IFLA WLIC 2013
SINGAPORE

Submitted on: July 22, 2013

物理学论文的关键词分类及定量主题检索工具开发

Translation of the original paper "*Classification of Keywords Selected from Research Articles on Physics and Development of a Quantitative Subject Access Tool*"

Bidyarthi Dutta
Assistant Professor
Dept. of Library & Information Science
Vidyasagar University, Midnapore
West Bengal, India
E-mail: bidyarthi.bhaswati@gmail.com

Krishnapada Majumder
Professor
Dept. of Library & information Science
Jadavpur University
Kolkata; India

Bimal Kanti Sen
80, Shivalik Apartments
Alaknanda
New Delhi; India

Translated by <纪姗姗，Shanshan Ji>, <中国科学院国家科学图书馆，National Science Library, Chinese Academy of Sciences>, <中国，China >

摘要：

所有的论文都有一个标题，多数含有摘要，一些含有关键词。以上三个特征详细描述了一篇文章的内容：标题直接反映论文的中心主题，摘要概述了研究内容，关键词指出了文章关注的主要领域或相关领域。利用关键词，研究人员与索引人员能快速方便的查找到感兴趣的特定

文献。关键词对抽取及索引服务非常重要，并且在信息检索中发挥着重要的作用。本文收集了《Chaos》、《Physics of Plasmas》以及《Low Temperature Physics》三种期刊 2006-2012 年发表的 2526 篇学术论文，对其中的 14221 个关键词进行分析。在这些作者定义的关键词中，不同的群组有 2571 个。采集之后，词法上相似近关键词形成了词簇。以此方式创建了几个词簇，发现在整个时间段内几乎所有词簇中的关键词组合都有所变化。

基于词簇中不断变化的关键词组合，本文定义了四个指标，分别是：稳定度指数、综合可见性指数、短时可见性指数以及能力指数。不同的词簇拥有不同的指标值，对其值域进行分类，可分为 5 组：非常高、高、中、低、非常低。在这些指标的基础上，本文提出了一个新的定量主题检索工具，它能够预测某一主题领域中可能的新关键词与废弃关键词。该工具命名为 keysaurus（keyword-based-thesaurus）。

关键词：主题检索工具；信息检索；信息检索叙词表；keysaurus；知识分类

1 引言

“关键词”是我们日常生活中不可分割的一部分，我们经常有意或无意的使用它。关键词通常来源于一个受控词表，或是被自由指配。来自于受控词表的关键词可以在特定主题领域中提高文献检索的精确度。因此，关键词的选择是一个信息系统的重要举措。索引人员通常阅读一篇文献或文本，定位叙词表中最佳的词汇，然后分配能最优描述文档内容的词汇。以这种方式收集的关键词存储在一个检索索引器中。索引的功能实际上取决于人类对某个题目或主题的分析。面对同样的主题或题目，不同的索引人员可能会分配不同的关键词 [1]。像冠词（a、an、the）、介词（by、with、for、to）、连词（and、or、but）等常用语并不能做关键词，因为它们不能反映一篇文档的本质。几乎所有的英文文档或网站都有冠词“the”，但检索它并没有任何意义。最流行的搜索引擎 Google 在它的索引器中去除了停用词，如“the”、“a”等。有时，新兴的主题或概念可能缺乏适当的关键词进行描述。Suraud 等人 [2] 发现若在新兴领域中缺乏良好定义的关键词，会增加书目检索的困难度。关键词有时被成为“主题描述符”，该词由 Calvin Mooers 在 1948 年创造。

主题检索工具的标准，例如主题词表（商业主题词表、美国国会图书馆主题词表）或者分类表（杜威十进分类法、冒号分类法）都是基于受控词表，而不是基于用户术语。对于受控词表的研究表明，当有公认的常见术语集描述所关注的主题领域概念时，受控词表可以发挥更好的作用 [3]。Solomon[4] 指出，分类表常常不能发挥很好的作用，因为它们不是建立在用户语言和知识的基础上，也不是基于任务或应用情境而建立。Hurt[5] 建议有必要对索引及分类系统进行更新和扩展。Soergel 等人 [6] 认为现有的分类框架与叙词表缺乏明确的语义及结构一致性。随着电子信息与互联网的出现，物品的物理位置信息已经不那么重要。由此产生了对分类框架的重新审查，更注重对知识的获取。Bates 等人 [7] 提出对叙词表结构与网络信息系统的设计进行改进。如果对分类框架的需求不再局限于将文档存储于某一地点，那么主题等级的制定可以更加灵活。还有一种更大的可能性，即对分类表进行定制，以满足特定用户的特定需求。在传统的图书馆系统中，用户需要利用文档的标题以及作者姓名来进行检索，然而在数字化环境下更关注于用关键词进行检索。来自于不同领域的用户使用不同的关键词，

