软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1714−1724 [doi: 10.3724/SP.J.1001.2011.03873]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

锚文本检索有效性分析*

周 博[1,2,3+], 刘奕群[1,2,3], 张 敏[1,2,3], 金奕江[1,2,3], 马少平[1,2,3]

[1](清华大学 智能技术与系统国家重点实验室,北京 100084)

[2](清华大学 清华信息科学与技术国家实验室(筹),北京 100084)

[3](清华大学 计算机科学与技术系,北京 100084)

Retrieval Effectiveness Analysis for Anchor Texts

ZHOU Bo[1,2,3+], LIU Yi-Qun[1,2,3], ZHANG Min[1,2,3], JIN Yi-Jiang[1,2,3], MA Shao-Ping[1,2,3]

[1](State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

[2](Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)

[3](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: b-zhou07@mails.tsinghua.edu.cn, http://www.csai.tsinghua.edu.cn

Zhou B, Liu YQ, Zhang M, Jin YJ, Ma SP. Retrieval effectiveness analysis for anchor texts. *Journal of Software*, 2011,22(8):1714−1724. http://www.jos.org.cn/1000-9825/3873.htm

Abstract: Anchor texts have been extensively used and have proven to be effective over the years in commercial Web search engines; however, anchor texts are created freely without any supervision, which causes there to be a large number of noises and spam in anchor texts. Moreover, for transactional queries, which need measurements of service quality, destination, Web pages of anchor texts are not always consistent with user experience. To overcome these problems, this paper focuses on the large scale of the user Web browsing behavior data. This paper first proposes a framework for a retrieval effectiveness measurement. Then, based on this framework, analyze the relation between the user Web browsing behavior data and retrieval effectiveness of anchor texts is analyzed. The properties of a user Web browsing behavior, which are useful for Web search, are mined and quantified. Experimental results show that the proposed method offers an accurate evaluation on the retrieval effectiveness of anchor texts.

Key words: user Web browsing behavior; anchor text; Web information retrieval

摘 要: 锚文本对网络信息检索性能的提升作用已经得到验证,并被广泛地应用于商用网络搜索引擎.然而,锚文本制作的不可控性导致其中蕴含大量与目标网页不相关或具有作弊倾向的无用信息.另外,对于需要衡量检索结果服务质量的事务类查询,原始锚文本推荐的目标网页也往往与真实的用户体验不一致.为了解决上述问题,基于大规模真实用户的互联网浏览行为日志展开研究.首先提出了锚文本检索有效性的评估框架,然后分析了用户网络浏览点击行为与锚文本检索有效性之间的联系,挖掘了用户网络浏览点击行为中有助于筛选高质量锚文本的特征.基于这些特征,提出了两种超链接文档生成方法.实验结果表明,基于用户网络浏览点击行为特征筛选出的锚文本,与原始锚文本相比,能够明显地提升网络检索的性能.

* 基金项目: 国家自然科学基金(60736044, 60903107); 高等学校博士学科点专项科研基金(20090002120005)
收稿时间: 2010-02-10; 修改时间: 2010-04-22; 定稿时间: 2010-05-14

关键词: 用户网络浏览行为;锚文本;网络信息检索
中图法分类号: TP391 文献标识码: A

现代社会网络信息极度丰富,网络信息检索工具已成为人们访问互联网资源的主要媒介.作为当前网络信息检索的主要工具,搜索引擎已成为人们访问互联网资源的有效途径.因此,如何改进搜索引擎系统的检索算法,提高检索性能,已成为研究界和产业界关注的主要话题.作为网页文本的补充,锚文本(anchor text)体现了网页制作者对目标网页主题或内容的一种描述.锚文本对搜索引擎性能的提升作用已经得到验证[1,2],并被广泛地应用于商用网络搜索引擎[3].目前,在网络搜索引擎中最常用也最有效的使用锚文本的方式是:将指向同一网页的锚文本整合为一个超链接文档,作为该目标网页的描述文档,与网页文本同时进行检索[1−5].这种方式有助于提升检索性能的原因是网页的超链接文档与用户检索该网页时所使用的查询在词汇上具有相似性[5].

然而,真实网络环境中锚文本的制作是不受控制的,这就导致锚文本中往往蕴含大量噪声,其中包括许多以作弊为目的的超链接[6].这些无用的信息不仅会浪费搜索引擎的计算资源,甚至会误导搜索引擎给出低质量的检索结果[2].另外,超链接文档中词汇的分布虽然体现了网页制作者的群体智慧,但是在用户检索事务类查询[7]时,这种网页制作者的群体智慧往往与用户对目标网页的真实体验偏差很大(第 3.2 节将详细分析讨论这种现象).为了解决上述问题,有必要通过评估锚文本对检索的有效性筛选出有助于检索的高质量锚文本,并剔除无用或不符合用户真实体验的锚文本.

本文提出的方法是基于用户网络浏览点击行为评估锚文本的检索有效性,这样做是因为用户通过浏览点击锚文本完成网页浏览,大部分用户在点击锚文本时一般会经过一个思维决策的过程,所以用户的网络浏览点击行为与锚文本的质量之间有一定程度的联系.在第 3 节中,我们会详细讨论二者之间的联系,并挖掘用户网络浏览点击行为中的有效特征评估锚文本的检索有效性.

