

网购平台搜索功能的质量评估方法研究

陈浩¹ 陶传奇¹ 文万志²

(南京理工大学计算机科学与工程学院 南京 210094)¹ (南通大学计算机科学与技术学院 南通 226019)²

摘要 随着大数据的迅速发展,大数据应用层出不穷,诸如网购零售平台、人脸识别系统、智能决策系统、自助客服、看病导医系统等典型的大数据应用使得人们的生活越发便捷。搜索系统是人们最常使用的大数据应用之一。然而,搜索系统在不同平台上的功能各有侧重,其标准尚且不完善,搜索的质量参差不齐,无法得到保障。与普通的文本搜索引擎相比,网购平台的搜索引擎增加了分类检索、筛选等特色功能,其质量的评价与保障更为复杂。通过对网络零售平台的搜索功能进行研究,针对网购平台搜索功能的质量评价提出了质量参考因素,针对质量因素提出了若干评价指标以及相应的实现算法,并通过实验来论证了质量指标的有效性。

关键词 搜索引擎,质量因素,质量指标,质量评估,网购搜索

中图分类号 TP311 文献标识码 A DOI 10.11896/j.issn.1002-137X.2017.11.020

Research on Quality Evaluation Approaches for Search Function of Online Shopping Platforms

CHEN Hao¹ TAO Chuan-qi¹ WEN Wan-zhi²

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)¹

(School of Computer Science and Technology, Nantong University, Nantong 226019, China)²

Abstract With the rapid development of big data, various big data applications are now in service in different fields. Typical big data applications, such as online shopping and retailing platform, face recognition system, intelligent decision system, self-help service system, medical treatment system make daily life more convenient. Search system is one of the most used big data applications. However, search system varies in different platforms, and there are few standards for it. The quality of search system is hard to assure and validate. Search engine for online shopping systems combines text search and classification-based retrieval comparing to common text search engines. It is harder to validate and evaluate quality of it. Through studying on search system of online shopping platform, some quality factors and relational implementation algorithm were provided in this paper to validate and evaluate shopping system search engines. Experiments were also carried out to assure the correctness of the quality index.

Keywords Search engines, Quality factors, Quality index, Quality evaluation, Online shopping search

1 引言

网购零售平台是指一类网络平台,它为用户提供如下服务:通过互联网搜索商品信息,通过电子订单发出购物请求,在商品提供商及购物者达成一致协议后,通过约定的方式付款,以快递或当面交易等方式进行交易。

通常,购物平台会将一小部分可能被广大用户购买的商品显示在购物网站的首页,因此这些商品得到了很大程度的曝光。然而剩下的很大一部分商品并没有显示在首页,需要用户通过关键字搜索,或者分类查看商品等行为才能获取到相关的商品信息。商品搜索功能是网络购物平台提供的最主要的功能之一,用户可以根据自己的需求设置关键字进行搜

索,同时可以根据一些如标签、筛选等类型的操作对搜索的结果进行分类查看,如根据类别、价格、发货地址等因素对商品进行筛选显示,根据商家信用或销量等对商品进行排序,从而满足自己寻找特定商品的需求。

因此,网购平台作为典型的大数据系统,其商品搜索功能的好坏对网购平台的质量有很大的影响。商品搜索功能的好坏不仅会影响用户体验、网购平台的形象,而且会在无形中改变用户在线购买商品的行为。一个存在质量问题的网购平台搜索引擎可能导致用户无法获取到自己所需要的商品,这也影响着网购平台的经济效益。

目前,针对网购平台搜索功能,尚没有人提出评估其质量的因素及具体的质量指标。对于网购平台质量的评估,目前

到稿日期:2016-10-11 返修日期:2016-12-12 本文受国家自然科学基金(61402229, 61502233),江苏省博士后基金(1401043B),南京大学软件新技术国家重点实验室开放式基金(KFKT2015B10),江苏省高校自然科学基金项目(15KJB520030)资助。

陈浩(1993—),男,硕士生,主要研究方向为大数据质量保障、云计算等,E-mail:11510600698@njust.edu.cn;陶传奇(1984—),男,博士,硕士生导师,主要研究方向为大数据质量保障、软件质量评估、软件测试与分析等,E-mail:taochuanqi@njust.edu.cn(通信作者);文万志(1982—),男,博士,副教授,主要研究方向为软件分析与测试、软件错误定位,E-mail:wenzhanzhi@ntu.edu.cn。

还没有可用的量化评估方法。

本文针对网购平台的搜索功能的质量评价,提出了可供参考的质量因素、评价指标及相关指标的计算方法,并通过实验进行了实例评价。本文所给出的质量因素及相关的质量指标可以用来对网购平台进行有针对性的搜索功能质量分析,针对不同质量因素及质量指标评价显示出的缺陷,可以指导网购平台进行搜索服务的优化,如优化搜索算法以提高搜索的准确性和稳定性,改善筛选条件可以使搜索更加人性化 and 便捷化,或者提供更加智能的关键词识别服务等。对于网购平台搜索功能的质量评估,最终可以使各网购平台不断优化自身搜索的用户体验,从而带来更多的经济效益。

本文第2节介绍了相关工作;第3节分析和探讨了网购平台搜索功能的质量,并提出了网购平台搜索功能的质量参考因素;第4节提出了网络购物平台搜索功能的质量因素评价指标及算法;第5节对实际应用的网购平台进行了实验,并分析了实验结果及其存在的问题;最后总结全文并展望未来。

2 相关工作

近年来,大数据的质量问题逐渐被认识和研究,大数据系统的质量问题往往会导致大量的经济损失^[1]。搜索引擎作为典型的大数据系统,其质量影响着搜索引擎的市场^[2]。搜索引擎的质量评价从提出到现在已经逐渐被广大从业人员接受,并逐渐成为搜索引擎相关领域的重要研究方向^[3-5]。文献^[6]提出了针对大数据系统的质量保障及其质量因素。

搜索引擎作为典型的大数据系统,其质量难以保障的原因有以下两点:1)搜索引擎的搜索结果与传统软件不同,没有一个预期的输出;2)搜索结果的评价应该与满足用户需求的程度有关,而满足用户需求的程度是一个难于量化的考核指标。

