基于MapReduce模式的NSTL用户热点检索词与中西文期刊热点关键词的对比分析*

□ 郝春云 / 中国科学技术信息研究所 北京 100038

摘要：文章简要介绍了MapReduce模式。基于2010年到2012年7月NSTL用户检索日志进行分析，采用MapReduce模式，针对用户的检索行为以及热点检索词进行分析，并与当年出版的文献的关键词进行比较，分析用户需求与文献提供的差异，旨在及时掌握用户的需求变化，为系统的功能完善、未来发展及文献采购提供参考依据。

关键词：NSTL，检索词，关键词
DOI：10.3772/j.issn.1673—2286.2012.11.005

1 背景

NSTL三期系统[1]自2010年4月正式运行以来，累积了大量的日志信息。针对这些日志进行分析，了解用户需求，能够及时掌握用户的需求变化，为系统的功能完善、文献采购及未来发展提供重要的参考依据。据笔者统计，从2010年4月1日到2012年6月30日，NSTL向用户提供检索服务7981580次。

由于数据量较大，在直接用数据库进行统计时，容易出现缓存溢出、结果集耗尽、IO写入错误等无法进行统计的情况，因此采用MapReduce模式进行了统计，先对数据进行切分，然后合并统计结果的方式来进行。

2 采用MapReduce算法进行统计分析

MapReduce是一种编程模型，用于大规模数据集的并行运算[2]。概念“Map（映射）”和“Reduce（化简）”，及它们的主要思想，都是从函数式编程语言借来的，还有从矢量编程语言借来的特性。当前的软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（化简）函数，用来保证所有映射的键值对中的每一个共享相同的键组[2]。

MapReduce通过把对数据集的大规模操作分发给网络上的每个节点实现可靠性；每个节点会周期性地把完成的工作和状态的更新报告回来。如果一个节点保持沉默超过一个预设的时间间隔，主节点记录下这个节点状态为死亡，并把分配给这个节点的数据发到别的节点。每个操作使用命名文件的不可分割操作以确保不会发生并行线程间的冲突；当文件被改名的时候，系统可能会把它们复制到任务名以外的另一个名字上去。化简操作工作方式很类似，但是由于化简操作的并行能力较差，主节点会尽量把化简操作调度在一个节点上，或者离需要操作的数据尽可能近的节点上[3]。

MapReduce能将大数据问题分解成多个子问题，将它们分配到成百上千个处理节点之上，然后将结果汇集到一个小数据集当中，从而更容易分析得出最后的结果。因此Google MapReduce模式被广泛应用于大数据的分析处理中。

本文中利用MapReduce模式的算法过程如下：

（1）检索词和关键词文件切分：MapReduce对检索日志文件按行进行自动切分，并将数据分发到每个Map任务，其中key值为ID，value值为countnumber，初始全部为1；

（2）Map任务的执行：接收key/value对，出现相同的key值

* 基金项目：国家高技术研究发展计划(863计划)子课题资源整合与知识组织技术研究（编号：2011AA01A206）。

时，value值相加，并舍弃之前的key/value值对，最后产生临时的key/value对集：

（3）临时结果的分组：对上面执行过程输出的临时结果进行分组，将相同的key值即ID号合并成同一组，并将其分发给空闲的Reduce：

（4）Reduce任务的执行：接收key/value对，对相同ID的value进行合并，得到当前该key的最高频次；

（5）MapReduce的迭代：每次Reduce后的结果又分发给下一轮的Map过程，通过多次迭代找到最终的无重复key值的key/value对集，算法结束。

3 分析结果

通过上面的算法，笔者分别对2010到2012上半年各年中用户的检索式，以及2010到2012年各年发表的中西文期刊的关键词进行了统计分析。

3.1 用户检索词分析

对用户提交的检索式经过提取检索词、切分和统计分析等操作，清洗了无意义的检索词，最终得到用户检索词及其出现频率的排行，其中2010年，共出现检索词457266条，2011年592218条，2012年上半年214797条。表1为各年中，出现最多的前30个检索词。

由上表可见，在这三年中“数字图书馆”都是最热检索词，同时，“气候变化”和“低碳经济”的热度也在逐年增加：2010年和2011年气候变化分别位于排行榜第557和第38，而在2012年，跃升至第2；低碳经济2010年和2011年分别位于排行榜第53和第11，而在2012年，跃升至第3。

