基于英文超级科技词表的文献主题标引系统设计与实现*

李军莲，夏光辉，王序文，李晓瑛，冀玉静，李赞梅
（中国医学科学院医学信息研究所，北京 100020）

摘要：针对海量英文文献信息自动化处理问题，构建了一个基于英文超级科技词表的文献主题概念自动标引系统，采用词典与规则方法相结合的术语提取机制，实现了英文文献术语提取、规范概念映射以及优选概念标引等功能，取得了较好的标引效果。

关键词：术语提取；术语匹配；主题标引
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.12.001

1 引言

日渐增长的科技文献数据为广大用户提供了丰富的知识资源，同时也带来了信息过载的压力。为了有效挖掘海量文献中存在的科技知识，促进文献信息内容的知识组织、关联及利用，进而支持用户日益膨胀的知识获取需求，国家科技文献信息中心组织实施了“面向外文科技文献信息的知识组织体系建设和示范应用”国家“十二五”科技支撑计划项目[1]。科技知识组织体系（Science and Technology Knowledge Organization System，简称STKOS）覆盖了理工农医领域大量专业术语、概念（超级词表）以及基于概念所形成的本体网络和科研本体知识库。这一领域全面、内容丰富的英文超级科技词表，也为面向海量文献的自动化信息处理任务提供了有力支撑[2]。

概念（Concept）是人类在认知过程中对特定事物的本质属性的抽象描述，其语言表达形式包括词语和词组。其中，领域概念是特定领域中具有特定语义的词汇集合，是领域知识的一种重要表现形式[3]。在各个学科领域知识不断推陈出新的背景下，利用计算机自动或者半自动地从文献中发现并标引领域概念的过程，是将非结构化的文本信息快速转变为知识单元的关键环节[4-6]。目前，概念标引的成果也已在信息检索、文本分类、机器翻译、本体构建[7]等研究领域得到了广泛的应用[8-9]。

本文基于STKOS超级词表，结合语言学规则以及文本统计信息，构建了面向海量外文科技文献的主题概念自动标引系统。这一工作既是STKOS超级词表的一个直接应用，也为进一步的知识对象关系计算以及知识网络构建打下了良好基础。

2 系统描述

2.1 系统结构与功能设计

本文面向海量数据加工任务，设计并实现了基于STKOS超级词表的交互式主题概念自动标引系统，系统的主要功能包括词典（知识库）管理、文献预处理、候选术语提取、规范概念映射以及概念标引等模块。系

* 本研究得到十二五国家科技支撑计划项目课题“信息资源自动处理、智能检索与STKOS应用服务集成”（编号：2011BAH10B01）资助。

统的基本结构设计如图1所示。

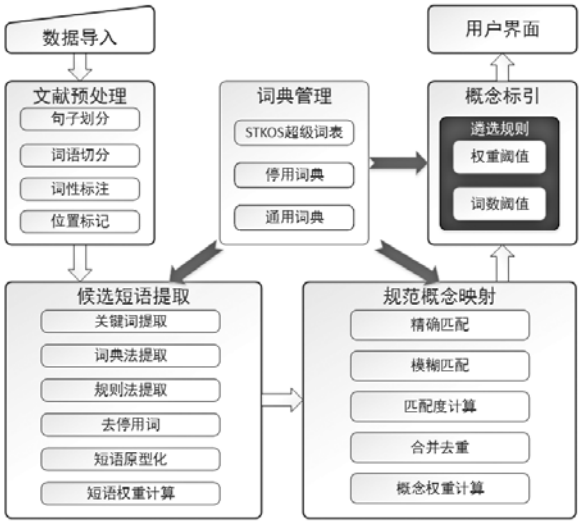

图1 主题概念自动标引系统结构

(1) 词典（知识库）动态管理

STKOS超级词表主要由基础词库、规范概念集合以及范畴体系构成，它是概念标引系统的重要知识基础，为自动标引过程中术语的匹配、概念的映射以及概念遴选提供了语言学依据。

为了便于对知识库进行动态维护、更新与扩展，将系统所使用的基于STKOS超级词表的切分词典、标引词典、停用词典、通用词典等内容映射到不同的类，通过增加类来实现词典的自动添加；通过增加或删除每个类下的实例，实现对词典下实例的修改操作。这一管理机制也有助于保证系统词表内容与持续更新版本的STKOS超级科技词表保持一致。