1 相关研究工作概述

在信息检索领域,文献[1]首先提出了使用超链接文档方式定位特定网站(即导航类查询的检索),在这种方式下,所有指向同一目标网页的锚文本被整合为一个超链接文档进行检索.文献[5]对超链接文档中的词汇进行了统计分析,比较了其与标题、网页内容的相似度.文献[4]在使用锚文本时同样使用了超链接文档的方式,与文献[1,4]不同的是,工作中没有使用 BM25 模型[8]而使用了语言模型[9].在上述研究中,超链接文档中不同词汇的权重被定义为该词汇在超链接文档中的词频.文献[2]提出了一种新的定义词汇权重的方式,这种方式使用了超链接的结构信息,不仅考虑了超链接是否来自同一网站,而且考虑了源网页与目标网页之间的关系.例如,是否是镜像站点,是否有合作关系等.

在与信息检索相关的其他领域中,锚文本也有不少应用.指向同一网页但不同语言的锚文本有助于跨语言信息检索[10].锚文本的目标分布可以作为查询分类的特征[11],该特征对导航类查询与其他查询有明显的区分度.文献[12]在研究中扩展了这种方法,进一步处理了查询与锚文本不完全匹配的情况.也有一些研究主要关注锚文本对查询优化的作用[13],这些研究表明,锚文本产生的查询优化结果可以作为查询日志产生结果的补充.文献[14]提出了利用锚文本的结构与内容自动描述网站的方法.

利用锚文本生成超链接文档的思想近年来也逐渐被应用于其他类似锚文本的信息源,如社会标签(social bookmark)[15]、Click-through 数据[16]等.文献[15,17,18]尝试将社会标签应用于网络信息检索.文献[16]提出了使用 Click-through 数据生成查询描述文档来提升检索性能.

本文的主要内容是:分析用户网络浏览点击行为与锚文本检索有效性之间的联系,挖掘用户网络浏览点击行为中的有效特征,评估锚文本的检索有效性,筛选有助于检索的高质量锚文本,并剔除无用或不符合用户真实体验的锚文本.

2 用户网络浏览点击行为分析

用户对互联网络的浏览是通过浏览并点击网页上的锚文本来完成的,用户在点击锚文本的同时一般伴随着思考选择的过程.因此,在本节中我们将详细讨论用户网络浏览点击行为与锚文本检索有效性之间的关系,挖掘出其中对评估筛选高质量锚文本有区分度的特征.

2.1 数据准备

为了避免混淆,在本节中我们首先对一些基本概念进行解释,并对实验中所使用的数据逐一进行说明.表 1 中给出了一个锚文本的具体例子.源网页为新浪首页,目标网页为介绍基金的网页,锚文本“基金”作为源网页对目标网页的描述.

Table 1 An example of anchor texts

表 1 锚文本举例

Item	Content
HTML script	〈a href=“http://finance.sina.com.cn/fund/”〉基金〈/a〉
Source page	www.sina.com.cn
Target page	finance.sina.com.cn/fund
Source site	sina.com.cn
Anchor text	基金

一个源网页一般包含多个超链接,分别指向多个不同的目标网页.一个超链接指向一个目标网页,源网页上的锚文本是源网页制作者对目标网页的描述,用于引导用户进行网页浏览.需要注意的是,在本文的实验中,如果一个源网页包含多个对同一目标网页的超链接,我们只对其计算 1 次.这样做是因为一个网页对同一目标网页的多次链接并不能表明目标网页的重要性,同时也是为了屏蔽一部份借助这种方式的超链接作弊[6].

2.1.1 网页超链接数据

本文研究实验的基础是超链接文本,所以实验中首先需要一个网页集合,并从中提取出超链接信息.目前有很多用于学术研究的网页集合,但是这些集合一般有两个主要缺点:第一,规模较小;第二,包含许多“坏死”的超链接.因此,本文的实验中采用了大规模的网页集合,其中包含来自 3 百万个不同网站的 1.3 亿个网页,我们从中提取约 6 千万个不同的锚文本.该网页集合已经公开并可以按照网址 http://www.sogou.com/labs/dl/t.html 中说明的方式免费获取.采用大规模网页集合进行实验的目的,是为了让本文的研究更加贴近真实网络规模,从而使得本文的实验结论能够对真实规模的网络检索有意义.

2.1.2 用户网络浏览点击行为数据

随着网络搜索引擎的发展,浏览器工具条(Web browser toobar)已经成为记录用户互联网浏览行为的主要工具之一.用户的网络浏览点击行为已经应用于一些研究工作[19].本文中使用的用户网络浏览点击行为数据是由一家商用搜索引擎公司提供的真实用户网络浏览日志.该日志中所记录的单条用户浏览事件的信息项见表 2.

