基于用户点击信息检索评价方法综述

肖冬青[1]，杨沐昀[1]，李 生[1]，齐浩亮[2]，赵铁军[1]

[1]哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001

[2]黑龙江工程学院 计算机技术系，哈尔滨 150050

E-mal: {dqxiao; ymy;tjzhao}@mtlab.hit.edu; lisheng@hit.edu.cn; haoliang.qi@gmail.com

摘 要：评价是信息检索研究长期关注的焦点，推动信息检索技术的进步。在简要分析 Cranfield 评价的优点和不足、基于检索日志进行检索评价的巨大潜力后，本文论述从搜索日志中获得可靠文档相关性估计存在的困难，分析了近年国内外研究人员提出的若干典型点击模型，并对其就可扩展性、增量可计算性、点击预测精度、模型的复杂性等方面进行讨论。

关键词：信息检索评价；Cranfield 框架；搜索日志；点击模型；位置偏置

A Survey of Information Retrieval Evaluation by Clickthrough Data

Xiao Dongqing[1], Yang Muyun[1], Li Sheng[1], Qi Haoliang[2], Zhao Tiejun[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

[2]Dept. of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050

E-mail: {dqxiao; ymy;tjzhao}@mtlab.hit.edu; lisheng@hit.edu.cn; haoliang.qi@gmail.com

Abstract: Evaluation methodology has been the focus of information retrieval siance it motivatesto the progress of information retrieval. After discussion of advantages and deficiencies of the prevailing Cranfield paradigm and potential of evaluation based search logs, this paper introduces and analyzes some current typical click models for explaining user behavior especially click rate. It also investigates their scalability, incremental computability, delicate balance between click prediction precision and model complexity.

Keywords: information retrieval evaluation; Cranfield paradigm; search log; click model; position bias

1 引言

评价是信息检索研究长期关注的焦点，“在信息检索中处于如此重要的地位，以至于任何一种新方法与他们的评价方法是融为一体”[1]。目前检索评价多采用 Cleverdon 提出的 cranfield 框架，即基于测试集评价方法。这种基于测试集的方法有良好的数学基础，直接针对检索系统性能的关键——检索出尽可能多的相关文档[2]。目前世界上著名的 TREC、NTCIR 和 CLEF 评测基本上都采用了这一框架，有力的推进信息检索科研、商业应用的发展。

这种基于测试集的评价是以牺牲部分真实性为代价，对检索任务的高度抽象，以期在可控的条件进行系统性能的比较。测试集被认为是对真实检索任务一个必要非充分的代表：真实的检索对象简化为相对静止、有限的文档集，将用户的信息需求简化为查询/主题集，将用户对检索结果的认知简化为主题相关性判断（二元/多元相关性判断），将用户行为决策、信息加工简化为以查准率、查全率为代表的评价指标后的行为模型。通过上述简化，基于测试集评价方法构造一个封闭世界，将“多变”的用户排除在评价过程外。而且，该方法依赖人工标注文档主题相关性，需要消耗大量人力、时间资源，无法大规模实时的开展。更为重要的是基于测试集评价中“相关性”。文档与查询的语义相关性并不等同于文档效用；文档效用与用户具体的信息需求有关、与用户的背景知识有关；文档效用还应包括对用户情感的影响，这些远非主题相关性所能涵括。此外用文档效用的判断天然处于一定上下文背景下，而基于测试集评价方法独立地考虑检索结果的主题相关

性。故该评价与用户检索体验存在距离。有研究表明，这种评价结果与用户检索效率、用户满意度并非显著相关[3-5]。

信息检索系统特别是近来的 Web 搜索引擎越来越重视满足用户的信息需求，使用户获得满意的检索体验。使用用户反馈进行系统评价天然贴近用户，并能把握相关性的丰富内涵。上个世纪 50、60 年代开展的基于用户反馈评价由于其实验无法重现、缺乏统计可靠性，深受诟病。随着检索引擎的流行和用户规模的扩大，大规模基于用户反馈进行系统性能评价成为可能。