(2) 文献预处理

科技文献的标题、作者关键词及摘要等内容是文献术语高频出现的区域，能够直接反映文章的主题内容，也是用户重点关注的“兴趣区域”。为了提高文献分析效率，分别提取每篇文献的标题、作者关键词以及摘要内容作为概念标引系统的分析对象。

从文本中提取概念之前，首先针对待分析的文本内容进行预处理。预处理过程主要包括：依据标点符号对标题和摘要进行句子划分；基于空格、标点符号、换行符等启发式规则将句子切分为独立的词语；对切分后的语言单元进行词性标注和位置标记，其中词性标注过程采用MetaMap的PhraseX工具实现，MetaMap已在美国国立医学图书馆（NLM）的相关工程化实践中取得了较好的应用效果[10]；进一步将词性标注结果转换成语法词性，以便基于语言学规则进行短语提取。

(3) 候选短语提取

这里所谓的“短语”，指名词性短语，即语法功能上相当于名词的短语，是反映文本内容的领域术语的主要来源。与通用术语不同，领域术语与某个特定领域具有较强的相关性，即在特定领域中出现频率较高，而在不相关领域中出现频率相对较低[11]。本文基于特定的语言学规则以及统计信息，从文本内容中提取候选短语，并将候选短语集合作为概念标引的基础。

对于关键词文本，依据指定分隔符直接提取关键词字段内容作为候选短语。对于标题或摘要文本中经过词性标注的内容，采取词典与规则相结合的提取策略，利用切分符号以及构词规则分别提取其中的简单短语以及复合短语。去除短语集合中的停用词，并执行短语原型化，统计标记每个短语的频次、位置及长度，合并去重后计算术语权重，从而获得候选术语集合。

(4) 规范概念映射

首先将候选术语原型与超级词表词表中的规范术语原型进行匹配，获得规范术语，术语匹配方式包括基于字符串的精确匹配以及模糊匹配。其中同形异义的术语需要借助超级词表的语义类型进行区分。根据不同的映射方式计算候选术语与规范术语之间的匹配度，依匹配度排序，将匹配度最大的术语保存至规范术语列表。

其次根据STKOS词表中已建立的术语与概念的对应关系，进一步将术语映射到规范概念。当一个术语对应多个概念时，按照文献的学科属性映射对应概念。根据术语对应的概念ID合并去重，并统计每个概念对应的术语集合。综合考虑概念词在文献中出现的频次、位置等统计信息，结合术语权重以及规范术语的匹配度，计算对应概念的权重，当多个术语对应同一个概念时，概念的权重为多个术语的权重累加，由此可以生成概念列表。

(5) 主题概念标引

基于STKOS超级词表中的术语-概念对应关系，可以直接获取每个规范术语所继承的概念。然而并非文献中出现的所有概念词都值得向用户推荐，因此，将文献中的术语映射为规范概念之后，需要进一步从文献概念列表中遴选与文献主题相关性较高的优选概念词[12-13]。

首先对概念词的相关性进行评估，根据权重大小对概念词进行排序。通过设置权重阈值和标引深度阈值筛选概念词，同时利用通用词表过滤掉领域相关性不高的概念词，从而降低通用概念对标引效果的影响，保留下来的概念词则作为能够表达每篇文献主题的优选概念，最终推荐给用户。