Table 2 Information items in a browsing event of user Web browsing logs

表 2 用户网络浏览日志中单条浏览事件的信息项

Item	Description
UsrID	User ID, unique label assigned for a user
SrcURL	Source URL which is visisted by a user
DstURL	Destination URL which is navigated to by a link
ClkAncText	Anchor text clicked by a user

在表 2 中,*UsrID* 是系统分配给用户的唯一标识.通过该 *UsrID*,可以对同一用户的日志信息进行聚集,实现对同一个用户搜索行为的分析研究.从表 2 中可以看出,日志中并没有记录任何涉及个人隐私的信息.本文使用了 2008 年 12 月 1 日~2008 年 12 月 31 日共 30 天的用户网络浏览数据,其中包含约 40 亿条用户网络浏览事件,共涉及来自 200 万个不同站点的 1 亿个不同的网页.

2.1.3 搜索引擎用户查询日志

为了研究用户网络浏览点击行为与锚文本检索有效性之间的关系,本文的实验中还需要获取用户在检索特定网页时所使用的查询.因此,我们采用了一家商用搜索引擎公司提供的用户查询日志.该日志中所记录的单条用户查询事件的信息项见表 3.

Table 3 Information items in a user searching event of query logs

表 3 用户查询日志中单条查询事件的信息项

Item	Description
Q*uery*	Query submitted by a user
URL	URL clicked by a user in the search res lt list
U*rID*	User ID, unique label assigned for a ser

从表 3 中记录的信息可以获取用户在点击特定 URL(检索结果)时提交搜索引擎的查询.在本文的实验中,我们使用 2008 年全年共 365 天的用户查询日志(2008 年单月的用户查询日志目前已经公开,并可以按照网址 http://www.sogou.com/labs/dl/q.html 说明的方式获得).

2.2 检索有效性评估框架

文献[5]的研究表明,锚文本有助于提升检索性能是因为网页的超链接文档与用户检索该网页时所使用的查询在词汇上具有相似性.二者的相似度是衡量锚文本检索有效性的关键,也是我们在之后衡量用户网络浏览点击行为特征的基础.在本节中,我们将量化二者的相似度,并提出评估检索有效性的基本框架.

首先,我们将指向同一网页的锚文本整合为该网页的超链文档.具体的整合方法是:对于网络中每一条超链接$\langle d_s,d_t,a\rangle$,

$$AncDoc(d_t)=AncDoc(d_t)\cup a,$$

其中,d_s 表示源网页,d_t 表示目标网页,a 表示锚文本,$AncDoc(d_t)$表示目标网页的超链接文档.在超链接文档中,每一个词汇的权重为$|IncomLink(d_t,a)|$,$IncomLink(d_t,a)$表示所有以 d_t 为目标网页并以 a 为锚文本的超链接构成的集合,$|\ldots|$表示集合中的元素个数.

然后,我们将用户在检索结果中点击同一网页时提交搜索引擎的查询整合为被点击网页的查询点击文档.具体的整合方法是:对于用户与搜索引擎交互日志中的每一条交互点击事件$\langle u,q,d\rangle$,

$$QClkDoc(d)=QClkDoc(d)\cup q,$$

其中,u 表示用户,d 表示被用户点击的文档,q 表示用户提交搜索引擎的查询,$QClkDoc(d)$表示被点击网页的查询点击文档.在查询点击文档中,每一个词汇的权重为$|IncomClk(d,q)|$,$IncomClk(d,q)$表示所有以 d 为被点击网页并以 q 为查询的交互点击事件构成的集合.

通过比较同一网页的超链接文档与查询点击文档中词汇的分布,我们可以得知网页的超链接本文与用户检索该网页时所用查询的相似度.这种相似度越高,则表明该网页的超链接本文对检索的帮助也越明显.在本文中,我们使用 KL Divergence 加 Laplace Smoothing 的平滑方式度量超链接文档与查询点击文档词汇分布的相似度.具体的公式如下:

$$KL(p_q \| p_a)=\sum_{w\in V_q} p_q(w)\log\frac{p_q(w)}{p_a(w)},$$

其中,p_q 与 p_a 分别代表词汇 w 在 $QClkDoc$ 与 $AncDoc$ 词汇分布中的相对频次,V_q 表示查询点击文档中所有词汇构成的集合.相对频次的计算公式如下:

$$p_q(w)=\frac{tf_q(w)}{\sum_{w\in V_q} tf_q(w)},$$

其中,$tf_q(w)$表示词汇 w 在某网页的 $QClkDoc$ 中出现的次数.为了处理 $p_a(w)$的取值可能等于 0 的情况,在本文中使用了 Laplace smoothing 的平滑方法.

$$p'_a(w)=\frac{tf_a(w)}{|V_q\cup V_a|+\sum_{w\in V_a}tf_a(w)}.$$

引入这种平滑方法是考虑到该方法不会引入额外的参数,为实验造成不必要的麻烦.

需要注意的是,由于采用 KL 距离与相似度成反比,所以在本文中,检索有效性越高,KL 距离取值越小.

