·理论探索·

应用 Autonomy 专题聚类方法揭示领域学科热点

岳 婷 张建勇
（中国科学院文献情报中心，北京 100190）

〔摘 要〕共词聚类分析是情报学中进行学科热点探测、掌握学科发展脉络的一种主要方法，目前已经比较成熟得到了广泛的应用。Autonomy公司开发的autonomy智能搜索系统也同样具备专题聚类的功能，本文对该系统专题聚类的原理以及功能进行了阐述，并用CSCD的试验数据对系统的聚类功能进行测试。通过对试验结果的分析和解释，证明了autonomy系统的专题聚类功能具有一定的应用价值，可以与其他聚类方法结合起来，对探测学科热点提供一定的帮助。

〔关键词〕autonomy系统；聚类；学科热点
〔中图分类号〕G353.1 〔文献标识码〕B 〔文章编号〕1008-0821（2009）08-0025-04

The Application of Autonomy on Exploring Discipline Hotspots

Yue Ting Zhang Jianyong
（Library of Chinese Academy of Sciences，Beijing 100190，China）

〔Abstract〕Co-word clustering analysis is a kind of main method to explore the discipline hotspots，and it is widely used in Information Science. Autonomy search system developed by autonomy company also has a function of clustering. The paper expounded the theory of autonomy's clustering and tested this function by some data in CSCD. By analyzing the clustering result，Autonomy's clustering is effective，and it could be used with other clustering method to find the discipline hotspots.

〔Key words〕autonomy；clustering；discipline hotspots

共词聚类分析是情报学中进行学科热点探测、掌握学科发展脉络的一种主要方法。它的主要原理是：选取一组文献的高频主题词，两两统计它们同一篇文献出现的频率，形成一个高频主题词的共词矩阵，以这个矩阵计算生成的相似矩阵为基础，利用聚类的方法来判断哪些主题词的关系紧密。这些关系密切的主题聚集在一起形成类团，表达某一领域分支的组成[1]。这种方法已经相对比较成熟，已经在学科领域热点的探测中得到了广泛的应用。

Autonomy系统是一个基于语义计算的智能搜索系统，专题聚类分析也是该系统的一个重要功能。系统的聚类分析是建立在香农信息论和贝叶斯概率论的基础之上，其原理与通常所使用的共词聚类分析不同。并且，Autonomy还具有对聚类结果进行可视化的功能。本文对Autonomy系统专题聚类的原理进行了分析和阐述，并尝试用Autonomy系统对中国科学引文数据库中图书情报领域的文摘数据进行聚类分析，旨在为揭示领域内的学科热点和研究结构提供一种新的思路和方法。

1 Autonomy系统的专题聚类原理分析

香农信息论和贝叶斯概率论的结合应用是Autonomy系统的特点之一。

香农对消息和信息进行了区分：消息由于具有不确定性而含有信息，对消息进行通信可以消除或部分消除这种不确定性。而信息是对事物运动状态或存在方式的不确定性的描述。也就是说，信源能够发出一系列的消息，消息经过通信消除了不确定性而变成信息。

香农的研究表明，如果信源$\{x1,x2\cdots,xn\}$所发生的概率分别是$\{p1,p2,\cdots pn\}$，那么每个信源消息xi发出后，产生的信息量为$I(xi)=-log(p(xi))$[2]。这个函数是一个负对数函数，说明一个信源消息发出的概率越大，它所产生的信息量越少。这是信息论的基本观点。

如果把一篇论文看作是一个信源，它含有若干个词语，

收稿日期：2009-02-11
作者简介：岳 婷（1981-），女，硕士研究生，研究方向：数据集成与网络服务系统，发表论文1篇。

一个词语重复的频率越多，其内容越不具有概括性，反之其包含的信息内容越丰富。

贝叶斯概率的计算公式为：

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{\sum_{\theta' \in \Theta} P(x|\theta') \cdot P(\theta')}$$

这一公式主要用于计算多个变量之间的概率关系，以及确定一个变量对另一个变量的影响程度。

贝叶斯概率论的本质是当一个事物的本质不能被准确知悉时，可以依靠与这一事物本质相关的事件出现的多少去判断其本质属性的概率。将这一理论应用到论文的聚类分析中：通过论文中一个词语出现的频率的多少和与其它词语之间的关系来决定其成分的重要性。论文中的每个词语的权重、论文间词语的相关度不仅由其本身出现的频率决定，还取决于与其他词语之间的关系。