一些针对传统搜索引擎(如百度、谷歌等关键词搜索引擎)的质量保障方法已被提出,例如将传统测试技术加以改进后用于评价搜索引擎质量^[7]。Zhou等人将蜕变测试用于评价及验证搜索引擎的质量,并提出了部分的蜕变关系。在文章中作者提出利用蜕变测试进行质量考核,定义了5个蜕变关系(包括2个考核搜索结果排名的蜕变关系 SwapJd 和 Top1Absent 和3个考核搜索结果缺失程度的蜕变关系 MP-SITE, MP-TITLE 和 MP-ReversJD)。通过考核是否满足以上蜕变关系来对百度、谷歌等一般的搜索引擎进行质量评价。但它并没有提出系统化的质量因素和参考指标。文献^[8]提出利用基于仿真的方法来进行搜索引擎质量评估。Shehata等人提出了一个仿真模型,通过对搜索引擎的性能及表现进行建模,利用模型来支撑性能驱动的系统演化,以评估一个分布式搜索引擎。文献^[9]提出利用页面级关键字来评估搜索引擎,他们提到的页面级关键字是在网站的各个页面上发现的关键词,可以用来评价搜索引擎结果的相关性。此外,在图像搜索引擎方面,亦有部分的质量评估办法^[10-11]。

表1简要地将上述针对搜索引擎质量的评估方法进行了比较。

表1 部分搜索引擎质量评估方法的比较

评价利用的方法	评价内容/方式	缺陷
文献 ^[7]	搜索页面完整性,排名稳定性	没有考虑相关性,缺少对智能处理关键字功能的评价
文献 ^[8]	建立模型整体评价搜索引擎	对于具体的功能性指标,没有细化的计算方法
文献 ^[9]	搜索结果的相关性	缺少对多关键字搜索时的相关性的验证
文献 ^[17]	检索数量、响应时间等性能指标	对于与用户有关的功能性指标,缺乏评价
文献 ^[18]	点击率、点击时间等	适用于一般的搜索引擎,对于商品搜索引擎不能很好地兼容

然而,作为复杂的大数据系统,如何系统化地评价一个搜索引擎(尤其是本文涉及的购物平台的搜索引擎),尚没有一套成型且有效的质量标准和指标。

商品搜索引擎需要面对不同用户的购物习惯和行为的差异,其往往需要提供个性化的搜索结果^[12-13],这使得其搭建和设计更为复杂^[14-15]。网购平台商品搜索服务的质量不仅关系着网购平台的用户体验,而且决定了网购平台的盈利^[16]。由于购物商品搜索引擎没有一个预期的输出结果,因此如何评价商品搜索系统的质量,有哪些质量因素可以作为评价的标准和指标,成为了目前商品搜索引擎质量保障的难点之一。

3 网购平台搜索功能的质量因素

购物商品搜索系统是一个典型的大数据系统,用户通过键入搜索的关键词以及选择相关的筛选条件来进行商品的搜索和展示。购物搜索系统的商品种类和数量较多,属性各异。

购物搜索引擎与一般的网页搜索引擎的质量也不尽相同,常见的搜索引擎(如百度等)主要通过用户输入关键字进行搜索;购物搜索引擎(如淘宝网等)则提供可供选择的筛选按钮,使用户可以进行条件筛选。同时,购物商品搜索引擎也提供多样化的商品的详细信息,以及诸如价格比较、邮费比较、商品评价、商家评价等功能。综合考虑网购平台的特性,网购平台的搜索功能与常用搜索引擎的质量保障既有共同点,又有相异之处。

搜索引擎的质量保障不仅需要验证程序的正确性,而且也需检测程序的可用性和实用性,即从面向用户的角度出发来对搜索引擎的质量进行评估。本文从满足用户搜索需求的角度,提出了商品搜索引擎的质量参考因素,如图1所示。

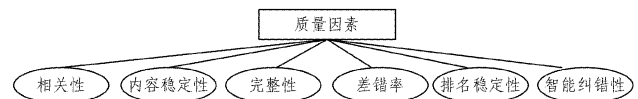


图1 网购平台搜索功能质量因素

下面是对提出的质量参考因素的解释。

(1)相关性:相关性是指搜索结果与搜索关键字之间的关联程度,如对关键字的匹配程度等,用于考察搜索引擎返回的结果与用户搜索关键字的相关程度。

(2)内容稳定性:最终搜索结果将某些数据呈现给用户,用户在搜索时,除去数据更新带来的影响,搜索结果应该保持一定程度的稳定。例如,同一用户短期内搜索同样的关键字

时搜索结果应该大致相同,或者用户在利用筛选开关进行商品搜索后,两次等价的筛选应该使搜索结果保持一定程度的不变。将这种质量因素定义为商品搜索内容的稳定性。

(3)完整性:完整性即商品搜索结果的完整程度,网购平台的搜索功能虽然没有一个可以预期的结果,但是搜索结果也应该满足最基本的要求,如搜索两个等价的关键字,出现的结果应该具有高度的一致性。完整性用于考核二次搜索的结果相比于一次搜索的结果时二者之间应该满足的关系。

(4)差错率:搜索结果的差错率即商品搜索结果不能准确地反映搜索关键字的程度,若出现与搜索关键字完全无关的搜索结果,则说明搜索引擎的正确性存在问题。它反映了不应该出现的商品出现的程度。

(5)排名稳定性:搜索结果都是以一定的排名呈现给用户的,排名稳定性用于衡量搜索引擎搜索结果排名机制的稳定性。

(6)智能纠错性:它反映了搜索引擎对于一些微小的错误能否自动识别并加以处理,如错别字、同义词、无用符号、简繁体等。

本文主要针对面向用户的功能质量因素,不对其他的因素(如搜索响应时间、搜索刷新时间等性能质量因素)作详细的讨论。

4 网络购物平台商品搜索功能的质量因素评价指标及算法

4.1 商品搜索引擎的质量评价指标

经过对商品搜索系统的分析,商品搜索系统质量对结果的影响主要体现在两方面:

