海量 Web 搜索引擎系统中用户行为的分布特征及其启示*

王建勇 单松巍 雷 鸣 谢正茂 李晓明
(北京大学计算机科学技术系网络与分布式系统研究室,北京 100871)

摘要 统计分析了大规模搜索引擎系统的用户行为的分布特征. 结果表明,用户查询内容和 URL 点击表现出明显的局部性;用户查询的分布符合幂函数的特征并具有良好的自相似性. 基于上述规律,设计了查询 cache,比较了 FIFO, LRU 及带衰减的 LFU 等 3 种 cache 替换策略. 然后,基于用户行为考察了海量网页信息的分布特征,并利用 URL 的入度、镜像度、目录深度等网页参数与用户行为反馈后的相关度的方差分析,阐明了其对优化搜索引擎系统定序算法(ranking algorithm)的启示.

关键词 万维网 搜索引擎 分布特征 网页 用户行为

随着 Internet 技术和应用的不断发展,Web 页面的数目已经超过 10 亿[1),且以不到半年翻一倍的速度增长. 由于搜索引擎系统可以避免用户的盲目"冲浪",已经成为从"爆炸式"增长的 Web 信息资源中进行信息发现的重要工具之一. 但是现有搜索引擎系统所提供的 Web 查询技术还不能完全满足人们的需求,主要体现在:查询速度较慢、对数以万计的查询结果缺乏准确、有效的相关度评价. 当用户面临包含成千上万篇文档的查询结果时仍然存在"迷航"的问题.

一个典型的搜索引擎主要维护了两类信息,即有关网页的信息和日志文件中记录的用户行为信息. 前者指的是机器人从网上抓取的网页经过分析后得到的信息,主要包括网页所包含的关键词、摘要信息、元信息(如网页作者、长度、修改时间等)以及 URL 超链信息. 而后一类信息主要包括用户输入的查询项、用户在输出页面中所点击的感兴趣的页面和 URL. 这两类信息的数据量都很大,比如在"天网"系统[1]中,它们都已超过百万量级. 本文使用"天网"系统的日志记录,对用户行为以及海量网页信息的分布特征进行了研究,分析了某些规律性的东西,这些结论可以用于搜索引擎系统的设计,以提高系统的查询速度和信息检索的服务质量(查准率、召回率、定序的合理性等).

2000-07-21 收稿,2000-12-12 收修改稿
* 国家"九七三"重大基础研究发展规划项目(批准号:G1999032706)

1) Sullivan D. Search engines: Looking back, looking forward. In: Fifth Annual Search Engine Meeting Report, Boston, MA, Apr 1999. http://websearch.about.com/library/weekly/aa041700a.htm

1 研究背景

1.1 相关研究

搜索引擎的主要功能是根据查询项为用户从 WWW 上找到所需要的网页．一些传统的 IR(information retrieval)技术如文档的向量空间模型[2]和 tf * idf[3]算法为提高搜索引擎的检索质量起到了关键的作用．但由于 WWW 上网页质量参差不齐，其组织性和结构性较差，且检索信息的用户缺乏相关的技能和知识，人们逐渐认识到原有 IR 技术已不能满足搜索引擎系统的要求，转而试图利用网页信息本身的特点和用户行为来弥补传统 IR 技术在处理 Web 查询时的不足．

与传统 IR 面对的信息相比，网页信息有一个很大的特点就是其包含了大量的超链信息．如果我们把网页看作节点，超链看作有向边的话，整个万维网就构成了一张巨大的有向图(图 1)．Stanford 大学的 Google 搜索引擎系统[4]和 IBM 的 Clever 系统[1]的研究人员基于对该有向图的理解，提出了各自的理论模型以改进搜索引擎系统的检索质量．Google 系统用“随机冲浪”模型来描述网络用户对网页的访问行为，并采用称为 PageRank 的技术来计算网页的权值．IBM 研究院的 Clever 系统选择了被称为权威型和目录型的两类网页，并使用称为 HITS (hyperlink-induced topic search)的相应技术来计算网页的权威型和目录型权值，以改善搜索引擎的检索质量．