这些大量的关键词形成了不同的词簇。为了构建灵活的、可定制的分类表，需要了解信息空间中的用户视图，而分析关键词簇是一种有效的方法。该分析基于对关键词各种特征的统计。簇（群组）分析在许多学科中都有广泛的应用，尤其是基于文档的聚类分析为自动化分类奠定了基础[8]。

现有信息系统的一个主要缺点就是不能表现关键词的行为特征，例如关键词在一个数据库中的出现模式。同时，没有系统以定量表的形式描述关键词属性。本文研究的模型具有许多优势，其一在于利用定量表描述关键词的行为特征。关键词簇是通过对关键词的索引而形成。本模型中定义的指标描述了关键词簇的定量特征。本文共定义了四种定量指标的趋势分析。

2 目标

本研究的主要目标包括：

1. 分析《Chaos》、《Physics of Plasmas》与《Low Temperature Physics》三种期刊发表的 2526 篇论文中的指定关键词。

2. 识别不同的关键词组，这些关键词通常含有一个共同的词，每个关键词组称为关键词簇。

3. 定义四个指标，它们描述了关键词簇的出现模式。

4. 提出一个由关键词簇组成的定量主题检索工具。

3 范围及方法

本研究在收集了《Chaos》、《Physics of Plasmas》与《Low Temperature Physic》s 三种期刊发表的 2526 篇论文中作者指定的关键词后展开。收集范围包括：2006-2012 年《Chaos》中的 1037 篇论文，2006-2010 年《Low Temperature Physics》中的 769 篇论文，2010-2012 年《Physics of Plasmas》中的 720 篇论文。每种期刊收集到的关键词数量如表 1 所示。

表 1：三种期刊的关键词数量

期刊	总关键词数	不同关键词数	关键词的平均频率
Chaos	4901	1155	4.2
Low Temperature Physics	5105	920	5.5
Physics of Plasmas	4215	496	8.5

收集关键词之后，需对其进行整理以识别不同的词簇，如利用共同的核心词（key-term）来发现关键词的群组。举例来讲，*crystal defect*、*crystal field interaction*、*crystal growth*、*crystal microstructure*、*crystal orientation*、*crystal structure* 以及 *crystal symmetry* 七个关键词形成了一个关键词簇，它们共同的核心词是 crystal。这种情况下，词簇的名称以它们共同的核心词来命名，如 crystal。本文考虑了与关键词簇相关的变量以定义相应指标，相关代表符号以引号注释，如下文所示。

1) 在一个关键词簇中的关键词总数为“ N ”；
2) 关键词簇 k_r 中所有关键词在整个时间段“ l ”中的出现率（Frequency of Occurrence）为 F；
3) 关键词簇在时间段“ l ”中的占有率（Occupancy）是“ A ”；

4) 同一个关键词簇的最高占有率（Highest possible Occupancy）是“ A_{Max} ”；

5) 关注的关键词出现时间段为“ l ”。

关键词簇的最高可能占有率等于关键词出现的时间段与词簇中关键词总数的乘积。如：$A_{Max} = l * N$

以表 2 词簇中的第三个关键词“ Semiconductor, elemental ”为例。该关键词的出现频率是 15，曾经在 15 篇不同的期刊论文中出现；占有率是 4，曾在 2006-2010 时间段中出现 4 次；最高可能占有率是 5，因为它在 2006-2010 特定时间段内最多出现 5 次。如果考虑整个“ Semiconductor ”词簇，整体出现频率与整体占有率分别是 129 和 63，词簇中的关键词总量是 28。以上变量的具体值请见表 3。

表 2：“Semiconductor”词簇及其所含关键词 5 年内的出现频率(2006-2010)

