

日本国立信息学研究所信息检索和 获取系统测试集及其专利挖掘任务*

——基于论文和专利生成技术动态图

□ 李颖 桂婕 乔晓东 / 中国科学技术信息研究所 北京 100038

摘要: 文章概要介绍NTCIR的整体情况, 重点介绍中国科学技术信息研究所本年度参与的NTCIR专利挖掘任务 (Patent Mining Tasks), 并阐述该任务的两个子任务, 即 (1) 以论文和专利统一分类、检索和分析为目的, 对学术论文按IPC分类号进行自动分类的“学术论文分类子任务”; (2) 基于某一领域的论文和专利的“生成技术动态地图子任务”。最后, 对参与NTCIR测评活动, 给出体会与建议。

关键词: NTCIR, NII信息检索和获取系统测试集, 专利挖掘任务, 论文IPC自动分类, 技术动态地图

DOI: 10.3772/j.issn.1673-2286.2010.09.009

1 NTCIR的整体情况^[1-2]

NTCIR项目的概要:

NTCIR (NII Test Collections for Information Retrieval and Access Systems) ——日本国立信息学研究所信息检索和获取系统测试集于1998年启动, 为促进信息获取技术的研究, 提供研究所需的大规模可再利用的试验数据。为了对研究观点及其合理性进行评价, NTCIR还主办研讨的论坛。具体地说, 是通过促进多个研究团队开发的共同“研究课题(任务)”, 在保持协调与竞争的同时, 实现研究者团队的发展。NTCIR目标是通过综合运用信息检索和自然语言处理技术, 实现信息获取技术的多样性。该测评活动周期为一年半, 每届都选择几个任务来促进发展。

NTCIR的目标主要有以下3点:

- ① 构建大规模及可复用的测试集;
- ② 举办论坛, 通过系统间的比较让研究者进行思想交流;
- ③ 促进信息获取技术评价手段、评价指标的研究。

NTCIR的特点:

NTCIR的最大特点是引入测评研讨会, 广泛提供预先设置的测试集及试验结果评测有关的统一流程。也就是说, 参加NTCIR的团队, 利用NTCIR规划小组准备的数据, 进行各种方法论的研究和试验。在信息检索和文本处理的研究中, NTCIR作为一种可反复利用的大规模标准数据集, 得到了广泛的认可, 它也是研究者的论坛。它通过提供数据集, 交流研究思想和技

术, 以及技术转移, 力图促进研究, 是一种新形式的共同研究。另外, 基于这些共同数据的各种研究和试验, 不同系统间的相互比较有了共同的基础, 这也是NTCIR的特征之一。

NTCIR的方向:

NTCIR最初的两个方向是: “传统的试验室型信息检索系统的评价测试”和“较新技术的评价”。在试验室型评价测试中, 特别把重点放在: (1) 日语及其他亚洲语言有关的信息检索(IR); (2) 跨语言检索, 在较新技术评价中, 在下述相关研究中倾注了力量; (3) 从文献检索到信息检索的趋势; (4) 比较现实的测评法, 比如: 文本自动摘要的测评手法、适用于多阶段适合性判定的评测指标、特定的文档类型及经常利

* 基金项目: 本文得到国家科技部“十一五”科技支撑计划(项目编号: 2006BAH03B03)、中国科学技术信息研究所重点工作项目(项目编号: 2009KP01-7-1)、中国科学技术信息研究所预研基金项目“领域专利和组织组织实现的关键问题研究”等项目的资助。

用这一类型文档的用户特性,以及适用于这一利用的测评法。以参加研讨会研究团队为中心的研究者论坛,支持着这两个方向性的活动。

历届NTCIR研讨会举办的基本情况:

NTCIR评测研讨会是通过多个研究团队之间的协调与竞争,就任务设定、测试集以及评价方法进行共同开发的机制。

NTCIR-1伴随1998年11月训练数据的发布拉开了帷幕。6个国家28个团队完成了任务并提交了结果。1999年8月30日-9月1日,在东京的KKR旅馆召开了成果报告会。在成果报告会上,以当时的邮政省通信综合研究所和自然语言处理的研究者为中心,联合另一测评研讨会IREX (Information Retrieval and Extraction Exercise, 信息检索和抽取训练),召开了会议,在具有类似研究的机构间建立了新型的合作关系。值得一提的是,IREX团队从2000年4月起,通过开展交流活动,实现了有关文本自动摘要的新任务,并与台湾大学的研究团队建立了合作关系,使得对象从日语和英语,扩大到了亚洲语言,实现了中文检索等新任务。