(1)商品搜索结果的排名是否合理。商品搜索之后以一定的排序显示给用户,商品的排名是否稳定、排名先后顺序是否合理等都是商品搜索质量的体现。

(2)商品搜索结果的内容是否合理。这里的合理不仅包括上节中的商品搜索结果应该满足的基本的完整性和内容稳定性,也包括了智能纠错是否实现、差错率的表现、相关程度等。

另外,从商品搜索的实际操作而言,我们考虑了可能影响搜索结果的行为,如关键字位置、关键字变异、是否开启筛选、关键字长度的影响(如等价类可划分为仅有一两个字符的短关键字和除此之外的长关键字)、关键字个数的影响(如等价类可划分为单个关键字和多个关键字)等。从一个商品搜索引擎满足用户搜索需求的角度出发,提出了 20 个商品搜索系统的质量评价指标。其中指标(2)、指标(17)、指标(20)体现了商品搜索结果排名的质量,其余指标体现了商品搜索结果内容的质量。

(1)多关键字搜索商品时关键字位置对搜索结果总体影响的评价指标:是指在用多个关键字搜索时,若关键字位置变化,其搜索结果变化与否及搜索结果产生变化的程度。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示多关键字交换顺序前后的搜索结果。

(2)多关键字搜索商品时关键字位置对搜索结果排名影响的评价指标:是指多个关键字搜索时,关键字位置的变化对搜索结果中每一个具体商品的排名变化的干扰程度。该指标的计算公式为:

$$result = Ave(\sum(index_{x_1} - index_{x_2}))$$

其中, $index_{x_1}$ 和 $index_{x_2}$ 分别表示搜索多关键字和改变关键字顺序时两次均出现在搜索结果中的同一个商品的位置。

(3)单关键字搜索商品时商品标题与商品搜索缺失情况发生时的相关性评价指标:是指单个关键字搜索时,其中有商品 A,其标题为 *title*,那么同时搜索关键字与 *title* 时商品 A 应该依然出现在新的搜索结果中。该指标的计算公式为:

$$result = \frac{I_1}{I_2}$$

其中, I_1 是商品 A 再次出现的实际次数, I_2 是商品 A 再次出现的理论最大次数。

(4)单关键字搜索商品时商品发货地与商品搜索缺失情况发生时的相关性评价指标:是指单个关键字搜索时,其中有商品 A,其发货地为 *loc*,那么同时搜索关键字与 *loc* 时商品 A 应该依然出现在新的搜索结果中。该指标的计算公式为:

$$result = \frac{I_1}{I_2}$$

其中, I_1 是商品 A 再次出现的实际次数, I_2 是商品 A 再次出现的理论最大次数。

(5)单关键字搜索商品时商品价格与商品搜索缺失情况发生时的相关性评价指标:是指单个关键字搜索时,其中有商品 A,其价格为 *price*,那么同时搜索关键字与 *price* 时商品 A 应该依然出现在新的搜索结果中。该指标的计算公式为:

$$result = \frac{I_1}{I_2}$$

其中, I_1 是商品 A 再次出现的实际次数, I_2 是商品 A 再次出现的理论最大次数。

(6)单关键字搜索商品时筛选选项与商品搜索缺失情况发生时的相关性评价指标:是指单个关键字搜索时,搜索结果中商品 A 的详情页中有部分属性,比如分类、款式、包邮等常见购物属性,那么在搜索商品关键字 A 时,开启商品搜索引擎提供的筛选功能,筛选相应属性,商品 A 应该依然出现在新的搜索结果中。该指标的计算公式为:

$$result = \frac{I_1}{I_2}$$

其中, I_1 是商品 A 再次出现的实际次数, I_2 是商品 A 再次出现的理论最大次数。

(7)多关键字搜索商品时关键字重复对搜索结果总体影响的评价指标:是指多个关键字搜索时,若其中有关键字 A, B,某个关键字重复多次时,系统应能对其进行识别并加以处理,即 AAB 与 AB 的搜索结果应该一样。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别是搜索 AAB 与搜索 AB 的结果。

(8)多关键字搜索商品时关键字粘连对于搜索结果总体影响的评价指标:是指多个关键字搜索时,如其中两个关键字

为“男”和“衣服”,那么将两个关键字间的空格去除使之变成一个关键字“男衣服”时对搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 AB 与搜索 AB 的结果。

(9)单关键字搜索商品时无用符号对于搜索结果总体影响的评价指标:是指单个关键字搜索商品时,关键字中出现的逗号、分号、括号之类的搜索无用符号对商品搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 A 与搜索 A 加上随机无用符号的结果。

(10)单关键字搜索商品时错别字对搜索结果总体影响的评价指标:是指单个关键字搜索商品时,关键字中出现错别字对商品搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 A 与搜索 A 的错别字的结果。

(11)长关键字搜索商品时部分缺失对搜索结果总体影响的评价指标:是指单个关键字搜索商品时,较长的关键字中某个字缺失(如“阿迪达斯”中缺失一个字后为“阿达斯”)对搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 A 与搜索 A 缺失一部分的结果。

(12)单关键字搜索商品时简繁体对搜索结果总体影响的评价指标:是指单个关键字搜索商品时,关键字中简繁体差异对商品搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 A 的简繁体转换的结果。

(13)单关键字搜索商品时与商品描述(标题、正文、详情信息)的匹配程度指标:是指单个关键字搜索商品时,搜索结果的标题与搜索所用关键字的匹配程度。该指标的计算公式为:

$$result = Ave(cmp(p_i, A))$$

其中, cmp 是字符串匹配函数, p_i 表示单条搜索结果, A 表示关键字。

(14)单关键字搜索商品时同义关键字对搜索结果影响的评价指标:是指单个关键字搜索商品时,关键字用同义关键字替换(如“凤梨”和“菠萝”)对商品搜索结果的影响。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示搜索 A 与搜索 A 的同义词的结果。

(15)多关键字搜索商品时关键字与商品描述(标题、正文、详情信息)的匹配程度指标:是指多个关键字搜索商品时,搜索结果的标题、正文、详情页等与搜索使用的关键字的匹配程度。该指标的计算公式为:

$$result = Ave(cmp(p_i, A))$$

其中, cmp 是字符串匹配函数, p_i 表示单条搜索结果, A 表示关键字。

(16)多关键字搜索商品时关键字匹配个数的衡量指标:是指多个关键字搜索商品时,搜索结果的标题、正文、详情页等匹配的关键字的数量,如 5 个关键字的搜索结果中每件商品各自满足的关键字的数量。该指标的计算公式为:

$$result = Ave\left(\frac{count(cmp(p_i, A) > e)}{count(FR_i)}\right)$$

其中, cmp 是字符串匹配函数, p_i 表示单条搜索结果, A 表示关键字, e 表示判定为匹配的阈值, $count$ 是计数函数。

(17)多关键字搜索商品时关键字匹配个数与排名的衡量指标:是指多个关键字搜索商品时,搜索结果的标题、正文、详情页等匹配的关键字个数与商品搜索结果排名的关系,例如,匹配了 5 个关键字的商品应该排在匹配了 4 个关键字的商品之前。

