基于共词分析的科技文献趋势挖掘①

吴潇泽，王小华，谌志群

(杭州电子科技大学 智能信息处理实验室，杭州 310018)

摘 要：采用共词分析方法对中文信息学报 2000-2009 年所发表的文献进行研究和分析，借助多元统计学中的聚类法，绘制出每两年为一时间段的主题战略坐标图，研究结果揭示了各时间段的研究热点分布以及主题演变情况，总结出新兴学科研究主题发展的一般规律，并探讨了中文信息处理领域的主题发展趋势。

关键词：趋势挖掘；共词分析；聚类分析；战略坐标图

Scientific Literature Trend Mining Based on Co-Word Analysis

WU Xiao-Ze, WANG Xiao-Hua, CHEN Zhi-Qun

(Intelligent Information Processing Laboratory, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: In this paper, co-word analysis was used for researching and analyzing the literature published in Journal of Chinese Information Processing in 2000-2009, using clustering method of multiple statistical, and map every two years' themes strategic diagram. The result reveals every period's research focus and themes' evolution and summarizes the general pattern of the research themes' development in emerging subject, and explores the themes of Chinese information processing trends.

Keywords: trend mining; co-word analysis; cluster analysis; strategic diagram

科技文献趋势分析对研究人员具有重要意义。它有助于研究人员把握学科领域研究热点，了解学科发展趋势，从而帮助研究人员做出决策。陈仕吉对目前科学研究前沿探测方法做出了总结，从引文分析和主题词两个角度探讨科学研究前沿的探测方法与技术，并分析各种方法的优缺点和应用环境[1]。

共词分析法最早在 20 世纪 70 年代中后期由法国文献计量学家提出，1986 年，法国国家科学研究中心的 M.Callon、J.Law 和 A.Rip 出版了《Mapping the dynamics of Science and Technology》[2]，此后，共词分析法被应用到多个领域。新加坡南洋理工大学的 Ying Ding, Gobinda G.Chowdhury, Schubert Foo 利用共词分析方法展现了信息检索领域 1987-1997 年的知识图谱，并对信息检索领域研究主题变化发展情况进行了一些预测[3]；我国学者蒋颖利用共词聚类方法，对 1995-2004 年全球文献计量学领域的主题内容进行分析，总结出文献计量学领域内部结构的变化以及主题的演变趋势[4]；张晗、崔雷利用共词聚类方法以及类团的战略坐标分析法，对生物医学的现状进行分析，总结出生物医学的研究热点以及变化趋势[5]。

1988 年 Law 等提出了用“战略坐标”来描述某一研究领域内部联系情况和领域间相互影响情况[6]；1995 年由 Kostoff 等提出基于数据库内容结构分析的共词分析方法，可以用于分析大量的数字化文本资源的系统[7]，然而这些共词分析方法的改进都没有涉及研究主题的演变规律。

本文以共词分析方法为理论基础，SPSS17.0 和 Eclipse 为工具，绘制出各个时间段的研究热点分布战略坐标图，并通过横纵向对比中文信息处理领域各阶段的研究热点变化，研究各个主题在不同时间段的参数变化，力图找出科技主题演变的普遍规律。

① 基金项目:教育部人文社会科学研究项目(08JC740011)
收稿时间:2010-08-10;收到修改稿时间:2010-09-11

1 数据和方法

共词分析方法基于这样一种假设：文献的关键词是关于文献内容的充分描述，如果两个不同的关键词出现在同一篇文献中，则认为这两个关键词之间有一定的联系[8]。基于这种概念，研究主题可以用几个特定的关键词来表示。通过计算所有选定关键词的两两共现频次，得到关键词的共现矩阵。再由聚类分析将相互联系大的关键词聚为一类，得到代表研究主题的关键词类团[9]。然后通过知识图谱展示出来，常用知识图谱可视化包括多维尺度图分析法[10]、战略坐标图分析法[6]。钟伟金等人详细分析了共词分析法的过程与方式[11]、共词分析法的类团分析[12]、共词聚类分析法的原理与特点[13]。

