计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130340
51(10):2239-2247,2014

搜索引擎索引网页集合选取方法研究

茹立云[1,2,3] 李智超[4] 马少平[1,2,3]
[1](智能技术与系统国家重点实验室(清华大学) 北京 100084)
[2](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)
[4](北京搜狗科技发展有限公司 北京 100084)
(lyru@vip.sohu.com)

Indexing Page Collection Selection Method for Search Engine

Ru Liyun[1,2,3], Li Zhichao[4], and Ma Shaoping[1,2,3]
[1](*State Key Laboratory of Intelligent Technology and Systems* (*Tsinghua University*), *Beijing* 100084)
[2](*Tsinghua National Laboratory for Information Science and Technology*, *Beijing* 100084)
[3](*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084)
[4](*Beijing Sogou Technology Development Co*, *Ltd*. *Beijing* 100084)

Abstract With the rapid development of the Internet, the number of pages is growing explosively. This presents a huge challenge for search engines which provide Web page search services. There are also lots of similar or even the exact same content pages and low-quality pages. In term of search engine, indexing such pages is no significant effect for retrieval results, but increases the search engine indexing and retrieval burden. A page selection algorithm is proposed to build indexing page collection from massive Web data for search engine. One hand, signature-based cluster algorithm is used to filter the similar pages to compress the size of the indexing page collection; on the other hand it combines a variety of features of the page dimensions and user dimensions, to ensure the quality of the collection. This algorithm is not only able to quickly cluster and select pages, but also achieve a higher compression ratio while still preserving the amount of information present in the indexing page collection. Experiments with actual page collections show that the size of indexing page collection selected by the proposed algorithm is about the entire page collection by 1/3, and can meet the vast majority of user click needs, with a strong practical.

Key words search engine; content signature; text clustering; machine learning; linear regression model

摘 要 随着互联网的快速发展,网页数量呈现爆炸式增长,其中充斥着大量内容相似的或低质量的网页.对于搜索引擎来讲,索引这样的网页对于检索效果并没有显著作用,反而增加了搜索引擎索引和检索的负担.提出一种用于海量网页数据中构建搜索引擎的索引网页集合的网页选取算法.一方面使用基于内容签名的聚类算法对网页进行滤重,压缩索引集合的规模;另一方面融合了网页维度和用户维度的多种特征来保证索引集合的网页质量.相关实验表明,使用该选取算法得到的索引网页集合的规模只有整个网页集合的约1/3,并且能够覆盖绝大多数的用户点击,可以满足实际用户需求.

收稿日期:2013-03-27;**修回日期**:2013-08-19
基金项目:国家"九七三"重点基础研究发展计划基金项目(2015CB358700);国家自然科学基金项目(60903107,61073071)

关键词 搜索引擎;内容签名;文本聚类;机器学习;线性回归模型

中图法分类号 TP391.3

随着互联网的快速发展,网页数量爆炸式增长.然而作为提供网页搜索服务的搜索引擎则面临着巨大的挑战.一方面,搜索引擎的索引能力有限,很难应对每天快速增长的新网页;另一方面,网页中存在大量的内容相同或相似的网页,同时索引这些网页对于检索结果并没有太大的意义,而且大量存在的低质量网页也会造成索引和检索资源的浪费[1].因此在面对海量的网页时,搜索引擎需要选择其中有价值的网页,而并不是将所有抓取到的网页都进行索引.从网页全集中选取索引网页的过程称为网页选取,其目的在于通过索引较小的文档集合来满足绝大多数的用户查询的相关性和多样性等需求[2-3].

文档选取有2个主要的策略:1)对网页进行滤重,在重复的网页中只选取一个进入索引;2)对网页质量进行评估,选取质量相对较高的网页进入索引.2种策略都能够降低索引网页的数量,只是着眼点不同,前者侧重于保留网页集合的信息量,而后者则侧重于保证网页集合的质量,在索引中去掉没有意义的网页.

在网页滤重方面,目前的研究方法主要有2种:一种是基于网页正文内容计算网页之间的相似度,再依据相似度来判断2个网页是否相同或相近的方法[4-7],这种方法的优点在于能够充分考虑网页内容的语义相似度,但是缺点在于要计算任意2个网页之间的相似度的复杂度太高[8].另一种是使用散列的方法对重复网页进行聚类[9-10],这类方法能够快速根据网页内容计算每一个网页的散列键值,并将键值相同的网页聚合在一起,选取其中一个网页以达到滤重的目的,但是这类方法并不便于将内容相似的网页进行聚类.本文将结合2种方法各自的优点,提出了一种基于网页签名的网页聚类算法,在网页中选取能够代表网页内容的重要词语和句子作为网页的签名来计算散列键值,这样既能够达到快速聚类的目的,又能够将具有相同重要句子的相似的网页进行聚合.