(18)单/多关键字搜索结果差错率的评价指标:是指用关键字搜索商品时,出现与关键字完全不相关的商品的概率。该指标的计算公式为:

$$result = Ave\left(\frac{count(cmp(p_i, A) < d)}{count(FR_i)}\right)$$

其中, cmp 是字符串匹配函数, p_i 表示单条搜索结果, A 表示关键字, d 表示判定为不匹配的阈值, $count$ 是计数函数。

(19)单/多关键字搜索结果的稳定性评价指标:是指用同一关键字多次搜索商品时,搜索结果是否保持一定程度的稳定。该指标的计算公式为:

$$result = \frac{FR_1 \cap FR_2}{FR_1 \cup FR_2}$$

其中, FR_1, FR_2 分别表示短时间内搜索同一关键字的结果。

(20)单/多关键字多次搜索排名的稳定性指标:是指用同一关键字多次搜索商品时,搜索结果的排名是否保持一定程度的稳定性。该指标的计算公式为:

$$result = Ave(\sum(index_1 - index_2))$$

其中, $index_1$ 和 $index_2$ 分别表示短时间内搜索同一关键字两次,在搜索结果中两次均出现在同一个商品的位置。

根据指标对应的考核的质量因素对其进行划分,考核相关性的指标有(13), (15), (16);考核搜索结果稳定性的指标有(19);考核完整性的指标有(1), (3) - (6), (8);考核搜索结果差错率的指标有(18);考核智能纠错性的指标有(7), (9) - (12), (14);考核排名稳定性的指标有(2), (17), (20)。将商品搜索引擎根据关键字个数进行等价类划分,结合 6 个质量因素、20 个质量指标提出的商品搜索引擎质量因素与指标评价对照情况如表 2 所列,表 2 中内容为质量指标及其在上下文中对应的编号。

表 2 商品搜索引擎质量因素与质量指标的对照情况

质量因素	单个关键字	多个关键字
相关性	(13)	(15), (16)
内容稳定性	(19)	(19)
完整性	(3), (4), (5), (6)	(1), (8)
差错率	(18)	(18)
智能纠错性	(7), (9), (10), (12), (14)	(11)
排名稳定性	(20)	(2), (17), (20)

4.2 商品搜索引擎质量评价指标的算法

针对上节的质量指标,通过对实际的搜索流程的研究,包括搜索-筛选-再搜索-搜索结果比对等,提出每个指标的相应的计算方法。算法的通用流程图如图 2 所示。

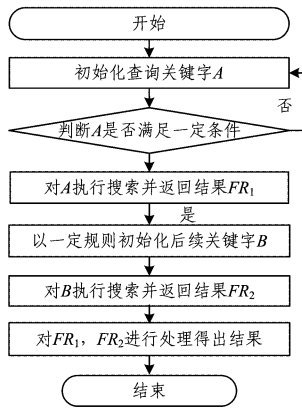


图 2 算法的通用流程图

算法 1 多关键字位置变化及关键字间粘连对搜索结果的影响算法

输入:初始多关键字数组 $A=A_1+A_2+A_3+\dots$

输出:对于每个关键字输入,指标的计算值

```

1. A.initial() //初始化单个关键字或多个关键字集合,记为 A
2. if A.respond.count >= 100 //返回结果应大于 100 条,若不满足,
   则重新初始化关键字 A(初始关键字搜索结果返回过少可能导致指
   标评价所使用的可用数据较少,影响其客观性):
   Continue;
   Else:goto 1
3. A1,A2=A.split(i) //将关键字 A 分为两部分
4. FR1=A.respond//进行搜索
5. FR1=FR1[0:100]//为了避免大量搜索结果带来的冗余影响,A 的
   搜索返回集合只取前 100 个,即 n=100
6. 指标(1),(2):B=A2+A1//构造新的关键字
   指标(8):B=A1A2
7. FR2=B.respond//用 B 进行搜索并返回结果
8. FR2=FR2[0:100] //B 的结果只取前 100 个
9. 指标(1),(8):result=jaccard(FR1,FR2)//根据两次搜索结果计算
   指标
   指标(2):
   FOR s in FR1 and s in FR2:{
       Count+=1
       Result+=(Index(s.FR2)-index(s.FR1))/FR2.count}
   Return Result/count

```

该算法主要适用于评价多关键字搜索结果的稳定性及完整性,可用于指标(1),(2),(8)。

对于一个良好的购物商品搜索系统,指标(1)应该满足关系:FR₁与FR₂有较大的jaccard相似系数。指标(2)应该满足:FR₁与FR₂中,相同的商品名次变化率应该尽可能小,即某件商品在前后两次搜索都出现,该商品的排名变化率((当前位置-原来位置)/总位置个数)应尽可能小。计算并返回相似系数或FR₁和FR₂中同时出现的商品的位置的变化率的平均值。其中,jaccard相似系数定义为:有两个集合X和Y,X与Y的交集占X与Y的并集的比例。

算法 2 单关键字商品标题、价格、产地等因素对搜索结果完整性的影响算法

输入:初始关键字 A

输出:对于每个关键字输入,指标的计算值

```

1-2. 同算法 1
3. FR1=A.respond
4. FR1=FR1[0:100]
5. for s in FR1:
6. 指标(3):B=A+s.title//对于 FR 中的每个搜索结果 pi,提取其商
   品标题并记为 title,发货地为 loc,价格为 price,并按此构造后续关
   键字
   指标(4):B=A+s.loc
   指标(5):B=A+s.price
7. FR2=B.respond
8. FR2=FR2[0:100]
9. if s in FR2:count=count+1
   Else:count=count
10. return count/FR1.count()

