

NSTL智能检索平台的扩展检索效果测评与分析*

王莉, 梁冰

(中国科学技术信息研究所, 北京 100038)

摘要: 本测评采用Pooling技术, 旨在考察NSTL智能检索平台的扩展检索效果。实验表明, 基于STKOS的概念扩展检索能够有效地提高检索召回率, 与此同时也导致整体检索准确率下降。实验结果中, 扩展检索在MAP和P@10两个指标上的表现均优于NSTL现有生产平台, 表明该功能能够给用户带来更好的检索体验。

关键词: NSTL; 智能检索平台; 扩展检索; 测评

中图分类号: G254

DOI: 10.3772/j.issn.1673-2286.2015.11.005

1 引言

NSTL智能检索平台是国家十二五科技支撑计划课题“信息资源自动处理、智能检索与STKOS应用服务集成”的重要成果产出, 重点解决自动标引、知识网络构建、多维索引、自动聚类、个性化服务等关键技术复杂应用下融会贯通的问题, 通过超级科技词表(以下简称STKOS)的支撑, 深入揭示NSTL海量外文科技文献资源, 建立知识网络, 为科技创新主体提供高效便捷的知识服务和科研文献信息支撑环境。为了了解NSTL智能检索平台的检索效果, 考察其是否达到课题预期目标, 课题组分别设计了针对自动标引和扩展检索两大核心功能的测试, 本文仅介绍扩展检索部分的测试工作。

为了测试扩展检索的效果, 课题组设计了针对NSTL生产平台和NSTL智能检索平台两大系统核心搜索引擎的检索效果对比测评实验。本文首先简要介绍NSTL智能检索平台的原理, 进而叙述本次测评的过程, 并针对测评结果对NSTL智能搜索引擎的扩展检索表现进行分析。

2 NSTL智能检索平台的原理

传统文献检索主要通过查询请求与文献特征之间的简单匹配来获得查询结果。这种简单的字符串匹配

方式不可避免地存在表达一致性问题, 一方面无法捕获用户的真实意图, 另一方面也无法满足用户在其知识不完整的情景下的文献检索需求。面对传统文献检索存在的诸多缺陷^[1], NSTL智能检索平台主要从改善检索效果与增强用户体验两个方面提出解决方案, 如图1所示。

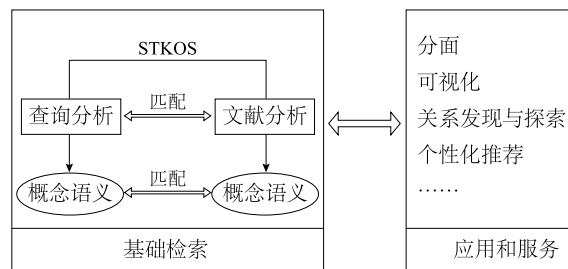


图1 NSTL智能检索平台的基本原理

改善检索效果的具体方法是: 利用STKOS对文献资源进行语义标注, 建立基于概念的文献索引, 同时, 在用户检索过程中, 借助STKOS及其他语料(如系统运行过程中积累起来的有效的历史检索词)分析用户输入的检索条件, 通过人机交互的方式获得准确的信息需求。通过分析“文献的语义”和“用户信息需求的语义”, 将查询过程自然地转换为具有语义关系的概念的匹配过程。这是NSTL智能检索平台的核心搜索引擎部分完成的工作, 是“智能检索”的基础, 其中, 查询分析

* 本研究得到国家十二五科技支撑计划课题“信息资源自动处理、智能检索与STKOS应用服务集成”(编号: 2011BHA10B05)资助。

由扩展检索实现(实时),文献分析由自动标引实现(预处理)。本次测评分析的主要对象是核心搜索引擎的扩展检索部分。增强用户体验是建立在搜索基础之上的应用和服务,主要包括检索结果的分面与可视化呈现、以关系发现为核心的探索式检索应用、以及个性化推荐,这部分内容不在本次测评范围内。

3 实验设计^[2-3]