在网页质量评估方面,目前比较成熟的算法是PageRank算法[11].PageRank是基于网页之间的链接关系来得到的可以用来评估网页质量的重要参数.一般认为一个网页有越多的其他网页有链接链向它,那么它的PageRank值越高,也就意味着它的质量越高.在最近的研究中,有很多在传统PageRank算法的基础上进行改进的算法,比如TrustRank算法等[12-14].此外,在搜索引擎中还引入了反作弊等策略[15]来识别垃圾网页和低质量网页,并在文档选取的过程中过滤掉这样的网页.目前在高质量网页选取的过程中所使用的特征都是基于网页集合本身的,包括网页之间的链接关系和网页本身的结构和内容等.本文在此基础上引入了用户维度的特征,并泛化到站点级别的特征来帮助选取索引网页集合.实验证明结合了用户维度的特征,能够使选取到的网页集合更大程度地满足用户的查询需求.

1 数据集合及评价方法

为了真实反映网页选取算法在真实互联网环境中的效果,本文使用搜狗实验室发布的SogouT①数据来进行实验.SogouT数据包含了一个真实的互联网网页数据集合、网页之间的链接关系图以及一个月的搜索引擎用户查询和点击日志.

SogouT的网页数据集合中包含了1.3亿个不同的真实网页,网页总大小超过5 TB.搜索引擎用户日志则记录了用户在搜索引擎中的点击行为,数据为了保护用户隐私已将用户信息去除.每条日志只记录了点击时间、用户提交的查询词、用户所点击的URL以及该URL在搜索结果页中的排序等.用户日志共记录了超过5 000万次的用户点击.

本文算法将在SogouT的网页数据集合中选取索引网页集合.使用压缩比和覆盖率2个指标来评价选取算法得到的索引网页集合的效果.

所谓压缩比 $Compress$,即原始网页集合 S 中网页数目和选取网页集合 I 中网页数目的比值,即

$$Compress=|I|/|S|.$$

压缩比越大说明选取的网页数目越少,搜索引擎的索引和查询负担越小.当不进行网页选取时,所有原始网页集合中的网页都将进入索引,这时的压缩比为1.

覆盖率则是衡量索引网页集合满足用户需求的

① http://www.sogou.com/labs/resources.html

程度.选取 10 d 的搜索引擎用户查询点击日志来构建覆盖率的评价集合,剩余的日志用来作为训练数据.抽取用户点击的 URL,构成评价集合 $E=\{\langle url,count\rangle\}$.集合中的每个元素为一个 URL 和该 URL 在点击日志中出现的次数.则覆盖率为

$$Coverage=\frac{\sum_{u\in E\&u.url\in P}u.count}{\sum_{u\in E}u.count},$$

其中,u 为 E 中的一个元素,分母为评价集合 E 中所有点击 URL 总的点击次数,而分子只计算在索引网页集合中的 URL 总的点击次数.覆盖率越高,说明索引网页集合越能够满足用户的查询点击需求,选取算法的效果越好.当不进行网页选取时,所有的用户点击的 URL 都将被索引,这时的覆盖率为 1.

网页选取的目的在于在保证高覆盖率的前提下尽可能提高压缩比,减少搜索引擎索引网页的数目.或者说在高压缩比的前提下,尽可能收录用户点击的网页,满足用户查询需求.

2 基于文本内容的签名算法

在网页选取的过程中,对内容相同或相近的网页进行聚类是关键的步骤.在传统的文本聚类算法中,大多以计算文本之间的相似度作为聚类的依据.一个文本通常被描述成以词语为特征的特征向量,通过计算向量之间的距离作为文本之间相似度的度量.这种方法的优点在于文本相似度有比较明确的物理意义,任意 2 个文本之间都可以进行相似度的计算.而基于相似度的文本聚类的缺点也正是在于算法要求计算任意 2 个文本间的相似度,使得计算复杂度较高.而以特征向量的方式描述文本,损失了词语的位置和所在区域等信息,计算过程中丢失重要的语义特征.本文提出了基于网页内容签名的聚类算法,一方面从网页文本中基于不同的区域抽取特征句子和特征词语来表征网页内容,进行签名,保留了原始网页的内容信息;另一方面将相同签名的网页认为是内容相同或相近的网页进行聚类,免去了计算相似度的过程,大大降低了聚类的计算复杂度,同时也保证了聚类的准确率.

