用向量之间的内积来计算：

$$\mathrm{sim}(D_1, D_2) = \sum_{k=1}^{N} w_{1k} \cdot w_{2k}$$

或用夹角余弦值来表示：

$$\mathrm{sim}(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{N} w_{1k} \cdot w_{2k}}{\sqrt{\left(\sum_{k=1}^{N} w_{1k}^2\right)\left(\sum_{k=1}^{N} w_{2k}^2\right)}}$$

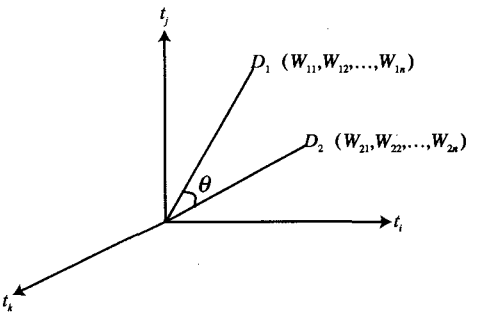

图4 文档的向量空间模型及文档间的相似度 $\mathrm{sim}(D_1, D_2)$

如图4所示，VSM的优点在于它把文档内容简化为特征项及其权重的向量表示，把对文档内容的处理简化为向量空间中向量的运算，使问题的繁杂性大为降低。

我们在已取得分词结果的基础上，对科技论文进行分句，把每句话看成是一个向量，每个向量由其所在的实词特征项来表示。下面首先介绍特征项的权重计算。

3.2 特征项的权重计算

用来表示文档内容的特征项可以是各种类别，有词、短语，甚至是句子或句群等更高层次的单位。特征项也可以是相应词或短语的语义概念类。

特征项的选择必须由处理速度、精度、存储空间等方面的具体要求来决定。选出的项越具有代表性，语言层次越高，项所包含的信息就越丰富，但分析的代价就越大，而且受分析精度（如句法分析的准确率）的影响就越大。其中，字、词、概念、短语等特征项，在文档中的出现频度较高，呈现一定的统计规律，更适用于信息检索、文档分类等应用系统；而由简单特征组合成的相对复杂的句子和段落特征，则更多地被摘要系统所采用。

给每个项赋权重时，应使得文本中越重要的项权重越大。第一种方法是由专家或用户根据自己的经验与所掌握的领域知识，人为地赋上权值。这种办法随意性很大，效率也不高，很难适用于大规模真实文本的处理。另一种办法是运用统计的方法，也就是用文本的统计信息（如词频、词之间的同现频率等）来计算项的权重。

事实上，如果将文档集中的词经过统计后按词频从高到低排列，则各个词的词频特征符合以下Zipf定律：

$$\text{Frequency} \times \text{rank} \approx \text{const}$$

就是说一个给定词的频数以这个词的排序号约等于其他词的频数乘以它本身的排序号。

目前被广泛采用的几个权重评价函数有：反比文档频数权重评价、信噪比、项的区分度、TFIDF、互信息量（Mutual Information）等方法。反比文档频数权重评价方法假设项的重要性正比于项的文档内频数（$FREQ_{ik}$），但反比于文档集中出现该项的文档的数量 $DOCFREQ_k$，称文档频数。信噪比是从信息论的角度来对索引项的重要性进行评价。一个项出现的频率越高，它所包含的信息量就越少。信息量的公式是：$I = -\log_2 P$，P 为项出现的频率。项的区分度的方法最初是由 H. P. Luhn 提出的。假设 D_i 和 D_j 表示两个文档，每个文档都是一组项组成的集合。用 sim（D_i，D_j）表示两个文档之间的相似度，相似度取值范围为［0，1］。互信息是一个信息论的概念，互信息衡量的是某个词和类别之间的统计独立关系。

TFIDF 方法的权重计算公式如下：

$$w_{ik} = tf_{ik} \cdot idf_k$$

tf_{ik}（Term Frequency）表示项 t_k 在文档 D_i 中的文档内频数，idf_k（Inverse Document Frequency）表示项 t_k 的反比文档频数，它们有很多种计算方法。其中最简单的公式是：

$$w_{ik} = tf_{ik}/df_k$$

其中的 df_k 表示出现 t_k 的文档数目，称为文档频数。

这种加权策略的直观解释是：若某个项在某一文档中出现的频率越高，其贡献越大；但若该项在整个文本集中出现的文档数较多时，它的贡献将会被减弱。如果再考虑文档长度等因素，可以将上式对文档长度进行归一化处理或者用相对频率代替绝对频数。