2.3 超链接的用户浏览特征

根据观察,用户在浏览网页时更加倾向于点击对目标网页描述更加客观、准确的锚文本.换句话说,用户在浏览网页时点击过的锚文本的检索有效性应该更高.为了验证用户网络浏览点击行为的此项性质,首先基于原始网页的锚文本生成超链接文档 *RawAncDoc*,再基于用户在网页浏览时至少点击过 1 次的锚文本生成另一种超链接文档 *ClkAncDoc*.之后,利用第 2.2 节中介绍的检索有效性评估方法对 *RawAncDoc* 与 *ClkAncDoc* 的检索有效性进行评估与比较.如果 *ClkAncDoc* 的检索有效性高于 *RawAncDoc*,则 *ClkAncDoc* 与 *QClkDoc* 的 KL 距离($KL(p_q||p_{ra})$)应该小于 *RawAncDoc* 与 *QClkDoc* 的 KL 距离($KL(p_q||p_{ca})$).

图 1 给出了 1 000 个随机抽样 URL 的 $KL(p_q||p_{ra})$与 $KL(p_q||p_{ca})$.这些 URL 根据 KL 距离取值由低到高排序.在图中,KL 距离的取值越高,说明锚文本与用户查询越不相似.从图中可以看出,代表 $KL(p_q||p_{ra})$的曲线失踪在代表 $KL(p_q||p_{ca})$曲线的上方,这说明被用户浏览过的锚文本对检索的作用比原始锚文本更加明显.

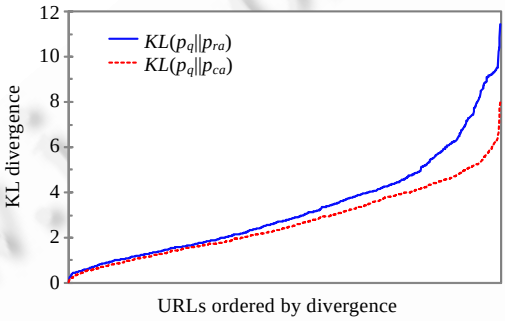


Fig.1 Kullback-Leibler divergence between the *ClkAncDoc*/*RawAncDoc* and *QClkDoc* distributions with Laplace smoothing applied to the query distribution

图 1 基于针对查询分布的 Laplace smoothing 平滑方法所得到的 *ClkAncDoc*/*RawAncDoc* 与 *QClkDoc* 的 Kulback-Leibler 距离

2.4 网页的浏览用户熵特征

如果以网页为粒度研究锚文本的检索有效性,根据我们的观察,某网页被越多不同的用户浏览,该网页上锚文本的检索有效性就越高.为了研究不同网页上锚文本的检索有效性,本文提出了网页的浏览用户熵(clicked user entropy,简称 CUE),用于度量某网页被不同用户浏览的倾向性.该特征的定义如下:

$$CUE(d_s)=-\sum_{u_i\in U}P([u_i,d_s])\log(P[u_i,d_s]),$$

其中,U 表示所有网络用户;$P([u_i,d])$表示网页 d_s 被用户 u_i 浏览的概率,计算公式如下:

$$P([u_i,d_s])=\frac{|BrwsEvent(u_i,d_s)|}{\sum_{u_i\in U}|BrwsEvent(u_i,d_s)|},$$

其中,$BrwEvent(u_i,d)$表示所有 *UsrID* 为 u_i 并且 *SrcURL* 为 d_s 的用户网页浏览事件构成的集合.

首先,将网页集合中的所有网页分为两个部分:$S1$ 与 $S2$,并保证$\forall d_i\in S1,\forall d_j\in S2, CUE(d_i)>CUE(d_j)$.然后,基于 $S1$ 与 $S2$ 分别构建两部分超链接文档 *CUEDoc*1 与 *CUEDoc*2.我们通过第 2.2 节中的评估方法对这两部分超链接文档进行检索有效性评估.

图 2 中,横坐标的数字表示 KL 距离的取值区间,纵坐标表示某取值区间内网页的分布.从图 2 中可以看出,

在 76%网页的 $KL(p_q||p_{cue1})$小于 1.5 的同时,有 56%网页的 $KL(p_q||p_{cue2})$小于 1.5;在 24%网页的 $KL(p_q||p_{cue1})$大于 1.5 的同时,有 44%网页的 $KL(p_q||p_{cue2})$大于 1.5.这说明,某网页越是倾向于被不同的用户浏览,该网页上的锚文本的检索有效性就越高.

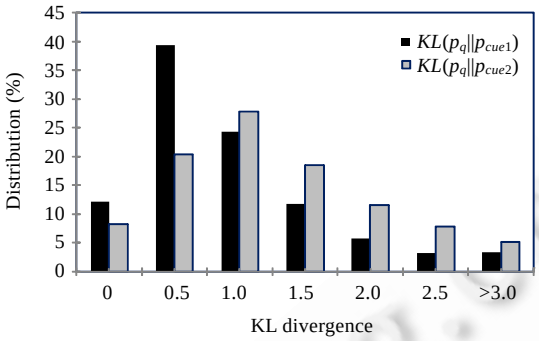


Fig.2 Distibution of Kullback-Leibler divergence between *CUEDoc*1/*CUEDoc*2 and *QClkDoc* distributions with Laplace smoothing applied to the query distribution

图 2 基于针对查询分布的 Laplace smoothing 平滑方法得到的 *CUEDoc*1/*CUEDoc*2 与 *QClkDoc* 的 Kulback-Leibler 距离的分布

