一种基于用户路径模型的搜索引擎检索性能度量方法

袁玉宇[1],罗学超[2]
(1.北京大学微处理器研究开发中心计算机系统结构研究所,北京 100871;
2.搜狐(北京)公司研发中心,北京 100084)

摘 要: 搜索引擎是信息时代人们获取所需信息的重要手段,搜索引擎性能度量方法也成为了一个重要的研究课题.本文分析了国内外搜索引擎检索性能度量的研究状况,提出了一种基于用户路径模型的性能度量方法.从用户搜索行为的角度研究了用户行为模型,通过抽象用户行为模型到路径模型的映射,给出基于用户路径模型的搜索引擎检索性能的评估步骤,针对度量数据给出了成功搜索度量方法,并通过度量平台的实现验证了该度量方法的实践意义.

关键词: 搜索引擎; 检索性能度量; 用户路径; 成功搜索
中图分类号: TP311 **文献标识码**: A **文章编号**: 0372-2112 (2008) 05-0969-05

A Measurement Method of Search Engine Retrieve Performance Based User Path Model

YUAN Yu-yu[1],LUO Xue-chao[2]
(1. *Department of Computer Science, Peking University, Beijing* 100871, *China*;
2. *Department of Research Center, SOHU, Beijing* 100084, *China*)

Abstract: Search Engine is an important means, which people looked for their requirement information. Then the measurement methods also turn to an important research course. This article deeply analyses the development status of search engine retrieve performance measurement in the native and abroad, then a measurement method based user path model was established. The user action model was researched form user retrieve action, and then the user path map rule is put forward. Furthermore, the evaluation process of search engine retrieve performance was described, and defined measurement method of successful search. Last, through the accomplishment of evaluation platform, the practice signification was proofed.

Key words: search engine; retrieve performance measurement; user path; successful search

1 引言

随着互联网产业的发展,搜索引擎在信息爆炸的时代发挥着巨大的作用,人们越来越习惯通过搜索引擎查找所需要的信息,人们对搜索引擎的期望值也就越来越高.因此搜索引擎检索性能度量也就随之成为搜索引擎研究的一个重要课题之一.对于搜索引擎这类为提供信息检索而设计的系统中,除了对时间和空间方面的度量以外,还有一些重要属性需要度量.事实上,由于用户查询请求本质上具有模糊性,检出的文献也往往不一定就是精确的答案,因此还需要将检出文献按照它们与查询之间的相关性进行排序分析.这种相关性排序是数据检索系统是无价值的,但对搜索引擎来说却具有重要意义.所以,搜索引擎需要对检索结果集的准确度进行度量,这种度量称为检索性能度量.

2 国内外研究现状分析

搜索引擎检索性能优劣,最终都要通过评价指标来体现,因此建立合理的、科学的评价指标体系至关重要.国外与国内众多学者都在致力于这方面的研究,侧重点各有千秋.国外学者重点了三个方面的研究,第一,如何将信息检索的评价指标转换为搜索引擎的评价指标;第二,针对不同的研究方法提出不同的评价指标体系;第三,优化现有提出的评价指标.而国内学者更加侧重于研究指标的数学模型建立与解释.

从研究思路来看,对搜索引擎的度量研究基本可以分为三大类:以系统为中心的(System-Centred)搜索引擎度量、以任务为中心的(Task-Centred)搜索引擎度量和以用户为中心的(User-Centred)搜索引擎度量.

以系统为中心的搜索引擎检索性能度量,是一种基

收稿日期:2006-12-28;修回日期:2007-12-22
基金项目:国家自然科学基金(No.60642008)

于实验环境来收集搜索引擎检索性能数据的度量方法.该度量方法的致命弱点就是基于了两个假设,一是用户的认知行为仅靠查询语句就能解决,二是结果文档的相关性是二元判断.以任务为中心的搜索引擎检索性能度量[1~4],更多地考虑了用户与搜索引擎进行交互的主要目的是为完成某项任务,因此检索的成功性取决于检索结果的"任务相关性",而且它从根本上仍然是基于实验集合,而不是基于现实的网络检索环境.以用户为中心的搜索引擎检索性能度量[5],通过识别搜索用户行为与用户满意度来确定系统的效果.该度量指标更加复杂,目前还没有形成统一的指标体系.虽然该方法已经不再是在实验环境中进行,但是该方法成本高、周期长、其结果稳定性差.因此,为了克服以上问题本文提出了基于用户路径模型的搜索引擎检索性能度量方法.