NTCIR-2从2000年6月开始募集参加者。来自8个国家的36个团队完成了任务并提交了结果,2001年3月7-9日,在东京的国立信息学研究所(NII)召开了成果报告会。

NTCIR-3从2001年8月开始募集参加者。在提交结果的9个国家的65个团队的参与下,2002年10月8-10日,在NII召开了成果报告会。

NTCIR-4从2003年3月开始募集参加者。在提交结果的10个国家74个团队的参与下,2004年6月2-4日,在NII召开了成果报

告会。研究成果的一部分,作为ACM Transactions on Asia Language Information Processing的NTCIR-4专刊予以发行。

在NTCIR-5中,测评任务为跨语言检索(CLIR)、跨语言问答(CLQA)、专利检索(PATENT)、问答(QAC)、WEB(基于标签和链接结构的大规模Web文本信息获取系统的研究)及动态信息的摘要和可视化(MuST)研究领域。作为最初的试行,设置了试验性任务,推出了动态抽取和可视化技术。2004年9月开始募集参加者,接受了共14个国家的102个团队的注册。在提交结果的12个国家87个团队的参加下,2005年12月6-9日在NII召开了成果报告会。

NTCIR-6从2006年4月开始募集参加者,测试任务有:跨语言信息检索(CLIR)、专利检索(PATENT)、问答(QAC日语)、多语言问答(CLQA)、观点分析(OPINION)、动态信息的摘要和可视化(MuST)。12个国家和地区的84个研究团队参与了此次活动,2007年5月15-18日在NII举办了成果报告会。从本届会议起,开设了EVIA 2007研讨会(The First International Workshop on Evaluating Information Access, 第一届信息获取技术评价国际研讨会)。

第7届研讨会2007年11月开始募集参加者,测评任务为比较前沿的跨语言信息访问(Advanced CLIA)、跨语言观点分析(MOAT)和Blog的跨语言信息检索(CLIR-B)、专利挖掘(PATMN)和翻译(PATMT)、趋势的多模态摘要(MuST, Multimodal Summarization of Trends),于2008

年12月16-19日在NII举行。来自15个国家82个团体参加了会议,与EVIA 2008一并召开。

第8届NTCIR成果报告会及第3届EVIA信息获取评价国际会议(EVIA 2010, The 3rd International Workshop on Evaluating Information Access)于2010年6月15-18日在东京召开,有17个国家的108个团队表示参加。NTCIR-8设定的任务有:处理复杂提问的跨语言检索问答系统(CCLQA)、地理和时间信息的检索(GeoTime)、面向问答体系的文本检索(IR4QA)、多语言意见分析(MOAT)、专利信息挖掘(PATMN)、专利信息翻译(PATMT)、实验性任务的社团QA(Community QA)。同时召开的EVIA 2010(第3届)是NTCIR-8的卫星研讨会。

NTCIR的测试集:

测试集的含义是“试验用数据集”。在信息检索中,用于信息检索系统的检索有效性评价,是具有正确数据的试验用数据集。测试集由数据库、描述用户检索需求的“检索课题”和满足检索课题的“全部相关文档一览”构成,对信息系统的研究和开发、评价起着不可欠缺的作用。为了信息检索及文本处理技术研发的利用,在逐步准备多样测试集的过程中,也促进了这些技术及其评价方法的研究。参见表“测试集一览”,是历届NTCIR所构建的测试集。

2 NTCIR-7&8专利挖掘任务(Patent Mining Tasks)^[4]

专利挖掘任务产生的背景

在NTCIR研讨会中,策划“专

表1 测试集一览 (2009-07-29更新) ^[3]