对网页进行签名,在选取网页内容特征时要保证所选特征能够体现网页内容的语义,这样才能将内容重复的网页聚成一类,达到网页滤重的效果.在选取网页特征时,不仅考虑词语和句子的重要程度,而且考虑区域的信息,将不同区域的内容进行区分,这样更能精确体现网页的原有特征,这些域包括"标题"、"正文标题"、"正文内容"、"出链接文本"等.

对于不同的域,基于域内文本长度的不同来决定选择词语或者句子作为内容特征来计算签名.在选择词语时,计算每个词语 t 在域内权重:

$$W(t)=\frac{\sum_{s\ contains\ t}length(s)+k}{\sqrt{f(t)}},$$

其中,s 为包含词语 t 的句子;$length(s)$ 表示句子 s 的长度;$f(t)$ 表示词语 t 在域内出现的次数;k 表示和域相关的常数.当词语 t 更多的出现在长句中,说明这个词语更重要,除以词语频度目的是降低停用词的权重.计算句子的权重:

$$W(s)=\frac{length(s)}{\sqrt{1+rank(s)/|S|}}\times\frac{|bigram(s)|}{bigramcount(s)},$$

其中,$length(s)$ 表示句子 s 的长度;$rank(s)$ 表示句子在域文本中的位置,句子越靠前,权重相对越大,说明句子越重要;$|S|$ 表示域中的句子数目;$bigram(s)$ 为句子中二元组的集合,所谓二元组即相邻两个字组成的字符串;$bigramcount$ 表示句子中二元组的总数.从所有的句子中选择权重最大的句子和权重最大的 N 个词语作为网页特征,计算签名.计算签名使用的散列算法是传统的 MD5 算法[16],MD5 算法可以极大程度地避免碰撞.

这种基于特征词语和特征句子的签名算法,能够对内容相近的网页计算得到相同的签名,能够将采集站在抄录其他网站时改动少数文字得到的网页和原始网页聚在一起,并通过类内选取,保留高质量的原始网页,达到网页滤重的效果.

3 索引网页集合选取算法

在网页选取过程中,除了对网页进行聚类并且滤重之外还要考虑网页的质量,在相同或相近的网页中尽量选取质量较高的网页进入索引.传统上评估网页质量的重要特征就是 PageRank,这是基于网页集合本身的链接关系计算得到的特征.而从用户搜索和点击的需求角度来讲,PageRank 的分布并不能完全拟合用户对搜索结果的点击.即用户也有搜索和访问 PageRank 较低的页面的可能,PageRank 只是从链接角度表征了用户访问到该页面的可能性,并不能完全表征网页质量,PageRank 较低的页面也有可能含有丰富的信息.所以仅从网页自身角

度的特征来作为网页选取的依据,具有一定的局限性.为了能够将更符合用户需求的页面选取进入索引,则需要引入更加贴合用户行为的特征.研究表明用户行为方面的特征在评估网页质量时也确实有着重要的意义[17-18].所以本文从搜索引擎用户日志中提取用户点击行为来计算网页质量,将页面及其站点被用户在搜索结果中点击的频度等作为特征,以弥补网页自身特征对用户需求拟合的不足,提高选取算法的效果.

因为选取要保证对用户点击的网页覆盖率,我们以网页被用户点击的概率来衡量网页的质量.这个概率通过训练得到的线性回归模型来计算.在训练线性回归模型时,所采用的网页特征包含网页维度的特征和用户维度的特征,这样能够将多方特征进行很好的融合.我们将没有用于覆盖率评价的搜索引擎用户日志平均分成两份,分别计算用户点击的网页得到2个网页集合,集合 C_1 用于统计网页的用户维度的特征,集合 C_2 作为训练集中的正例来学习线性回归模型.

我们用向量来描述一个网页 $\boldsymbol{p}=(x_1,x_2,\cdots,x_n,y)$,其中 x_i 为网页的特征,y 表示网页是否在集合 C_2 中,即这个网页是否被用户点击过.从SogouT的网页集合中随机挑选500万网页和集合 C_2 共同构成训练集合.线性回归模型也可以描述成向量 $\boldsymbol{M}=\langle c_1,c_2,\cdots,c_n\rangle$,其中 c_i 表示特征 x_i 的权重.一个网页 $\boldsymbol{p}$ 的点击概率为

$$prob(\boldsymbol{p})=\sum_{i=1}^{n}c_ix_i.$$

点击概率越高说明网页被用户点击的可能性越大,越需要被选入索引网页集合.