序号	年 / 关键词	2006	2007	2008	2009	2010	总计
1	Semiconductor (cluster name)					1	1
2	semiconductor, amorphous			1			1
3	semiconductor, elemental	4	5	3	3		15
4	semiconductor, ferroelectric		1			1	2
5	semiconductor, III-V	1	7		4	1	13
6	semiconductor, III-VI					2	2
7	semiconductor, II-VI	1	6	1	11		19
8	semiconductor, IV-VI			1			1
9	semiconductor, magnetic		2		1		3
10	semiconductor, narrow band-gap	1		1			2
11	semiconductor, piezoelectric			1			1
12	semiconductor, semimagnetic		2		3	1	6
13	semiconductor, superconducting			1			1
14	semiconductor, ternary		1				1
15	semiconductor, wide band-gap	2			5	1	8
16	semiconductor-doped-glass				1		1
17	semiconductor-doping		2	1	3	1	7
18	semiconductor-epitaxial-layer			1	1		2
19	semiconductor-growth			1	1		2
20	semiconductor-heterojunction	1	3	2	3		9
21	semiconductor-laser		1				1
22	semiconductor-material		3	1	1	1	6
23	semiconductor-metal boundary		1	1			2
24	semiconductor-nanotube			1			1
25	semiconductor-quantum-dot		2	1	1		4
26	semiconductor-quantum-well	3	6	1	4	1	15
27	semiconductor-quantum-wire	1				1	2
28	semiconductor-superlattice			1			1
	All						

表 3：“ Semiconductor ”词簇的一些变量值

变量	代表符号	数值
关键词总量	N	28
出现率	F	129
占有率	A	63
最高可能占有率	A_{Max}	l* N = 5*28 = 140

一个关键词在每年都有某种出现频率，可能在较短时间段内出现率很高，或者在较长时间段内出现率很低。这种在某个时间段内的出现率称为占有率（Occupancy）。因此，“ 出现率 ”与“ 占有率 ”是与关键词或关键词簇相关的两个重要变量。

这两个变量象征了一个关键词或关键词簇的两个基础维度。高“ 占有率 ”代表在某个时间段内具有较高的稳定性或者较高的临时稳定性。高“ 出现率 ”说明关键词簇在期刊论文中覆盖度较高。期刊论文可以看作是知识空间，因此高“ 出现率 ”代表了更高的空间稳定性。

另一个重要的变量时关键词簇中的关键词数量，用 N 来表示，它说明了一个词簇的强度。以上三个基础的变量反映了主题领域中的三个基础特征，如表 4 所示。

表 4：关键词/关键词簇的三个基础变量

变量	代表符号	说明在论文组成的主题空间内具有的特征
出现率	F	空间稳定性
占有率	A	时间稳定性
关键词总量	N	能力
最高占有率	A(max)	最大可能的时间稳定性

关键词特征指标：以下四个指标基于从一个关键词簇中识别出的四个变量，如表 5 所示。

表 5：关键词特征指标

序号	指标	代表符号	定义
1	整体可见性指数	v	F / N
2	短时可见性指数	m	F / A
3	能力指数	p	ln(N*F)
4	稳定度指数	s	(A/A(max))*100

1) 整体可见度指数，用 v 表示，反映了某个时间段内一个关键词簇在期刊论文空间的曝光度。定义为整个时间段内单个关键词覆盖的期刊论文数量。
2) 短时可见度指数，用 m 表示，反映了单个时间一个关键词簇在期刊论文空间的曝光度。定义为单一时间单个关键词覆盖的期刊论文数量。
3) 能力指数，用 p 表示，反映了一个关键词簇的能力，定义为关键词总数与出现频率乘积的自然对数。
4) 稳定度指数，用 s 表示，反映了一个关键词簇的临时稳定性，定义为词簇的占有率与最大占有率的比例乘以 100。

总体来讲，这四个指数定义了关键词簇的五个基本属性，如表 6 所示，分别为：（1）可见度，（2）分散度，（3）强度，（4）稳定度，（5）密度。

表 6：基本属性与说明的相应趋势

基础属性		相应指标	高指标值说明的趋势
可见性	综合	v	高可见性关键词，可能是特定主题、通用主题或支持性词汇
	短时	m	高可见度，数量多但较孤立。通常关键词属于一个领域，并且支持此分类下的研究领域。
强度		p	含有大量关键词及高占有率的关键词簇，意味着高度相关、主题中心的关键词。
稳定度		s	实际占有率与最大可能占有率的比率，说明了时间段内的平均占有率。值越高代表稳定度越高。

