中 文 信 息 学 报
第 20 卷 增刊 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 SUPPLEMENT

文章编号:1003 - 0077(2006)增刊 - 0019 - 06

2005 年度 863 信息检索评测方法研究和实施

张俊林[1], 刘 洋[2], 孙 乐[1], 刘 群[2]

(1. 中国科学院软件研究所,北京 100080; 2. 中国科学院计算技术研究所,北京 100080)

摘要:本次 863 中文信息检索评测的目的是检测互联网环境下大规模数据的中文信息检索技术的研究现状和系统有效性,中文与接口技术评测组综合考虑了目前信息检索面临的难点以及中文信息检索具有的特点设计了本次信息检索评测,本文详细描述了本次评测的组织过程,包括查询条件设计,语料库情况,标准答案查找方法以及评价指标和评测软件的介绍,通过对参评队伍的结果数据进行分析并结合查询条件的类型,本文还讨论了现有检索技术的优点以及存在的不足。

关键词:863;中文信息检索;评测

中图分类号:TP391 **文献标识码**:A

Research on the 863 Chinese Information Retrieval Evaluation (2005)

ZHANG Jun-lin[1], LIU Yang[2], SUN Le[1], LIU Qun[2]

(1. Institute of Software, Chinese Academy of Science, Beijing 100080, China;
2. Institute of Computation, Chinese Academy of Science, Beijing 100080, China)

Abstract: The objective of 863 Chinese information retrieval evaluation is to investigate the current research status and the system validity of Chinese IR system under the circumstance of mass data of WEB. The organizers consider both the difficulties the IR technology is facing and the characteristic of Chinese IR to design the evaluation. In this paper, we describe the following main procedures of the evaluation: query designing, corpus composition, relevant document set finding and the evaluation tools. We also discuss the advantages and disadvantages the current IR technology shows in the test by analyzing the performance of all the submitted runs .

Key words: 863; Chinese information retrieval; evaluation

1 引言

为了进一步了解国内外在中文信息处理和智能人机接口技术领域的现状,检查 863 计划信息领域计算机软硬件技术主题中相关课题的进展情况,促进交流和提高,推动技术进步和成果的应用与产业化,并为 863 计划课题验收和下一轮课题评选打下基础,计算机主题专家组举行了 2005 年度 863 计划中文信息处理与智能人机接口技术评测——信息检索评测。

本次检索的目的是检测互联网环境下大规模数据的中文信息检索技术的研究现状和系统有效性,中文与接口技术评测组对参加信息检索评测的六个单位的信息检索系统进行了测试,其中一个单位因为某些原因没有提交结果,所以本次评测有效参加队伍为五个。

收稿日期:2005 - 11 - 05 **定稿日期**:2006 - 01 - 13

基金项目:国家 863 计划资助项目(2004AA114010;2003AA111010)

作者简介:张俊林(1974—),男,博士,助理研究员,主要研究方向为信息检索,自然语言处理等.

在本次评测,我们针对信息检索面临的难点以及中文信息检索面临的独特问题设计了评测实验,希望能够通过大规模数据评测来评价现有检索技术。本文以下内容分别叙述本次评测的组织过程以及针对评测结果对现存技术的优点和不足作初步的分析。

2 实验设计

研究表明[1]英文信息检索导致效果差的主要原因包括:查询条件词汇权重无法正确设定;词汇的歧义以及查询条件和文章选择表达某一主题词汇不同导致的不匹配等,另外对于中文来说,由于相对英文一般在索引或者查询条件分析阶段增加了分词过程,所以中文分词中存在的一些问题比如命名体识别,新词语,缩略语等也会影响检索的准确性。中文与接口技术评测组针对信息检索面临的难点以及中文信息检索面临的独特问题设计了评测实验,期望能够通过大规模数据评测来推动中文信息检索技术的发展。

2.1 查询条件

此次评测拟订了50个中文查询条件。查询条件(Topic)模拟用户需求,由若干字段组成,采用规范格式描述用户希望检索的信息。每个查询条件统一由以下4个字段组成:编号(num)、标题(title)、描述(desc)和叙述(narr)。下面是本次评测所采用的查询条件示例:

<top>

<num> 编号:020

<title>奥兰多·布鲁姆

<desc> 描述:奥兰多·布鲁姆参与演出的影片的相关介绍

<narr> 叙述:自从出演了《魔戒》中的"精灵王子"以后,奥兰多·布鲁姆受到了越来越多观众的喜爱。查询奥兰多·布鲁姆参与演出的影片的名字,剧情等相关报道,奥兰多·布鲁姆的个人生活不在检索范围之内。