3.3 实验结果及分析

笔者随机选择了 CNKI 期刊文献库中的科技论文进行了测试分析，下面是对《向知识标引进军》一文进行的测试结果，如图 5 所示。

从图 5 中可以看出，《向知识标引进军》一文可看成由 62 个向量构成的文献空间，每一个向量又是由不同权重的名词组成，可以用这些名词来线性表示。如第 2 句："最近又见到的《知识元挖掘》一书，是温有奎、徐国华等同志在以往论文梳理和提炼的基础上，进一步系统地总结和阐述了知识标引的基本原理、理论、方法等，实际上全书各章论述的内容都涉及信息服务走向知识服务不可回避的基础建设，涉及今后信息资源和知识资源建设的发展道路，涉及未来情报学理论研究的方向等，正是当前国内

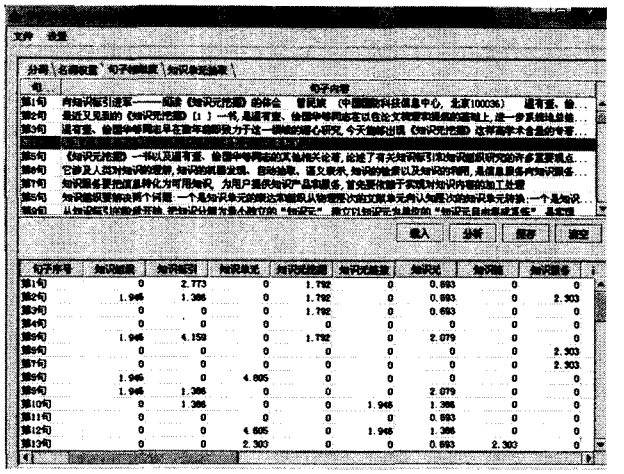

图 5 中文科技文献的名词权重矩阵

知识组织研究的热点课题"。可以表示成"1.946 × 知识组织 + 1.388 × 知识标引 + 2.303 × 知识服务 + 2.708 × 知识组织研究 + 1.792 × 知识元挖掘 + 0.693 × 知识元"。图 5 中的名词权重矩阵中的每一列表示该词在每句话中的权重，每一行表示每个词在该句中的权重值。因此，我们通过每一列中的纵向求和，就可以得出每个词在该篇文献中的权值，如表 1 所示。

表 1 前 20 个词在《向知识标引进军》一文中的权重值

词名	词权	词名	词权	词名	词权	词名	词权
知识元	45.748	知识元链接	19.459	知识链	13.816	知识网络	12.425
创新点	29.789	知识组织	19.459	知识网格	13.816	通过知识元链接	11.983
知识标引	29.112	知识结构	17.918	计算机	13.816	信息科学基本方程	11.983
知识元挖掘	23.293	知识组织研究	16.248	基本方程	13.54	客观知识	10.832
知识单元	20.723	知识服务	16.118	Brookes	13.54	有关知识	10.832

由此可见，能表征《向知识标引进军》一文的关键词应该是："知识元"、"创新点"、"知识标引"、"知识元挖掘"、"知识单元"等词。

我们选取了 CNKI 文献库中的近 50 篇科技文章进行了实验测试，结果能识别出 45 篇绝大多数的关键词，如"条件互信息/new、特征选择算法/new、信息增益/new、数字指纹/new、句子相似度计算/new"等。但也出现了一些没有任何意义的词，如"具有最/new、出文档/new、主要内/new"等，主要是没有对新词进行过滤。在有些文章中，如《基于多元判别分析的文本分割模型》一文中的关键词"遗传算法"，被识别成"遗传/v 算法/n"，是由于"遗传算法"在该文中出现的次数很少，故未被算法识别，这也从一个侧面说明作者提出的关键词有时不能很好地反映中文科技文献的语义和主旨。

由于笔者提出的算法是基于词频统计而不是词库来发现新词，所以科技文献篇幅越长，分词得到的结果相对越准确，对于短篇文献，虽然能通过降低阈值的方法来识别出新词，但也带来了一些副作用，主要包括一些没有意义的组合词。