	Ad Hoc/ CLIR [Scientific Abstracts] (日语和 英语检索)	Chinese IR (中文检 索)	CLIR [News] (跨语言检 索)	CLQA (跨语言 问答)	MuST (「动态信息 的摘要和可 视化」)	OPI- NION (观点 分析)	PATENT (专利检索 和分析)		QAC (问答)	TMREC (用语 抽取)	TSC (摘要)	WEB
			ACLIA (先端跨语 言信息获 取)		MOAT (跨语言 观点分析)	PATMN (专利 挖掘)	PATMT (专利 翻译)					
NTCIR-1	NTCIR-1 Ad Hoc/ CLIR	—	—	—	—	—	—	—	—	NTCIR-1 TMREC	—	—
NTCIR-2	NTCIR-2 Ad Hoc/ CLIR	CIRBO10	—	—	—	—	—	—	—	—	NTCIR-2 SUMM	—
NTCIR-3	—	—	NTCIR-3 CLIR	—	—	—	NTCIR-3 PATENT 专利检索	NTCIR-3 QA	—	—	NTCIR-3 SUMM	NTCIR-3 WEB
NTCIR-4	—	—	NTCIR-4 CLIR	—	—	—	NTCIR-4 PATENT 专利检索	NTCIR-4 QA	—	—	NTCIR-4 SUMM	NTCIR-4 WEB
NTCIR-5	—	—	NTCIR-5 CLIR	NTCIR-5 CLQA	—	—	NTCIR-5 PATENT 专利检索 分类	NTCIR-5 QA	—	—	—	NTCIR-5 WEB
NTCIR-6	—	—	NTCIR-6 CLIR	NTCIR-6 CLQA	NTCIR-6 MuST	NTCIR-6 OPINION	NTCIR-6 PATENT 专利检索 分类	NTCIR-6 QA	—	—	—	—
NTCIR-7	—	—	NTCIR-7 ACLIA	NTCIR-7 MuST	NTCIR-7 MOAT	NTCIR-7 PATMN	NTCIR-7 PATMT	—	—	—	—	—

注：NTCIR-8为如下的任务准备了测试集，它们是：ACLIA、GeoTime、MOAT、Patent Mining、Patent Translation、Pilot Task (Community QA)

利挖掘任务”，向国内外募集参加团队始于2007年。这来源于大学研究者已不仅仅局限于论文，也有检索相关专利信息的需求，他们申请专利的机会也在不断增多。2009年6月日本政府知识产权战略总部发表了“知识产权推进计划2009”，在继2006、2007及2008之后，持续强调专利信息在大学研究中的重要性。这一计划中，包含了以供大学研究者利用的专利和论文信息综合检索系统，这种倾向被认为今后会持续增强。

综合检索论文和专利也不限

于大学的研究者。例如，专利厅的审查官为了判断申请的技术是否可以取得专利权，要进行先行技术的调查，需要调查过去是否有过同样的专利申请或论文的发表。除此之外，所谓检索者（Searcher）这一专门担当要对审查官审查的申请技术进行再度调查。另外，为了使竞争对手的权利无效，在私企进行可证明权利无效的资料调查，也需要以论文和专利为检索对象。

从NTCIR-3开始，就有了专利检索任务PATENT，伴随信息挖掘技术成为热点话题，专利挖掘

成为了PATENT的重要任务。难波英嗣 (NANBA Hidetsugu)、藤井敦 (FUJII Atsushi) 等人从2007年6月-2008年12月NTCIR-7，作为构筑研究论文和专利处理对象的测试集，启动了专利挖掘的任务。下节阐述NTCIR-7及NTCIR-8专利挖掘任务。

2.1 专利挖掘任务的路线图

NTCIR的专利挖掘任务是：开发各种自然语言处理技术，对某一

领域论文和专利进行检索和趋势分析，自动生成技术动态图。为实现这一目标，在专利挖掘任务中实施的两个子任务是：（1）学术论文分类（采用国际专利分类体系之一的IPC对学术论文自动分类，使论文与专利可以统一分类、检索和分析）；（2）基于论文和专利，生成技术动态地图。

