

查询会话中的用户行为分析

王晓春, 李 生, 杨沐昀, 赵铁军

(哈尔滨工业大学 计算机科学与技术学院, 150001 哈尔滨, xcwang@mtlab.hit.edu.cn)

摘 要: 为了更好的理解搜索用户的检索行为,在查询会话基础上分析了用户修改查询的行为规律及用户、查询和点击三者之间的关系.从查询会话的角度,综合考虑了会话中修改的次序、用户、查询和点击等多方面因素,发现了用户修改查询的规律,证明了查询推荐的必要性.统计结果显示:用户修改查询的方式和修改次数有关,而对应的点击数量相对稳定;最终查询一般要比初始查询长,通常是部分替换了初始查询;用户使用搜索引擎的频率越高,修改查询的可能性就越大.

关键词: Sogou 搜索引擎;日志;查询会话分析;用户行为分析;查询

中图分类号: TP242

文献标志码: A

文章编号: 0367-6234(2011)05-0076-03

Research on user behavior based on session analysis

WANG Xiao-chun, LI Sheng, YANG Mu-yun, ZHAO Tie-jun

(School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China, xcwang@mtlab.hit.edu.cn)

Abstract: The session in search engine log denotes a sequential of queries from a user when he is searching for certain information during a period of time. Session analysis is a fundamental work for user behavior analysis. In contrast to current query log analysis focusing on the query modification and the relationship among user, query and clickthrough, this paper presents user behavior analysis based on session, which includes the times of reformulation and the relationship among user, query and click-through. Experimental results show that the way users reformulate queries relates closely to the times they reformulate, while the number of click-through is steady; the last query often gets longer than the first, what is more, the main variation lies in some changed characters; the more often the users use the search engine, the greater the likelihood that they reformulate queries.

Key words: sogou search engine;log;session analysis;user behavior analysis;query

网络搜索引擎逐渐成为人们搜索信息的重要途径,在查找信息的时候大约 70% 的人会求助于搜索引擎^[1].然而,用户提交的查询绝大多数是短查询,这些查询不可避免的存在歧义或者语意模糊现象,检索系统往往不能准确地返回用户所需的信息.为了进一步掌握用户的查询意图和搜索规律,从而将用户满意的结果返回给用户,很多研究者通过分析搜索引擎日志来分析用户行为.

国内外用户查询行为的研究方法主要包括查

询和点击的分布、高级搜索、相似查询、查询语言和词性、查询重复和查询主题等^[2-12].然而,这些都是对单个的查询进行分析.通常一个用户在一段时间里,带着一个固定的搜索意图而提交了一系列查询,该过程称为查询会话.在查询会话上进行用户行为分析,一方面可以更好的理解用户如何修改查询,另一方面可以更准确的把握用户的查询意图.

1 基于查询会话的用户行为分析

90 年代中期开始有研究者开始分析用户检索行为,90 年代末期进行日志分析时开始采用大规模的搜索引擎日志.

Silverstein 等对 AltaVista 用户检索行为进行

收稿日期: 2010-04-01.

基金项目: 国家自然科学基金重点资助项目(60736044);国家高技术研究发展计划资助项目(2006AA010108).

作者简介: 王晓春(1982—),女,博士研究生;
李 生(1943—),男,教授,博士生导师;
赵铁军(1962—),男,教授,博士生导师.

分析,发现 63.7% 的查询会话只有 1 个查询词,平均每个查询会话中用户提交 2.02 个查询词. Jasen 等^[13]发现 Excite 用户也有类似的行为表现. 文献[12]发现 Sogou 用户超过 80% 的查询会话的搜索次数 <5 次,但是约 10% 的查询会话搜索次数超过 10 次. 绝大多数的查询修改区分了增加 term、减少 term 和完全替换 term 这 3 种类型. Silverstein 认为 AltaVista 用户 12% 的查询修改是通过加词、删词或者其他操作符完成的,1/3 的查询词在查询会话中完全被修改了. 余慧佳等认为 Sogou 用户加减词的比例达到了 9%,完全改变查询词的数量占总数的 83.27%. 这些研究并没有充分体现出查询会话中的重要信息. 本文在查询会话的基础上,从查询修改、查询差异以及不同频率用户的行为特点这 3 个角度,对用户行为的规律进行总结,一定程度上弥补了现有查询会话分析中用户行为分析中的不足.