4 结束语

随着时代的发展，汉语中新词语的不断涌现是一个客观规律。而当今新词语发现的研究还不能很好地满足人们的现实需求。

本文提出了一种基于向量空间模型和特征项权重计算的中文科技文献关键词抽取算法，首先用改进的方法对文本进行分词，然后用 TFIDF 权重方法对文献空间的每个词进行权重计算，最后根据计算结果选取能表征该科技文献语义的关键词。通过自编软件进行的实验测试表明该算法对中文科技文献的关键词自动抽取成效显著。今后的工作主要是在词权矩阵的基础上，通过计算句子间的相似度来分析文章间的语义关系，并用模式识别的方法获取科技文献的知识单元。□

参考文献

[1] 夏迎炬. 文本过滤关键技术研究 [D]. 上海：复旦大学，2003

[2] 索红光，刘玉树，等. 一种基于词汇链的关键词抽取方法 [J]. 中文信息学报，2006，20 (6)

[3] 刘佳宾，陈超，等. 基于机器学习的科技文摘关键词自动抽取方法 [J]. 计算机工程与应用，2007，43 (14)

[4] Turney P D. Learning to extract keyphrases from text [R]. National Research Council, Canada, NRC Technical Report ERB-1057, 1999

[5] Witten I H , Paynter G W, Frank E, et al. KEA: practical automatic key-phrase extraction [C] //Proceedings of the 4th ACM Conference on Digital Libraries, Berkeley, California, US, 1999: 254 - 256

[6] Hulth an improved automatic keyword extraction given more linguistic knowledge [C] //Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003: 216-223

[7] Yang Wenfeng. Chinese keyword extraction based on max-duplicated strings of the documents [C] //Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002: 439 - 440

[8] 李素建，王厚峰，俞士汶，等. 关键词自动标引的最大熵模型应用研究 [J]. 计算机学报，2004，27 (9)：1192 - 1197

[9] 王军. 词表的自动丰富——从元数据中提取关键词及其定位 [J]. 中文信息学报，2005，19 (6)：36 - 43

作者简介：徐文海，男，1982 年生，硕士生。

温有奎，男，1951 年生，教授。

收稿日期：2007 - 08 - 07

(上接第 316 页)

Korea: a micro analysis [R]. Working Paper, International Labour Office In Geneva, 1982: 216-224

[12] 金匡烁，金素锦. 韩国竞争情报 [EB/OL]. [2005-12-26]. http: //www. scip. org

[13] Nakagawa J. Asia, FTA and intelligence [EB/OL]. [2006-05-25]. http: //www. Cireline. com

[14] 2007 中国创投融资发展高层论坛部分嘉宾简介：李镇樟 [EB/OL]. [2007-07-12]. http: //finance. sina. com. cn

[15] 梁战平，我国科技情报界显现的新视点 [J]. 中国信息导报，2006 (7)：13-18

[16] 彭靖里，等，对当前我国竞争情报产业化发展的思考 [J]. 中国信息导报，2007 (1)：12-15

[17] 朱东华，袁军鹏，雷静，等. 论技术监测的对象 [J]. 科研管理，2006，27 (1)：23-28

[18] 陈劲，余芳珍. 技术创新审计模型及其应用研究 [J]. 研究与发展管理，2006，18 (5)：9-14

[19] 刘平，张静. 专利地图制作及应用例析——以激光信息存储技术 [J]. 管理学报，2005，2 (5)：555-558

[20] 赵刚，等. 技术创新与企业竞争 [M]. 北京：华夏出版社，2003：64-72

[21] 陈峰. 开展竞争情报与技术预见交叉研究的若干发现 [J]. 图书情报工作，2007，51 (2)：26-29

[22] 仪德刚，齐中英. 从技术竞争情报、技术预见到技术路线图 [J]. 科技管理研究，2007，27 (3)：13-14，18

[23] 罗时凡，刘书兰. 我国实施“市场换技术”战略的再思考 [J]. 对外经贸实务，2006 (11)：65-69

[24] Wayne A, Krans R. 技术情报的过去、现在和未来，竞争情报应用战略——企业实战案例分析 [M]. 普赖斯科特，等编. 包昌火，等译. 长春：长春出版社：303-313

[25] 吴晓伟，宋文官，徐福缘. 竞争情报软件发展现状和趋势研究 [J]. 情报杂志，2007，26 (4)：70-72

[26] Chiesa V, et al. Development of technical innovation audit [J], IEEE Engineering Management Review, 2003: 64-84

作者简介：彭靖里，男，1959 年生，研究员，硕士生导师。发表论文 40 余篇。

李建平，男，1963 年生，副教授。

杨斯迈，男，1948 年生，教授，博士生导师。

张伟，男，1968 年生，工程师，硕士生。

收稿日期：2007 - 10 - 08