2.5 网页的浏览分散度特征

根据我们的观察,用户在浏览某网页时越是倾向于点击不同的超链接(即用户的点击越分散),该网页上锚文本的质量就越高.为了量化该特征,本文提出了网页的浏览分散度(CAE)特征,用于度量用户对某网页浏览时点击的分散程度.该特征的定义如下:

$$CAE(d)=-\sum_{a_i\in A}P([a_i,d_s])\log(P[a_i,d_s]),$$

其中,A 表示所有锚文本;$P([a_i,d_s])$表示网页 d_s 上的锚文本 a_i 被用户浏览的概率,计算公式如下:

$$P([a_i,d_s])=\frac{|BrwsEvent(a_i,d_s)|}{\sum_{a_i\in A}|BrwsEvent(a_i,d_s)|},$$

其中,$BrwEvent(a_i,d)$表示所有 *ClkAncText* 为 a_i 并且 *SrcURL* 为 d_s 的用户网页浏览事件构成的集合.

首先,将网页集合中的所有网页分为两个部分 $S1'$与 $S2'$,并保证$\forall d_i\in S1',\forall d_i\in S2',CUE(d_i)>CUE(d_j)$.然后,基于 $S1'$与 $S2'$分别构建两部分超链接文档 *CAEDoc*1 与 *CAEDoc*2.我们通过第 2.2 节中的评估方法对这两部分超链接文档进行检索有效性评估.

图 3 中,横坐标的数字表示 KL 距离的取值区间,纵坐标表示某取值区间内网页的分布.从图 3 中可以看出,在 52%网页的 $KL(p_q||p_{cae1})$小于 1 的同时,有 31%网页的 $KL(p_q||p_{cae2})$小于 1;在 48%网页的 $KL(p_q||p_{cae1})$大于 1.5 的同时,有 69%网页的 $KL(p_q||p_{cae2})$大于 1.这说明,用户浏览某网页时越是倾向于点击不同的超链接,该网页上的锚文本的检索有效性就越高.

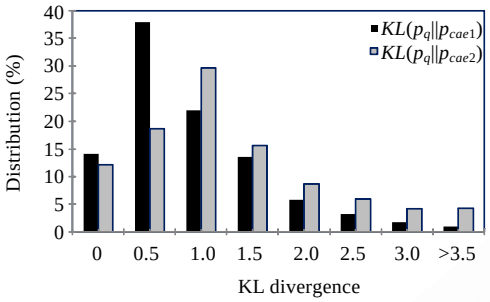


Fig.3 Distibution of Kullback-Leibler divergence between *CAEDoc*1/*CAEDoc*2 and *QClkDoc* distributions with Laplace smoothing applied to the query distribution

图 3 基于针对查询分布的 Laplace smoothing 平滑方法得到的 *CAEDoc*1/*CAEDoc*2 与 *QClkDoc* 的 Kulback-Leibler 距离的分布

2.6 用户体验与网页制作者推荐的不一致性

原始网页的超链接体现了网页制作者对网页的推荐,这种推荐在一定程度上存在合理性.特别是当目标网页是导航性质的网页(如门户或官方网站的首页)时,这种作用非常明显[1].然而,当目标网页是在线服务性质的网页时,这种网页制作者的推荐往往与用户实际对目标网页服务的体验有很大的偏差.例如,锚文本“机票预订”,其所有的目标网页中约有 70%指向“www.airtofly.com”,而只有 1%指向“www.qunar.com”,但是,如果从网页实际提供在线服务的质量、用户体验以及用户数量的角度来考虑,“www.qunar.com”在检索结果中应该排在“www.airtofly.com”的前面.这个例子说明了原始锚文本的一个缺陷:体现网页制作者推荐的原始锚文本不能客观地反映目标网页在线服务的实际质量以及用户的实际体验.如果我们只考虑被用户浏览过的超链接,则会发现,被用户浏览过的锚文本“机票预订”有 30%的目标网页指向“www.qunar.com”,而只有指向 2%指向“www.airtofly.com”.

为了量化说明网页制作者推荐与用户偏好之间的差异,本文定义了两个倾向性因子:(1) $RScore_{pa}(a,d)$,用于度量网页制作者通过锚文本 a 指向网页 d 的倾向性;(2) $RScore_u(a,d)$,用于度量用户通过锚文本 a 浏览网页 d 的倾向性.这两个倾向性因子的公式如下:

$$RScore_{pa}(a,d_t)=\frac{|IncomLink(a,d_t)|}{\sum_{d_t\in D}|IncomLink(a,d_t)|},$$

$$RScore_u(a,d_t)=\frac{|BrwsEvent(a,d_t)|}{\sum_{d_t\in D}|BrwsEvent(a,d_t)|}.$$

其中,$IncomLink(a,d)$表示所有以网页 d_t 为目标网页并以 a 为超链接本文的超链接构成的集合,D 表示所有网页集合,$BrwEvent(a,d)$表示所有 *ClkAncText* 为 a 并且 *DstURL* 为 d_t 的用户网页浏览事件构成的集合.