1.1 数据来源

本文以 noteexpress 为检索工具，选择《中国期刊全文数据库》为数据源，检索期刊为中文信息学报，检索年限为 2000-2009 年。经去除重复、会议通知等对表达研究主题无意义的文献后共得 815 篇文献。将 10 年的文献分成 5 个时间段，2000-2001；2002-2003；2004-2005；2006-2007；2008-2009。其中 2000-2001 年文献 107 篇，关键词 409 个；2002-2003 年文献 112 篇，关键词 554 个；2004-2005 年文献 161 篇，关键词 850 个；2006-2007 年文献 207 篇，关键词 1170 个；2008-2009 年文献 228 篇，关键词 1325 个。

1.2 数据处理及方法

1.2.1 关键词标准化

816 篇文献中共计 1910 个独立关键词，这些关键词中，有些不同的关键词表达了相同或相近的意思，需要进行关键词人工合并处理。

例如：手写汉字识别=汉字识别；空间向量模型=向量空间模型；等等。

另有部分关键词所表达范围过大，在关键词处理中应将其剔除。

例如：人工智能、自然语言处理、中文信息处理等词。

1.2.2 关键词排序

标准化后对关键词进行词频统计并排序。此项操作可于 access 数据库中用 SQL 语言实现。

SELECT keyword, count(*) AS 频次 FROM 表名 GROUP BY keyword ORDER BY count(*) DESC；

1.2.3 高频词选定

关键词经标准化、排序后选定每个时间段中出现频次大于 1 的关键词作为研究对象。

1.2.4 构造关键词共现矩阵

两两统计关键词对在同一篇文献中出现的次数，n 个关键词则形成一个 n*n 的关键词共现矩阵。

1.3 聚类分析

以关键词共现矩阵为基础进行聚类分析，将相似度大的关键词聚为一类。相似度大小以距离远近度量，两词距离越大，相似度越小；距离越小，相似度越大。两个关键词之间的距离采用平方欧氏距离来度量。类团间距离用 ward’s method[15].

$$d_{xy}=\sum(x_i-y_i)^2 \quad i=1,2,\cdots..n$$

1.4 战略坐标图

战略坐标图建立在关键词共现矩阵和聚类分析的基础上，用可视化图谱方法来表示产生的结果。战略坐标图有两个指标：向心度、密度。

向心度表示研究主题间的联系强度，可以用代表该研究主题的关键词与其它研究主题的所有关键词共现频次之和来表示[15]，向心度用来量度一个类团与其它类团的联系程度。一个学科领域与其它学科领域联系的数目和强度越大，这个学科领域在整个研究工作中就越趋于中心地位[16]。

密度是用来度量使关键词聚合为一类的联系的强度，可以用代表该研究主题的所有关键词两两共现频次总和来度量[15]。以向心度为横轴，密度为纵轴，坐标原点为两个轴的平均数。则可以将空间划分为四个象限。

基于战略坐标图的思想，以研究主题出现频次代替密度，向心度定义为代表该研究主题的关键词与所有外部主题的关键词共现频次之和。可以得到一个二维坐标图,如图 1 所示：

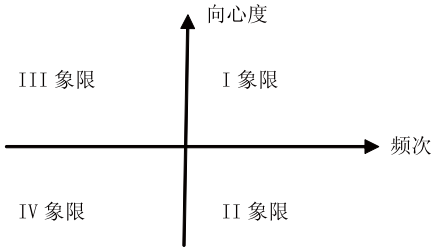

图 1　战略坐标图

处于 I 象限的主题，出现频次高，向心度大，和其它研究主题间的联系密切，处于整个研究领域的核心阶段。处于 II 象限的主题，出现频次高，向心度小，说明处于该领域的主题受到越来越多的关注，具有很大的发展潜力。处于 III 象限的主题，出现频次低，向心度大，说明处于该领域的主题受到过学者的关注，并已经发展为成熟主题。处于 IV 象限的主题，出现频次低，向心度小，说明处于该领域的主题处于萌芽状态。