(6) 用户界面

为更好地满足用户的知识获取及研究需求，系统通过交互式用户界面向用户提供了标引方式管理、标引结果展示以及文献浏览等服务。

2.2 系统流程

综合上述功能设计，系统进行概念标引的具体流程如下：

Step1：从数据采集层中自动导入并存储外文文献资源，文献类型包括期刊、会议、标准、专利等内容，并将不同来源的数据转换为系统支持的标准格式。

Step2：自动提取每篇文献的题名、作者关键词、摘要等字段内容，并逐一进行句子划分、词语切分、词性标注以及位置标记等预处理操作。

Step3：基于STKOS超级词表及构词规则，从待分析文本中自动提取短语，过滤其中的停用词，并对候选术语进行原型化，加入候选术语集合。

Step4：按术语原型进行合并去重，根据术语位置、频次以及词长计算权重。为每个术语词条保存的信息包括源术语、术语原型、术语词频、术语长度、位置以及术语权重等。

Step5：将候选术语原型与STKOS规范术语原型进行匹配，对匹配成功的术语匹配度进行评估。

Step6：基于STKOS超级词表中术语与概念的对应关系实现术语到规范概念的映射，获得文献概念列表。

Step7：结合术语权重以及概念映射匹配度，计算概念权重。为每个概念保存的信息包括文献ID、规范概念ID、规范概念名称以及概念权重评分等。

Step8：根据权重对概念进行排序，设置权重阈值及词数阈值，筛选概念词。

Step9：输出概念标引结果至用户界面。

3 关键技术

文本中术语的识别提取以及概念映射遴选是本系统的重要环节。

经过反复测试，课题最终制定先识别提取后映射过滤的实现思路，首先广泛获取文本内容中潜在的短语；基于短语在文献中的相关统计信息计算其权重；评估候选术语与规范术语的匹配度；结合术语权重及其规范匹配度计算概念的权重，并据此对概念进行遴选。

在本系统实现过程中，重点在短语识别提取、术语匹配、概念权重计算三个环节开展研究和算法优化，现简要介绍如下。

3.1 短语识别提取

短语提取是对文本中的名词性短语进行自动化的识别和提取。已有的研究方法主要包括基于规则的方法、基于统计的方法以及规则与统计相结合的方法[14]。

为了灵活高效地应对增量式大规模文献数据的标引任务，本系统分别采用基于词典和基于语言学规则的短语提取方法，对经过预处理后的文本内容（标题、摘要、关键词）进行短语提取。

其中词典匹配方法是基于STKOS超级词表，依据正向最大匹配原则对每个句子中的短语进行提取。该方法简单直观，对词典中长度较长的短语匹配效果较好，而对单词性术语或未登录新术语的发现能力相对有限。

基于规则的匹配方法则利用短语的语言学特征（词性标注信息），分别对简单短语以及复合短语进行识别，识别规则如下：

（1）简单短语提取规则：首先过滤停用词，再按照切分符号直接提取切分符号之间的片段作为名词短语。切分符号由两种类型组成：一是非名词短语组成成分的单词，如“conj”（连词）、“prep”（介词）、“verb”（动词）等；二是能正确切分名词短语的标点符号，如“，”（逗号）、“.”（句号）、“?”（问号）等。

（2）基于构词规则的复合短语提取规则：首先统计STKOS规范术语的构词形式，从中遴选出常见的构词形式作为提取复合短语的构词规则（见表1）。按照表1中的四种复合短语构词规则，提取复合短语，不需过滤停用词。其中noun代表名词，prep代表介词，adj代表形容词，det代表定冠词。

表1 复合短语构词规则

构词规则	示例
noun+prep+noun	quality of life
adj+noun+prep+noun	kupffer cells from pigs
noun+prep+det+noun	heterogeneity of art protocols
noun+noun+……	lymph node tissue

3.2 术语匹配

本系统采用如下原则进行候选短语与STKOS词表术语进行匹配：所有匹配均基于原型进行；当组成术语的单词数小于等于2时，执行精确匹配；当术语中所包含的单词数大于2时，先执行精确匹配，匹配不成功时，则执行模糊匹配。

模糊匹配过程只在术语单词数[-1，+1]的范围内进行，即术语在增加一个单词、减少一个单词或者替换一个单词的情况下进行匹配。模糊匹配时，从匹配结果中选取评估值最高的短语为最终匹配结果，当多个短语的评估值一样时，选取多个结果。

计算规范术语匹配度的评估值参数包括向心度、覆盖度和内聚度：

（1）向心度（CEN）：考察待匹配词串是否包含原短语的核心词，若包含核心词，取CEN=1，否则CEN=0。

（2）覆盖度（COV）：考察短语与STKOS词串在匹配过程中被覆盖或包含的程度，见公式1。其中，MML代表超级叙词表匹配字串长度，ML代表超级叙词表字串长度；PML代表短语匹配字符串长度；PL代表短语字串长度。

$$COV = \frac{2}{3} * (MML / ML) + \frac{1}{3} * (PML / PL) \quad (1)$$

（3）内聚度（COH）：考察短语与STKOS词串在匹配过程中的连续字串匹配的程度，见公式2。其中，MCL代表超级叙词表匹配连续字串长度，PCL代表短语匹配连续字串长度。