图 4 中给出了对锚文本“机票预订”与“在线地图”$RScore_{pa}(a,d_i)$与 $RScore_u(a,d_i)$的取值.图中的横坐标表示不同的 URL,并根据 $RScore_{pa}(a,d_i)$由高到低排序;纵坐标则是 $RScore_{pa}(a,d_i)$与 $RScore_u(a,d_i)$的取值.以“在线地图”为例,根据 $RScore_{pa}(a,d_i)$,大约有 60%的目标网页指向“www.51ditu.com”与“www.52maps.com”;但是根据 $RScore_u(a,d_i)$,23%的用户倾向于浏览指向“map.google.com”锚文本,10%的用户倾向于“map.sogou.com”,15%的用户倾向于“map.baidu.com”.这些用户倾向性较高的目标网页的综合服务质量与用户体验都优于“www.51ditu.com”与“www.52maps.com”.

原始超链接推荐与用户体验的明显差异在目标网页为在线服务性质的网页时尤其明显.这也说明引入用户网络浏览点击行为特征可以提升锚文本对在线事务类查询的检索作用.第 2.7 节的实验也验证了这一点.

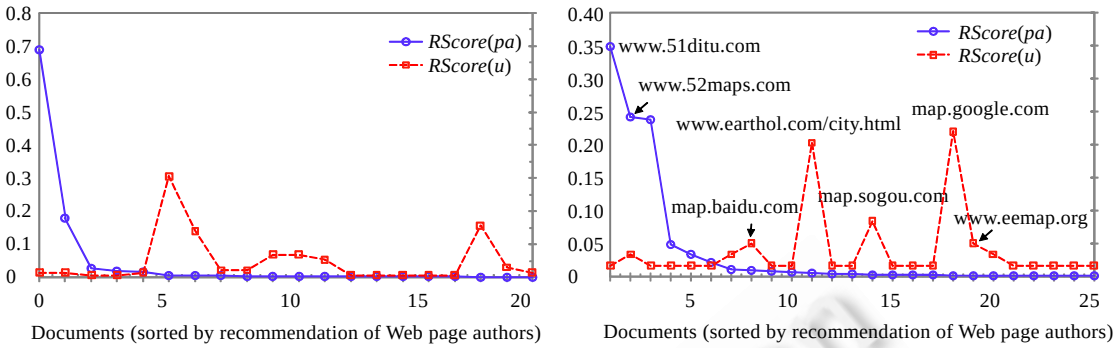


Fig.4 RScore of different destination Web pages with the anchor text “air ticket booking” and “map online” calculated using raw anchor data and anchor texts with Web users’ clicks respectively

图 4 对于超链文本“机票预订”与“在线地图”,基于原始锚文本数据与包含用户点击锚文本数据分别计算得到的各目标网页的 RScore 取值

2.7 实验与结果分析

2.7.1 实验设置和方法简述

在实验中,我们首先基于不同方式生成不同种类的超链接文档,再使用 Okapi BM25 模型[8]对不同种类的超链接文档进行检索,通过评价检索效果来评测不同超链接文档生成方式的优劣.本文使用 Okapi BM25 模型而不是其他检索模型是为了便于与其他相关研究进行比较.也是出于这个原因,在本文的实验中我们使用了与其他相关工作中同样的参数($k1$=2.0,b=0.75)[1,2].下面给出不同的超链接文档生成方式:

(1) 网页独立模型(PAM),使用了原始锚文本数据,认为来自不同网页的超链接彼此独立[1,4].

(2) 网站独立模型(SMM),使用了超链接的结构信息,认为来自不同网站的超链接彼此独立[2].

(3) 浏览链接模型(BLM),根据第 2.3 节与第 2.6 节的分析设计,只使用用户浏览过的锚文本数据.

(4) 浏览网页模型(BPM),根据第 2.4 节与第 2.5 节的分析设计,只使用满足条件网页的锚文本数据,网页需要满足的条件:$CUE(d)\times CAE(d)>c$,c 在本文实验中的取值为 37.4.

为了评测检索结果,我们使用了 3 000 个从查询日志中按查询频次分桶随机抽样的查询,这样做是为了使查询可以覆盖真实网络检索中各种频次的查询.实验中,3 000 个查询的查询频次如图 5 所示.可以看出,查询的选取对不同查询频次的查询都有很好的覆盖,并且各种频次的查询也符合真实查询日志中 Power-Law 的规律.对于每一个查询,我们收集了 4 家商用搜索引擎(百度、谷歌、搜狗、必应)前 20 位的检索结果,采用 Pooling 的方法形成 Pooling 池,并采用人工标注的方法对其进行标注.在检索性能评测时,我们采用了 TREC 评测中主流的评测指标 Mean Average Precision(MAP).

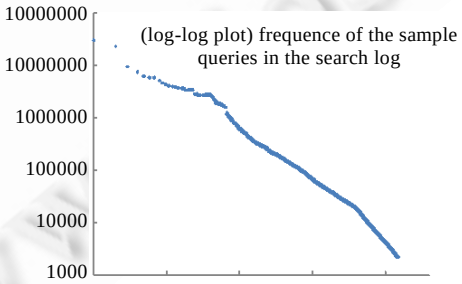


Fig.5 (log-log plot) frequence of the sample queries in the search log

图 5 抽样查询在日志中的查询频次(log-log 画法)

2.7.2 实验结果与分析

图 6 给出了不同模型对不同类型查询的检索性能,图中的纵坐标代表 MAP 的取值,横坐标表示不同类型查询.其中,All queries 表示所有查询,Navigational queries 表示导航类查询,Informational queries 表示信息类查询,Transactional queries 表示事务类查询.为了更加清晰地比较不同模型检索性能的不同,MAP 的具体数值以及相对增幅见表 4.