2 实验结果

对各时间段关键词进行以上步骤处理后，可以得到各自聚类结果。以 2000-2001 年为例。

2000-2001： 聚类结果

A：汉字识别、版面分析、神经网络、分类器

B：自动分词、中文姓名识别、支持向量机

C：信息检索、汉字编码、全文检索、数据压缩

D：文本分类、向量空间模型、特征提取、机器学习

E：语料库、语言模型、语音识别、词性标注

F：关系数据库、ER 模型

G：机器翻译

H：句法分析

I：少数民族文字处理

J：地形图

K：汉字输入法

在聚类结果的基础上，计算各主题的频次、向心度，可以绘制出各时间段的战略坐标图，见表 1：

表 1 2000-2001 年研究主题频次向心度(部分)

	A	B	C	D
向心度	4	9	3	4
频次	15	7	7	5

战略坐标图原点定义为所有主题的出现频次均值，向心度均值。

战略坐标图原点：（频次 向心度） （5.4 2.1）

绘制各主题所处战略坐标图位置，可得 2000-2001 年中文信息处理领域研究主题分布情况，如图 2 所示。

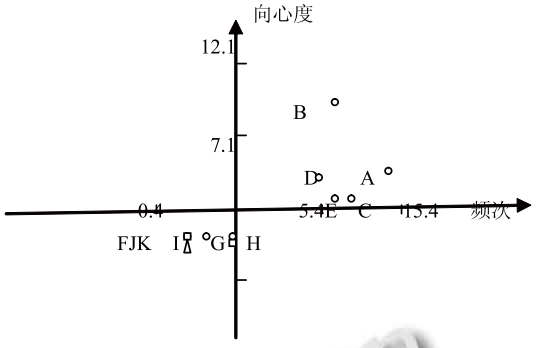

图 2 2000-2001 各研究主题的战略坐标分布

3 实验结果分析

通过对比不同时间段战略坐标图的变化情况可以得出各个时间段研究热点的变化。另外通过对比不同时间段同一研究主题的频次、向心度变化可以预测该研究主题的走势。

3.1 文献数量变化

对比各时间段中文信息学报所发表文献数量，可知国内学者对中文信息检索领域的研究兴趣日益增长，见表 2。

表 2 各时间段所发表文献数量

2000-2001	2002-2003	2004-2005	2006-2007	2008-2009
107	112	161	207	228

3.2 中文信息处理领域研究热点变化情况

统计各时间段中文信息处理领域研究热点，并进行比较分析可知机器翻译、语料库、文本分类一直处于中文信息处理领域的研究热点。中文字体识别处理经过近年来的发展已经由汉字识别处理转向少数民族文字处理。一些经典的概率图模型如条件随机场、最大熵模型得到越来越多的关注，被广泛应用在中文信息处理的各个领域，等等。