Autonomy 系统的聚类分析方法是这样的[3]：

(1) 系统的聚类分析是抽样进行的，首先根据论文集合的数量计算出抽样的次数和每次抽样的文档数量，并开始抽样。

(2) 对于每次抽取出的样本论文，利用香农信息论抽取论文中的信息内容最丰富的重要词汇（系统称之为“概念”）作为聚类的主题来源（系统支持学科专业词表，基于词表抽取的概念能够更加规范化）。

(3) 基于贝叶斯概率论计算每个概念在单篇论文中的权重以及在系统中所有论文中的权重，根据这两个权重计算概念之间的相关度，相关度大于某个阈值的聚为一类。当系统中论文集合有所变化时，概念在所有论文中的权重会随之变化重新进行计算。因此，系统对概念之间的相关度计算不仅依赖于概念在单篇论文中的出现的词频，更加依赖于其所在的上下文环境以及与其他概念之间的关系。

表 1 中对 Autonomy 使用的聚类分析方法与共词聚类方法的主要特点进行了对比。

表 1 Autonomy 使用的聚类分析方法与传统共词聚类方法的对比

	共词聚类分析	Autonomy 聚类
聚类主题来源	高频主题词或关键词，由信息标引者或作者直接给出	基于香农信息论，动态抽取含有信息量最多的概念，由系统自动完成
计算相关度的方法	计算词对在同一论文中出现的频率形成共词矩阵，词与词之间的相关度仅与共同出现的频率有关	基于贝叶斯概率论计算概念权重，概念之间的相关度计算与上下文环境以及其他概念具有关联性

2 基于 Autonomy 的信息聚类试验

2.1 数据来源与试验方法

本研究所采用的试验数据来自中国科学引文数据库(CSCD)[4]，按照中国图书馆分类法分类号为 G250 进行检索，共得到 1997 - 2006 年图书情报类中文文摘数据 1 316 条。将1 316条文摘数据按照 Autonomy 系统规定的格式转化成 XML 文档导入系统。

抽取1 316篇论文中的关键词，写入到系统 userdic.txt 文件中，作为系统概念抽取时的专业词表。

系统中检索词设置为空值，即对所有数据进行聚类，设定出版时间为“2000 年 1 月 - 2008 年 1 月”，相关度为“60”。

2.2 试验结果与分析

2.2.1 试验结果

聚类结果如表 2 和图 1 所示。

表 2 图书情报类文摘数据聚类分析结果

序号	标题	文档数
1	企业/图书馆知识，本体	37
2	企业竞争情报，wto	27
3	用户兴趣，个性化，网页	27
4	集成服务，个性化，决策	23
5	语义检索，xml，查询	23
6	xml，档案，检索，着录	22
7	企业信息化，cio，cko	21
8	图书情报，期刊论文	21
9	情报/情报学学科，文献学	20
10	主题，标引，检索，词	19
11	rdf，发布，检索，语义	17
12	专利，内涵，建设，高校图书馆	17
13	共享，共建，图书馆，情报	17
14	传统图书馆，图书馆，职高，馆员	14
15	网站链接，影响力，期刊	14
16	信息搜索行为，搜索，用户，科技数据库网站	13
17	wto，信息化，工业化，战略	13
18	图书馆，版权，知识产权保护，观念	13
19	主题，引文分析，情报学，期刊	13
20	实施 erp，cims，企业	13
21	图书情报，兴起，网格技术	10
22	jsp，发布，查询，申报	9
23	compendex，实证，数据库，期刊	8
24	xml，信息结构，导航，超文本	7
25	ei compendex，embase，检索	6