在信息检索过程中存在大量用户与检索系统的交互。一个完整的检索过程为，用户基于一定信息需求，根据自身经验构造查询关键词，并将其提交给在线搜索引擎。在线搜索引擎根据查询关键词与一定的检索策略返回可结果列表。用户观察结果列表，如标题、摘要、URL、前后结果文档等，选择点击其认为可能相关文档。如果该文档能满足信息需求，用户可能结束此次检索；反之，用户将返回检索结果页面，继续寻找可能相关的结果等。点击信息可视为用户群体智慧，蕴含着丰富文档相关性信息。从点击信息中挖掘文档相关性，已然成为信息检索研究的热点问题。

2 用户行为模式

所有试图从搜索日志中挖掘文档相关性信息的研究，面临着同一个问题—如何正确解读用户点击行为。

在这方面，Jochims 与他的同事们开展一项名为“眼睛跟踪”的基础性研究研究，使用传感器记录用户眼球的转动，以获得用户注意力的集中区域、集中时间、移动路径等[7-9]。通过“眼睛跟踪”与调查问卷与搜索日志的比较，Jochims 等人发现若干用户行为规律：多数用户顺序浏览搜索引擎返回结果列表，用户多通过 URL、摘要、标题评估返回结果的相关性，评估后用户选择点击该文档、以返回结果每页 10 个为例，用户注意力集中在前部，此后，用户注意力涣散，6-10 位置无明显差异；用户使用滚动条。

Jochims 等人设计若干规则从用户点击序列获得文档相对相关性，与“眼睛跟踪”和用户调查比对后，证实在较短位置距离内，可获得可靠的相对相关性。这样 Jochims 使用文档相关性偏序对训练 Ranking SVM，以优化搜索结果排序。Jochims 实验还发现当文档距离较远时，被点击文档相关性不总是大于略过文档，故该方法仅能达到局部优化，而非全局优化。

点击频度是否等同文档绝对相关性？“眼睛跟踪”实验表明除文档相关性外，检索结果排序、内容展示等多方面的因素影响用户点击，即文档相关性与点击没有绝对的相关性。

文档相关性与其他因素共同作用的结果为点击分布。对点击建模，即考虑文档相关性、其他因素如何联合作用产生点击，获得点击模型后，对给定点击序列，使用贝叶斯公式，使用最大后验概率估计可获得文档相关性序列，用于检索系统评价。问题就转化为对点击行为建模。除部分工作[8]外,多数点击模型基于浏览假设，即观察到一次点击事件当且仅当用户浏览到该位置，并认为该文档为相关文档，分开考虑用户浏览模式，文档相关性对点击的影响。问题进一步转化成对用户浏览行为建模。

3 基于点击模型的主要评价方法

本文将介绍近年来国内外几个典型的点击模型。根据浏览独立性的不同理解，点击模型可分为四类：浏览独立于先前的检索决策、浏览依赖于先前的点击决策、浏览依赖于先前的点击决策与文档相关性、浏览依赖当前用户状态。

使用下列符号描述检索过程，点击序列 $C_1, C_2 \ldots C_t, C_i$ 为二元随机变量，表示是否点击第 i 个返回结果；t 表示用户最后一个点击的位置。浏览序列 E: $E_1, E_2 \ldots E_m$，m 表示浏览最后一个文档位置；r_{d_i} 表示第 i 篇文档相关性。文档序列 D: $D_1, D_2 \ldots D_n$，表示检索系统返回结果。

3.1 浏览独立于先前浏览序列

浏览独立模型最早由 Richardson 等人提出，Craswellge 给出形式化定义。该模型基于强假设：先前检索过程不影响当前浏览、点击决策。Richardson 假设用户浏览第 i 篇文档的概率仅与位置 i 有关。该模型处理位置因素影响，但过分僵化，假设存在一个固定不变的完全由位置决定的浏览模式；同时将一次连续的检索过程拆分离散单元，损失点击浏览之间的序列信息。

3.2 浏览条件依赖于先前的点击决策

该类模型主要包括级联模型（Cascade Model）[10]，依赖点击模型（Dependent Click Model, DCM）[11]，用户浏览模型（User Browsing Model, UBM）[13]，贝叶斯浏览模型（Bayesian Browsing Model，BBM）[18]。