```

该算法主要用于单关键字情况下部分商品因素导致的搜索结果完整性相关的指标,可用于指标(3)-指标(5),与算法 1 重合的注释不再备注。

一个良好的商品搜索系统应当满足蜕变关系:任意的 p_i均属于对应的 FR_i。该算法对于同关键字搜索结果中每件商品的返回结果为布尔值 T 或 F,同关键字搜索结果中所有商品的平均值为 0~1 间的实数。比如,关键字 A 为电脑,初始搜索结果如图 3 所示,则在构造 B 时,指标(3)构造为电脑+Asus/华硕 V VM510LF5500 VM510L VM510LF I7 15.6 英寸笔记本电脑,指标(4)构造为电脑+上海,指标(5)构造为电脑+3769。

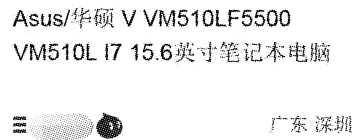


图 3 算法 2 的第一次搜索结果图

算法 3 单关键字商品可筛选属性对于搜索结果完整性的影响算法

输入:初始关键字 A

输出:对于每个关键字输入,指标的计算值

```

1-4. 同算法 2
5. for s in FR1://对于 FR 中的每个结果 pi,解析对应的网页
6. for d in s.character[]//提取每一个商品详情页页面中商品属性
   部分的内容,判断筛选属性是否提取完毕,如还可提取,则提取解
   析后商品筛选属性 d 并判断 d 是否是搜索页面中可以开启筛选
   的属性,如可开启筛选,则继续,否则退出算法
   If d is checked || d is not enable;Goto 6://若属性 d 是可以
   开启筛选且未被计算过的,则加入计算
7. B=A+d.check()//这里是指模拟搜索关键字 A 的同时开
   启筛选属性 d 的操作
8. FR2=B.respond
9. FR2=FR2[0:100]
10. if s in FR2:Count=count+1
   Else:Count=count

```

```

11. result += count/s.character[...].count()
12. return result/FR1.count()

```

该算法适用于指标(6)。

算法结果=开启筛选后商品仍在搜索结果中的次数/可筛选属性的总筛选次数。例如,搜索关键字“衣服”,出现的某条结果为 p_i , 打开 p_i 详情页后提取到的属性如图 4 所示,发现 4 个属性,且它们全部可以在淘宝筛选开关中启用。则每次分别启用一个开关,并且搜索关键字 A , 将结果记为 FR_i , 如果 p_i 仍然在 FR_i 中,则计数 $N+1$, 否则不变。然后计算 $N/4(0 \leq N \leq 4)$, 最后计算所有商品的平均值。

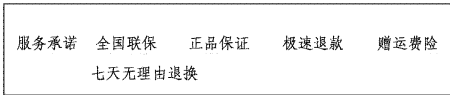


图 4 提取到的属性

算法 4 智能纠错性计算算法

输入:初始关键字 A

输出:对于每个关键字输入,指标的计算值

1-4. 同算法 2

5. 指标(7): $B=A+A$

指标(9): $B=A+\text{symbol.random()}$ //令 $B=B+A+$ 任意杂质符号,如‘?’,’,’,’.

指标(10): $B=A.\text{convert(错别字)}$

指标(11): $B=A.\text{remove(random)}$ //去除 A 中任意一个字

指标(12): $B=A.\text{convert(繁体)}$

指标(14): $B=A.\text{convert(同义词)}$

6. $FR_2=B.\text{respond}$

$FR_2=FR_2[0:100]$

7. $\text{result}=\text{jaccard}(FR_1,FR_2)$

该算法用于考察搜索引擎对明显错误的纠正能力,可用于指标(7),(9)-(12),(14)。

该算法应当满足关系: FR_1 与 FR_2 有较大的相似系数,计算并返回其相似系数。该算法返回结果应为 0~1 间的实数。例如, A 为凤梨,则指标(9)令 $B=$ 凤梨+“?”,指标(10)令 $B=$ 凤梨,指标(14)令 $B=$ 菠萝。

算法 5 搜索结果相关性计算算法

输入:初始关键字或关键字数组 A

输出:对于每个关键字输入,指标的计算值

1-4. 同算法 2

5. 指标(13):for s in FR_1 :

Result += compare(s.title,A)

Return Result/ $FR_1.\text{count}$ //指标(13)将标题与单关键字 A 进行字符串匹配算法,计算每个商品的标题与 A 的字符串匹配程度(百分比),并计算所有商品该指标的平均值。

指标(15):for s in FR_1 :

for d in A :

Result += compare(s.title,A[i])

$d+=\text{Result}/A.\text{count}$

Return $d/FR_1.\text{count}$ //指标(15)将标题与多关键字数组 A 进行字符串匹配算法,并将标题与 A 中每个子关键字字符串进行匹配的同时计算其匹配程度,然后计算该商品与 A 的平均子关键字匹配程度,最后计算所有商品该指标的平均值

指标(16):for s in FR_1 :

for d in A :

If compare(s.title,A[i])>阈值

Count += 1

$d+=\text{count}/A.\text{count}$

Return $d/FR_1.\text{count}$ //指标(16)计算与 A 中每个关键字完全匹配或高度匹配(设置阈值,如 80%)等的关键字匹配个数占关键字个数的比例,并计算所有商品该指标的平均值

指标(17):for s in FR_1 :

for d in A :

If compare(s.title,A[i])>阈值 MAX

Count += 1

C.append(Count)

Sum($C(i)-C(i+1)$) //指标(17)对于 A 的每一个返回结果 p_i ,提取其标题信息,并与关键字 A 中每个子关键字 A_i 进行字符串匹配,计算标题与 A_i 的字符串完全匹配的个数,并将计算结果作为该搜索结果的特征值,对于 100 个搜索结果,考核其特征值是否按降序排列,若不是,则计算结果与将结果进行降序排列后数组的数组差

指标(18):for s in FR_1 :

If compare(s.title,A)<阈值 MIN:

Count += 1

Else count += 0

Return $\text{count}/FR_1.\text{count}$ //指标(18)计算标题与关键字的字符串完全匹配度是否低于一定阈值(如 10%),若是则认为该商品是不正确的,计算 100 个商品中出现不正确商品的概率。该算法返回结果应为 0~1 的实数。