2 实验方案以及实验数据

根据中文 Sogou 搜索引擎一个月的日志进行了查询会话分析,为了研究查询会话中紧邻有序查询对的变化,定义了 8 种查询修改,如表 1 所示. 首先把查询对按照是否含有重复的字符,分为完全替换、首尾替换和部分替换. 其中,完全替换是指前后 2 个查询完全没有重复的字符;首尾替换是指相邻的 2 个查询字符串其中的一个是另一个的子串. 与此同时,本文以前一个查询长度为基准,相对上一个查询分别对长度变长、变短或者不变进行统计.

表 1 基于紧邻有序查询对的 8 种查询修改

长度	完全替换(D)			首尾替换(S)		其他替换(P)		
	增加 (M)	减少 (N)	不变 (P)	增加 (A)	减少 (B)	增加 (X)	减少 (Y)	不变 (Z)

针对最初的查询和最终的最佳查询之间的差异,本文定义了 9 种类型,前 8 种类型与紧邻有序

查询对的变化种类相同,本文又补充了查询完全不变的情况(用字母 Q 表示). 此外,统计了不同修改次数下的对于返回结果的平均点击数.

实验数据取自于大规模 Sogou 中文搜索引擎 2007 年 3 月份日志. 本文按照 Daniel^[13]提出的 session 划分方法,根据用户提交查询的时间间隔以及相邻查询之间的相似度,最后划分出 17 117 304 个查询会话.

3 查询会话中的用户行为分析

用户行为分析指检索用户使用搜索引擎系统来满足搜索需求的整个行为的分析. 本文从查询会话的角度,分析不同搜索频率的用户的查询和点击行为.

3.1 查询修改分析

将查询会话内总共 3 129 660 个相邻有序查询对,按照提出的 8 种查询修改类型进行统计,结果如表 2 所示.

表 2 紧邻有序查询对的 8 种查询修改 %

长度	D			S		P		
	M	N	P	A	B	X	Y	Z
百分比	35 *	51 *	14 *	21	10	21	24	23

注: * 数字为相对于完全替换的,其他数字为相对于总量的.

实验数据表明:完全替换占总数的 0.3%,在 D,S 和 P 这 3 种变化中比重微小;部分替换方式占了大约 2/3;首尾替换大约占了总数的 1/3. 在完全替换和部分替换中,查询长度减少的情况最多. 在首尾替换中,长度变长的情况要比缩短的数量多. 进一步对首尾替换的查询对进行研究,发现无论是长度增加还是减少,几乎全部(99%)的用户同时在首尾两侧对查询进行修改.

为了研究进行不同查询修改次数的查询会话的特点,定义修改次数*i*表示在 1 个查询会话中用户恰恰修改了*i*次. 字母串表示用户的多个查询修改的有序组合,如表 3 所示.

表 3 不同修改次数的查询会话中的修改

序号	1		2		3		4		5		6	
	修改	频率	修改	频率	修改	频率	修改	频率	修改	频率	修改	频率
1	A	378 483	XY	39 166	YXY	7 115	ZZZZ	1 898	ZZZZZ	622	ZZZZZZ	256
2	Y	301 320	YX	38 194	ZZZ	6 986	YXYX	1 467	XYXYX	385	YXYXYX	99
3	Z	300 393	ZZ	35 322	XYX	6 151	YXXY	1 454	YXXYX	377	YXXYXY	93
4	X	253 280	AB	27 014	ABA	4 856	ABAB	741	ABABA	174	YZZZZZ	46
5	B	134 006	AY	26 213	AXY	3 862	AYXY	671	YXXXY	150	AZZZZZ	44
6	N	203	AZ	22 288	AYX	3 659	YXXY	654	AZZZZ	144	YXYXXY	40
7	M	187	AX	21 614	XYZ	3 024	AZZZ	608	AXYXY	143	ABABAB	39
8	P	153	BA	20 079	AZZ	2 997	YXYZ	573	XYZXY	142	YXYXYZ	39
9			YZ	17 676	ZXY	2 994	AXYX	563	ZZZXY	136	YXXYXY	37
10			XZ	15 794	XXY	2 875	XYXZ	509	XXYXY	132	YXZZZZ	34

表 3 充分说明查询修改动作和查询修改的次数有关. 如果用户进行 2~3 次查询修改, 那么最常见的修改就是改变查询长度. 倘若用户修改 3 次以上, 那么最有可能保持查询长度的前提下, 仅仅进行部分替换. 仅仅修改 1 次的行为中, 最广泛的修改动作就是在原始查询的两边同时进行替换, 从而构造更长的查询. 现有的查询推荐主要关注用户查询和返回结果的语意相似度, 倘若将用户的修改行为规律补充到查询推荐当中, 可以更加贴近用户真实的查询行为, 缩短用户修改查询的次数, 让用户更快的找到所需的信息.

