基于搜索引擎的中文文本检索效果比较研究

李　萌　孙　甲　贾　佳　孙济庆

(华东理工大学科技信息研究所　上海　200237)

摘　要　提出了对搜索引擎比较研究的方法,并设计了一组实验,对4大搜索引擎(百度、谷歌、雅虎、MSN)进行中文文本检索的实验,通过数据分析,评价各个搜索引擎的检索效果。

关键词　搜索引擎　信息检索　中文文本

中图分类号:G354.2　　**文献标识码**:A　　**文章编号**:1005-8095(2009)02-0049-03

1　引言

随着因特网信息量的急剧膨胀，如何全面有效地把高质量的信息资源提供给用户已成为各大搜索引擎亟待解决的问题。对于中文文本检索来说,词与词之间没有明确的分隔标记,而是连续的汉字串,为汉语自动分词过程带来了很大的困难，从而造成了歧义等一系列严重影响检索效果的问题。因此全面有效地提高中文文本的检索效率，提高检索的相关性是各大搜索引擎努力的目标。

关于搜索引擎的比较研究一直以来备受人们关注,目前相关的研究大体可分为3类:(1)从搜索引擎的工作原理出发，进而探讨或对比各搜索引擎的功能、特点等。(2)从搜索引擎的应用角度进行对比研究。(3)以搜索引擎整体作为研究对象,分析搜索引擎的发展进程以及预测发展走势。

在搜索引擎检索效果研究方面，绝大多数相关文献均是从宏观角度对搜索引擎进行的定性分析，对其收录信息量、搜索范围、更新频率、标引项目、检索方式、检索符、扩展功能、检索界面设计、检索结果显示格式等方面进行了全面的比较。而对于搜索引擎最重要的检索相关性、定量统计分析的文献非常少。因此本文运用测试实验法,采用较大样本空间,完善样本的全面性，细化相关性等级等定量研究方法,对搜索引擎的研究加以完善和调整,以期得到相对科学、准确的数据,进而定量地分析总结4大搜索引擎在检索效果方面的差异及可能产生的原因,以提高人们对搜索引擎的认识。

2　研究方案

2.1　研究对象的选择

本研究项目选择了Google中文(谷歌)、雅虎中国、MSN和百度作为评估对象。此选择来自于2005年美国加利福尼亚的网页性能监测公司Keynote进行的名为“搜索引擎行业消费者体验排行榜”的调查,其研究样本取自2 000名18岁以上的美国消费者,采访他们的上网体验,得到的结果为:Google、雅虎和MSN按次序位列三甲,Google被受访者选为“以后最有可能选用的搜索引擎”。百度是目前最常用的中文搜索引擎之一。

2.2　检索研究的样本

本研究主要从微观角度通过检索样本进行定量分析。在搜索引擎定量分析研究中,我们根据《中图法》的26个大类,再加上日常生活中检索需求较大的社会信息、企业信息、娱乐、常识、新闻等方面的内容,共设计选取了30个检索课题作为检索研究的样本,所有课题均为随机抽样选定。

2.3　检索效果研究的基本思路

本次评估分为4大搜索引擎横向比较以及个别搜索引擎纵向比较两方面进行。

(1) 横向比较指在同一课题相同检索式的条件下,对各个搜索引擎的检索结果从检索条数、响应时间、相关性、死链接个数以及重复率等5方面进行比较。检索条数、响应时间是从宏观上反映搜索引擎的总体检索性能,相关性、死链接个数是从微观上考察搜索引擎的检索效果，重复率是比较各搜索引擎的排序效能。

(2)纵向比较是指对检索结果进行聚类分析,以评测不同搜索引擎对于不同类别的课题，其检索效果有无显著差异。

2.4　抽样检索评价方案设计

由于搜索引擎对于每一个课题的搜索结果条数均可能数以万计,甚至百万计,因此对检索结果的考察只能选择其中的一小部分检索记录进行分析。根据2004年3月搜索引擎营销公司iPropect的调查研究:81.7%的搜索引擎使用者在看完前3页之前就停止阅读搜索结果。本文研究4个搜索引擎前3页检索结果取前30个检索记录。

相关性的评价是根据每个课题前3页进行评价，逐条根据其检索结果内容与课题检索的相关程

收稿日期:2008-07-05

作者简介:李萌(1982—),女,2006级情报学硕士研究生;孙甲(1984—),女,2006级情报学硕士研究生;贾佳(1981—),女,2006级情报学硕士研究生;孙济庆(1952—),男,研究员,硕士生导师。