该算法用于考察搜索结果与搜索使用的关键字之间的相关程度,可用于指标(13),(15)-(18)。

例如,搜索关键字为“联想 电脑 笔记本 散热器”,结果中一个商品 p_i 如图 5 所示,标题为诺西笔记本散热器 14 寸 15.6 寸联想华硕戴尔电脑散热底座支架垫。则对于指标(15),4 个关键字中“联想”的匹配程度为 100%，“笔记本”的匹配程度为 100%，“电脑”的匹配程度为 100%，“散热器”的匹配程度为 66.7%，则该商品对关键字的平均匹配程度为 $(1+1+1+0.667)/4=0.91675$,指标(16)完全或高度匹配了 3 个关键词,剩余 1 个,结果计算为 $3/4=0.75$,指标(18)的计算为 $0/4=0$ 。

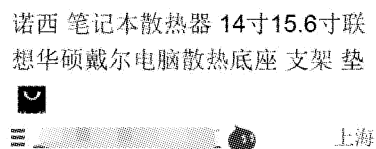


图 5 算法 5 的第一次搜索结果

算法 6 搜索结果稳定性计算算法

输入:初始关键字或关键字数组 A

输出:指标单次关键字输入的计算值

1-3. 同算法 2

4. time.sleep(time)

5. $FR_2=A.\text{respond}$

6. $FR_2=FR_2[0:100]$

7. 指标(19):

result=jaccard(FR_1,FR_2)

指标(20):

```

For s in FR1 {
  If s in FR2 :
    count += 1
  Result += (Index(s, FR2) - index(s, FR1))/FR2.count()
  Return Result/count//指标(19):短期内重复执行 step3 -
step4,计算返回的各个 FR 的相似系数。指标(20):短期内两次执
行 step3-step4,计算返回的两个 FR 中,计算每件两次均出现在搜索
结果里的商品的名次变化率,并计算所有商品的平均名次变化率
  该算法用于考察与搜索结果的稳定性相关的指标,可适用
于指标(19),(20)。

```

5 实验及分析

5.1 实验设计

本文第 2-4 节讨论了网络购物平台的服务与质量、质量因素划分、质量因素中具体可用来进行计算的指标等,初步描述了一个比较完善的网络购物平台所应满足的质量要求,提出了一套基于商品搜索功能的质量评估方法。

本节针对提出的 6 个质量因素和 20 个质量指标进行简单的实验验证,以淘宝网为实验对象。实验的关键字与普通的搜索引擎不同,因为是商品搜索系统,所以将任意的字符作为输入并没有意义,也不能体现商品的排名及推荐的算法。

从淘宝每日更新的搜索关键字排行榜中选择实验使用的关键字,解析品类排行榜数据并提取到本地作为关键字使用。对各品类均匀选择若干数量,共选择不少于 1000 个关键字,部分结果如图 6 所示。为了避免数据冗余对结果造成的影响,在实验搜索结果中只取前 100 个作为有效结果。每个指标的实验均取不少于 1000 个搜索关键字,对每个关键字进行多次实验并求均值。

针对不同的关键字均进行多次迭代,并定义每个关键字的常见错别字形式和同义词形式。实验以多线程批处理的方式运行,根据网速进行批量的搜索提交和结果返回,从而保证了搜索结果不会因过长的时间间隔而发生较大的变化。



图 6 使用的淘宝排行榜数据举例

对于指标(1),实验中为了方便计算,设置两个关键字 A1 和 A2。先根据关键字 A1+A2 和 A2+A1 进行搜索,然后针对结果进行处理,再比较二者的搜索结果,计算出相似系数。对于指标(2),实验中为了方便计算,设置两个关键字 A1 和 A2。先根据关键字 A1+A2 和 A2+A1 进行搜索,然后针对结果进行处理,对两个搜索结果计算单个商品排名的变化率,即位置差/总位置数,最后计算平均值。对于指标(3)-(5),获取单件商品的标题、发货地、价格,并将其加入到关键字中进行搜索。对于指标(6),通过解析商品的详情页,获得商品属性,然后在筛选开关里利用机器模拟进行仿人工筛选选择。

对于指标(7)-(12),(14),通过将关键字变异进行搜索性能判定,如错别字、无用符号、关键字重复、繁体字等。对于指标(7),采用重叠关键字方法实现;对于指标(8)采用关键词去除空格的方法实现;对于指标(9)通过关键字掺杂任意非法字符实现;对于指标(10)和(12),分别通过用字典进行错别字置换和繁体字置换实现;对于指标(11)通过随机去除一个字实现。按上述实现后续关键字 B 的生成;对于指标(14),通过同义词置换实现。指标(13),(15),(16),(18)通过获取搜索结果的标题、正文、详情页,与关键字进行字符串匹配,计算匹配程度、完全不匹配的商品个数、多个关键字匹配的关键字个数。指标(17)是指多关键字搜索商品时关键字匹配个数与排名衡量指标,指标(19)和指标(20)指在同一台主机上连续间隔若干秒发送相同的搜索请求命令,并根据返回结果进行计算。

5.2 实验结果

经过对千余组关键字进行搜索和对 20 个指标进行计算,得出了一系列能初步评价购物网站搜索系统的结论。各指标的详细结果如表 3 所列。

表 3 实验结果表

指标序号	平均值	最大值	最小值
1	68.4%	100%	0
2	3.12%	33.98%	0/4.3%
3	96%	100%	93%
4	≈0	≈0	≈0
5	≈0	≈0	≈0
6	88.5%	100%	35%
7	73%	82%	38%
8	>80%	100%	≈0
9	>80%	100%	45%
10	≈0	≈0	≈0
11	>80%	96%	23%
12	>70%	100%	0
13	>95%	100%	89%
14	30%	>50%	0
15	>95%	100%	89%
16	>95%	100%	>95%
17	匹配度极高	匹配度极高	匹配度极高
18	≈0	≈0	≈0
19	99%	100%	97%
20	99%	100%	96%

实验结果的整体箱型分布图如图 7 所示,横坐标为指标编号,纵坐标为实验结果。

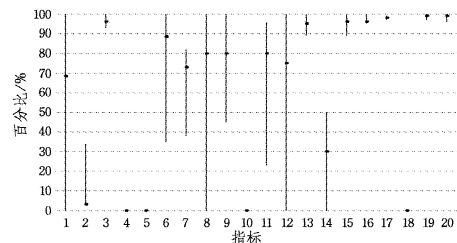


图 7 实验结果的整体箱型分布图

从图 7 可以看出指标评估的整体情况,每个指标的最上端和最下端分别表示该指标在千余个关键字上的最大值和最小值,中间的横向点表示该指标在千余个关键字上的平均值。部分指标的评估(如指标(7),(11)等)均没有出现极端值为 0 的情况,即不存在指标完全不满足的情况;而部分指标(如(1),(8)等)出现了最小值为 0 的极端情况,指标完全不满足,