3.1 基本思想

本次测评的基本思想是采用Pooling^[4]为大型文献集构建相关集合,分别考察NSTL生产平台的核心搜索引擎(简称NSTL)和NSTL智能检索平台的扩展检索(简称NSTL+)在检准和检全两个方面的表现。

首先,不同检索系统性能的可比较是建立在统一的数据集合和统一的测试方法基础之上的。本次测评选择NSTL生产平台作为对比,一方面因为NSTL智能检索平台的外文期刊文献数据和外文会议文献数据来源于NSTL生产平台,数据源一致;另一方面,课题预期目标是对改进NSTL现有生产平台的检索效果,将二者做对比测评更有针对性。

其次,NSTL生产平台和NSTL智能检索平台均表现为完整应用,需要将各自核心搜索引擎部分剥离出来进行测试,而不是直接在用户页面上进行检索。从查询主题的输入到检索结果的输出完全由搜索引擎完成,没有人工干预。因此,经过剥离的NSTL+不具备NSTL智能检索平台提供的人机交互方式调整检索条件的功能,同时NSTL+将NSTL智能检索平台原有的扩展查询提示,改为自动扩展1-5个查询词,与原始查询词以“或”的方式自动生成新的查询条件进行检索,封装成一个完整的扩展检索引擎。

因为硬件平台的差异,本次测评仅关注准确率和召回率,而不记录检索速度。

3.2 实验过程

实验包括准备查询条件、构建测试集、测评3个基本步骤。

(1) 查询条件

本次测评查询条件并没有从NSTL平台的用户历史查询中选取,主要原因是历史查询数据量大,且质量良

莠不齐,在短时间内很难筛选出有代表性且有普遍意义的查询条件;而且这些在平台使用过程中积累起来的查询条件仅仅以“字符串”的形式存在,没有更多关于其内涵的说明,不利于后期人工判断检索结果的相关性。因此,课题组从TREC国际检索大会历年题目中筛选出50个主题构成查询条件集合。

TREC查询主题^[5]示例如下:

```
<top>
<num> Number: TD4
<title>wireless communications</title>
<desc>Description: Information on existing and
planned uses, research/technology, regulations and
legislative interest
</top>
```

TREC查询主题筛选的基本原则是,用<title>内容直接在NSTL中检索至少能够命中1篇文献。这一原则确保能够获得有效测试集,同时筛选过程中并不检查STKOS是否覆盖查询主题,一定程度上兼顾了有限样本中尽可能的普遍意义。

为了便于后期的测评分析,对筛选出的主题全部重新编号,形如“S01”、“S02”……“S50”。采用<title>内容作为查询词执行全字段搜索。原主题中的描述信息<desc>并不参与检索过程,而是为人工判断检索命中文献与查询主题之间是否相关提供参考。

(2) 测试集

分别针对NSTL和NSTL+,用50个查询条件依次检索期刊论文和会议论文,每次查询选取前20篇命中文献(记录以下关键信息:查询编号、文献唯一标示符、题名、作者、关键词、摘要、分类号、出处、文献在检索结果集中的排序),共计4000篇文献,经过查重合并处理之后得到3901篇文献,作为本次测试文献集(Pooling)。

人工对测试集中每篇文献进行检索相关性判断,采用二元评判,只关心相关或不相关,如果文献和查询主题相关,则赋值1;如果不相关则赋值0。

(3) 评价指标

本次测评考察NSTL和NSTL+的Recall、R-Precision、MAP和P@10四个评价指标。

对于单个查询主题*i* (*i*=1..50),检索结果图示如下:

其中,A:检索到的、相关的文献数量;B:检索到

	相关	不相关
检索到	A	B
未检索到	C	D

图2 检索结果图示^[6]

的、不相关的文献数量; C: 未检索到的、但却相关的文献数量; D: 未检索到的、同时也不相关的文献数量。

$$\bullet \text{ Recall} = \frac{A}{A+C}$$

召回率是最基本的评测指标, 通过检索出的相关文献数和集合中所有的相关文献数的比率, 衡量检索系统的查全情况。一般来说, 召回率和准确率没有必然联系, 但是在实际应用中, 两者相互制约, 当一个检索系统通过命中更多结果来查到更多相关信息, 同时也会查到更多不相关信息, 导致召回率提高, 而准确率降低。