$$COH = \frac{2}{3} * (MCL^2 / ML^2) + \frac{1}{3} * (PML^2 / PL^2) \quad (2)$$

（4）术语匹配度的评估函数见公式3，其中精确匹配的术语评估值为1。将每个候选术语对应的规范术语词条按匹配度评估值排序，取评估值最大的术语加入文献规范术语集合。

$$E = (CEN + 2 * COV + 2 * COH) / 5 \quad (3)$$

3.3 概念权重计算

已有研究表明，术语在文献中出现的频次是评估其重要性的一个依据。此外，在文献中不同位置出现的词语对文章内容的反映程度也不同，例如出现在科技文献的标题、摘要、关键词等位置的短语成为术语的可能性较大；又如大部分医学领域文献中，出现在摘要首末句中的短语与出现在中间句子中的短语相比，前者与文献主题相关的可能性更大。

有鉴于此，本文对传统的TF.IDF算法（见公式4）加以改进，设计了短语权重计算函数，见公式5，综合考虑了短语的频次、出现位置、词长等因素，对于处于不同位置的特征词分别赋予不同的权值，即关键词权值>标题权值>摘要首末句权值>摘要中间句权值。

$$TF.IDF_i = \frac{tf_i * \log(N / n_i)}{\sqrt{\sum_j (tf_i * \log(N / n_i))^2}} \quad (4)$$

$$W_i = \frac{(\sum_{j=1}^{5} f_{i,j} * \lambda_{i,j} + L * \lambda_l) * \log(\frac{N}{n_i} + 2)}{\sqrt{\sum_{i=1}^{m} [(\sum_{j=1}^{5} f_{i,j} * \lambda_{i,j} + L * \lambda_l) * \log(\frac{N}{n_i} + 2)]^2}} \quad (5)$$

$$L = \begin{cases} 0 & l \le 3 \\ l - 3 & l > 3 \end{cases} \quad (l\text{为词长}) \quad (6)$$

其中，$f_{i,j}$ 分别表示特征词 W_i 在文档集合中的标题（j=1）、关键词（j=2）、摘要首句（j=3）、摘要中间句（j=4）、摘要末句（j=5）等位置出现的频数，$\lambda_{i,j}$ 分别表示特征词 W_i 出现在上述位置时的权重系数，L 为词长取值，λ_l 为词长权重系数，n_i 为特征词 W_i 出现的文档频数；N为文档集合中的文档数量；m为全部特征词数。

结合上述术语的权重以及规范术语匹配度，可以

计算每个概念 C_i 的权重 T_i ，见公式7。

$$T_i = W_i * E_i \quad (7)$$

4 用户界面设计

本文构建的是一个交互式文献概念自动标引系统，其用户界面如图2所示。用户可以通过该界面选择标引方式（单篇或批量标引）、设置相关参数、进行标引处理、浏览待标文献和标引结果，并进行词典管理等。

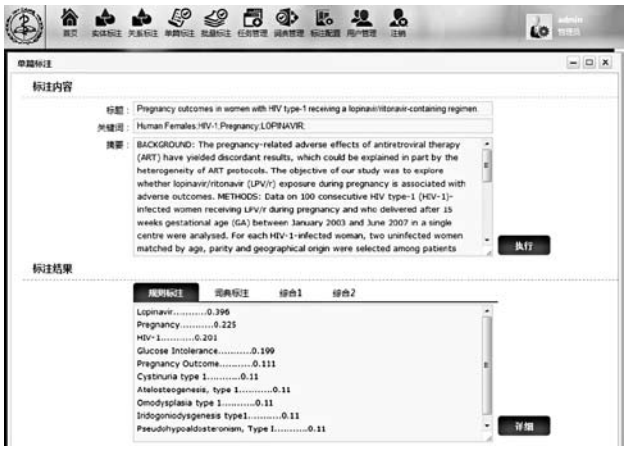

图2 文献概念自动标引系统界面

批量标注能够支持常见的文献数据格式，包括XML、Excel等文本格式以及Access、DBF、MySQL、SQL Server等数据库格式，系统能够自动将不同格式的输入数据转变成系统支持的固定格式，并且针对不同的数据源可以选择不同的标引方案。词典管理可以支持用户针对特定任务选择指定的词典进行短语提取和概念映射。参数配置可以为标题、关键词、摘要首末句、摘要中间句等设置不同的位置权重，还可以设置词长阈值以及词长权重。此外在概念遴选时，还支持通过词数和权重阈值两个参数的布尔逻辑组配进行。