3.1 网页维度的特征

网页维度的特征都是通过网页集合本身计算得到的,主要包括网页自身的特征以及网页所在站点的特征2种.网页自身的特征包括PageRank、网页URL长度、URL中参数的个数、聚类之后网页在类内按PageRank排序得到的排名等.此外使用网页所属站点的特征,一方面同一站点的网页可能具有相似的重要程度,另一方面能加强模型的泛化能力.站点特征包括SiteRank、站点内所包含的网页数目等.

站点的SiteRank的计算方法和网页的PageRank计算方法类似,在构建链接关系图时,把站点作为关系图的节点,站点之间的链接作为关系图的边,站点 A 和站点 B 之间边的权重为

$$W(A,B)=|\{[a,b]|a\in A,b\in B,link(a,b)\}|,$$

其中,a 为站点 A 中的网页,b 为站点 B 中的网页,$link(a,b)$ 表示 a 中有链接指向 b.

在这个链接关系图中,使用传统的PageRank算法计算得到每个站点的Rank值就是这个站点SiteRank.网页所属站点的SiteRank值也可以一定程度表征网页的质量.

3.2 用户维度的特征

用户在搜索中点击的网页不一定都是PageRank很高的网页.我们将页面的PageRank值归一化到区间[1,20]以方便对比,图1给出了原始网页集合中PageRank的分布和用户点击的网页集合的PageRank的分布,可以看出即使网页的PageRank较低也会有很多用户点击,有将近10%的点击页面PageRank不大于3.这也进一步说明了PageRank并不能完全表征用户的检索点击的需求,所以在网页选取时引入用户维度的特征是很有必要的.

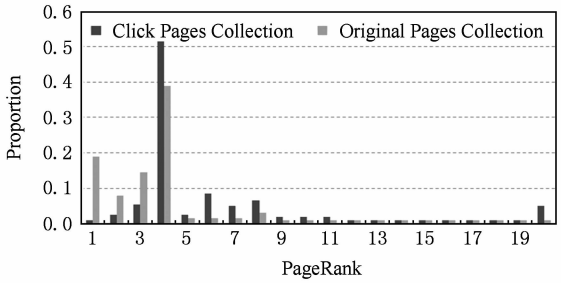


Fig. 1 PageRank distribution of original pages collection and click pages collection.

图1 原始网页集合和点击网页集合的PageRank分布

首先从搜索引擎用户日志中的集合 C_1 中可以得到每个网页每天被用户点击的次数.可以由此得到网页 $\boldsymbol{p}$ 的UserRank值:

$$UserRank(\boldsymbol{p})=\sum_{i=1}^{n}\frac{uv_i}{\sigma+1},$$

$$\sigma=n-\sum_{i=1}^{n}I(uv_i=0)I\Big(\prod_{j=1}^{i}I(uv_j=0)\Big),$$

其中,n 为日志的总天数,uv_i 表示在第 i 天一天内用户点击网页 $\boldsymbol{p}$ 的次数,当参数中的条件 $uv_i=0$ 成立时,函数 I 取值为1,否则 I 取值为0.σ 表示从有用户点击开始的第1天到最后一天的总天数.

和网页维度的特征类似,用户维度的特征也可以考虑网页所在站点的特征.将站点中所有网页的平均UserRank以及站点内被点击过的网页的总数作为特征加入到网页用户维度的特征中.

考虑搜索引擎日志的数目有限,我们对用户维度的特征进行了扩充.扩充方法主要考虑了搜索引

擎展现结果,因为在搜索引擎中排名靠前的网页被用户点击的可能性更大.所以可以通过展现结果来扩充用户维度特征,从而达到更好的训练效果.将页面在搜索结果中出现的频度以及页面所在站点中所有页面在搜索结果中出现的频度作为页面的特征.在实验中,从搜索引擎日志中随机抽取若干条查询词,在真实的搜索引擎中进行搜索,使用搜索结果的前10条来生成特征.

图2给出了抽取不同规模查询词后站点展现频度大于0的站点数,可以看出展现的站点数和查询规模基本呈线性关系.而图3给出了不同规模查询词所展现的站点对于用户搜索点击行为日志中所点击的网页的覆盖率情况,可以看出当查询词增加到10 000时,虽然其展现的站点数还在线性增长,但是对于用户点击来说,这些站点已经能够覆盖超过80%的用户点击,而且增长缓慢.一方面用户点击的站点相对比较集中,另一方面展现数据也能比较好地表征用户点击行为.所以在计算这个特征时选取了10 000个查询词展现结果作为基础数据.