度进行主观性打分,并分为 4 个等级(见表 1)。

表 1　课题得分等级

范畴等级	范畴定义
一等(8~10 分)	内容符合用户需求,信息明显有用;毋须点击二级链接即可进入;页面干净,查找方便
二等(5~7 分)	内容较符合,信息有潜在有用性;可能经由点击二级链接才能进入;页面混乱,查找困难
三等(2~4 分)	内容与用户需求有一定联系,但无法解决用户需求
四等(0~1 分)	内容与用户需求毫无关联;死链接(死链接的确定为不同时间段内刷新至少 3 次)或重复链接

在搜索引擎中相关性排序也是搜索引擎检索效能的重要方面。为了能正确反映检索结果的相关性排序,根据记录的排序为每条记录赋以不同的权重:

第 1 页 0.6;第 2 页 0.3;第 3 页 0.1。每页的前 3 条 0.45;中间 3 条 0.32;最后 4 条 0.23。每条信息的排序权重=页数权重 * 条数权重。

相关性分数=sum(分数 * 信息权重)。

检索条数、响应时间直接取自各个搜索引擎给出的数据。死链接个数根据各课题前 3 页中不能实现链接的个数统计。

重复率是比较各搜索引擎的排序效能按各课题前 3 页中 4 大搜索引擎均有相同记录的条目数统计,首页重复率对比 4 大搜索引擎各课题第一页中均有的记录条目数。

3　数据分析及结果讨论

3.1　4 大搜索引擎平均重复率

我们分别依据 30 个课题,采用同样的检索词对 4 大搜索引擎进行检索,同一链接在 4 大搜索引擎检索结果中同时出现的条数(见表 2)。

表 2　4 大搜索引擎重复条数　　(单位:条)

课题序号	1	2	3	4	5	6	7	8	9	10
重复条数	0	1	0	0	0	0	0	0	0	0
课题序号	11	12	13	14	15	16	17	18	19	20
重复条数	1	0	0	0	0	1	0	0	3	0
课题序号	21	22	23	24	25	26	27	28	29	30
重复条数	4	0	0	1	0	0	0	0	0	0

可以看出,在对某一课题进行检索时,4 大搜索引擎提供给用户的检索结果是大相径庭的,重复率仅为 1.22%。我们又对同一链接同时出现在 4 大搜索引擎检索结果第 1 页的条数进行了统计,仅有 3 条重复,重复率仅为 0.33%。

这一结果与基于英文文本搜索的结果一致。根据搜索网站 Dogpile.com 联合美国匹兹堡大学和宾夕法尼亚州立大学的研究人员所进行的一项调查结果,Windows Live、Ask Jeeves、Google 和 Yahoo 4 大搜索引擎的搜索结果存在明显的不同。在对 19 332 个查询进行调查后,他们发现,Google、Yahoo、Windows Live 和 Ask 搜索的顶部结果相同率仅为 3.6%,即使不考虑排序问题,四者呈现在顶端的 3 个结果也从来没有相同过,所有站点第 1 页搜索结果相同的几率不足 1%。

在我们每个课题的搜索中,也呈现类似的搜索结果,平均有 85%的搜索结果是各搜索引擎所独有的。这些数字表明,各大搜索引擎的搜索结果差异很大,都提供了不同的结果。这也表明每个搜索引擎有其各自的特点,都有不同的组织排序方法和不同的索引方法,因而用户的检索结果也是不同的,这也是我们研究评价搜索引擎的基础。

3.2　搜索引擎课题相关性

搜索引擎相关性是指用户输入的查询内容与搜索引擎提供的文件内容或相关网站资源之间的相关关系。搜索引擎的检索相关性关系着用户的搜索效率和满意程度,相关性好的搜索引擎可以节省用户的时间。

3.2.1　平均课题相关性

依据我们设定的相关性评价方案,对 30 个课题,共计 3 600 个链接页面的内容、排序等方面按打分表考核评分后,计算出 4 大搜索引擎的平均课题相关性分数(如表 3)。

表 3　平均课题相关性分数　　(单位:分)

搜索引擎	百度	谷歌	雅虎中国	MSN
平均相关分数	17.69	20.08	20.29	16.35

注:平均相关性分数满分(即每个链接都是 10 分的情况)为 32.3 分。

从表 3 可以看出,雅虎中国、谷歌的平均相关性分数相对较高,MSN 分数最低,百度的成绩也差强人意。

在总体评价上,可以看出,相关性分数满分为 32.3 分,而 4 大搜索引擎的分数均在 21 分之下,如按百分制考核计算不足 65 分。这说明搜索引擎对于中文文本的检索,在效果和质量上还存在一定的问题,主要表现为检索结果数量巨大,但是有用信息排在前 3 页的比例较小。

3.2.2　分类课题相关性

针对不同搜索引擎可能存在的不同的检索特性,我们把 30 个课题分为了自然科学、社会科学、综合 3 个类别,分别计算了四者的平均相关性分数(见表 4)。

表 4　分类课题平均相关性分数　　(单位:分)