5 系统测试

5.1 实验设计

本文基于NSTL英文文献资源，以医学领域为例，对概念自动标引系统的标引效果进行了初步评测。

首先以“aids”作为检索词在PubMed数据库中进行检索，从检索结果中随机选取标题、关键词、摘要信息都完备的50篇发表于2014年的英文文献作为实验数据，并请3位领域专家对其进行人工概念标引，按照概念与文献的相关程度，将概念词划分为密切相关和比较相关两个等级。同时，通过主题概念自动标引系统处理相同的50篇文献，将自动标引结果与专家人工标引的结果进行对比，按照密切相关、比较相关以及弱相关划分自动标引结果。

评估方式：将系统标引出的概念与专家标引概念进行对比，采用准确率（Precision）、召回率（Recall）以及综合评分F值作为概念标引结果的主要评价指标。

实验分别考察了基于不同短语提取方法的标引方案，包括词典提取标引法（DictMatch，词典法）、基规则提取标引法（RuleMatch，规则法）以及词典与规则相结合标引法（Merge，综合法），并对三种标引效果进行了对比和分析。

5.2 实验结果

首先，以某篇医学文献的概念标引结果作为展示实例，具体标引内容见表2。表3则展示了应用不同标引方案从测试文献中遴选相关主题概念的情况。

从上述表格中可以看出，基于词典匹配方案所遴选的相关主题概念词数量最多，但同时也误标引了很多弱相关或者不重要的概念，因此，词典方法适合于注重查全率的标引需求，如果能够结合人工审查修正结果将会取得比较好的效果。

基于规则匹配的标引方案主要是为了弥补领域词典的不足，以发现词典中的未登录术语。在本文的测试中，通过规则的方法在短语提取阶段也能够自动发现大量的候选短语，然而经过与STKOS词表进行术语匹配及概念映射后，该方案在概念标引环节最终遴选出来的相关概念词数相对较少，但是其误标引的弱相关或者不重要的概念也比较少。因此，本文将词典方法和规则方法结合起来，以规则方法修正词典方法。最终实际结果表明，将词典与规则相结合的标引方案既保留了较为重要的相关概念，也能够适度减少弱相关或不相关概念的数量。

表4 标引方案综合效果评价

	P	R	F
规则法	0.6260	0.6243	0.6252
词典法	0.6167	0.7735	0.6863
综合法	0.6530	0.7486	0.6976

表2 概念标引实例

	密切相关概念	比较相关概念	弱相关概念
规则法	hiv infections; chemoprevention	jamaica; barbados	human rights; health resources
词典法	anti-retroviral agents; hiv infections; chemoprevention	jamaica; barbados	magic; human rights; hiv; treatment; health resources
综合法	anti-retroviral agents; hiv infections; chemoprevention	jamaica; barbados	human rights; hiv; health resources
人工标引	Chemoprevention; HIV Infections; Anti-Retroviral Agents; Antiretroviral treatment	Barbados; Jamaica; HIV programming; Response	

表3 测试文献自动标引概念遴选统计结果

文献ID	规则法				词典法				综合法			
	密切相关	比较相关	误标	漏标	密切相关	比较相关	误标	漏标	密切相关	比较相关	误标	漏标
24378514	2	1	1	5	3	3	1	2	3	2	1	3
24378515	5	0	1	6	6	1	0	4	6	1	0	4
24378516	2	2	2	4	3	2	5	3	2	2	3	4
24378517	2	3	3	2	2	4	3	1	2	4	2	1
24378518	3	2	1	2	3	2	2	2	3	2	1	2
24379301	4	2	3	1	4	2	3	1	4	2	3	1
24379507	1	1	0	4	3	2	0	1	3	2	0	1
24379752	2	2	2	3	2	2	2	3	2	2	2	3
24379851	2	3	1	2	2	4	3	1	2	4	1	1
24380016	1	2	0	3	2	4	4	0	2	3	1	1
24380669	2	1	4	3	4	1	0	1	3	1	2	2
……	…	…	…	…	…	…	…	…	…	…	…	…
24470893	2	3	3	1	3	2	2	1	3	3	2	0
合计	109	117	135	136	136	144	174	82	132	139	144	91