3.2 中西文期刊的关键词分析

同时笔者对2010到2012年NSTL西文期刊和中文期刊中的关键词进行了统计及排行。统计方法为从文献数据库中抽取当年的关键词，并进行统计。

由上面排行榜可见，中文期刊

表2 2010到2012年NSTL中西文期刊不重复关键词数

	2010	2011	2012
中文	434256	391591	793107
西文	536105	383229	281839
合计	972371	776831	1076958

中，教育、基本医疗是最受关注的问题，西文期刊中，关注最多的是医学、生物以及发达国家和发展中国家的经济发展等问题。

3.3 用户检索词和期刊关键词排行榜前10000条的匹配分析

由于NSTL系统用户检索词和关键词数量众多，笔者分别截取了排行榜前100、1000和前10000词条进行对比分析。统计分析结果如下，TOP100排行榜中，在检索词和关键词榜单中同时出现的词条，2010年6个，2011年10个，2012年7个；TOP1000排行榜中，在检索词和关键词榜单中同时出现的词条，2010年118个，2011年257个，2012年239个；在TOP10000排行榜中，在检索词和关键词榜单中同时出现的词条，2010年有1269条，2011年为2934条，2012年2675条。

由于篇幅所限，本文只列出了同时出现在关键词和检索词的TOP100排行榜中的词汇，如表6所示。

由上面的分析可见，NSTL用户检索词与期刊关键词的匹配度在20%以下，差异还是比较大的。一方面，这与NSTL的用户群以及NSTL收藏的文献类型有关，另外，尖端科研人员的人数比起大多数用户来说还是少数，因此这些用户的检索词在分析时往往被一些无意义的大众流行词所淹没。

4 结语

相对来说，期刊的关键词比用户输入的检索词更规范。用户信息素质培养还有很长的路要走。

从系统设计的角度，如果把用户的检索词按照规范词表进行规范，命中率会增加，不容忽视的一个问题是要考虑对系统性能的影响。

从数字图书馆受到的关注度来看，NSTL的用户群中图书情报行业的用户较多，针对用户的分析也正印证了这一点。

真正有价值的检索词可能被大量无意义的随意检索词淹没。

用户使用中文检索词比较多，这直接导致了相匹配的检索词和关键词中中文占了绝大多数。

MapReduce模式在处理大数据的时候，可以分解为多个子问题，从而更容易得出分析结果，本文中的分析验证了其有效性。

表1 2010到2012年上半年用户检索词top30排行榜

序号	2010年		2011年		2012年	
	出现频次	检索词	出现频次	检索词	出现频次	检索词
1	9830	数字图书馆	14656	数字图书馆	10562	数字图书馆
2	6859	信息检索	5511	计算机	3577	气候变化
3	5428	中国图书馆学报	3870	治疗	4363	低碳经济
4	3685	计算机	3439	类黄酮积累	2288	计算机
5	3523	数字化电视节目	2583	信息检索	1656	电子商务
6	3023	博士	2501	电子商务	1474	治疗
7	2896	电子产品	2387	数字化电视节目	1127	糖尿病
8	2758	submarine	2318	管理	1108	雷达
9	2432	life	2166	高血压	1086	Eric McArthur
10	2305	治疗	2107	糖尿病	1064	校园
11	2033	络活喜	1805	低碳经济	1018	辣椒色素
12	3645	CAD	1763	应用	993	会计
13	1816	Laser Printers	1732	会计	980	博士
14	1752	电子商务	1690	建筑	894	数字化电视节目
15	1745	设计	1607	解放军护理杂志	882	陶瓷
16	1708	高血压	1519	steel	878	纳米技术
17	1692	steel	1506	设计	871	单片机
18	1609	管理	1495	二氧化碳	869	云计算
19	1606	阿司匹林	1432	护理	868	汽车
20	1595	汽车	1422	单片机	858	食品
21	1581	computer	1293	汽车	855	高血压
22	1486	建筑	1290	中国图书馆学报	810	管理
23	1482	satellite	1286	中国化学会	795	science
24	1481	糖尿病	1278	食品	739	护理
25	1471	超声波废水处理	1263	人力资源	734	steel
26	1379	laser	1228	物流	723	设计
27	1330	硕士	1227	博士	709	CAD
28	1293	CMOS	1197	复合材料	702	数据库
29	1223	ultra-heavy steel plate	1196	猪	698	机械
30	1215	bee	1173	通信	689	胰腺炎