3.2 查询差异分析

用户所构造第 1 个查询的时候没有参考任何信息, 接着用户通过返回结果来修改上一个查询, 直到满足查询需求. 统计结果显示, 只有极少数 (大约 6%) 的查询会话中最终查询和初始查询相同, 而查询差异主要集中在用户进行了首尾替换同时长度增加 (大约 24%), 大约 22% 的用户进行了长度减少的部分替换, 用户查询差异分析证明了查询推荐的必要性.

3.3 查询会话内的用户、查询和点击

3.3.1 用户和查询

从查询和用户来看, Sogou 用户搜索频率主要集中在 80 次以内, 其中: 使用 10 次以内 (包括 10 次) 搜索引擎的用户占全部用户的 97.97%; 仅使用了 1 次的用户占不超过 10 次的用户的 83.30%. 可见, Sogou 用户平均每天使用 2.67 次搜索引擎, 绝大多数用户每次不修改查询.

此外, 按照用户使用搜索引擎的习惯区分为每次都修改查询、每次都不修改查询和不固定等 3 种情况. 实验数据表明, 越是频繁使用搜索引擎的用户, 越是不断修改查询的频率越高. 另一方面, 从一贯修改查询的用户分布上来看, 这部分用户的搜索频率集中在 10 次以内. 通过对修改习惯不固定的用户进行统计, 发现不修改的要比修改的数量多, 最常见的修改动作是部分替换. 可见, 查询推荐适合那些频繁使用搜索引擎的一类用户.

3.3.2 查询和点击

为了研究在一个会话中进行不同修改次数的查询对应用户点击数量分布, 本文对含有多个查询的会话进行分析. 结果表明, 各个修改次序上的查询平均对应 2~3 个点击. 可见, 随着修改次数的增加, 用户查看结果的兴趣和耐心并没有衰减, 每次平均点击 2 个左右的结果.

3.3.3 用户、查询和点击

实验中对查询会话中的最后一个查询和点击

进行统计, 分别取前 100 个常见的查询、用户, 点击及其之间的任意组合. 实验结果表明重复查询的数量最多, 其次是查询和点击完全相同的这一类, 最后是相同的用户. 可见, 不同的用户提交相同的查询, 不同的用户提交相同查询并且点击结果也相同的情况很多, 但是同一个用户提出不变的历史查询或者点击同一个结果的情况就很少.

经过去重处理, 发现查询的数量最少, 其次是点击, 最后是相同的查询和点击. 平均一个查询由 2.27 个不同的用户提交, 平均一个查询对应 1.47 个不同的点击. 可见, 一个月内用户关注的查询和点击的数量是相对稳定和集中的, 因而重复的情况时有发生. 用户的查询意图的重复性分析, 再次证明了查询推荐的必要性.

4 结 论

1) 用户修改查询的时候, 常常在原有查询的基础上进行部分替换, 修改 3 次以上时查询的长度保持不变.

2) 用户提交的最初查询和最终查询的主要差异在于长度增加, 往往对原来查询的首尾同时进行修改.

3) 用户的查询和点击经常发生重复, 用户并不会随着修改次数的增多而对点击结果失去耐心, 经常使用搜索引擎的用户他们更愿意多次修改查询.

参考文献:

- [1] SPINK A, BATEMAN J, JANSEN B J. Searching the web: A survey of excite users [J]. Internet Research: Electronic Networking Applications & Policy, 1999, 9(2): 117-128.
- [2] SILVERSTEIN C, MARAIS H, HENZINGER M, et al. Analysis of a very large web search engine query log [C]//ACM SIGIR Forum. New York, NY: ACM, 1999: 6-12.
- [3] JANSEN B J, SPINK A, SARACEVOC C T. Real life, real users, and real needs: A study and analysis of user queries on the web [C]//Information Processing and Management: An International Journal. Tarrytown, NY: Pergamon Press, 2000: 207-227.
- [4] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析 [J]. 中文信息学报, 2007, 21(1): 109-114.
- [5] HOLSCHER C, STRUBE G. Web search behavior of Internet experts and newbies [C]//Proceedings of the 9th International World Wide Web Conference on Computer Networks: The International Journal of Computer and Telecommunications Netowrking. New York, NY: Elsevier North-Holland, 2000: 337-346.

(下转第 105 页)