表4展示了对所有标引方案的综合结果对比，其中词典与规则结合的标引方案准确率最高，达到65.3%，其综合评分F值也比较显著，说明该方法能够取得较好的标引效果，应该作为自动标引后续优化研究的主要方法。此外，如果标引任务对数据处理的时间复杂度有着比较严格的要求，而且比较注重标引结果的全面性，则可以采取单一的词典方法。

6 结论与展望

本文以STKOS超级词表作为知识源，构建了一个交互式文献概念自动标引系统，实现了面向多个

领域的大规模英文文献主题概念的自动标引。一方面，本系统是STKOS知识组织体系的直接应用，另一方面，本系统的工程化实践又为NSTL数据加工服务提供了有力支撑，为进一步的深层知识关系计算奠定了基础。

本文以医学领域为例，检验了英文文献主题概念标引的效果。在后续的研究及工程化实践中，还将对该系统继续进行完善。例如，面向更多领域、更大规模文献开展标引实践，在保证系统运行效率的基础上，对概念的遴选策略进行优化，进一步降低不相关概念（噪声）的影响，提高概念标引结果的文献相关性。此外，针对低频概念以及STKOS超级词表未登录术语的获取问题也将是提升概念标引系统性能的一个重要因素，值得深入探索和研究。

参考文献

[1] 孙坦,刘峥.面向外文科技文献信息的知识组织体系建设思路[J].图书与情报,2013(1):2-7.

[2] 王波.面向STKOS的概念映射与关联算法研究及其实现[D].杭州电子科技大学,2012.

[3] 姚贤明.领域概念自动抽取研究[D].昆明理工大学,2010.

[4] SHAMSFARD M, BARFOROUSH A. Learning ontologies from natural language texts [J]. International Journal Human-computer Studies, 2004, 60(1): 17-63.

[5] MOLDOVAN D, GIRJU R, RUS V. Domain-specific knowledge acquisition from text [C]// Proc. of the Sixth Conference on Applied Natural Language Processing, 2000: 268-275.

[6] MICHAEL B, MIRIAM E, DONALD E, et al. Concept annotation in the CRAFT corpus [J]. BMC Bioinformatics, 2012(13): 161.

[7] 陈珂,姚天昉.构造领域本体概念关系的自动抽取[M].上海交通大学出版社,2008.

[8] 余蕾,曹存根.基于Web语料的概念获取系统的研究与实现[J].计算机科学,2007,34(2):161-165.

[9] 钱庆,洪娜,李勇,等.中文非相关知识发现系统CmedLBKD构建[J].情报理论与实践,2012,35(4):109-113.

[10] MetaMap. MetaMap - A Tool For Recognizing UMLS Concepts in Text [EB/OL]. [2014-11-20]. http://mmtx.nlm.nih.gov.

[11] 李丽双.领域本体学习中术语及关系抽取方法的研究[D].大连理工大学,2012.

[12] 邓本洋.电子病历中的概念抽取研究[D].哈尔滨工业大学,2013.

[13] 黄利强.面向文本的领域概念筛选算法研究[D].重庆大学,2013.

[14] 祝清松,冷伏海.自动术语识别存在的问题及发展趋势综述[J].图书情报工作,2012,56(18):104-109.

作者简介

李军莲，女，1972年生，中国医学科学院医学信息研究所副研究馆员，研究方向：信息组织与系统，E-mail：li.junlian@imicams.ac.cn。

Research and Design of a Subject Indexing System Based on STKOS Super-thesaurus

LI JunLian, XIA GuangHui, WANG XuWen, LI XiaoYing, JI YuJing, LI ZanMei

(Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: This paper describes the STKOS super-thesaurus-based automatic subject indexing system for processing large-scale English literature. A dictionary-based method combined with linguistic rules was used for term matching. The system has implemented automatic term extraction, standard concept mapping, and concept indexing, and has achieved a good performance in English concept indexing tasks.

Keywords: Term extraction; Term matching; Subject indexing

（收稿日期：2014-12-04）