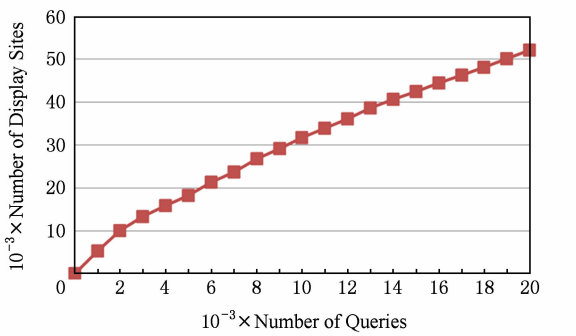


Fig. 2 Number of display sites on different size of queries.

图2 查询词数量和展现站点数的关系

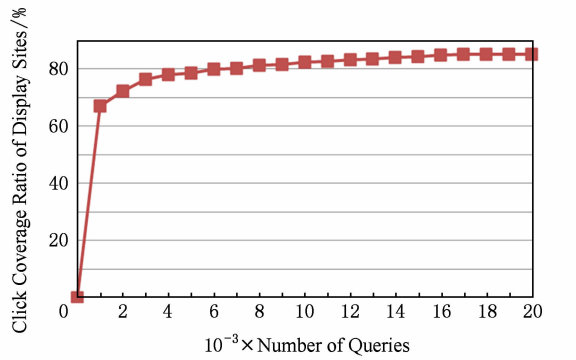


Fig. 3 Click coverage ratio of display sites on different size of queries.

图3 查询词数量和展现站点对用户点击数覆盖率

3.3 网页选取算法

在网页选取时,首先对网页进行聚类.传统的文本聚类算法依赖于文本之间相似度的计算,无论是K-means算法还是层次聚类算法等,都需要计算文本集合中两两实例之间的相似度,对于数据规模为N的文本集合来说,计算复杂度至少是$O(N^2)$.这对于数以亿计的互联网网页集合来说是不现实的.而基于内容签名的聚类算法则无需计算网页之间的相似度,只需将签名相同的网页认为是同一内容的聚在一类即可达到网页选取的效果.这样聚类过程可以近似看作是一个数据散列的过程,在时间复杂度上相当于$O(N)$.

聚类之后在每个类中选取点击概率最高的网页进入索引网页集合.算法1中给出了网页选取算法的描述,输入数据为整个网页集合,输出数据为从输入数据中选取得到的索引网页集合.

算法1. 文档选取算法.

输入:网页集合$P=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m\}$;

输出:选取集合S.

1)对P中每个网页$\boldsymbol{p}_i$

计算签名$\boldsymbol{p}_i.sig=Sig(\boldsymbol{p}_i)$;

在类别Hash表C中查找键值key为$\boldsymbol{p}_i.sig$的元素c_k;

if 不存在c_k then

生成一个新类别c_k,将c_k以$\boldsymbol{p}_i.sig$为键值加入到C中;

endif

将$\boldsymbol{p}_i$加入到c_k,并按照PageRank排序;

2)计算网页维度特征和用户维度特征,加入到特征向量$\boldsymbol{p}_i.\boldsymbol{V}$中;对$C$中每个类$c_j$中的网页排名,加入到特征向量$\boldsymbol{p}_i.\boldsymbol{V}$中;

3)使用线性回归模型对P中网页$\boldsymbol{p}_i$计算点击概率;

4)对C中每个类c_j中的网页按点击概率排序,将排在首位的网页$\boldsymbol{p}_i$加入到选取集合S中.

在整个过程中算法一方面去掉的是内容相同或相近的网页,从而不会因为缺少必要的数据而减低检索结果的效果和多样性,另一方面选取的是点击概率最高的网页,而不会使用户需要的重要网页过滤掉.

3.4 选取算法优化

在基于聚类的选取算法中,可能会出现页面质量都比较高的网页由于内容相似被聚在一起的情况,而对于不同用户来说,由于偏好不同而可能会倾向于不同的网页.比如不同新闻站点对同一新闻的报道虽然被聚在一类,但是都有用户点击的需求.如

果在一个聚类严格只选择一个页面,则会造成一定量用户需求不能满足.因此在基于聚类的选取算法中进行了改进,在一个聚类中选取多个页面.

在选取多个页面时,有2种策略:1)对于一个聚类选取点击概率最高的 K 个页面进入选取网页集合;2)如果一个聚类中非点击概率最高的页面如果点击概率超过一定阈值 t,则进入选取网页集合.实验表明第2种方法中当点击概率阈值选取得当的情况下,能够达到比较好的效果.