</top>

本次评测拟订查询条件遵循如下原则:

(1)由于是互联网环境下大规模数据的中文信息检索技术评测,所以评测的查询条件尽量模拟搜索引擎用户的真实信息需求。标题查询域(title)尽量简短,一般长度为2~5个词汇。描述查询域(desc)一般为一句到两句自然语言语句。叙述查询域(narr)进一步详细描述用户的信息需求。部分题目出题时参考目前互联网搜索引擎统计出的比较热门的用户需求。

(2)题目涉及领域尽可能全面。包含政治,经济,文化,娱乐,体育等多个不同领域,每个领域题目在3~10个不等。

(3)题目整体难易程度适中。所谓适中,指的是简单题目大约占总题目数量的50%。

(4)题目的标准答案数目不应太少也不宜过多。因为这样的题目对于参评队伍来说没有区分能力。在没有得到最终标准答案前靠人工搜索大致评估答案数目,对于答案过多的问题,在NARR查询域增加限制条件以减少相关文章数目。

遵循以上命题原则,根据测试数据,按照大纲上规定的格式制作了题目,利用检索工具对测试数据进行索引,人工对题目进行评价并按照查询条件拟订原则进行调整。

参加评测的单位通过自动方式和人工方式根据查询主题构造查询。自动方式是指在没有任何人为因素的影响下根据主题构造查询的方式;除此之外的方式均为人工方式。本次评测不对构造查询和建索引的技术做任何限制。主题的字段中的信息全部可以利用。

2.2 评测语料

本届信息检索评测采用的评测集合是由北京大学计算机网络与分布式系统实验室提供的以中文为主的 Web 测试集 CWT100g(Chinese Web Test collection with 100 GB Web pages),CWT100g 是根据天网搜索引擎[2]截止 2004 年 2 月 1 日发现的中国范围内提供 Web 服务的 1,000,614个主机,从中采样 17,683 个站点,在 2004 年 6 月搜集获得 5,712,710 个网页,包括网页内容和 Web 服务器返回的信息,容量为 90GB。其中每个网页对应的服务器返回信息中的 MIME 类型都是"text/html"或者"text/plain"。

2.3 标准答案

在确定了评测语料以及查询条件,在各队提交了检索结果后,结合各队的结果利用 Pooling 方法形成初步标准答案集合,因为参赛队伍比较少,为了增加标准答案的有效性,通过人工草拟查询条件利用搜索引擎来查找相关答案对初步答案进行补充,最后合并形成最终的标准答案。

本次采取的 Pooling 作法为:针对每个查询主题,从参与评比的各系统所送回的测试结果中抽取出前 100 篇文档,合并形成一个 Pool,视之为该查询主题可能的相关文档候选集合,将集合中重复的文档去除后,再送给该查询集的构建者进行相关判断。这种做法的一个局限是,当参赛队较少的时候,可汇集的结果权威性较差。本次评测设计了弥补这一局限性的方法,即:人工对于查询主题草拟若干个查询条件,提交针对 CWT100g 语料构建的搜索引擎,对于检索结果进行人工判断与筛选,将两个方法所获得的答案进行结果集合成。这样,即使参加队数量不多,也能形成质量较高的结果集,达到检验参与系统检索质量的目的。从效果看,两者确实有很好的互补作用。

本次评测采用二元评判,即一个网页或者与主题相关,或者不相关。一个网页与主题相关,必须同时满足以下两个条件:

[1]网页的内容切合主题;

[2]网页的内容符合主题的 desc 域(描述)和 narr 域(叙述)提出的约束条件。

2.4 评价指标

1. MAP (Mean Average Precision)

单个主题的平均准确率是每篇相关文档检索出后的准确率的平均值。主题集合的平均准确率(MAP)是每个主题的平均准确率的平均值。MAP 是反映系统在全部相关文档上性能的单值指标。系统检索出来的相关文档越靠前(rank 越高),MAP 就可能越高。如果系统没有返回相关文档,则准确率默认为 0。

2. R-Precision

单个主题的 R-Precision 是检索出 R 篇文档时的准确率。其中,R 是测试集中与主题相关的文档的数目。主题集合的 R-Precision 是每个主题的 R-Precision 的平均值。