级联模型：级联模型假设用户总是从第一个检索结果开始，有序的浏览每一个返回结果，用户获得一个相关文档后，便停止此次浏览。级联模型无法解释真实一个检索过程中存在多个点击与用户中途放弃此次检索的现象，适用面窄。

依赖点击模型：DCM 扩展级联模型，描述一次检索过程中存在多个点击和用户中途放弃检索现象。用户点击文档后仍有可能继续浏览更多检索结果，而用户的耐心、时间有限，故其可能性与文档所在位置有关，表示为用户行为模式参数 λ_1。实验[11]表明 DCM 点击预测能力优于级联模型。

用户浏览模型：UBM 认为用户随时可能停止此次检索过程，或是修改查询、放弃此次检索。UBM 认为用户继续浏览的可能性，与当前位置、与最近一次点击位置的距离有关。Chao li 实验证明 UBM 多引入参数 β 后，较 DCM 的性能提升仅 5%，计算量大幅度提高。

贝叶斯浏览模型（bayesian browsing model，BBM）：BBM 与 UBM 使用相同用户假设。不同的是，BBM 使用贝叶斯网络形式描述浏览过程。BBM 将文档相关性视为一个概率分布形式未知的随机变量，而不是 UBM 中一个确定、未知的数值。

上述模型主要考虑用户浏览习惯，而非考虑整个检索过程中用户检索体验和信息量的累积变化。一连串点击不相关文档，使得用户沮丧，更可能放弃此次检索；一连串点击相关文档，用户信息需求得到满足，用户可能结束此次检索。

3.3 浏览条件依赖于先前的点击序列、文档的相关性

该类模型主要包括点击链模型（click chain model，CCM）[19]，一般点击模型（general click model，GCM）[12]。

点击链模型（click chain model，CCM）CCM 引入参数 a_2，a_3，考虑最近一次点击文档的相关性对于当前浏览决策的影响。这两个参数分别用户检索体验和短期内收集信息过程。假设点击文档高度相关，一方面表明返回结果质量较高，鼓励用户继续浏览；另一方面，用户的信息需求得到进一步满足，用户较可能终止浏览。

$$p(E_i = 1 \mid E_{i-1} = 1, C_{i-1} = 0) = \alpha_1$$

$$p(E_i = 1 \mid C_{i-1} = 1) = \alpha_2(1 - r_{d_{i-1}}) + \alpha_3 r_{d_i}$$

一般点击模型（general click model，GCM），GCM 将文档相关性、浏览、点击都视为概率分布形式未知的随机变量。Z. Zhu 等人证明其他模型都视为该模型特例。

$$p(E_i = 1 \mid E_{i-1} = 1, C_{i-1} = 0) = \Pi(A_i > 0)$$

$$p(E_i = 1 \mid C_{i-1} = 1) = \Pi(B_i > 0)$$

$$p(C_i = 1 \mid E_i) = \Pi(r_{d_i} > 0)$$

上述点击模型描述能力、预测精度得到进一步增强；更多参数的引入，带来计算复杂度大幅度增长。事实上先前文档的相关性，点击序列是通过影响当前的用户状态，影响当前浏览决策。

3.4 浏览条件依赖于用户状态

动态贝叶斯网络点击模型（dynamic Bayesian model，DBM）[20]，DBM 中通过引入一个隐变量S_{d_i}，表示点击后阅读文档内容后用户状态变量。用户满意度影响用户浏览下一个文档的决策。

$$p(E_i = 1 \mid E_{i-1} = 1, C_{i-1} = 0) = \gamma$$

$$p(E_i = 1 \mid C_{i-1} = 1) = r(1 - s_{d_i})$$

会话效用模型（session utility model, SUM）[ ]，SUM 引入节点用户当前获得信息总量 U（c），并将 U（c）与用户信息需求得到满足，停止浏览的函数形式确定为 sigmod 形式。

4 主要评价方法的对比分析

上文介绍若干典型点击模型，模型之间的差异在于对浏览独立性的不同处理，不同的模型考虑不同影响用户浏览决策因素，如下表所示。

点击模型 \ 浏览独立性	固定，未知的行为模式					变化的用户状态	
	位置独立	线性浏览	单个点击	当前位置	与最近点击的距离	先前检索经历	用户当前状态
浏览独立模型	√						
级联模型		√	√				
依赖点击模型		√		√			
用户浏览模型		√		√	√		
贝叶斯浏览模型		√		√	√		
点击链模型		√				√	
一般点击模型		√				√	
动态贝叶斯网络模型		√					√
会话效用模型		√					√