4 实验分析

本文使用SogouT作为实验数据集,SogouT的网页数据集合中包含1.3亿个不同的真实网页,网页总大小超过5 TB.搜索引擎用户日志则记录了用户在搜索引擎中的点击行为,用户日志共记录了超过5 000万次的用户点击.选取最初20 d的搜索引擎用户查询点击日志作为训练数据,余下10 d的日志用来构建覆盖率的测试集合.我们将搜索引擎用户日志训练数据平均分成2份,分别计算用户点击的网页得到2个网页集合,集合 C_1 用于统计网页的用户维度的特征,集合 C_2 作为训练集中的正例来学习线性回归模型.从SogouT的网页集合中随机挑选500万个网页和集合 C_2 共同构成训练集合,用于学习线性回归模型的参数.

4.1 压缩比

使用本文描述的网页选取算法,在SogouT上随机抽取的不同规模的数据集合上进行实验,得到的压缩比如图4所示.可见当文档集合越大时,文档选取算法的压缩比也就越大,能够节省的空间的比例也就越多.在SogouT的网页全集上,可以得到3.12的压缩比.这是由于在越大量的网页集合中,存在越多的相似网页,信息越趋于饱和.而如果用整个网页内容作为签名,只能得到1.09的压缩比,几乎对网页过滤没有什么贡献,因为内容完全一样的页面在整个集合中所占比例并不多.

Fig. 4 Compression ratio of selection algorithm on different number of pages collection.

图4 选取算法在不同规模网页集合的压缩比

基于内容签名的聚类算法的聚类准确率和网页在签名计算时选取的特征词语的数目密切相关.图5给出了选取不同规模特征词语对聚类准确率和对整体选取压缩比的影响.在评估准确率时,随机抽样1 000对被聚在一起的网页进行了人工标注.可见特征词语增加时,聚类准确率在增加,这说明增加特征词语数目之后,内容签名表征页面内容更加完整,更能将内容相同的页面计算得到相同的签名从而提高准确率.另一方面,对于压缩比,特征词语数量增加使得压缩比在降低,内容相近的网页有可能被判为不同的网页被重复选取,造成数据冗余.在实际应用时需要权衡对2方面的影响来选择合适的阈值.

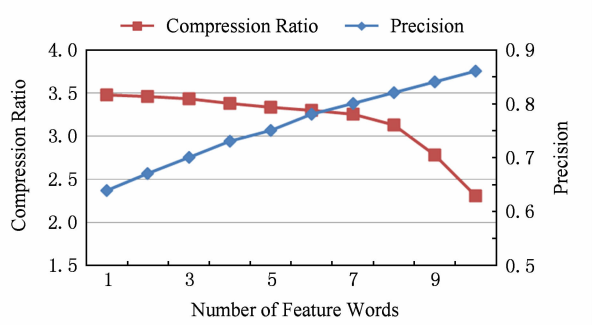


Fig. 5 Cluster precision and compression ratio on different size of feature words.

图5 特征词语数和聚类准确率及压缩比的关系

4.2 覆盖率

我们首先讨论不对网页集合进行聚类,而是直接依据网页质量进行选取,在固定压缩比为3.12时,比较不同选取算法的覆盖率指标.表1给出了在不进行聚类时,使用不同选取算法得到的索引网页集合的覆盖率.

Table 1 Coverage Ratio of Different Selection Algorithm without Cluster

表1 不聚类时不同选取算法覆盖率

Algorithm	Coverage Ratio
No Selecting	1.000
Randomly Selecting	0.322
Selecting Based on PageRank	0.778
Selecting Based on PageRank+UserRank	0.802
Selecting Based on Click Probability	0.836

可以看出在选取时,PageRank 能够表征页面的重要程度,相比随机选取的覆盖率提升 1 倍多,可以达到 0.778.而引入了用户维度的特征 UserRank 之后,能够在相同压缩比的情况下覆盖率达到 0.802.引入更多网页维度的特征和用户维度的特征,并通过线性回归模型计算得到的点击概率则更能体现用户查询和点击的倾向性,覆盖率能够达到 0.836.

表 2 给出了通过网页内容签名聚类后在每一类中选取一个网页的选取方法得到的覆盖率.可以看出相比不使用聚类算法覆盖率得到的比较明显的提升.这说明即使在点击概率比较高的页面中,也存在大量的重复内容,而索引这些重复的内容并不能为用户提供更多的信息量,不能提高对用户点击的覆盖率有所帮助.当对内容相似的网页聚类之后,可以过滤掉大量的重复网页,从而选取一些点击概率相对较低的页面,而这些页面的引入将提升选取网页集合的信息量,满足更多的用户查询和点击的需求.所以能够得到更高的覆盖率.这说明在网页选取算法中网页滤重起到了非常重要的作用.