Keysaurus：定量的主题检索工具：可定义为一个分类工具，用以图书馆、档案馆或其他文档中心用以管理它们的数据及其他信息。该工具的设计目标是便于用户识别对数据进行分类和命名的优先（或授权）术语，并且为获取这些术语提供一些途径。分类表、主题词表及叙词表是广为人知的主题获取工具。同时，主题工具也为文档检索提供了便利措施，并且增加了检索的成功率。该项功能是通过建立关键词之间的关系来实现。

主题检索工具的设计与开发基于对知识的分类。以下是在对领域知识进行分类是主要考虑的一些方面：

1. 信息组织与主题关系显示的分类原则；

2. 受控词表的特征，尤其是对同义词和同形异义词的控制，以提高召回率和精确度；

3. 搜索战略的指定和为优化检索结果的预先存储。

在本文的研究中，提出了一个便于信息检索的主题检索工具，它基于对关键词簇的分析，对信息进行了自下到上的处理与组织。该工具命名为 Keysaurus，它基于关键词簇的分析，并且描述了关键词必要的定量特征。该主题检索工具显示了关键词一些定量参数的数值，关键词的参数命名为“关键词簇轨迹指标(Keyword cluster locus indicator ，KCLI)”，因为这些是对关键词/关键词簇未来发展方向的预测。为了描述关键词簇的状态，工具包含了四个指标：稳定性指数、综合可见性指数、短时可见性指数以及能力指数。

稳定性指数描述了在规定时间段内关键词簇的临时稳定性。如果关键词簇在长时间内有规律的出现，说明它具有较高的稳定性指数。可见度指数描述了在期刊论文中

的出现度。若关键词簇在大量期刊论文中出现，说明它有较高的可见性指数。综合可见性指数针对的是单个关键词（独立或群组）在整个时间内的可见性，而暂时可见性指数针对的是某一时间多个关键词（或单个独立关键词）的可见性。能力指数代表了一个关键词簇的重要程度。若关键词簇包含了大量种类的关键词，它的权重就很高，也就具有较高的能力指数。能力指数高的关键词簇反映了重要的研究领域。现有研究表明，以下四个关键词簇核心指标互相独立，如表 7 所示。

表 7：指标及其说明的相关现象

关键词簇核心指标(KCLI)	词簇中关键词具有的特定现象	该关键词簇相关的属性	说明的研究趋势
稳定度指数	关键词的持久度（短暂或长期）	稳定度	研究的持久性
整体可见性指数	单个关键词覆盖的平均期刊论文数量	整体可见度	研究潜力
短时可见性指标	单个时间多个关键词覆盖的平均期刊论文数量	短时可见度	研究强度
能力指标	一个词簇中的关键词种类	权重	研究的重要领域

表 8：指数数值的等级

关键词簇轨迹指标(KCLI)	说明的关键词簇属性	KCLI 数值的等级				
		++ (非常高)	+ (高)	0 (中)	(-) (低)	(-)(-) (非常低)
稳定度指数	稳定度	完美的持久度	很强的持久度	中等的持久度	较弱的持久度	短暂的
综合可见度指数	整体可见度	很高潜力	高潜力	中等潜力	低潜力	非常低的潜力
短时可见度指数	短时可见度	非常高的强度	高强度	中等强度	低强度	很低强度
能力指数	权重	非常高的研究领域	高研究领域	中等研究领域	弱研究领域	不相容的

利用最高值与最低值的差额除以 5，每个关键词簇的指标值都被平均分成了 5 个域。不同区域的名称可见表 8。最高值区域用“++”表示，较高值区域用“+”表示，以此类推。

表 9 至表 12 汇总了关键词/关键词簇的三个基础变量（见表 4）、四个关键词特征指标（见表 5）与指标数值的等级（见表 8），并利用从《Low Temperature Physics》（表 9 与表 10）、《Chaos》（表 11）、《Physics of Plasmas》（表 12）期刊中收集