3.3 关键词类团向心度分析

向心度用以度量各学科主题间的相互联系强度。计算各时间段研究主题平均向心度，见表 3。

表 3 各时间段研究主题平均向心度

2000-2001	2002-2003	2004-2005	2006-2007	2008-2009
2.1	1.4	1.9	7.3	7.6

通过对比可见中文信息处理领域研究主题间相互

联系的强度越来越大，代表了各主题之间相互渗透的趋势愈来愈强。

3.4 主题演变趋势分析

通过对比分析单个主题在整个研究领域内的频次、向心度变化，进而预测该主题的发展趋势。

以少数民族文字处理和文本分类为例,分别计算其各时间段的频次、向心度，然后绘制其主题变化战略坐标图。见图 3、图 4。

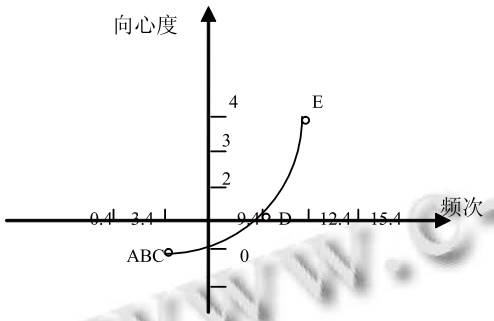

图 3　少数民族文字处理主题变化战略坐标图

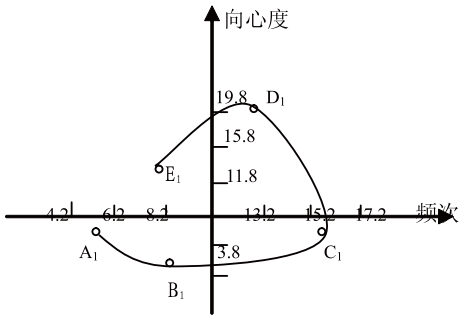

图 4　文本分类主题变化战略坐标图

综合各主题随时间段的战略坐标图变化情况可以得出：研究主题总是遵循着战略坐标图中的象限变化规律：

IV→II→I→III

其物理意义解释：当某一个研究主题出现时，受到的关注度较低，且与其它主题联系不强，处于其发展历程战略坐标图的 IV 象限；随着研究的深入，该主题受到的关注度加强，处于战略坐标图的 II 象限；当该主题研究达到一定程度后，与外界主题的联系加强，处于战略坐标图的 I 象限；最后，该主题研究进入 II 象限，受到的关注度逐渐降低，发展成为成熟主题。

3.5 趋势预测

绘制 2008-2009 年各研究主题战略坐标图，统计各主题所处战略坐标图象限位置，见表 4。

表 4　2008-2009 年研究主题所处各自主题发展战略坐标图象限位置

I	II	III	IV
机器翻译	垃圾邮件过滤	自动分词	
语料库		语音识别	
少数民族文字处理		文本分类	
信息检索、聚类分析		问答系统	
机器学习		信息抽取	
句法分析		语义消歧	
垃圾邮件过滤		词法分析	
		手机输入法	

通过各个时间段主题战略图分析可知：

未来一段时间内，机器翻译、语料库、少数民族文字处理、信息检索、机器学习、句法分析、垃圾邮件过滤处于其自身发展历程战略坐标图的 I 象限，处于此象限的研究主题趋于成熟，在一段时间里仍是整个中文信息处理领域的研究热点。

自动分词、语音识别、文本分类、问答系统、信息抽取、汉字识别、词法分析、手机输入法处于其自身发展历程战略坐标图的 III 象限，已经成为成熟主题，研究关注度逐渐减少。

文本挖掘处于其发展历程战略坐标图的 II 象限，受到的关注度将越来越高。

4　结束语

本文通过关键词共现分析，分时间段绘制了中文信息处理领域近 10 年的战略坐标图，通过各时间段战略坐标图研究主题的横纵向对比，总结了中文信息处理领域研究主题的变化发展情况和研究主题自身发展规律，并对今后中文信息处理领域内主题走势做出预测。与其它挖掘方法相比较而言，本文所描述挖掘过程总结出了研究主题变化的一般规律，为该主题的趋势预测提供更为充分的依据，使研究人员能够清晰把握到某一学科领域研究主题所处位置，是否有必要进行深入研究等。本文的后续工作是对表达研究主题发展情况的战略坐标图进行改进，用更为精确的参数来描述主题的发展情况，为科技文献的趋势预测提供更为有利的依据。

参考文献

1 陈仕吉.科学前沿探测方法综述.现代图书情报技术,2009,(9):28-33.