对于多个查询而言, 召回率是所有查询召回率的均值。

$$\bullet \text{ R-Precision} = \frac{A}{A+B}$$

单个查询的R-准确率是检索出R篇文献时的准确率, R是当前查询返回的文献总数, 即A+B。多个查询的R-准确率是所有单个查询的R-准确率的均值。

R-准确率可用于快速比较两个检索算法的性能。用 RP_A 和 RP_B 分别表示使用检索算法A和检索算法B执行查询时得到的R-查准率, 它们之间的差值记作: $RP_{A-B} = RP_A - RP_B$ 。如果 $RP_{A-B} = 0$, 则表明两个算法性能相同; 如果 $RP_{A-B} > 0$, 则表明A算法具有较好的性能; 如果 $RP_{A-B} < 0$, 则表明B算法有较好的性能。

$$\bullet \text{ MAP} = \frac{1}{A} \sum_{i=1}^A \frac{i}{\text{第}i\text{篇相关文献的位置}}$$

平均准确率进一步考虑检索出的相关文献的排序情况, 检索结果列表中相关文献越靠前, MAP值就可能越高; 如果检索结果列表中没有相关文献, 则MAP值为0。

多个查询的平均准确率是所有单个查询的MAP值的均值。

$$\bullet \text{ P@10} = \frac{\text{系统返回前10篇文献中相关文献数量}}{10}$$

NSTL生产平台和NSTL智能检索平台均默认检索结果页面每页显示10篇文献, 因此采用P@N方法计算查询结果中前10篇文献的准确率。多个查询的P@10是所有单个查询的P@10的均值。

4 评测结果及分析

实验输出结果如表1所示。

表1 文献检索测评结果

	NSTL	NSTL+
Recall	0.518062	0.55764
R-Precision	0.801604	0.739009
MAP	0.422003	0.501807
P@10	0.693	0.768

从整体检索效果来看, NSTL+召回率较NSTL有所提高, R-准确率则有所下降。与此同时, NSTL+在MAP和P@10两个指标上的表现均优于NSTL。

课题组对每个查询主题和对应的检索结果, 结合NSTL+对查询主题的扩展(扩展词相关、扩展词无关)进行分析, 旨在对NSTL+的四个评价指标给出合理的解释。

(1) 召回率的提高与基于STKOS的概念扩展有直接关系。

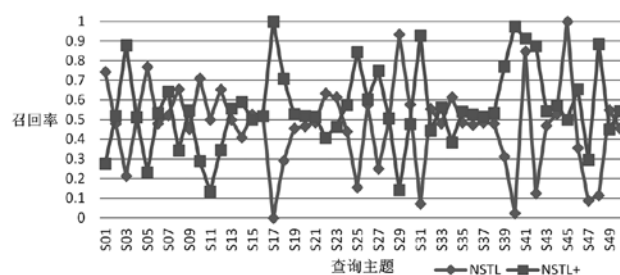


图3 50个查询主题下NSTL和NSTL+召回率对比

对50个查询及其结果进行分析, 其中34个查询召回率有所提升, 16个查询召回率有所下降, 总体表现为小幅上升。所有召回率上升的查询主题都经过扩展得到1-5个相关扩展词, 其中上升幅度较大的扩展效果尤为突出, 例如S40、S17、S42等。这种扩展突出表现为两个方面, 一是通过“原始查询词-相关术语-规范概念”的路径从STKOS中获得更多词或短语。这些词或短语和查询主题非常接近, 一般在同一个范畴类目下, 是原始查询词的不同表现形式, 并且根据可推荐的扩展词数量动态缩小原始查询词的内涵(下位词), 或适当扩大原始查询词的外延(上位词)。例如S40“weather hazards and extremes”, 通过STKOS扩展