的关键词簇，来举例说明 Keysaurus 主题检索工具的设计。

表 9：利用《Low Temperature Physics》期刊中收集的关键词簇示说明 Keysaurus 主题检索工具

关键词簇名称	N	F	A	A(max)	v = F/N	m = F/A	p = ln(N*F)	s = (A/A(max)) *100
Alloy	54	168	114	270	3.11 (-)(-)	1.47 (-)(-)	9.11 (+)(+)	42.22 (-)
Antiferromagnetism	3	73	14	15	24.33 (+)(+)	5.21 (+)	5.39 (-)	93.33 (+)(+)
Compound	70	448	191	350	6.4 (-)	2.35 (-)	10.35 (+)(+)	54.57 (0)
Crystal	11	56	28	55	5.09 (-)	2 (-)	6.42 (0)	50.91 (0)
Dislocation	9	24	15	45	2.67 (-)(-)	1.6 (-)	5.38 (-)	33.33 (-)
Doping	2	24	10	10	12 (0)	2.4 (-)	3.87 (-)(-)	100 (+)(+)
Electricity	12	56	25	60	4.67 (-)(-)	2.24 (-)	6.51 (0)	41.67 (-)
Electron	34	139	72	170	4.09 (-)(-)	1.93 (-)	8.46 (+)	42.35 (-)
Exchange-interaction	2	41	6	10	20.5 (+)(+)	6.83 (+)(+)	4.41 (-)(-)	60(0)
Exciton	2	20	6	10	10 (0)	3.33 (0)	3.69 (-)(-)	60 (0)
Fermion	3	9	5	15	3 (-)(-)	1.8 (-)	3.3 (-)(-)	33.33 (-)
Ferrimagnetism	3	12	7	15	4 (-)(-)	1.71 (-)	3.58 (-)(-)	46.67 (-)
Ferroelectricity	3	10	6	15	3.33 (-)(-)	1.67 (-)	3.4 (-)(-)	40 (-)
Ferromagnetism	5	100	23	25	20 (+)(+)	4.35 (+)	6.21 (0)	92 (+)(+)
Helium	11	95	28	55	8.64 (-)	3.39 (0)	6.95 (0)	50.91 (0)
Impurity	7	55	20	35	7.86 (-)	2.75 (-)	5.95 (-)	57.14 (0)
Laser	5	7	7	25	1.4 (-)(-)	1 (-)(-)	3.56 (-)(-)	28 (-)(-)
Lattice	5	28	11	25	5.6 (-)	2.55 (-)	4.94 (-)	44 (-)
Magnetism	47	352	122	235	7.49 (-)	2.89 (-)	9.71 (+)(+)	51.91 (0)

Metal	10	29	19	50	2.9 (-)(-)	1.53 (-)	5.67 (-)	38 (-)
Nanostructured-material	12	63	26	60	5.25 (-)	2.42 (-)	6.63 (0)	43.33 (-)
Optics	14	19	17	70	1.36 (-)(-)	1.12 (-)(-)	5.58 (-)	24.29 (-)(-)
Organic-compound	2	38	7	10	19 (+)	5.43(+)(+)	4.33 (-)(-)	70 (+)
Paramagnetism	4	35	14	20	8.75 (-)	2.5 (-)	4.94 (-)	70 (+)
Phonon	8	46	18	40	5.75 (-)	2.56 (-)	5.91 (-)	45 (-)
Plasma physics	6	7	7	30	1.17 (-)(-)	1 (-)(-)	3.74 (-)(-)	23.33 (-)(-)
Plasmon	6	6	5	30	1 (-)(-)	1.2 (-)(-)	3.58 (-)(-)	16.67 (-)(-)
Quantum physics	12	40	28	60	3.33 (-)(-)	1.43 (-)(-)	6.17 (-)	46.67 (-)
Semiconductor	28	129	63	140	4.61 (-)(-)	2.05 (-)	8.19 (+)	45 (-)
Spin dynamics	16	81	40	80	5.06 (-)	2.03 (-)	7.17 (0)	50 (0)
Superconductivity	30	297	90	150	9.9 (-)	3.3 (0)	9.09 (+)(+)	60 (0)
Surface physics	9	19	14	45	2.11 (-)(-)	1.36 (-)(-)	5.14 (-)	31.11 (-)(-)
Thin film	9	62	24	45	6.89 (-)	2.58 (-)	6.32 (0)	53.33 (0)
Tunnelling	3	31	11	15	10.33 (0)	2.82 (-)	4.53 (-)(-)	73.33 (+)
X-ray	4	27	11	20	6.75 (-)	2.45 (-)	4.68 (-)(-)	55 (0)

(第一行的符号解释请见表 4 和表 5)