3. P@10

单个主题的 P@10 是系统对于该主题返回的前 10 个结果的准确率。主题集合的 P@10 是每个主题的 P@10 的平均值。

2.5 评测软件

这次 863 信息检索评测所使用的程序基于 TREC[3]提供的标准程序 trec_eval.7.0。主要是对文档的输入(trec_eval.c)输出(print.c)部分进行了修改。此次的 863 评测提出的 3 个评分标准完全包含于 TREC 评测指标中,此外 TREC 评测还给出了其他的一些评分标准。该软

件读取文本检索系统生成的结果,对比标准答案,输出大纲所要求的各项参数的指标。完成自动评测软件后,进行多次调试和校验,确保与大纲保持一致。

3 评测结果以及分析

本次评测分为两组:自动构造查询条件,手工构造查询条件。使用自动评测软件对参评单位信息检索系统的检索结果进行评估。对于评估的结果进行反复校验,确保和大纲一致,没有差错。结果记录在以下两表中:

表1 信息检索评测结果—自动组

指 标	System 1	System 2	System 3	System 4	System 5
MAP	0.2727	0.1862	0.3107	0.3175	0.2858
P-PRECISION	0.3320	0.2554	0.3672	0.3605	0.3293
P@10	0.5300	0.5180	0.6240	0.5540	0.6280

表2 信息检索评测结果—手工组

指 标	System 1	System 2	System 3	System 4	System 5
MAP	0.3257	0.1705	0.3538	0.2673	0.3671
P-PRECISION	0.3826	0.2327	0.4078	0.3185	0.4140
P@10	0.5580	0.4640	0.6840	0.4800	0.7040

从所有参评队伍的整体检索效果看,与2004年的评测结果相比,各个指标都有了很大提高。导致本次评测效果提高有以下几个可能原因:

1. 由于本次评测语料扩大为90G,同时提供了URL等链接信息,所以参评队伍可以利用更多的相关评价因素,比如链接分析技术,链入链出文字等信息,这样可以更加准确的搜索正确结果;

2. 由于提供去年的评测数据作为训练语料,所以参赛队伍能够针对评测采取有效的技术手段来克服中文检索中的某些特点,比如实体名称识别等,同时可以根据训练语料来获得比较稳定有效的检索模型系统参数;

3. 部分参评队伍采用了有效的相关反馈或者重排序技术对于提高检索效果也有一定的帮助。

从自动组和手工组检索结果比较来看,参评队伍中有两个队伍(System 2 和 System 4)的手工组结果与自动组结果相比性能有所降低,而其它三个参评队伍的手工组结果与自动组结果相比有大幅度的提高。System 4 所造成手工组结果降低的原因在于没有象自动组进行相关反馈,System 2 手工组结果相对自动组结果偏低的原因可能是所采用的检索模型差异以及伪相关反馈造成的。而其它队伍手工组结果的大幅度提高从一个方面说明,如果用户草拟的查询条件能够比较全面准确的表达用户需求的话,现有的中文检索技术一般能够提供比较好的检索结果。

从采用的检索模型来说,参评队伍采用了向量空间模型[4],概率模型[5,6],语言模型[7]等基本检索模型或者是几种检索模型混合模型[8],同时利用了PageRank,链入分析等应用在互联网信息检索的链接分析或者页面分析技术来有效的提高检索效果。

中文检索相对英文等其它语种检索来说,如何正确分词对于检索结果的效果有所影响,尤其是命名体,缩略语以及新词等一些词典未登录词的正确识别对于某些查询来说影响比较大。本次评测绝大部分队伍(除了System 5以外的4个队伍都采用了命名体识别,System 5则采用了大词典和双向分词技术)在索引以及查询条件分析阶段采用了命名体识别,从结果来看,取

得了比较好的效果，大部分未登录词都能够正确识别。即使对于未能正确识别的未登录词，有些队伍也通过“词对”[5]或者“短语”[8]来进行有效的处理。而对于中英混合的查询来说，大部分队伍也作了相应的处理，System 2 在索引阶段没有将英文以及数字进行索引，导致对于中英混合性的查询条件处理效果很差，这对于整体效果影响也比较大。

通过对自动组结果的具体分析，我们来探讨现存系统的优点以及存在的问题。

首先，根据 5 个参评队伍结果每个查询主题的情况进行分析，我们将 50 个查询条件分为以下三类：所有（大部分）参评队伍检索效果都不好的查询主题；部分参评队伍检索效果不好的查询主题（至少两个队伍效果比较差）；大部分参评队伍检索效果都比较好的查询主题（都比较好，或者只有一个队伍检索效果比较差）；