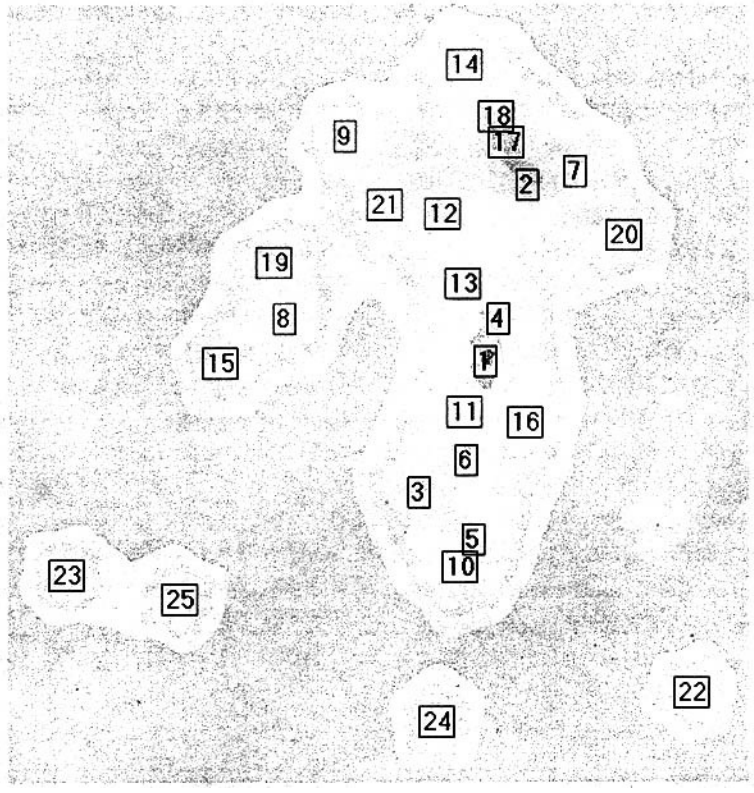

图 1 图书情报类文摘数据聚类分析信息岛图

(在同一“信息岛”上颜色越深表明该主题研究越为热点；同一信息岛中的类簇或信息岛之间的距离越近说明主题之间的关联越大)

2.2.2 专题聚类分析

对表 2 的聚类分析结果进行进一步分析，可以看出：

(1) 企业/图书馆知识，本体

“企业/图书馆知识，本体”是聚类结果中最热点的主题。近几年来，无论是企业还是图书馆，都开始重视知识管理和知识服务的研究和探索。企业对自己的资源和知识进行有效的组织和管理，能够不断挖掘企业自身的创新点，给企业的生存和发展带来更大的空间。对图书馆来说，其业务发展正在逐渐的由“信息管理”向“知识管理”转变，更加注重隐性知识的搜集、整理、存储和应用，为用户提供更加深层次的服务，这是图书馆未来发展的趋势，这方面的研究自然成为近几年图书馆学研究的热点之一。

图书馆界有关本体的研究已经开始了一段时间，近几年研究的重点在于本体的构建和应用方面，如何把本体应用到图书馆的信息组织和检索系统中去，实现其真正的功能。对照类簇名称中“本体”阅读相关的文摘，《面向知识处理的领域本体及其应用研究》[5]，《数字图书馆领域本体构建研究——以数字参考咨询领域为例》[6]，《VISION：集成分类法、主题词表和语义元数据的概念网络》[7]，都是本体的应用实例研究，而不再仅仅局限在理论上的探讨。

(2) 企业竞争情报，wto

对这一类的论文进行进一步分析，大多属于关于“获取企业战略竞争情报的方法，系统的构建”以及“人际网络分析”方面。企业竞争情报主题一直以来就是是情报学的热点问题。值得关注的是，近年来对人际网络的研究逐渐多了起来，用“人际 情报 网络”在维普数据库进行检索，有 38 篇文章，发表时间都在 2005－2007 年的。秦铁辉等人在 2007 年发表的文章《竞争情报与人际网络研究述评》[8]中指出，“近年来，随着人际网络理论在各个领域的广泛应用，竞争情报活动中的人际网络也引起了国内外学者的关注。”

(3) 用户兴趣，个性化，网页

这一类簇中的文章，主要包括这样两类：通过网页日志或是一些算法对网页中的用户行为进行分析，实现搜索引擎或是信息检索系统的个性化推荐服务；网页信息的抓取和组织，如《搜索引擎检索结果的组织技术》[9]、《网站频道关键词选择方法研究》[10]等。

随着用户信息素质的不断提高，他们的信息需求越来越趋向多样化。利用数据挖掘、数据推送、网页跟踪、协同过滤等信息技术为用户提供个性化服务，对庞大的信息进行有效的组织和呈现，不仅是搜索引擎开发商们未来发展的关注的热点，同时也是数字图书馆不断努力的方向。

(4) 语义检索，xml，查询

语义网环境下，对“语义检索”的研究自然成为研究的焦点。XML、RDF 等信息组织的语言和框架如何真正的应用到信息检索系统中去，也是图书情报领域研究者比较关注的问题。这一类簇中的论文正是体现了这一特点。

此外，其他类簇所出现的“知识产权保护”等词也属于目前图书情报领域比较热衷的话题。

2.2.3 主题之间关联分析

图 1 中类簇 1，3，4，5，6 所包含文献的研究内容都是有关图书馆服务，信息集成，个性化服务等，主题之间有一定的联系，因此在它们属于同一个信息岛中的一个热点区域内；而类簇 2 同样属于热点研究内容，却与 1，3，4，5 联系相对较少，它与类簇 7，17，18 的主题相关，主要研究企业的信息化，企业的战略以及 WTO 等。

第 23 和 25 个类簇形成一个小的“信息岛”，与大的“信息岛”有一定的距离，说明这两个主题的研究内容相对比较独立。类簇 23 是有关“compendex，实证，数据库，期刊”，类簇 25 为“ei compendex，embase，检索”，这两个类簇的研究内容都是有关某个特定数据库的分析和试验，与大的“信息岛”的各个研究主题相关性较小。而这两个类簇的研究内容之间却有较高的相关性。