3 用户路径模型的建立

3.1 用户行为模型

用户行为模式可以简单地理解为用户在使用浏览器进行关键词搜索,以找到个人所需结果的过程中,用户通过浏览器与搜索引擎进行交互的步骤特征.Tingshao Zhu[6]提出了一个非常简单的搜索用户行为模型,Yvonne Rogers[7]在研究用户行为模式中引入认知论,提出了一个可以根据用户浏览搜索结果页提取用户行为习惯的用户行为模型.本文要考虑的用户搜索行为是在多搜索引擎之间进行比较的,因此重新定义了用户行为模型如图1.该模型概括了用户与搜索引擎之间的交互行为,模型的左边描述了用户为达到用户目标需求所必须要完成的行为,右边描述了用户为达到目标需求可能要完成的行为.

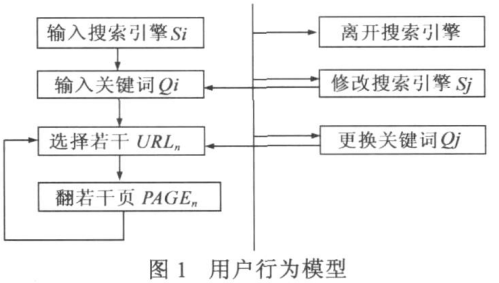

图1 用户行为模型

3.2 用户的路径模型

根据路径分析法,行为模型中所定义的行为导致不同结果所经过的路径,称之为用户路径模型.首先,由用户行为模型直接导出原始用户路径模型;其次,对原始用户路径模型的每条路径的含义进行完备性分析;最后,得出自顶向下和自底向上两类用户路径模型.

自顶向下路径模型包含的路径集如下:

*P*1:输入 *Si* \ 输入 *Qi* \ 离开搜索引擎;其目的表示无结果搜索.

*P*2:输入 *Si* \ 输入 *Qi* \ { …*URLi* …*URLj* …*URLk* :选择的 *url* 超过一半是自顶向下} \ 离开搜索引擎;其目的表示首页结果搜索.

*P*3:输入 *Si* \ 输入 *Qi* \ { …*URLi* …*URLj* …*URLk*:选择的 *url* 超过一半是自顶向下} \ 翻页 \ 离开搜索引擎;其目的表示翻页结果搜索.

*P*4:输入 *Si* \ 修改 *Sj*;其目的表示更换引擎后进一步搜索.

*P*5:输入 *Qi* \ 修改 *Qj*;其目的表示更换关键词后进一步搜索.

自底向上路径模型包含的路径集如下:

*P*1:输入 *Si* \ 输入 *Qi* \ 离开搜索引擎;其目的表示无结果搜索.

*P*2:输入 *Si* \ 输入 *Qi* \ { …*URLi* …*URLj* …*URLk* :选择的 *url* 超过一半是自底向上} \ 离开搜索引擎;其目的表示首页结果搜索.

*P*3:输入 *Si* \ 输入 *Qi* \ { …*URLi* …*URLj* …*URLk*:选择的 *url* 超过一半是自底向上} \ 翻页 \ 离开搜索引擎;其目的表示翻页结果搜索.

*P*4:输入 *Si* \ 修改 *Sj*;其目的表示更换引擎后进一步搜索.

*P*5:输入 *Qi* \ 修改 *Qj*;其目的表示更换关键词后进一步搜索.

4 搜索引擎检索性能度量步骤

4.1 获取用户行为原始数据及表示

用户在使用多搜索引擎进行搜索时产生的用户行为原始数据表示如下:

(1)用户 *ID* 或 *IP*,用于区分不同用户即 *UID* 或 *UIP*.

(2)用户原本想访问的搜索引擎 *Se*1,数据表示为:*UID*-*TIME*-*Se*1

(3)用户输入的关键词 *Key*1,数据表示为:*UID*-*TIME*-*Se*1-*Key*1

(4)用户在 *ORes*[1…*m*,4]和 *NOrd*[1…*m*,2]构成的新结果页中所做出的选择,包括所选择文档对应的链接 *ChURL*、所选择文档对应的标题 *ChTi*、所选择文档对应的摘要 *ChSu*、所选择文档在 *NOrd*[1…*m*,2]中对应的编号 *ChNOrd*、所选择文档在 *Se*[1…*m*]中对应的原搜索引擎 *ChSe*、该请求对应的页号 *ChPa* 和选择所对应的时间 *ClickURLTime*.数据表示为:*UID* *ClickURLTime*-*Se*1-*Key*1-*ChPa*-*ChSe*-*ChURL*-*ChTi*-*ChSu*-*ChNOrd*