2 Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology：Sociology of Science in the Real World. Macmillan, 1986.

3 Ding Y, Chowdhury GG, Foo S. Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing and Management, 2001,(37):817-842.

4 蒋颖.1995-2004 年文献计量学研究的共词分析.情报学报,2006,25(4):504-512.

5 张晗,崔雷.生物信息学的共词分析方法研究.情报学报,2003,22(5):613-617.

6 Law, et al. Policy and the mapping of scientific change：a Co-Word analysis of research into environment acidification. Scientometrics, 1988,14(3-4):251-264.

7 Kostoff RN, et al. Database tomography for information retrieval. Journal of Information Science, 1997,23(4):301-311.

8 Cambrosio A, Limoges C, et al. Historical scientometrics Mapping over 70 Years of biological safety research with coword analysis. Scientometrics, 27(2):119－143.

9 黄咏梅.读者需求分析中的数据挖掘技术.大学图书情报学刊, 2006,24(4):48-50.

10 Noyons ECM, van Raan AFJ. Bibliometric cartography of scientific and technological development of an R&D field. Scientornetrics, 1994,30(1):157-173.

11 钟伟金,李佳.共词分析法研究(一)共词分析的过程与方式.情报杂志,2008,5:70-72.

12 钟伟金,李佳.共词分析法研究(二)类团分析.情报杂志,2008,6:141-143.

13 钟伟金,李佳.共词分析法研究(三)共词聚类分析法的原理与特点.情报杂志,2008,7:118-120.

14 Arabie P, Carroll JD, Desarbo WS. Three-way Scaling and Clustering. Newbury Park: Sage Publication.1987.

15 Callon M. Courtial JP, Laville F. Co-word analysis as a tool for describing the netword fo interactions between basic and technological research: the case of polymer chemistry. Scientornetrics, 1991,22(1):155-205.

16 邵峰晶,于忠清.数据挖掘——原理与算法.北京:中国水利水电出版社,2003.91-98.

（上接第 160 页）

5 结论

在对矿权管理问题研究基础上，借助 Arc Engine 以及 C#软件的基本功能，以及关系数据库的理论，开发了矿权管理系统，实现了对矿区开采等问题的数据管理以及矿区的查询、统计、分析等功能。在某种程度上，可以规范矿产权的管理市场，避免资源浪费和生态破坏等问题。具有广泛的应用前景。

参考文献

1 王贵山.最新矿业权、探矿权、采矿权与矿产资源宏观调控及规划管理.北京:中国矿大出版社,2008.

2 李玉龙.ArcGIS 地理信息系统教程.北京:电子工业出版社,2004.

3 韩鹏,王泉.地理信息系统开发——ArcEngine 方法.武汉:武汉大学出版社,2008.

4 冯克忠,崔纪锋.ArcObjects 开发指南.第 2 版.北京:电子工业出版社,2003.

5 张正祥,张洪岩.ArcObjects 组件在地理信息系统二次开发中的应用.遥感信息,2004,2:34-45.

6 Huddleston J, et al.杨浩,等译.C#数据库入门经典.第 2 版.北京:清华大学出版社,2006.

7 什么是.NET Framework? [2010-06-30]. http://www.microsoft.com/china/msd

8 汤国安.地理信息系统空间分析实验教程.第 2 版.北京:科学出版社,2007.

9 邬伦,刘瑜.地理信息系统原理与方法应用.北京:科学出版,2003.

10 温晓蕾.基于 ArcGISEngine 的历史街区保护管理信息系统的研究与开发[硕士学位论文].重庆:西南大学,2008.

11 聂小波,吴北平,何保国.基于 ArcGIS Engine 的专题图模块的设计与实现.地理空间信息,2006,4(l):12-14.

12 吴玮,李小帅,张斌.基于 ArcGIS Engine 的 GIS 开发技术探讨.科学技术与工程,2006,6(2):176-178.