专利挖掘任务的最终目标是：



图1 从特定领域的一组论文和专利中生成技术动态地图的实例

为了自动生成这种地图，需要 某一领域各个语种的论文和专利。
以下2步工作： ②从步骤1收集的论文和专利
①对于某一给定的领域，收集 中，抽出要素技术和功效对。

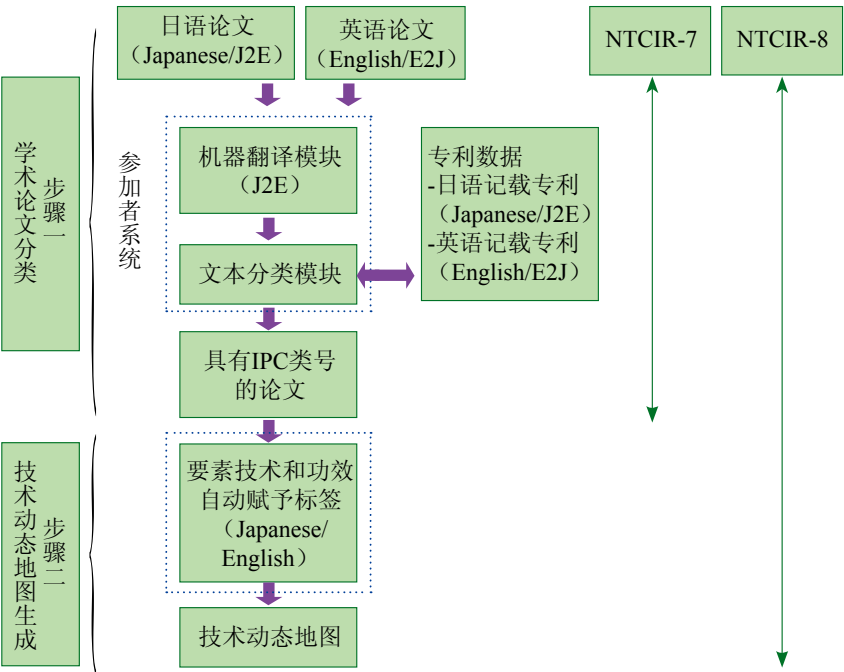


图2 专利挖掘任务路线图

上述2步工作，都要执行专利挖掘任务中的两个子任务。即学术论文分类子任务和技术动态地图生成子任务。其中，有关“学术论文分类子任务”，在NTCIR-7上已经实施，技术动态地图生成是NTCIR-8的实验型子任务。NTCIR-8研讨会中，实施这两项子任务。图2表示了专利挖掘任务的路线图。下节阐述各子任务的概要。

2.2 专利挖掘任务的子任务概要

2.2.1 学术论文分类子任务

如图2所示，在专利挖掘任务中，对日语或英语论文的文摘，自动赋予国际专利分类IPC号。IPC基于专利文献的技术内容，从大到小算起，由“部”、“大类”、“小类”、“主组”、“分组”5级构成。本任务是赋予论文摘要到“分组”级别的IPC类号。图3是日语论文的例子。这里<TOPIC-ID>是论文的题目和文摘。将图3所示加了标签的输入数据发布给任务参加者，要求其构建自动对应IPC类号的输出。以上的子任务可以用图4表示。其中：

- 日语文本分类 (Japanese)：日语论文用日语专利数据分类。
- 英语文本分类 (English)：英语论文用英语专利数据分类。
- 跨语言文本分类 (J2E)：日语论文用英语专利数据分类。
- 跨语言文本分类 (E2J)：英语论文用日语专利数据分类。

[Japanese]
PM磁束制御用コイルを設けて<Technology>閉ループフィードバック制御</Technology>を施すため、<Effect><Attribution>電力損失</Attribution>を<Value>最小化</Value></Effect>できる。
[English]
Through <Technology>closed-loop feedback control</Technology>, the system could <Effect><Value>minimize</Value> the <Attribution>power loss<Attribution> </Effect>.