3 讨论与结语

通过对 CSCD 数据库图书情报领域中文期刊文摘数据聚类结果的初步分析，结合其他的综述性文献的阐述，可以看出，利用 Autonomy 对这一领域专题聚类的效果基本符合实际情况，能够初步揭示图书情报领域近年来的主要研究热点。这也证明了 Autonomy 系统所使用的香农信息论和贝叶斯概率论相结合的聚类分析方法对于判断领域热点来说是有效的。

同时，Autonomy 的可视化显示功能相对比较强大，不仅能通过同一信息岛内研究点颜色的深浅揭示研究热点，还可以根据信息岛之间距离的远近的变化来观测主题与主题之间的关联程度。这一方面比其他的聚类方法更加直观、清晰。

Autonomy 的专题聚类是对样本论文抽样进行的，聚类能否进行与样本量的大小有一定的关系，如果要对某一学科中某一具体主题进行热点分析，可能由于样本量不够而无法进行。因此，系统对某一大的学科领域的热点分析的效果还比较理想，但是当热点探测范围缩小到某一小的主题领域，对某一学科热点进行进一步的深层次挖掘时，还存在一定的局限性。Autonomy 的专题聚类分析只适用于对某一学科热点的初步揭示，而不适用于对学科热点进行更加深度的分析。在真正的实际应用中，可以把 Autonomy 的专题聚类与共词聚类分析或其他聚类方法结合起来，为情报人员对领域内热点的进一步分析提供帮助。

参 考 文 献

[1] 钟伟金，李佳，杨兴菊. 共词分析法研究（三）——共词聚类分析法的原理与特点 [J]. 情报杂志，2008，(7)：118－120.
[2] 李亦农，李梅. 信息论基础教程 [M]. 北京：北京邮电大学出版社，2004.
[3] Autonomy 核心技术说明 [S]. 2008.
[4] 中国科学引文数据库 [EB]. http://sdb.csdl.ac.cn，2008－11－03.
[5] 曾庆田，段华，杨红梅，等. 面向知识处理的领域本体及其应用研究 [J]. 情报学报，2006，(6)：713－719.
[6] 肖洪，余锦凤. 数字图书馆领域本体构建研究——以数字参考咨询领域为例 [J]. 大学图书馆学报，2006，(6)：26－29.
[7] 王军. VISION：集成分类法、主题词表和语义元数据的概念网络 [J]. 情报学报，2003，(4)：412－418.
[8] 秦铁辉，刘宇，杨蕴薇. 竞争情报与人际网络研究述评 [J]. 情报科学，2007，(12)：1761－1768.
[9] 赵荣，黄燕云，张露. 搜索引擎检索结果的组织技术 [J]. 情报学报，2004，(1)：69－72.
[10] 索红光，刘玉树. 网站频道关键词选择方法研究 [J]. 情报学报，2007，(2)：249－252.

(上接第 24 页)

4 结 语

总之，绩效考核是人力资源管理中最为重要的环节之一，它表现为对员工日常表现的考核，而不单单是对员工业绩和效率的考核，是人员晋升、任用、薪酬、培训、惩戒等人事决策的重要依据[7]。绩效考核工作是一项技术性的工作，是每位管理者必备的管理技术，虽然不是每一项绩效考核标准都可以量化的，但必须都是可衡量的，使绩效考评更全面、客观和公正，为避免一方考评的主观性和片面性，从不同的角度来考评，全方位、准确的考评人员的工作绩效，真正做鼓励先进，鞭策后退，提高图书馆的服务水平和质量。

参 考 文 献

[1] 徐建华. 现代图书馆管理 [M]. 天津：南开大学出版社，2003. 10：224－225.
[2] 蒋知义. 基于模糊综合评价的图书馆 HRM 绩效评估 [J]. 情报杂志，2006，(10)：79－81.
[3] 邱薇. 构建高校图书馆绩效考评指标体系的研究 [J]. 情报杂志，2006，(8)：135－137.
[4] 王茂梅. 浅议高校图书馆员工的业绩考核 [J]. 武汉职业技术学院学报，2007，(5)：113－117，120.
[5] 李安泽. 推行绩效管理 强化行政效能 [J]. 江西政报，2006，(17)：47－48.
[6] 王红. 高等学校图书资料人员绩效考评体系研究 [D]. 大连理工大学.
[7] 黄江瑛. 绩效考核的实践与思考 [J]. 交通企业管理，2006，(7)：14－15.