(5)用户请求翻页时的选择,包括原搜索引擎 *Se*1、用户输入关键词 *Key*1、新请求页号 *NPa* 和请求该页的时间 *ResultClickPageTime*.数据表示为:*UID*-*ResultClickPageTime*-*Se*1-*Key*1-*Npa*

4.2 将用户行为原始数转换为用户路径模型

用户行为原始数据尽管从发生的次数来说很多,但

最多有如下四种可能性的分类:

A 分类: *UID-TIME-Se*1

B 分类: *UID-TIME-Se*1-*Key*1

C 分类: *UID-ClickURLTime-Se*1-*Key*1-*ChPa-ChSe-ChURL-ChTi-ChSu-ChNOrd*

D 分类: *UID-ResultClickPageTime-Se*1-*Key*1-*Npa*

用户路径模型研究的是某一用户在一段时间的一连串点击行为,当将一条一条独立存在的原始行为数据按照用户路径模型描述的规则去标识时,这些原始数据就被表示成了该用户在某一时间段的用户路径模型数据.

4.3　正则表达式标识用户路径模型

应用正则表达式将四类用户行为原始数据标识为用户路径模的映射(表 1). 在描述中规定一些符号的含义:

位于"/ "定界符之间的部分就是将要在目标对象中进行匹配的模式;

" + "元字符规定其前导字符必须在目标对象中连续出现一次或多次;

" * "元字符规定其前导字符必须在目标对象中出现零次或连续多次;

"() "符号包含的内容必须同时出现在目标对象中.

表 1　正则表达式标识用户路径模型

用户路径模型分类	原始用户行为数据检测串	原始用户行为数据标识串
*P*1(无结果搜索)	*Bnull* 或者 *BA*	*B*
*P*2(首页结果搜索)	*B/ C + / B* 或者 *B/ C + / A* 或者 *B/ C + / null*	*B/ C + /*
*P*3(翻页结果搜索)	*B/ (C * D +) + / B* 或者 *B/ (C * D +) + / A* 或者 *B/ (C * D +) + / null*	*B/ (C * D +) +*
*P*4(更换搜索引进一步搜索)	*AB/ (C * D +) + / AB* 或者 *ABAB*	后 1 个 *AB*
*P*5(更换关键词进一步搜索)	*B/ (C * D +) + / B* 或者 *BB*	后 1 个 *B*

5　度量结果分析与评价

用户的每一次操作都会记录为用户行为原始数据,这些用户行为原始数据首先标记为 *A*、*B*、*C* 和 *D* 类. 在某一时间段内对已分好类并按时间戳升序排好序的用户原始行为数据从前往后扫描,按照用户路径模型正则表达式标识串被标识为 *P*1、*P*2、*P*3、*P*4 和 *P*5. 判断用户搜索成功与否不仅需要从前往后扫描已路径化的用户数据,而且需要考察该用户某一路径化数据的前 *i* 个或后 *j* 个用户路径模型类型,甚至判断该用户搜索成功与否还与其他用户的搜索行为相关. 具体分析如下:

5.1　搜索成功判定

用户的一个完整搜索中包含 *P*1、*P*2、*P*3、*P*4 和 *P*5 的可能性都存在,如何判断用户找到自己满意的结果了呢? 一个完整搜索中若仅包含 *P*1、*P*4、*P*5 三种路径属性的任何一种组合,都是无效搜索,而若包含了 *P*2 和 *P*3,搜索成功否就变得复杂了.

与路径相关的判断,如下(表 2)中列举的常见路径组合都不能仅从构成上完全做出已成功搜索的结论,却能得到最后无论搜索成功与否的搜索转换成本和搜索成本.