得到“weather hazards”、“weather-related hazards”和“climatic hazards”3个短语,都与原始查询词含义非常接近,扩展前(NSTL)命中相关文献1篇,扩展后(NSTL+)命中文献直线上升为20篇,且均为相关文献。扩展的另一个突出表现是有效地拓宽了检索结果的学科覆盖范围。例如S17“Emergency and disaster preparedness assistance”,通过STKOS扩展得到“Emergency preparedness”,“Emergence planning, preparedness, and response program”,“Disaster preparedness”,“Disaster Readiness”和“US Emergency preparedness act”5个短语,与原始查询词含义接近。尤其是扩展得到的“Emergence planning, preparedness, and response program”具有重要的检索意义。扩展前(NSTL)检索结果集中在医药卫生和工业技术两个类目下,工业技术下只涉及一般工业技术和自动化技术、计算机技术两个子类;扩展后(NSTL+)检索结果中增加了环境科学、安全科学类目下的文献(包括环境污染及其防治,废物处理与综合利用,环境质量评价与监测),工业技术类也扩展了石油、天然气工业,水利工程,原子能技术。

在本次实验中有16个查询NSTL+召回率下降,其中5个查询经扩展得到的词完全无关,其他11个查询,每个查询的扩展词有的相关,有的无关。扩展词不相关的情况,对召回率没有贡献,这是合理的。扩展词相关的情况下,召回率也表现为下降,这可能是与文献集合的覆盖度有关,例如,虽然得到1个相关扩展词,但是该词在现有文献集合中没有查到符合的文献,自然不能提高召回率。当然,这只是一个假设,在当前测试集规模下无法进行验证。在本次实验中,这种局部表现出的召回率下降与扩展功能之间无明显关系。

(2) R-准确率的下降与基于STKOS的概念扩展有关,且局部影响较大。

对50个查询及其结果进行分析,其中21个查询

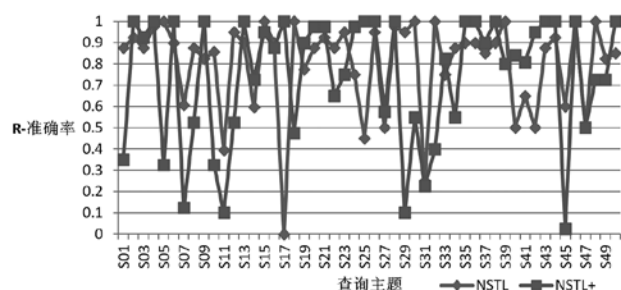


图4 50个查询主题下NSTL和NSTL+R-准确率对比

R-准确率下降,2个查询R-准确率不变,27个查询R-准确率上升,总体表现为小幅下降。所有R-准确率下降的查询都经过扩展得到1-5个无关扩展词,整体表现为无关扩展词越多,R-准确率下降越明显。例如S01“affirmative action”,通过STKOS概念扩展得到“action”,“biochemical action”,“solvent action”,“shear action”。其中,“action”将原始查询词的外延拓展得过宽,已经没有检索意义;另外3个扩展词则和原始查询词完全无关,导致R-准确率直线下降。

在27个R-准确率上升的查询中有17个查询每个均得到3-5个有检索意义的相关扩展词,其他10个查询的每个查询的扩展词有的相关,有的无关。在本次实验中,这种局部表现出的R-准确率上升与扩展功能之间无明显关系。

(3) MAP值的提高与基于STKOS的概念扩展有关。

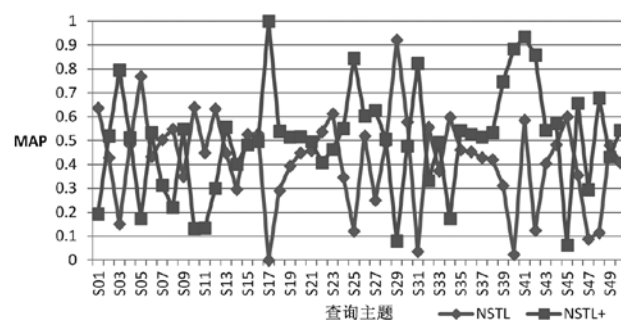


图5 50个查询主题下NSTL和NSTL+MAP值对比

对50个查询及其结果进行分析,其中33个查询MAP值上升,所有MAP值上升的查询都经过扩展得到1-5个相关扩展词,整体表现为相关扩展词越多,MAP值上升越明显。例如前文中列举的查询S40,扩展效果较好,MAP值从原来的0.125迅速攀升到0.883159。