表 10：利用《Low Temperature Physics》期刊中收集的单个关键词示例说明 Keysaurus 主题检索工具

单个关键词	N	F	A	A (max)	v = F/N	m = F/A	p= ln(N*F)	s = (A/A(max)) *100
Bose-Einstein-condensation	1	39	5	5	39 (+)(+)	7.8 (+)(+)	3.66 (+)(+)	100.00
Specific-heat	1	30	5	5	30 (+)	6 (+)	3.4 (+)(+)	100.00
Conductivity, thermal	1	25	5	5	25 (+)	5 (+)	3.22 (+)(+)	100.00
Band-structure	1	24	5	5	24 (0)	4.8 (+)	3.18 (+)	100.00
Carbon nanotube	1	24	5	5	24 (0)	4.8 (+)	3.18 (+)	100.00
Quasiparticle	1	21	5	5	21 (0)	4.2 (0)	3.04 (+)	100.00
Argon	1	18	5	5	18 (0)	3.6 (0)	2.89 (0)	100.00
Flux-pinning	1	18	5	5	18 (0)	3.6 (0)	2.89 (0)	100.00
Cryogenics	1	16	5	5	16 (-)	3.2 (-)	2.77 (0)	100.00
Fermi-level	1	15	5	5	15 (-)	3 (-)	2.71 (0)	100.00
Fullerene	1	15	5	5	15 (-)	3 (-)	2.71 (0)	100.00
Ab-initio-calculation	1	14	5	5	14 (-)	2.8 (-)	2.64 (-)	100.00
Fermi-surface	1	11	5	5	11 (-)	2.2 (-)	2.4 (-)	100.00
Boson-system	1	9	5	5	9 (-)	1.8 (-)	2.2 (-)(-)	100.00
Fermi-liquid	1	7	4	5	7 (-)(-)	1.75 (-)	1.95 (-)(-)	80.00

(第一行的符号解释请见表 4 和表 5)

有趣的是，除关键词“Fermi liquid”的稳定性指数为 90 之外，其他所有单个关键词的稳定性指数都是 100。但是在关键词簇中，只有两个词簇的稳定性指数超过了 90。

表 11：利用《Chaos》期刊中收集的关键词簇示说明 Keysaurus 主题检索工具

关键词簇名称	N	F	A	A(max)	v = F/N	m = F/A	p = ln(N*F)	s = (A/A(max))*100
Atmospheric science	3	7	6	21	2.33 (-)(-)	1.17 (-)(-)	3.04 (-)(-)	28.57 (-)
Biomedical	6	14	11	42	2.33 (-)(-)	1.27 (-)(-)	4.43 (-)	26.19 (-)
Cellular biophysics	4	45	15	28	11.25 (-)	3 (-)	5.19 (0)	53.57 (+)
Circuit theory	7	26	15	49	3.71 (-)(-)	1.73 (-)(-)	5.2 (0)	30.61 (-)

Crystal	4	5	5	28	1.25 (-)(-)	1 (-)(-)	3 (-)(-)	17.86 (-)(-)
Image processing	7	9	9	49	1.29 (-)(-)	1 (-)(-)	4.14 (-)	18.37 (-)(-)
Laser	7	9	9	49	1.29 (-)(-)	1 (-)(-)	4.14 (-)	18.37 (-)(-)
Magnetism	4	5	5	28	1.25 (-)(-)	1 (-)(-)	3 (-)(-)	17.86 (-)(-)
Nonlinear dynamics	9	417	33	63	46.33 (+)(+)	12.64 (+)(+)	8.23 (+)(+)	52.38 (+)
Numerical analysis	2	110	9	14	55 (+)(+)	12.22 (+)(+)	5.39 (0)	64.29 (+)(+)
Optics	22	53	37	154	2.41 (-)(-)	1.43 (-)(-)	7.06 (+)	24.03 (-)(-)
Pattern formation	3	58	12	21	19.33 (-)	4.83 (-)	5.16 (0)	57.14 (+)(+)
Plasma physics	13	16	16	91	1.23 (-)(-)	1 (-)(-)	5.34 (0)	17.58 (-)(-)
Polymer	5	6	6	35	1.2 (-)(-)	1 (-)(-)	3.4 (-)(-)	17.14 (-)(-)
Quantum physics	10	20	17	70	2 (-)(-)	1.18 (-)(-)	5.3 (0)	24.29 (-)(-)
Semiconductor	5	9	6	35	1.8 (-)(-)	1.5 (-)(-)	3.81 (-)(-)	17.14 (-)(-)
Surface science	5	11	11	35	2.2 (-)(-)	1 (-)(-)	4.01 (-)	31.43 (-)
Telecommunication	6	15	10	42	2.5 (-)(-)	1.5 (-)(-)	4.5 (-)	23.81 (-)(-)