分类	百度	谷歌	雅虎中国	MSN	平均
自然科学	15.60	17.94	18.37	15.09	16.15
社会科学	17.28	20.85	19.84	15.71	18.42
综合	19.61	20.86	22.12	17.87	20.11

注:平均相关性分数满分(即每个链接都是 10 分的情况)为 32.3 分。

从表 4 可以看出,在总体表现上,4 大搜索引擎对于综合课题的检索成绩均在 19.5 分之上,高于自然科学和社会科学类的平均水平,而社会科学和自然科学相比,则前者较为优势。每个搜索引擎的检索能力均按照综合、社会科学、自然科学的顺序逐级递减。雅虎中国在综合和自然科学类课题的检索中效果较为突出,而谷歌在社会科学和自然科学类课题检索中

也表现不俗。MSN 在这 3 类中,分数均为最低。百度的表现比较中庸。

同时我们还通过 SPSS 软件对每个搜索引擎 30 个课题的相关性分数、死链接等数据进行了系统自动聚类分析(见表 5),得到了两个较为明显的结论。

表 5　SPSS 对百度的聚类结果

类别	相关性均值(分)	死连接个数均值(个)	检索条数均值(条)
第 1 类	24.89	0.8	3 108 000
第 2 类	16.25	1.48	222 996

①百度的相关性分数较高、死链接个数较少。通过对这一类课题特性的聚类分析,我们看出其大多为当今社会的热门话题。

②雅虎中国有一小部分课题的搜索结果的相关性分数明显低于其他课题。观察得出,这一小部分课题多属于某领域的规则、程序、基本概念类。

3.2.3　检索条数与响应时间

表 6　搜索引擎检索条数统计表　　(单位:条)

搜索引擎	百度	谷歌	雅虎中国	MSN
检索条数均值	703829.70	423431.33	445099.23	52881.27

由表 6 可以看出,MSN 的检索条数明显低于其他 3 类搜索引擎。

从具体数据可以看出,检索条数的分布跟课题有很大的关系。

另外,4 个搜索引擎检索时的响应时间均小于 1 秒钟,对用户的使用没有造成影响。

3.2.4　搜索引擎死链接分析

表 7　死链接个数统计

项目	百度	谷歌	雅虎中国	MSN
平均死链接个数(个)	1.37	2.2	1.3	3.07
最大死链接个数(个)	5	8	4	6
最小死链接个数(个)	0	0	0	0
死链接个数总和(个)	41	66	39	92
死链率(%)	4.5	7.3	4.3	10.2

注:平均死链接个数指每个课题前 3 页(共 30 条记录)出现死链接的个数;最大、最小死链接个数指 30 个课题前 3 页(共 30 条记录)的死链接个数和的最大、最小值;死链接个数总和为每个搜索引擎 30 个课题 900 条记录中死链接个数总和。

如表 7 所示,雅虎中国和百度在死链接问题上处理较好,死链率均小于 5%。而 MSN 的检索死链率则高达 10%以上,平均每页都会有死链接。谷歌的死链接介于前两者之间。

死链接除了反映搜索引擎的内容更新质量外,还将受时间、地域、网络状况、课题等多种因素影响,因此我们所统计的数字仅表示本次试验环境和时段中的结果。

4　总结

从以上 4 大搜索引擎的比较研究可以看出:百度作为"全球最大的中文搜索引擎",虽然有中文信息覆盖面广、信息量大、更新快,能快速追踪热门话题等方面的优势,但是其对中文文本的检索质量并不突出;谷歌搜索的总体检索效果较好;MSN 检索效果相对最差,死链接问题严重,且个别课题存在检索无结果的情况。总体来说,与国外对英文搜索引擎研究的数据相比,中文搜索引擎的检索效果目前较为差强人意,究其原因是由中文歧义引起的。

对于中文检索来讲,如何更好地解决分词、歧义问题,怎样实现语境分析,如何实现智能检索将是推动中文搜索引擎前进的突破口。以雅虎、谷歌、百度为代表的中文搜索引擎如果能够各取所长,相互学习,相信一定能够更加完善,更加人性化,满足用户更深层的需求。

本次研究也存在一些不足:

(1)课题的样本量有待提高。本次研究共设计了 30 个课题,但对于数据庞大的 Internet 来说,这只是沧海一粟,因此样本在数量和类别上都不尽完善。

(2)关键词选取权威性还可商榷。本次研究均由评估人员进行了反复研究考量选取检索词,但难以保证其恰当与否,因此对总体检索水平可能有一定程度的影响。

(3)有些搜索引擎在检索结果排序中采用了竞价排名。这对部分检索结果会产生一定的影响,如何评价还有待考量。

参考文献

1　陈海龙. 搜索引擎的评价标准及方法研究. 情报杂志,2004(9)

2　贾红英. 搜索引擎检索功能的比较研究. 现代情报,2003(11)

(责任编辑:赵日珑)