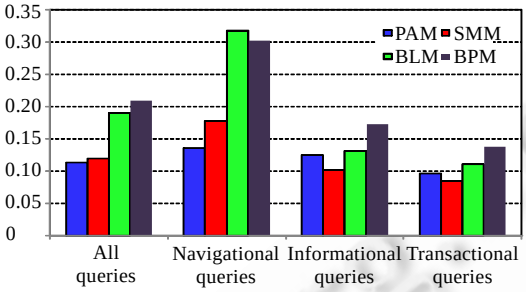


Fig.6 Ranking results using BM25 scoring over different types of search queries

图 6 用 BM25 模型评分后对于不同类型查询的排序结果

Table 4 Ranking results using BM25 scoring over navigational queries, informational queries and transactional queries

表 4 用 BM25 评分后导航类查询、信息类查询、事务类查询的排序结果

Models	All queries	Navigational queries	Informational queries	Transactional queries
PAM	0.113 (–)	0.1 (–)	0.125 (–)	0.096 (–)
SMM	0.12 (+6.2%)	0.178 (+30.9%)	0.102 (–18.4%)	0.085 (–11.5%)
BLM	0.19 (+68.1%)	0.318 (+123.8%)	0.131 (+4.8%)	0.111 (+15.6%)
BPM	0.209 (+85.0%)	0.302 (+122.1%)	0.173 (+38.4%)	0.138 (+43.8%)

从图 6 与表 4 中的数据可以看出:

(1) 基于第 2.3 节~第 2.7 节中介绍的用户网络浏览点击行为特征设计的锚文本生成模型(BLM 与 BPM)与 baseline 模型(PAM 与 SMM)相比,在检索性能上有明显的优势(对全部查询的性能提升分别为 68.1%与 85.0%),并且这种优势对于各种类型的查询都有体现(BLM 对各种类型查询的性能提升分别为 123.8%,4.8%,15.6%,BPM 对各种类型查询的相对性能提升为 122.1%,38.4%,43.8%),对导航类和在线服务类查询性能的提升更加明显.

(2) BLM 与 BPM 对导航类查询的性能比传统的 PAM 模型相对提升了 123.8%与 122.1%,表明用户网络浏览点击行为特征可以筛选出高质量的锚文本.这些锚文本对目标网页的描述更加客观、准确.

(3) SMM 对导航类查询的性能也有一定的提升(30.9%),但是与 BLM 与 BPM 相比,对检索性能的提升幅度还有差距,这也说明了本文提出的用户网络浏览点击行为特征相对于超链接结构特征的优点.

(4) BLM 与 BPM 对事务类查询的性能提升作用明显,相对于 PAM 的检索性能提升了 15.6%与 43.8%.这表明用户网络浏览点击行为特征筛选出的锚文本可以体现用户对目标网页服务质量的评价.

(5) SMM 对事务类查询的检索性能的提升基本没有贡献,反而对检索性能有所影响.这表明基于超链接结构信息的锚文本筛选特征并不能很好地提升事务类查询的检索性能.

为了验证模型对检索性能的提升作用具有统计有效性,针对文中检索性能的提升进行了 T 检验,其中,α取值为 0.05,具体的检验结果见表 5.

按照 T 检验的性质,如果 t 值(t Stat)小于临界值(t critical)或者 P 值小于 0.05,则说明检索性能的提升具有统计有效性.表 5 中的 T 检验数据无论是单尾(one-tail)还是双尾(two-tail),都已经满足了 T 检验的性质.这说明本文提出的基于用户网络浏览点击行为的特征筛选出的锚文本不仅能够提升网络检索的性能,而且性能的提升

具有统计有效性.

Table 5 T-Test results

表 5 T 检验的结果

Models	BLM-PAM	BPM-PAM
t Stat	−17.90	−19.42
P (*T*≤*t*) one-tail	1.25E−69	4.04E−81
t Critical one-tail	1.65	1.65
P (*T*≤*t*) two-tail	2.50E−69	8.09E−81
t critical two-tail	1.96	1.96

3 结论和未来的工作

虽然锚文本已经被证实是一种对网络信息检索非常有效的信息源,但是原始锚文本有两个难以解决的主要问题:第一,原始锚文本中蕴含着大量噪声与无用信息;第二,原始锚文本仅体现了网页制作者对目标网页的推荐,这种推荐往往与用户实际对目标网页的体验不一致.本文通过挖掘用户网络浏览点击行为中的有效特征,对锚文本的检索有效性进行评估,并根据挖掘的有效特征特征设计了两种超链接文档生成方法,基于新方法生成的超链接文档与原始超链接文档以及用其他方法生成的超链接文档相比,在检索性能上有明显的优势.本文的主要结论有:

(1) 用户浏览过的锚文本比其他锚文本质量要高,在生成超链接文档之后,对检索的作用更加明显.