(第一行的符号解释请见表 4 和表 5)

表 12：利用《Physics of Plasmas》期刊中收集的关键词簇示说明 Keysaurus 主题检索工具

关键词簇名称	**N**	**F**	**A**	**A (max)**	**v = F/N**	**m = F/A**	**p = ln(N*F)**	**s = (A/A(max))*100**
Acoustics	4	4	5	12	1 (-)(-)	0.8 (-)(-)	2.77 (-)(-)	41.67 (-)(-)
Astrophysical plasma	3	43	7	9	14.33 (0)	6.14 (-)	4.86 (-)	77.78 (+)
Cyclotron	3	6	4	9	2 (-)(-)	1.5 (-)(-)	2.89 (-)(-)	44.44 (-)(-)
Dielectric function	3	3	3	9	1 (-)(-)	1 (-)(-)	2.2 (-)(-)	33.33 (-)(-)

Dispersion	3	58	6	9	19.33 (0)	9.67 (0)	5.16 (-)	66.67 (0)
Doppler effect	4	6	5	12	1.5 (-)(-)	1.2 (-)(-)	3.18 (-)(-)	41.67 (-)(-)
Electricity	5	11	8	15	2.2 (-)(-)	1.38 (-)(-)	4.01 (-)	53.33 (-)
Electromagnetism	4	4	4	12	1 (-)(-)	1 (-)(-)	2.77 (-)(-)	33.33 (-)(-)
Electron	12	30	17	36	2.5 (-)(-)	1.76 (-)(-)	5.89 (-)	47.22 (-)
Magnetism	15	99	24	45	6.6 (-)(-)	4.13 (-)	7.3 (0)	53.33 (-)
Microwave	4	7	5	12	1.75 (-)(-)	1.4 (-)(-)	3.33 (-)(-)	41.67 (-)(-)
Numerical analysis	2	68	5	6	34 (+)(+)	13.6 (+)(+)	4.91 (-)	83.33 (+)(+)
Optics	5	7	7	15	1.4 (-)(-)	1 (-)(-)	3.56 (-)(-)	46.67 (-)
Plasma physics	63	2670	165	189	42.38 (+)(+)	16.18 (+)(+)	12.03 (+)(+)	87.3 (+)(+)

(第一行的符号解释请见表 4 和表 5)

4 结论

以上结果表明，与其它两种期刊相比，《Low Temperature Physics》中词簇信息的评价更高。对《Low Temperature Physics》中的单个关键词进行分析，发现几乎所有关键词都具有最高的稳定性指数（如 100）。三种期刊中都只有少数词簇含有很高的指数值(+)(+)。具有高指数值的关键词簇可被视为主题领域中有潜力的关键词或内容描述符。因此，Keysaurus 工具可以定位到正确的关键词或内容描述符，使得检索结果的准确度与相关性更高。

参考文献

1) Bertrand A, Cellier J M, Psychological approach to indexing: effects of the operator's expertise upon indexing behaviour, *Journal of Information Science*, 21 (6) (1995) 459-472.

2) Suraud M G et al, On the significance of databases keywords for a large-scale bibliometric investigation in fundamental physics, *Scientometrics*, 33 (1) (1995) 41-63.

3) Voorbij H J, Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences, *Journal of Documentation*, 54 (1998) 466-476.

4) Solomon P, Use-based methods for classification development. *Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop*. Washington DC (1991).

5) Hurt C D, Classification and subject analysis: looking to the future at a distance, *Cataloguing and Classification Quarterly*, 24 (1-2) (1997) 97-112.

6) Soergel D et al, Re-engineering thesauri for new applications: the AGROVOC example, *Journal of Digital Information*, 4 (4) (2004).

7) Bates M J, Wilde D N and Siegfried S, An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report, No.1, *Library Quarterly*, 63 (1) (1993) 1-39.

8) Willett P, Recent trends in hierarchical document clustering: A critical review, *Information Processing & Management*, 24 (5) (1988) 577-597.