图3 已标引输入数据

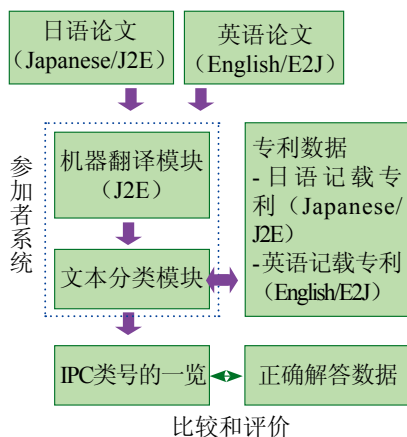


图4 学术论文分类子任务的描述

2.2.2 技术动态地图生成子任务

这一子任务是从NTCIR-8开始的课题，目的是从论文和专利中自动抽出表现要素技术和其功效的文本。例如，输入“由于设置PM磁束控制线圈，实施闭环反馈控制，可使电力损失最小化”这一句话，如图5所示，在要素技术和功效之处，自动赋予“Technology”及“Effect”标签。在“Effect”中，进一步给出

“Attribute”（属性）和“Value”（属性值）两类下位标签。有关技术功效的表现多种多样，如果将其全部作为处理的对象，目前的语言处理技术非常困难。为此，对于像“处理速度（Attribute）提高（Value）”及“噪音（Attribute）减小（Value）”这样的，仅仅将技术功效用“属性（Attribute）”和“属性值（Value）”对来表现的内容作为处理对象。近年的自然语言处理领域中，抽取文本中出现的属性和属性值对是研究的热点，相关的技术积累也在增多。采用论文和专利中属性及属性值对，抽取表现技术功效的部分，可以期待这种现行技术的利用。从某一领域的论文和专利中，如果抽出如图5所示的要素技术和功效对，可以认为能实现图1所示的技术动态地图的自动生成。

在技术动态地图作成的子系统中，实施如下的课题：

- 日语信息抽出（Japanese）：从日语论文和专利中，自动抽出有关要素技术的部分。
- 英语信息的抽出（English）：从英语论文和专利中，自动抽出有关要素技术的部分。

NTCIR-8专利挖掘任务中，还追加了实施技术动态可视化的课

题。这一课题，是技术动态地图生成子任务的一种。图6是在技术动态可视化子任务中所用的数据一例。

比如从“形态素解析”及“机器翻译”等特定领域的论文及专利中，可以抽出图6所示的精度值，以精度值为纵轴，论文著作年及专利申请年为横轴，可以画出精度值随时间推移的图形。图7是从有关“形态素解析”一组论文中，实际手工抽出的精度值，绘制成的图形。从此图看出，在形态素解析领域，从1994年左右，精度达到了95%以上，可以知道，此领域技术逐步成熟。在此，2000年精度有若干的下降，这是因为评价用数据不同之故。本来评价用数据及实验条件如果不同，不能直接比较评价价值，但是，对于新加入该领域的企业，可以作为参加可能与否的判断材料，针对此目的，图7的图形是十分有用的。

```
<Technology>CRF</Technology>を用いた手法では、
<Effect><Value>0.935</Value>の
<Attribution>精度</Attribution></Effect>が得られた。
```

图6 技术动态可视化子任务中所用数据的一例

3 NTCIR参会体验及建议

为使读者系统地了解在国际信息检索测评领域中具有重要地位的NTCIR，本文对其整体情况进行了介绍，并重点介绍了NTCIR-7和8中实施的专利挖掘任务，即论文的IPC自动分类和基于论文和专利技

PM磁束制御用コイルを設けて<Technology>閉ループフィードバック制御</Technology>を施すため、<Effect><Attribution>電力損失</Attribution>を<Value>最小化</Value></Effect>できる。