17个MAP值下降的查询中有5个查询经扩展得到的词完全无关,其他12个查询的每个查询的扩展词有的相关,有的无关。在本次试验中,这种局部表现出的MAP值下降与扩展功能之间无明显关系。

(4) P@10的提高与基于STKOS的概念扩展有关。

对50个查询及其结果进行分析,其中31个查询P@10值上升,所有P@10值上升的查询都经过扩展得到1-5个相关扩展词,但是相关扩展词的数量与P@10值上升幅度之间没有明显关系。

19个P@10值下降的查询中有5个查询经扩展得到的词完全无关,其他14个查询的每个查询的扩展词有的相关,有的无关。在本次实验中,局部表现出的

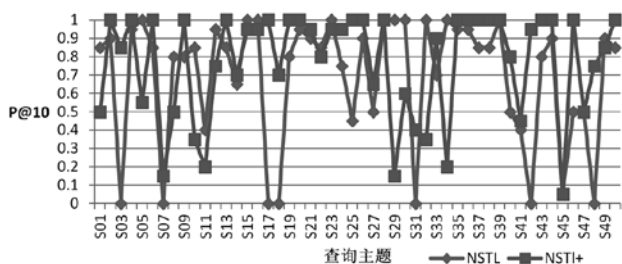


图6 50个查询主题下NSTL和NSTL+P@10对比

P@10值下降与扩展功能之间无明显关系。

综上所述, 基于STKOS的概念扩展能够有效地提高检索召回率, 与此同时也导致整体检索准确率下降; 而扩展词与原始查询词是否相关对四个测评指标均有不同程度的影响。在本次实验条件下, 无法进一步分析得到这种相关性影响的规律。

5 结语

为了测试NSTL智能检索平台中扩展检索的效果, 课题组采用Pooling技术, 选择Recall、R-Precision、MAP和P@10四个评价指标, 分别考察NSTL生产平台的核心搜索引擎和NSTL智能检索平台的扩展检索功能的表现。实验表明, 基于STKOS的概念扩展检索能够有效地提高检索召回率, 与此同时也导致整体检索准确率下降。召回率与准确率两个指标虽然没有必然联系, 但是在实际应用中存在相互制约的关系。在文献检索过程中, 当用户的检索需求不清晰, 或者难以明确表达的时候, 一般希望被检索到的内容越多越好, 这是追求“召回率”; 反之, 当用户对自己的需求非常了解, 能够明确地用查询条件表达出来, 总是希望被检索到的相关的内容越多越好, 不相关的内容越少越好, 这是追求“准确率”。NSTL智能检索平台扩展检索功能在“检全”和“检准”两个方面的表现不能简单地判断为更好或更差, 这是需要与用户的具体场景相结合的。正因为这个原因, NSTL智能检索平台对扩展检索的处理采用的是“系统主动提示, 用户自主扩展”的

方法, NSTL+的自动扩展模式只是为了测评需要而进行的改造。

虽然用户对“检全”和“检准”的追求不同, 但是对MAP和P@10的体验是相同的, MAP值越高表明检索结果列表中相关文献越靠前, P@10值越高表明用户最关注的第一个检索结果页面包含的相关文献越多。实验表明, NSTL智能检索平台扩展检索功能能够给用户带来更好的检索体验。

本次对比测评实验结果显示, 扩展词与原始查询词是否相关对四个测评指标有影响, 但是并没有找出这种影响的规律。一方面需要扩大测试数据集的规模, 另一方面需要引入专家对扩展词与原始查询词的相关性进行更细致的分析, 不仅仅是相关与不相关二值判定, 而应该包括同义、上位、下位等更多的相关关系。这种细致的相关关系判断分析, 能够帮助发现“如何扩展才是更好的扩展”, 从而指导扩展检索功能的优化。这也是课题组下一步要做的工作。