(2) 浏览用户熵值与浏览分散度取值较高的网页,锚文本具有较高的可靠性.

(3) 用户浏览过的锚文本反映了用户对目标网页服务质量与实际体验的评价,这种评价弥补了原始锚文本仅体现网页制作者推荐的缺陷.

通过挖掘用户网络浏览点击行为中的有效特征,可以对锚文本的检索有效性进行评估,并筛选出高质量且体现用户体验的锚文本来帮助搜索引擎提升检索性能.未来的工作会沿着该方向展开,继续挖掘用户网络浏览点击行为中的其他有效特征,改进搜索引擎的检索性能.另外,未来的工作中会深化研究描述用户网络浏览点击行为的理论模型,进一步提升搜索引擎的检索性能.

References:

[1] Craswell N, Hawking D, Robertson S. Effective site finding using link anchor information. In: Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2001). New York: ACM, 2001. 250−257. [doi: 10.1145/383952.383999]

[2] Dou ZC, Song RH, Nie JY, Wen JR. Using anchor texts with their hyperlink structure for Web search. In: Proc. of the 32nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2009). New York: ACM, 2009. 227−234. [doi: 10.1145/1571941.1571982]

[3] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and Isdn Systems, 1998,30(1-7): 107−117. [doi: 10.1016/S0169-7552(98)00110-X]

[4] Westerveld T, Kraaij W, Hiemstra D. Retrieving Web pages using content, links, urls and anchors. In: Proc. of the 10th Text Retrieval Conf. 2001. 663−672.

[5] Eiron N, McCurley KS. Analysis of anchor text for Web search. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval (SIGIR 2003). New York: ACM, 2003. 459−460. [doi: 10.1145/860435.860550]

[6] Castillo C, Donato D, Gionis A, Murdock V, Silvestri F. Know your neighbors: Web spam detection using the Web topology. In: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2007). New York: ACM, 2007. 423−430. [doi: 10.1145/1277741.1277814]

[7] Broder A. A taxonomy of Web search. In: Proc. of the ACM SIGIR Forum. New York: ACM, 2002. 3−10. [doi: 10.1145/792550.792552]

[8] Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. In: Proc. of the TREC-3. 1995. 109−126.

[9] Ponte JM, Croft WB. A language modeling approach to information retrieval. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98). New York: ACM, 1998. 275−281. [doi: 10.1145/290941.291008]

[10] Lu WH, Chien LF, Lee HJ. Anchor text mining for translation of Web queries: A transitive translation approach. ACM Trans. on Information Systems (TOIS), 2004,22(2):242−269. [doi: 10.1145/984321.984324]

[11] Lee UC, Liu ZY, Cho JH. Automatic identification of user goals in Web search. In: Proc. of the 14th Int'l Conf. on World Wide Web (WWW 2005). New York: ACM, 2005. 391−400. [doi: 10.1145/1060745.1060804]

[12] Fujii A. Modeling anchor text and classifying queries to enhance Web document retrieval. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). New York: ACM, 2008. 337−346. [doi: 10.1145/1367497.1367544]

[13] Kraft R, Zien J. Mining anchor text for query refinement. In: Proc. of the 13th Int'l Conf. on World Wide Web (WWW 2004). New York: ACM, 2004. 666−674. [doi: 10.1145/988672.988763]

[14] Amitay E, Paris C. Automatically summarising Web sites: Is there a way around it? In: Proc. of the 9th Int'l Conf. on Information and Knowledge Management. New York: ACM, 2000. 173−179.

[15] Carman MJ, Baillie M, Gwadera R, Crestani F. A statistical comparison of tag and query logs. In: Proc. of the 32nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2009). New York: ACM, 2009. 123−130. [doi: 10.1145/1571941.1571965]

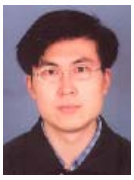
[16] Xue GR, Zeng HJ, Chen Z, Yu Y, Ma WY, Xi WS, Fan WG. Optimizing Web search using Web click-through data. In: Proc. of the 13th ACM Conf. on Information and Knowledge Management (CIKM 2004). New York, 2004. 118−126. [doi: 10.1145/1031171.1031192]

[17] Yusuke Y, Adam J, Satoshi N, Katsumi T. Can social bookmarking enhance search in the Web? In: Proc. of the JCDL 2007. New York: ACM, 2007. 107−116.

[18] Bao SH, Xue GR, Wu XY, Yu Y, Fei B, Su Z. Optimizing Web search using social annotations. In: Proc. of the 16th Int'l Conf. on World Wide Web (WWW 2007). New York: ACM, 2007. 501−510. [doi: 10.1145/1242572.1242640]

[19] Bilenko M, White RW. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). New York: ACM, 2008. 51−60. [doi: 10.1145/1367497.1367505]

周博(1982－),男,安徽安庆人,博士,主要研究领域为信息检索,机器学习.



金奕江(1970－),男,博士,工程师,主要研究领域为自然语言处理,网络信息检索.

刘奕群(1981－),男,博士,助理研究员,主要研究领域为信息检索.

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别与后处理,中文古籍数字化.

张敏(1977－),女,博士,副教授,主要研究领域为机器学习,信息检索.