Table 2 Coverage Ratio of Different Selection Algorithm Based on Content Signature Cluster

表 2 基于内容签名聚类时不同选取算法覆盖率

Algorithm	Coverage Ratio
Randomly Selecting One Page from Each Group	0.532
Selecting the Page with Largest PageRank from Each Group	0.861
Selecting the Page with Largest PR+UR from Each Group	0.891
Selecting the Page with Highest Click Probability	0.938

另一方面,无论是否使用网页聚类,在引入用户维度的特征之后,都对覆盖率有较为明显的帮助.而点击概率的效果也要优于只单纯使用 UserRank 和 PageRank 的方法,这是由于 UserRank 是针对页面级别的特征,如果页面没有在训练集的日志中被点击,那么这个页面就没有 UserRank,在进行选取时没有任何优势.而点击概率的计算模型中引入的站点级别的特征,即使一个页面没有被点击过,但是它同站点的很多其他页面被用户点击过,说明这个站点的质量比较高,那么这个页面也将在点击概率上得到加强,从而提升了选取的覆盖率.所以聚类之后,在每一个类中选取点击概率最高的网页进入索引的方法,能够得到最优的选取覆盖率 0.938.

当在一个聚类中选取多个网页时,选取覆盖率还能得到进一步的提升.图 6 给出了类内选取数 k 的变化对选取覆盖率和压缩比的影响.可以看出当选取数增加时压缩比在减小,选取页面数据在增加,同时提高了选取覆盖率.当类内选取数达到 10 时选取覆盖率达到 0.958,但这时压缩比只有 1.70 左右,该方法并不能得到很好的过滤效果.图 7 给出了点击概率阈值 t 对覆盖率和压缩比的影响.同样地,当 t 降低时,被选取的页面数增多,选取覆盖率增加,压缩比降低.当 t 选择在 0.7 时选取覆盖率增加到 0.973,而压缩比依然保持在 3 以上,这说明在达到同样覆盖率的条件下,使用点击概率阈值的方法要比使用选取数阈值的方法能够得到更高的压缩比,这也说明了点击概率对用户点击的预测具有比较好的效果.

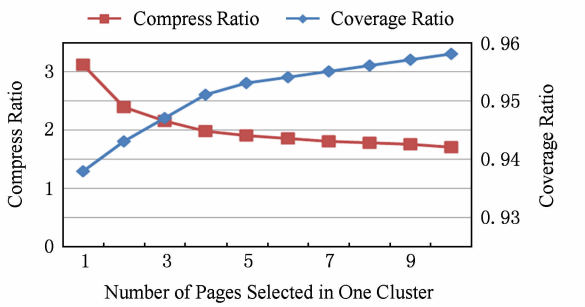


Fig. 6 Compress ratio and coverage ratio on different number of pages selected in one cluster.

图 6 类内选取数和压缩比、覆盖率的关系

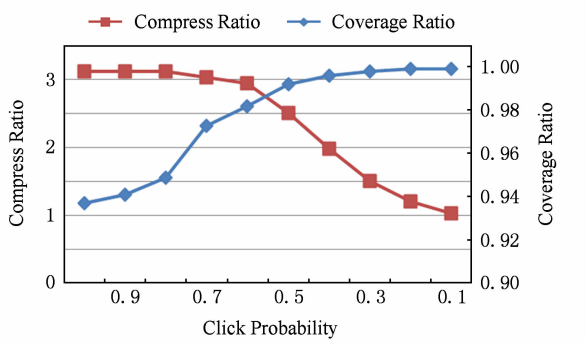


Fig. 7 Compress ratio and coverage ratio on different click probability.

图 7 点击概率阈值和压缩比、覆盖率的关系

在一个聚类中选取多个网页的方法,一方面解决了用户不同需求的高质量相似页面的选取问题,另一方面也从一定程度上缓解了由于签名聚类错误造成的信息量损失的问题.从而提高了选取覆盖率.

实验表明本文提出的网页选取算法是非常有效的,能够在仅选取 1/3 左右的网页集合的同时满足 90%以上的用户点击需求. 尤其是在对聚类之后的类内选取多个页面的策略对于覆盖率的提升也很明显.