简单的点击模型无法描述真实点击的多样性，点击预测精度也较低；复杂的模型带来的是计算上巨大的时间和空间开销。点击模型应在精确性与复杂性之间进行取舍。Yin Hc 和 Kuansan Wang 将点击贝叶斯斯网络简化为两个部分可观察的马尔可夫时序模型[17]，做了模型复杂度与预测精度之间的折衷；Yi Min Wang 等人考虑查询所类别不同，用户检索决策出发点不同[14]。

除了保证文档相关性推理的可靠性外，一个理想的点击模型应能处理大规模、TB 级点击数据，同时能根据数据更新动态更新模型。点击链模型、依赖点击模型、动态贝叶斯点击网络兼具三个特性。上述所有研究工作或是基于可控实验环境中，与真实网络用户行为有一定差距。比如检索日志可能包含大量非真实用户—网络爬虫行为，用户在检索过程受到广告干扰等等。上述所有研究工作或是基于大规模群体用户的点击行为分析，尤其是对同一查询，需要大量的用户点击信息，难以处理用户访问频度低的长尾查询词。

5 小结

评价为信息检索研究的一个重点，推动信息检索技术的进步。在简要分析 Cranfield 评价的优点和不足与基于检索日志进行评价的巨大潜力后，本文论述从检索日志中获得可靠文档相关性估

计存在的困难，同时分析、比较近些年典型的点击模型，对其可扩展性、增量可计算性、点击预测精度、模型的复杂性进行讨论。

参 考 文 献

[1] Evaluation in information retrieval.Lecture Notes in Computer Science, 2001, Volume 1980/2001, 81-92, DOI: 10.1007/3-540-45368-7_4.

[2] Donna Harman. Is the cranfield paradigm outdated? SIGIR'10, July 19-23, 2010, Geneva, Switzerland.

[3] Ellen M. Voorhees*. On test collections for adaptive information retrieval In proc.44th Information Processing and Management, Pages1879-1885, 2008.

[4] Mark Sanderson, Monica L. Paramita, Paul Clough, Evangelos Kanoulas. Do User Preferences and Evaluation Measures Line Up? SIGIR 10 (2010).

[5] Mark D. Smucker, Chandra Prakash Jethani. Human performance and retrieval precision revisited SIGIR '10.

[6] Thorsten Joachims. Optimizing search engines using clickthrough data. In KDD '02.

[7] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In SIGIR '05.

[8] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Trans. 2007.

[9] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In SIGIR '06: 2006.

[10] G. E. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using click-through data and a user model. WWW '07, 2007.

[11] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models .In WSDM '08: Proceedings of the international conference on Web search and web data mining, pages87-94. ACM, 2008.

[12] Z. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen. A novel click model and its applications to online advertising. In ACM Conference. Web Search and Data Mining (WSDM), 2010.

[13] Ramakrishnan Srikant, Sugato Basu, Ni Wang, Daryl Pregibon. User Browsing Models: Relevance versus Examination. KDD'10.

[14] Yisong Yue, Rajan Patel, Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. WWW'10.

[15] Ramakrishnan Srikant, Sugato Basu, Ni Wang, Daryl Pregibon. User browsing models: relevance versus examination. KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

[16] Chao Liu, Fan Guo, Christos Faloutsos. Bayesian Browsing Model: Exact Inference of Document Relevance from Petabyte-Scale Data. Transactions on Knowledge Discovery from Data (TKDD), Volume 4 Issue 4.

[17] Yin He, Kuansan Wang. Inferring search behaviors using partially observable markov model with duration (POMD). WSDM '11.

[18] C. Liu, F. Guo, and C. Faloutsos. Bbm: Bayesian browsing model from petabyte-scale data. In ACM Conf. on Knowledge Discovery and Data Mining (KDD), 2009.

[19] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In World Wide Web Conference,WWW'09, 2009.

[20] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In World Wide Web Conference, WWW'09..