1. 所有参评队伍检索效果都不好的查询主题；

在这个类别的查询条件包括：

（1）005：NBA 全明星赛

（2）014：电脑游戏的危害

（3）019：“三农”问题

（4）026：汽车招回

（5）030：美朝关系

（6）040：量子物理

（7）044：军演意外事件

（8）046：街舞

（9）049：心理疾病

对于查询条件 005 和 014，现有技术失效的原因可能是由于查询条件叙述域（narr）部分内容较多，导致自动构造的查询主题无法获得主题重点内容造成主题偏移；

对于查询条件 019，现有技术失效的原因可能由于查询条件并未指出“中国现任国家主席”的名称，这样导致查询条件和文档词汇失配（mismatch）导致检索效果下降，另外一种可能是检索系统对于缩略语“三农”无法正常解析索引导致；

对于查询条件 026 和 044 以及 049，现有技术失效的原因可能是查询条件和文档词汇失配（mismatch）导致检索效果下降；

对于查询条件 030 和 046，现有技术失效的原因可能是检索系统未能正确识别缩略语或者新词（街舞）导致检索效果下降。

2. 部分参评队伍检索效果不好的查询主题

下列查询条件属于此类：006/007/009/012/013/027/028/034/037/038/047/048。这个类别的查询条件有些队伍检索效果比较好，而有些查询条件有些队伍检索效果比较差甚至很差，导致这个现象的原因可能与各个队伍采用的不同检索算法和技术手段相关，需要结合具体查询以及各个队伍采用的技术方法进行具体分析才能得出结论。可以看出问题主要集中在以下几个方面：

（1）中英文以及数字等混合查询条件处理；

（2）中外人名识别；

（3）查询条件与文档词汇失配。

之所以有些参评队伍效果很好而有些参评队伍效果较差，这可能与参赛队伍具体采用的索引方式，命名体识别系统效果以及是否采用相关反馈技术等有密切关系。除了上面提及的

一些问题，另外一个比较常见的问题是无法正确设定查询词汇权重也在一定程度上影响检索效果，比如本次评测005号查询条件“NBA全明星赛”，我们发现参评队伍提交答案中很大比例的前列相关文档是讲述关于CBA全明星赛的内容。

3. 大部分参评队伍检索效果都比较好的查询主题

其它不在上述两个类别的查询一般检索效果都比较好，我们发现除了个别队伍外，大部分检索系统对于中外人名，地名机构名称以及新词，缩略语或者是中英混合查询条件处理效果都不错。这说明随着中文信息处理技术的提高以及与现有检索技术的融合有效的缓解了中文信息检索面临的一些独特的问题。

综合以上分析，可以看出目前中文检索系统对于大部分实体名称，新词，缩略语以及多语言混合查询处理效果比较好，但是对于以下方面还存在着一些问题：

- 查询条件与文档词汇内容失配；
- 部分命名体，新词以及缩略语识别还存在着一些问题；
- 在计算相似度时，查询词汇权重的设定正确与否也在一定程度上影响检索效果。

这些问题的存在导致现有检索系统处理类似内容的失效，针对这些问题，现有的检索技术还有很大的改善空间来获得比较满意的检索结果。

4 结论

为了进一步了解国内外在中文信息处理和智能人机接口技术领域的现状，计算机主题专家组举行了2005年度863计划中文信息处理与智能人机接口技术评测——信息检索评测。本次检索的目的是检测互联网环境下大规模数据的中文信息检索技术的研究现状和系统有效性。本文通过对于本次评测的组织过程以及结果数据分析讨论了现有技术的优点以及存在的不足，我们相信通过863信息检索评测，对于中文信息检索技术的研究会有非常好的促进作用。

参 考 文 献：

[1] Shah, C and Croft, WB, Evaluating High Accuracy Retrieval Techniques[C]. In: Proceedings of SIGIR 04, (2004).

[2] 北大天网，http://e.pku.edu.cn[EB].

[3] Text REtrieval Conference, http://trec.nist.gov[EB].

[4] 程羽心，等. 2005年863网页检索ICST评测报告. 863技术报告[C]. 2005.

[5] 赵乐，等. 清华THUIR 2005年863信息检索评测. 863技术报告[C]. 2005.

[6] 徐蔚然，等. PRIS信息检索技术报告. 863技术报告[C]. 2005.

[7] 吕碧波，等. 863信息检索评测—自动化所. 863技术报告[C]. 2005.

[8] 张志昌，等. 2005年863信息检索评测哈尔滨工业大学信息检索研究室技术报告. 863技术报告[C].