5 结论和未来工作

本文提出了一种搜索引擎索引网页集合的选取算法. 算法对网页不同域的文本进行特征词语和特征句子的选取来表征网页内容,计算签名并进行聚类,在聚类的基础上引入了用户维度的特征,并泛化到站点级别的特征,通过线性回归模型来计算点击概率,选取点击概率最高的网页来构建索引网页集合. 算法不但能够快速地对网页集合进行聚类并进行选取,还能够在保证选取网页集合的信息量的同时得到较高的压缩比. 对实际网页集合进行的选取实验可以表明算法的有效性. 尤其是在对超大规模的网页集合进行处理时更能够得到优秀的压缩比.

参 考 文 献

[1] Agrawal A, Husain M, Tiwari R G, et al. A novel technique for database selection and document selection [J]. International Journal of Computer Applications, 2011, 17(8): 22-26

[2] Lin H H, Zhang Y, Davis J. Best document selection based on approximate utility optimization [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 1215-1216

[3] Welch M J, Cho J, Olston C. Search result diversity for informational queries [C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 237-246

[4] Broder A Z. Identifying and filtering near-duplicate documents [C] //Proc of the 11th Annual Symp on Combinatorial Pattern Matching. Berlin: Springer, 2000: 1-10

[5] Broder A Z, Glassman S C, Manasse M S, et al. Syntactic clustering of the web [J]. Computer Networks and ISDN Systems, 1997, 29(8): 1157-1166

[6] Mathew M, Shine N D, Lakshmi T R, et al. A novel approach for near-duplicate detection of web pages using TDW matrix [J]. International Journal of Computer Applications, 2011, 19(7): 16-21

[7] Dubes R C, Jain A K. Algorithms for Clustering Data [M]. Englewood Cliffs: Prentice Hall, 1988

[8] Salloum M, Tsotras V J, Srivastava D, et al. Selection and ordering of candidate documents for effective query answering in XML databases [C] //Proc of the 5th Int Workshop on Ranking in Databases. New York: ACM, 2011: 201-207

[9] Ding Zhenguo, Wu Baogui, Xin Youqiang. Research of large-scale URL filter based on bloom filter [J]. New Technology of Library and Information Service, 2008, 3(3): 45-50 (in Chinese)
(丁振国, 吴宝贵, 辛友强. 基于 Bloom Filter 的大规模网页去重策略研究[J]. 现代图书情报技术, 2008, 3(3): 45-50)

[10] Charikar M S. Similarity estimation techniques from rounding algorithms [C] //Proc of the 34th Annual ACM Symp on Theory of Computing. New York: ACM, 2002: 380-388

[11] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web [OL]. [2013-03-01]. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[12] Wei Chao, Chen Fei, Xu Danqing, et al. A framework for web page quality evaluation [J]. Journal of Chinese Information Processing, 2011, 25(5): 3-8 (in Chinese)
(魏超, 陈飞, 许丹青, 等. 网页质量评价体系的研究[J]. 中文信息学报, 2011, 25(5): 3-8)

[13] Wang Canhui, Liu Yiqun, Zhang Min, et al. Topic-independent web high-quality page selection based on *K*-means clustering [G] //LNCS 3689: Proc of the 2nd Asia Conf on Asia Information Retrieval Technology. Berlin: Springer, 2005: 516-521

[14] Spirin N, Han J. Survey on web spam detection: principles and algorithms [J]. ACM SIGKDD Explorations Newsletter, 2012, 13(2): 50-64

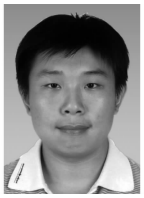
[15] Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating web spam with trustrank [C] //Proc of the 30th Int Conf on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2004: 576-587

[16] Rivest R. RFC1321: The MD5 message-digest algorithm [EB/OL]. 1992 [2005-08-12]. http://tools.ietf.org/html/rfc1321

[17] Singh D. Improving web search ranking through user behavior information [J]. International Journal of Information Technology and Knowledge Management, 2011, 4(2): 635-638

[18] Chen M, Yamada S, Takama Y. Investigating user behavior in document similarity judgment for interactive clustering-based search engines [J]. Journal of Emerging Technologies in Web Intelligence, 2011, 3(1): 3-10

Ru Liyun, born in 1979. PhD candidate of Tsinghua University. Member of China Computer Federation. His main research interests include Web information retrieval and machine learning.



Li Zhichao, born in 1985. Received his PhD degree from Tsinghua University in 2011. His main research interests include Web information retrieval and natural language processing (lizhichao@sogou-inc.com).

Ma Shaoping, born in 1961. Received his bachelor and PhD degrees in Tsinghua University. Currently professor and PhD supervisor in the Department of Computer Science and Technology in Tsinghua University. His current research interests include Web information retrieval, Web user behavior analysis and machine learning (msp@tsinghua.edu.cn).