图5 技术动态地图生成子任务所用数据的一例

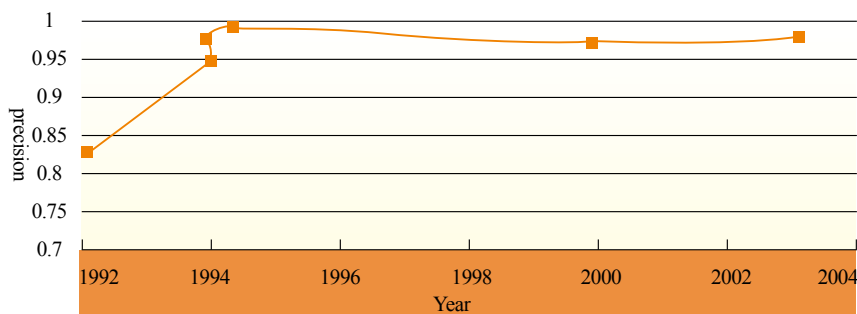


图7 技术动态可视化子任务的输出例

术动态地图的自动生成,这是基于目前研究者的需求,利用自然语言处理技术,通过综合分类、检索及分析论文和专利信息,制作反映领域技术动态地图的热点研究课题。为此,中信所也启动了类似的研究课题,并在首次尝试参与NTCIR-8的活动中,取得了在专利挖掘任务中第2名的成绩。我们的体会与建

议是:

- 贵在参与: NTCIR是研究者之间竞争与协作发展的平台,比起排名,其主要目的是通过参与,客观地审视自己的系统;通过交流,发现问题,从而改善系统;

- 反复验证: 目前自然语言处理技术的各种手段都是复杂的研究课题,各有利弊,没有适用于各种

情景的解决方案。只有通过反复验证,才能研发出最佳的实用技术,没有捷径可走;

- 基于实用: 信息检索是一个非常实用的课题,离开实际应用,片面地追求某一算法的理论排名,就等于脱离了孕育生命力的土壤,理论必须与实际相结合。

未来的计划: 中国科学技术信息研究所本次参赛的数据处理对象为美国专利文献,通过本届的交流与实践,我们期望在未来的研发中,把可处理文献种类进一步扩大到中国专利、英语及中文论文等,最终实现自动生成基于领域论文和专利的动态地图,从而达到利用先进的技术手段支持国家创新的发展的目的。

参考文献

- [1] 神門典子(NTCIR执行委员长). The 8th NTCIR Workshop Meeting NTCIR-8 成果报告会NTCIR. 情报检索システム評価用テストコレクション構築プロジェクト [EB/OL]. [2010-7-14]. <http://research.nii.ac.jp/ntcir/index-ja.html>.
- [2] NTCIR Flyer [EB/OL]. [2010-06-10]. <http://research.nii.ac.jp/ntcir/ntcir-ws8/meeting/ntc8-mtg.flyer.j.pdf>.
- [3] NTCIR Project [EB/OL]. [2010-06-10]. <http://research.nii.ac.jp/ntcir/data/data-ja.html>.
- [4] 难波英嗣,藤井敦,岩山真,等. 論文と特許を対象にした技術動向分析—第7回,第8回NTCIRワークショップ特許マイニングタスク[J]. 情報管理, 2009(6):334-342.
- [5] 李颖,赵蕴华,霍翠婷,等. 日本专利检索体系中主题分类“FI/F-term”的理论与应用研究[J]. 数字图书馆论坛, 2008(11):11-15.

作者简介

李颖, 日本筑波大学信息学博士, 信息系统专业。主要研究方向是: Web信息知识系统、基于XML的跨媒体数字出版。最近研究课题: 基于XML的数字出版、基于DOI的文献链接系统、跨语言检索、专利分析、数字版权保护等。通讯地址: 北京市海淀区复兴路15号 中国科学技术信息研究所 信息技术支持中心 100038。E-mail: liying@istic.ac.cn

桂捷, 博士。研究方向: 专利分析和科技创新管理。通信地址同上。E-mail: guij@istic.ac.cn

乔晓东, 硕士, 研究员。中国科学技术信息研究所 信息支持中心主任。研究方向: 信息服务和信息资源管理。通信地址同上。E-mail: qiaoxd@istic.ac.cn

Overview of NTCIR and Its Patent Mining Task: Science and Technology Trends Analysis Based on Papers and Patents

Li Ying, Gui Jie, Qiao Xiaodong / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: This paper gives the overview of NTCIR; focuses on its Patent Mining Task that ISTIC participated in on 15-18 June 2010, introduces its two sub-tasks. They are Research Papers Classification and Technical Trend Map Generation (a pilot subtask). Finally, it gives our experiences and suggestions for participation in NTCIR.

Keywords: NTCIR, NII Test Collection for IR Systems, Patent mining task, Research papers classification, Technical trend map generation

(收稿日期: 2010-08-15)