这有可能是因为该网站的网页布局导致的结果提取错误,或者较为严重的网页 bug,或者搜索机制导致的。另外,如指标(17)–(20)等均处在较高水平,这说明该指标的评测较为稳定。

5.3 实验结果的解释及分析

搜索结果展示了在购物搜索网站搜索商品时各搜索操作和关键字的选择、关键字改变或部分改变对于搜索结果的影响。下面结合具体指标说明导致其质量问题的可能的原因。

指标(1)和(2)的质量原因可能在于多个关键字搜索结果的重叠率和排名的差异,购物搜索引擎在面对多个关键字时会选取其中一个作为优先级更高的搜索关键字,其余的优先级较低,那么当关键字交换时,搜索结果因为优先级对应的关键字不同而不同。

指标(3)表现一般,原因有可能是存在过长的商品标题,而重名商品名发生的情况比较少,使得搜索结果通常比较稀少,从而导致搜索缺失的可能性较小。

指标(4)与指标(5)均没有很好表现,如搜索“衣服”与“衣服 苏州”,前者的搜索结果为“衣服”,后者的搜索结果中出现了诸如“衣服制作”之类的书籍和工具,而不是产自苏州的衣服。有可能购物平台在关键词搜索机制中不会自动识别发货地或价格并对其进行筛选,而在对搜索结果进行地址筛选操作时,才会显示与发货地对应的商品。

对于指标(6),同样的商品 A 的发货地为广州,在搜索 A 的关键字时 A 排名靠前,而同时搜索 A 和广州时排名靠后。可能是因为设置了这样的策略,即商品 A 在不限地域范围内搜索时,排名系统给它的排名是较靠前的,但是在地域排名内综合排名系统给了它一个较低的 rank 值。

指标(7)的表现效果不佳,可能的原因是购物网站搜索引擎对于多关键字不会采取默认去重,而是将两个完全一样的关键字进行 OR 处理,从而便能搜索到部分商品名中有重复字段的商品。而根据实际观察,确实存在部分商品的名称中有重复的关键字。

指标(8)–(12)表现一般的原因是:1)购物网站的搜索引擎对繁体字的处理效果好,但实际可能是所选取的关键字的简繁体一样,导致未产生差别;2)关键字中缺少一个字,有部分程度的模糊查询,处理效果一般;3)关键字粘连指标偶尔有极端为 0 的值,可能是关键字分开和连在一起产生了歧义;4)部分符号不能被识别并处理,可能是在搜索中该符号有特殊含义或无法识别,从而导致无法被搜索引擎智能处理。

指标(10)的命中率约近似为 0。如搜索“凤梨”与“凤梨”,前者的结果全部为水果凤梨,后者的结果大多为与梨花烫、吹风机有关的美发产品;对于错别字而言,由于搜索引擎可能没有储备错别字库,没有纠正错别字,导致错别字搜索结果几乎为 0。在实验中,淘宝会提示可能的错别字,并给出可能的正确字,但是它依然会按照错别字进行搜索,并不会像通常的互联网搜索引擎那样对错别字进行默认纠正处理。

指标(13)的平均匹配程度极高,发生不匹配时常常是因为关键字缺失头部第一个字或末尾最后一个字,这表明购物网站应具备应有的搜索功能。

指标(14)的平均结果较低,但并不为 0,有可能是部分卖家将一个商品的关键字及其同义字均写在商品标题中导致的。但经过实验证明,该购物网站并不能识别同义词,在进行搜索时,自动进行同义词转换。通过它们依然能正确搜索出所需要的商品,如通过“凤梨”和“菠萝”均能搜索出相同水果。

指标(15)–(17)的搜索结果几乎都满足所有关键字,而不满足关键字的常常不作为结果显示,可能是淘宝网的搜索引擎在关键字较少时采用严格搜索的方法不会默认地进行模糊搜索;当关键字较多时,会将一部分关键字作为首要关键字,而其他关键字则有可能开启模糊搜索。

指标(18)的结果较低,出现不正确商品的可能性较小。指标(19)–(20)在短期内均有很高的 jaccard 相似系数与较低的排名变化率,可能的原因是在极短的时间内购物搜索网站的搜索索引不会有太大变化,因此结果也有可能存在很大程度上的重复。

造成搜索结果不一致的原因也有可能是其他情况,比如搜索算法问题、搜索关键字本身的歧义问题、排名机制问题、政策原因等。

5.4 实验结论——实例评价淘宝网的结果

对于淘宝网,我们进行了如上 6 个质量因素和 20 个质量指标的简单测试和分析,针对该网站的搜索引擎质量做相应评估,得出对于该网站的质量评价如下。

(1)相关性:该网站的搜索功能对于用户搜索目的的贴合性较好,搜索结果符合用户要求的程度较高。搜索结果与用户选取的搜索关键字之间的匹配程度较高,较好地反映了用户的搜索需求。

(2)搜索结果中内容的稳定性:该网站搜索功能的稳定性较好,在不考虑相关的推荐策略和商业因素的情况下,进行同样的搜索时关键字位置的变更对于结果的影响在合理范围内。

(3)完整性:该网站搜索功能的完整性较差,对同样的商品等价地变更搜索条件,再次对其进行搜索时,商品发生缺失的概率较高。同时,对于搜索时的筛选选项的开关不能智能化处理,对于关键字中典型的搜索要素(如地址、价格等)也不能智能识别。

(4)搜索结果差错率:该网站搜索功能的搜索结果差错率指数较低,简单的商品搜索反映了用户搜索目的的程度较好,出现错误商品的概率较低。

(5)排名稳定性:该网站搜索功能的排名稳定性在合理范围内,除去商业因素影响,排名相对较稳定。

(6)智能纠错性:该网站搜索功能的智能纠错性表现一般,具体表现为:1)不能识别重复关键字;2)可以识别大部分无用符号,部分不能识别;3)无法识别错别字,可能是网站没有设置对应的常用错别字处理库,因此表现较差;4)简繁体和同义词表现不好,系统没有针对简繁体或者同义词做识别与处理。