参考文献

- [1] 王莉, 梁冰, 白海燕等. 基于本体的科技文献检索框架与技术实现[J]. 数字图书馆论坛, 2012(7):37-44.
- [2] 张俊林, 刘洋, 孙乐等. 2005年度863信息检索评测方法研究和实施[J]. 中文信息学报, 2006, 20(21):19-24.
- [3] 信息检索评测技术[EB/OL][2015-10-13]. <http://baike.baidu.com/view/10504202.htm>.
- [4] 02信息检索评价[EB/OL] [2015-10-13]. http://wenku.baidu.com/link?url=x6YNro_I2fJEzxmR9YltGvORvWkrcuZmoOJFvMTMH0LDYijqho542MHZhLplORAlh2aM97xzOCMuGlzG8jU6BorWISG-eSFwJ105urdPHmi.
- [5] Data-English Test Questions (Topics) Files List [EB/OL] [2015-10-13]. http://trec.nist.gov/data/topics_eng/index.html.
- [6] 召回率[EB/OL] [2015-10-13]. http://baike.baidu.com/link?url=T6MJjE1J-z0gks7L2Ji39FqVpHJC7bVNy03TFF3NyBxSVANKPz_Mi-ZMoAEJj-0czf3gX7FR0HP-89zWYew9AYa.

作者简介

王莉, 女, 硕士, 中国科学技术信息研究所研究员, 研究方向: 数字图书馆, E-mail: wangli@istic.ac.cn。

梁冰, 男, 博士, 中国科学技术信息研究所高级工程师, E-mail: liangb@istic.ac.cn。

Evaluation and Analysis of the Expand Retrieval of NSTL Intelligent Retrieval Platform

WANG Li, LIANG Bing

(Institute of Scientific & Technical Information of China, Beijing 100038, China)

Abstract: The objective of this retrieval evaluation which uses Pooling technology is to investigate the effect of the expand retrieval of the NSTL intelligent retrieval platform. The results show that, the expand retrieval based on STKOS can effectively improve the recall, but also lead to the decrease of the precision. In the results, MAP and P@10 of the expand retrieval are better than those of NSTL's current platform. This shows that this function can boost users' experience.

Keywords: NSTL; Intelligent Retrieval Platform; Expand Retrieval; Evaluation

(收稿日期: 2015-10-29)

2015年全国知识组织与知识链接学术交流会 会议通知

为了探讨知识组织与知识链接的研究与发展方向,尤其是网络环境下知识组织与知识链接的新理念、新思路、新方法,中国科学技术信息研究所、国家科技图书文献中心、中国科学技术情报学会定于2015年12月11日在北京召开第六届“全国知识组织与知识链接学术交流会”,由《数字图书馆论坛》承办。欢迎国内图书情报界及相关领域的专家学者参加。会议相关事项如下:

一、会议主题

知识组织

- 知识组织研究进展
- 大数据中的数据清洗、组织和分析
- 事实型数据识别与分析
- 叙词表、本体等知识组织体系的构建和应用研究

知识评价

- 科学计量与评价
- 引文分析、主题分析
- 专利分析与利用
- Web科技资源评价

知识链接

- 科研实体关系揭示
- 网络资源链接及其关联分析
- 面向项目研发产出的关联分析
- 数据关联挖掘和揭示

知识服务

- 知识服务、知识管理研究进展
- 数字科研环境与开放共享服务
- 用户分析与个性化用户服务
- 知识图谱及可视化分析

二、时间地点

会议时间: 2015年12月10-12日。12月10日全天报到,12月12日疏散。

会议地点: 北京市海淀区复兴路15号中国科学技术信息研究所一层多功能厅(大厅报到)。

参会费用: 不收取会议费。11日中午提供工作餐,其他费用自理。

会议报名: 报名截至12月5日,每单位限报2人。请填写“报名回执”发至会议邮箱报名。

三、联系方式

联系人: 赵莹莹、王立学, 010-58882061

会议邮箱: KOLink@istic.ac.cn

会议网址: <http://168.160.16.186/conference>