表3 2010年中西文期刊关键词TOP30排行榜

西文		中文	
频次	关键词	频次	关键词
2856	animals	5891	中国
2828	eukaryotes	4291	小学
2004	plants	3725	作文
1530	internet resource	3676	护理
1498	spermatophyta	3112	课外阅读
1471	temperature	2996	语文教学
1457	angiosperms	2937	学生
1312	chordata	2901	中学
1299	vertebrates	2461	应用
1225	DNA	2390	文学作品
1166	adsorption	2365	阅读材料
1147	mammals	2215	治疗
1124	nanoparticles	2174	金融危机
1095	stability	2100	现代文学
1091	simulation	2044	孩子
1090	synthesis	2001	对策
1086	microstructure	1742	科学发展观
1072	humans	1680	文学
1045	developing countries	1609	世界
1036	dicotyledons	1601	老师
1007	structure	1499	企业
955	Asia	1460	美国
940	mechanical properties	1442	个人
935	developed countries	1374	大学生
934	kinetics	1366	生活
927	crystal structure	1353	阅读
913	optimization	1347	管理
892	genes	1311	故事
862	oxidation	1228	同学
862	growth	1143	教师

表4 2011年中西文期刊关键词TOP30排行榜

西文		中文	
频次	关键词	频次	关键词
4077	eukaryotes	4426	学生
3085	animals	3717	护理
2787	plants	3680	中国
2262	spermatophyta	2673	应用
2172	angiosperms	2559	对策
1951	vertebrates	2178	培养
1947	chordata	2001	语文教学
1580	mammals	1917	教学
1545	dicotyledons	1792	课堂教学
1527	developed countries	1790	教师
1260	developing countries	1701	创新
1126	oecd countries	1696	治疗
895	microstructure	1659	大学生
895	Asia	1533	教学方法
880	mechanical properties	1505	管理
835	bacteria	1481	小学
826	internet resource	1450	文学
825	nanoparticles	1372	问题
816	ungulates	1370	素质教育
792	america	1354	孩子
790	temperature	1280	企业
787	invertebrates	1263	高校
769	monocotyledons	1190	现状
767	adsorption	1111	设计
759	crystal structure	1077	文学作品
758	DNA	1071	中学
752	commonwealth of nations	1070	小学生
732	oxidation	1042	分析
731	north america	1017	生活
712	stability	988	学习

表5 2012年中西文期刊关键词TOP30排行榜

西文		中文	
频次	关键词	频次	关键词
1118	eukaryotes	11154	护理
872	animals	8691	对策
709	plants	8558	学生
607	apoptosis	8044	应用
601	microstructure	6472	中国
593	spermatophyta	5516	教学
582	vertebrates	5080	管理
577	angiosperms	5043	问题
567	chordata	4911	创新
566	inflammation	4872	培养
545	nanoparticles	4823	教学方法
524	mechanical properties	4595	治疗
516	oxidative stress	4247	语文教学
483	adsorption	4146	大学生
466	simulation	3902	课堂教学

西文		中文	
频次	关键词	频次	关键词
458	stability	3668	现状
440	mammals	3584	教师
435	optimization	3505	分析
412	climate change	3428	高校
403	cancer	3320	设计
400	temperature	3309	诊断
396	developed countries	3151	企业
392	depression	3086	教学改革
385	obesity	2918	措施
372	dicotyledons	2822	疗效
372	breast cancer	2816	影响因素
365	children	2784	小学生
362	kinetics	2762	语文学习
361	gene expression	2705	素质教育
355	epidemiology	2694	发展

表6 2010年到2012年同时出现在关键词和检索词的TOP100排行榜中的词汇

年 份	关键词和检索词TOP100排行榜中同时出现的词条
2010	高血压管理设计糖尿病应用治疗
2011	发展高血压管理护理技术设计糖尿病应用诊断治疗
2012	高血压管理护理设计糖尿病应用治疗

参考文献

[1] 关于我们[EB/OL]. [2012-08-22]. http://www.nstl.gov.cn/NSTL/nstl/facade/aboutus.jsp.
[2] MapReduce [EB/OL]. [2012-09-01]. http://baike.baidu.com/view/2902.htm.
[3] MapReduce [EB/OL]. [2012-09-01]. http://zh.wikipedia.org/wiki/MapReduce.

作者简介

郝春云（1973-），高级工程师，主要研究方向：信息系统管理、信息系统用户分析、数字图书馆技术等。E-mail: chyhao@istic.ac.cn

Comparative Analysis of NSTL Users Hot Search Terms and Journal Hot Keywords Based on the MapReduce Mode

Hao Chunyun / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Based on NSTL users search log of 2010 to 2012, using MapReduce algorithm, this article compared the users' hot search term and the journal hot keywords in the same year, designed to hold the users' changes in demand, for promoting the system function and providing the reference for the future documentation order.

Keywords: NSTL, Search term, Keyword

（收稿日期：2012-09-22）