表 2　路径相关的搜索成功判断

典型完整搜索路径合	搜索成功判断	备注
*P*1、*P*4、*P*5 的任意组合	失败	例如, *P*1 *P*5 *P*4
*P*2 +	需进一步确认	例如, *P*2 …*P*2
*P*3 +	需进一步确认	例如, *P*3 …*P*3
*P*2 + *P*5 *P*2 +	第 1 个 *P*2 + 至少还不能完全满足用户需求,第 2 个 *P*2 + 需进一步确认	例如, *P*2 …*P*2 *P*5 *P*2 …*P*2
*P*2 + *P*4 *P*2 +	同上	例如, *P*2 …*P*2 *P*4 *P*2. . *P*2
*P*3 + *P*5 *P*3 +	第 1 个 *P*3 + 至少还不能完全满足用户需求,第 2 个 *P*3 + 需进一步确认	例如, *P*3 …*P*3 *P*5 *P*3 …*P*3
*P*3 + *P*4 *P*3 +	同上	例如, *P*3 …*P*3 *P*4 *P*3 …*P*3

用户一次完整搜索的搜索转换成本 *TSeSwCost* 按如下公式进行计算: *TSeSwCost*(*Pi*) = 已完整搜索标识的路径化用户数据中 *Pi* 的个数,其中 *i* = 4,5. 用户一次完整搜索的搜索成本 *TSeCost* 按如下公式进行计算: *TSeSwCost*(*Pi*) % = 已完整搜索标识的路径化用户数据中 *Pi* 的个数/已完整搜索标识的路径化用户数据中 *P* 的个数,其中 *i* = 1,2 - 5.

与相似搜索相关,相似搜索分类组 *ESiSe*[1 …*K*][2]生成后,一次用户完整搜索,若包含 *P*2 或者 *P*3,其成功命中率计算如下;若不包含,则成功命中率直接为 0. 判断该用户这一次完整搜索 *Q* 的已二元化分词 *Key*1 集、*ChURL* 集在相似搜索分类组 *ESiSe*[1 …*K*][2]的哪个分类组号 *Kj* 中;若这个分类组号 *Kj* 在 *EsiSeMap*[*K*]所对应的原样本 *EsiSeMap*[*Kj*]仅包含一个样本,即就是该用户这一次完整搜索 *Q* 的数据,则该用户的这一次完整搜索 *Q* 的成功命中率为 - 1,即可疑成功命中;否则取出 *ESiSe*[1 …*Kj*][2]数据,则 *Q* 该用户的这一次完整搜索 *Q* 的成功命中率 *SeSucc* % 计算如下: (*Q* 中 *ChURL* 集的数量/ *ESiSe*[1 …*Kj*][2]中 *ChURL* 集的数量) * (*EsiSeMap*[*Kj*]所对应的原样本数/ *max*(*EsiSeMap*[*K*]所对应的原样本数))

这样,一次用户完整搜索又多了一个属性,即成功命中率. 当一次用户完整搜索 *Q* 中包含了 *P*2 或者 *P*3 路径时,类似 *Q* 的相似搜索样本数越多, *URL* 公共集越多,成功命中的可能性越大,即满足用户的搜索需求越

大.

5.2 用户搜索成功概率评价

含义:单个完整搜索平均成功搜索概率

公式:

$$USeSucc1\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SeASucc\ \%/\ nj))/\ n$$

含义:单个完整搜索平均可疑成功搜索概率

公式:

$$USeSucc2\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SeASucc\ \%/\ nj))/\ n$$

(仅当 $SeSucc\ \% = -1$ 时才参与计算)

含义:单个完整搜索平均非可疑成功搜索概率

公式:

$$USeSucc3\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SeASucc\ \%/\ nj))/\ n$$

(仅当 $SeSucc\ \% <> -1$ 时才参与计算)

5.3 引擎搜索成功贡献度评测

含义:单个完整搜索中每个评测搜索引擎为成功搜索所贡献的价值

进一步定义:

用户一次完整搜索中搜索引擎成功命中贡献度 $SSejjSucc\ \%$ 按如下公式计算: $SSejjSucc\ \% =$ (已完整搜索标识的路径化用户数据中路径为 $P2$ 或 $P3$,且此数据的 $ChSe$ 为 $Sejj$ 的数据条数/已完整搜索标识的路径化用户数据中路径为 $P2$ 或 $P3$ 的数据条数) * 该完整搜索的 $SeSucc\ \%$;

$$ESejjSucc1\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SSejjSucc\ \%/\ nj))/\ n$$

含义:单个完整搜索中每个评测搜索引擎为可疑成功所贡献的价值

公式:

$$ESejjSucc2\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SSejjSucc\ \%/\ nj))/\ n$$

(仅当 $SSejjSucc\ \%$ 中计算 $SeSucc\ \% = -1$ 时才参与 $ESejjSucc2\ \%$ 计算)