5.5 存在不足

实验和论文尚且存在不足,如使用的关键字数量为千级,一定程度上对实验结果的客观性会造成一部分影响,当关键字数量不断增多,直至覆盖绝大多数关键字时,实验结果会越

精确。另外,实验的网络环境、带宽、淘宝网自身更新商品的速度也会对结果造成影响。对于存在极端值的指标,要进一步研究存在极端值的原因,根据情况将实验结果进行清洗,以得到更为客观、准确的结论。本文提出的质量因素和质量指标还可以继续改进,针对不同的个性化的网购零售平台,还可以提出更多个性化的、自适应的质量因素和指标。

结束语 本文针对网络购物平台的搜索功能进行了研究,针对搜索功能提出 6 个质量因素和 20 个相关的质量指标以及相关的质量指标表,同时结合这些质量指标进行了实例的实验和验证,实例评价了一个购物网站的搜索功能,本文工作对于各个网购零售平台评估其搜索策略的质量有一定的帮助。结果证实了网购平台搜索引擎的质量可以被量化评价和考核,并给出了评价的范例。

本文提出的质量因素和相关的质量计算指标与其他针对一般的搜索引擎(如百度谷歌等)的评价的不同之处在于:

(1) 本文是针对网购平台的搜索功能提出质量因素和指标,其与其他评价一般搜索引擎的方法相比,对评价网购平台的搜索功能具有更好的适应性。

(2) 本文不侧重于搜索性能方面,如搜索时间、页面返回时间等,而是从网购搜索功能满足用户需求的角度提出了面向用户需求的质量因素和指标,这更能反映设计搜索功能的初衷是为了便于用户检索信息。

(3) 本文提出的对于购物推荐系统的质量因素和计算指标同样可以拓展和应用到一般的搜索引擎(如分类检索、图像搜索、文本搜索引擎)中。

本文针对大数据的网购平台搜索功能的质量因素进行了研究,在相关研究领域大数据质量保障是一个急切且重要的研究课题。随着大数据数据量的爆炸式增长及数据应用的广泛普及,人们将从传统数据中解放出来,利用质量有保障的大数据从事生产科研活动。对于搜索引擎的质量评价方法,可以将其拓展到其他大数据系统的质量及质量评价中,大数据软件的质量保障会在未来被广泛地研究与应用。

参 考 文 献

- [1] WIGAN M R, CLARKE R. Big Data's Big Unintended Consequences[J]. Computer, 2013, 46(6): 46-53.
- [2] LIANNOS I, MOTCHENKOVA E. Market dominance and search quality in the search engine market[J]. Journal of Competition Law & Economics, 2013, 9(2): 419-455.
- [3] HAWKING D, CRASWELL N, BAILEY P, et al. Measuring Search Engine Quality[J]. Information Retrieval, 2001, 4(1): 33-59.
- [4] LOSEE R M, PARIS L A H. Measuring search-engine quality and query difficulty: Ranking with target and freestyle[J]. Journal of the American Society for Information Science, 1999, 50(10): 882-889.
- [5] † E V C. Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality[J]. Journal of Computer-Mediated Communication, 2007, 12(3): 866-887.
- [6] GAO J, XIE C, TAO C. Big Data Validation and Quality Assurance - Issues, Challenges, and Needs[C]// IEEE, IEEE International Symposium on Service-Oriented System Engineering. 2016: 433-441.
- [7] ZHOU Z Q, XIANG S, CHEN T Y. Metamorphic Testing for Software Quality Assessment: A Study of Search Engines[J]. IEEE Transactions on Software Engineering, 2016, 43(3): 264-284.
- [8] PHELAN T, PATEL A, CIARDHUÁIN S O. Simulation Based Approach to Evaluate a Distributed Search Engine[C]// Iadis International Conference Wwww/internet 2003 (Icwi 2003). Algarve, Portugal, 2003: 347-354.
- [9] GOEL S, YADAV S. Search engine evaluation based on page level keywords[C]// Advance Computing Conference. 2013: 870-876.
- [10] LAZARINIS F, EFTHIMIADIS E N. Measuring search engine quality in image queries in 10 non-English languages: an exploratory study[C]// Proceeding of the, ACM Workshop on Improving Non English Web Searching (Inews 2008). Napa Valley, California, USA, 2008: 89-92.
- [11] TIAN X, LU Y, YANG L. Query Difficulty Prediction for Web Image Search[J]. IEEE Transactions on Multimedia, 2012, 14(4): 951-962.
- [12] WANG N. Design and Implementation of a Crawling System in Shopping Search Engine[C]// Second International Workshop on Computer Science and Engineering, 2009 (WCSE'09). IEEE, 2009: 212-216.
- [13] MOORE T J, KELLEY W C, WILSON A R. Online Shopping Search Engine for Vehicle Parts: US, US20090019008 [P]. 2009.
- [14] PARK S, CHO K, CHOI K. Information Seeking Behavior of Shopping Site Users: A Log Analysis of Popsheoes, a Korean Shopping Search Engine[J]. Journal of the Korean Society for Information Management, 2015, 32(4): 289-305.
- [15] LI S W, QIAN X D. Model of personalized online shopping search engine[J]. Application Research of Computers, 2010, 27(16): 2176-2180. (in Chinese)
李世威, 钱晓东. 一种构建个性化网络购物搜索引擎模型研究[J]. 计算机应用研究, 2010, 27(6): 2176-2180.
- [16] LIHAILONG L, LIUYU L. Shopping search engine technology based on services[C]// Conference on Applied Computer Science. 2008: 420-423.
- [17] TAO Y H, SUN M S, WANG X G. Evaluation system of Internet search engine [J]. computer engineering and science, 2001, 23(3): 25-27. (in Chinese)
陶跃华, 孙茂松, 王锡钢. 因特网搜索引擎评价系统[J]. 计算机工程与科学, 2001, 23(3): 25-27.
- [18] CEN R W. Research on the evaluation of search engine based on user behavior analysis [D]. Beijing: Tsinghua University, 2010. (in chinese)
岑荣伟. 基于用户行为分析的搜索引擎评价研究[D]. 北京: 清华大学, 2010.