含义:单个完整搜索中每个评测搜索引擎为非可疑成功所贡献的价值

公式:

$$ESejjSucc3\ \% = (\sum_{1}^{m}\sum_{1}^{nj} SSejjSucc\ \%/\ nj))/\ n$$

(仅当 $SSejjSucc\ \%$ 计算中 $SeSucc\ \% <> -1$ 时才参与 $ESejjSucc3\ \%$ 计算)

6 搜狐搜索引擎度量平台介绍及运行效果评价

搜狐搜索引擎度量评价平台主要包括数据收集和数据分析两个过程.数据收集过程在依照体系中测试参考集所规定的收集策略和收集流程.当用户进行查询实例搜索时,即刻完成两两比较搜索引擎结果页面抓取与结果混合后呈现给用户,若用户在混合后的结果页面做出选择,记录下来测试参考集所规定的搜索数据和评价方法中所规定的原始行为数据.数据分析过程对数据收集过程所记录的原始用户行为数据进行用户行为建模,转换为用户模型数据,接着根据用户路径映射规则转换成路径化用户行为数据,按照完整搜索判断规则和成功搜索命中率计算方法对已路径化的用户行为数据进行数据成功命中率化,最后扫描所有已经成功命中率化的用户行为数据按照平台中所定义的评价指标计算出评价指标.

评价平台分三层:限流转发层、混合器加工与显示层,以及用户日志分析与指标计算层.限流转发层 RequstLimiterTransfer 是个 ApacheModule,负责混合器搜索引擎请求流量限制和转发.混合器加工与显示层包含 RequestResultSpiderService、ResultMixer、MixSearchEngine 和 UserLogPingback.用户日志分析与指标计算包含 UserLogReceiverService、UserLogAnalyser、LogClustor 和 RPEComputing.

应用该度量方法建立起的搜狐搜索引擎度量平台,完成了搜狐搜索引擎与 baidu、google、yisou 的检索性能两两比较,每天可采集行为数据按发生次数 1000 条,连续 2 月下来分八周对接近 60000 条的用户行为数据进行分析,发现该度量方法对检索性能的度量具有以下特点:首先,现在收集器对于用户的任何输入都可以实时采样,这样查询实例在时效性、输入习惯、数量规模等因素都能更真实地反映了两个参与度量的搜索引擎检索性能.其次,现在平台的收集器中的抓取器实时根据查询实例抓取两两比较的搜索引擎结果页面,这些结果页面一方面完全依赖原搜索引擎的文档集,同时实时反映原搜索引擎文档的变化.而且,该平台中的分析器可以定期执行,也可以实时执行,完全取决于使用方的需求,它都能对收集器中用户行为原始数据接收器接收的数据某个时间段的数据进行分析.

参考文献:

[1] Ryen White. Query-biased web page summarisation: A tasori-ented evaluation[J]. Information Processing and Management, 2001,25(5):357 - 379.

[2] Jane Reid. A task-oriented non-interactive evaluation methodology for information retrieval systems[J]. Information Retrieval, 2000,2(1):115 - 129.

[3] Louise T Su. Value of search results as a whole as the best single measure of information retrieval performance[J]. Information Processing and Management,1998,34(5):557 - 579.

[4] Howard Greisdorf, Amanda Spink. Median measure: An ap-

proach to IR systems Evaluation[J]. Information Processing and Management, 2001, 37: 843 - 857.

[5] Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems[J]. Information Research, 2003, 8(3): 152.

[6] Tingshao Zhu, Russ Greiner, Gerald Haeubl, Kevin Jewell, Bob Price. Using learned browsing behavior models to recommend relevant web pages[A]. Nineteenth International Joint Conference on Artificial Intelligence[C]. Edinburgh, UK, 2005. 1589 - 1591.

[7] Yvonne Rogers. Cognitive strategies in web searching[A]. 5th Conference on Human Factors & the Web[C]. USA: NIST, June 3, 1999.

作者简介:

袁玉宇　女, 1971 年 10 月生于内蒙古, 博士, 北京大学微处理器研究开发中心计算机体系结构研究所博士后; 现任国家软件标准化推广中心副主任; ISO/IEC JTC1 SC7 国际软件工程标准委员会 WG6 工作组成员; 中国电子学会数据库专家委委员; 中科院研究生院软件学院兼职教授与硕士导师. 目前主要研究方向为系统测试、软件工程、软件认知学、软件质量、软件性能工程.

E-mail: yuanyy@cesi.ac.cn

罗学超　女, 搜狐(北京)公司研发中心高级测试经理.