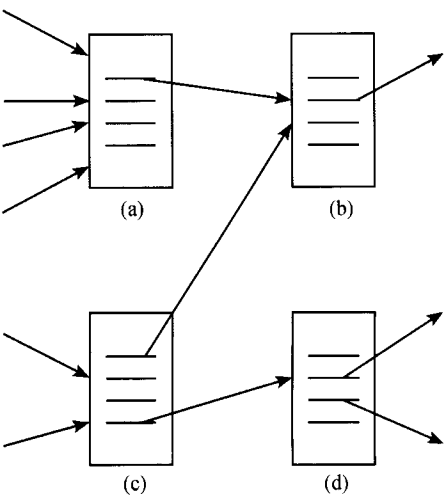

图 1　网页的链状结构

与传统 IR 的用户群相比，搜索引擎系统的用户经验少但数量巨大．国际上某些搜索引擎能够跟踪用户行为，来获取大量的有用信息，以便提高检索质量[2)]．Gray Cullis 将搜索引擎使用的 4 种信息——网页本身信息(author)、超链信息(other author)、人工编辑产生的目录信息(editor)和用户行为信息(user behavior)进行了比较[2)]，发现用户行为信息的利用对提高检索的准确率和召回率最有优势．这是可以理解的，因为用户是搜索引擎的直接使用者，也是服务质量好坏的最终评判者．

1.2 研究方法

虽然人们已认识到传统 IR 技术在处理网页信息检索时的不足，并开始利用网页信息和用户行为来改善搜索引擎的服务质量，但是到目前为止，尚未发现有相关的研究能够从深层次上来挖掘网页信息和用户行为信息，特别是通过分析海量网页信息以及用户行为的分布特征，来发现其潜在的某些规律，以更好地设计和评价一个搜索引擎，提高其服务质量，如系统的性能和检索质量．

举一个例子．Web 信息具有异质性和动态性，受时间和存储空间的限制，即使是最大的搜索引擎也不可能将全球所有的网页全部搜集过来．一个好的搜集策略是优先搜集重要的网

1) Members of the Clever Project. Hypersearching the Web. http://www.sciam.com/1999/0699issue/0699raghavam.html
2) Culliss Gray. User Popularity Ranked Search Engine. http://www.infomortics.com /searchengines/boston1999/culliss/index.htm

页,以便能够在最短的时间内把最重要的网页抓取过来. Junghoo 等人对网页的重要度作了假设[1),即包含了较多超链的网页是"重要"的. 他们以 Stanford 大学校园网内的网页信息为例,通过实验证明了如果依此假设来制定搜索导向,获取"重要"网页的速度要比没有搜索导向的快得多. 这种对网页重要度的定义是否合理,有没有更好的度量网页"重要度"的方法,都是本文要回答的. 我们认为某个网页是否"重要"最终是由用户决定的,并且将使用用户行为信息来对影响网页重要度的潜在因素进行评判.

为此,我们首先根据"天网"系统维护的日志数据,统计分析了用户行为的分布特征,主要包括以下内容:

(ⅰ)用户查询词的分布情况,

(ⅱ)雷同查询词的统计,

(ⅲ)相邻 N 项查询项的偏差分析,

(ⅳ)用户点击 URL 的分布情况,

(ⅴ)用户在输出结果中的翻页情况.

在分析上述统计结果时,发现用户行为表现出强烈的局部性,这启发我们采用查询 cache 和热点击 cache 来提高系统性能. 我们以日志中的用户行为作为输入,模拟测试了 FIFO, LRU 以及带衰减的 LFU 等 3 种 cache 替换策略的缓存命中率,比较了它们的优劣. 另外,由相邻 N 项查询项的偏差分布是稳定的,我们猜想用户查询项的分布过程符合自相似性,进而采用数学方法从理论上对此进行了验证. 最后,我们根据"天网"系统所搜集的网页信息统计分析了网页信息的一些重要参数的分布特征,并利用 URL 的入度、镜像度等参数与用户行为反馈后的相关度的方差分析来度量这些参数对网页重要度的影响,以挖掘其对搜索引擎系统定序算法(Ranking algorithm)的一些启示.

2 用户行为的分布特征及其启示

2.1 用户行为分布特征的统计分析

2.1.1 用户查询词的分布情况

我们以"天网"系统于 1999 年 4 月 15 日至 1999 年 6 月 10 日期间所做的日志记录作为分析对象,首先统计了用户查询词的分布情况. 假设用户的查询词序列为

$$S_1 = \{q_1, q_2, \cdots, q_n\},$$

其中这 n 项查询词中共有 m 个不同的查询词,按其查询次数进行降序排序得到序列

$$S_2 = \{Q_1, Q_2, \cdots, Q_m\},$$

而

$$S_3 = \{C_1, C_2, \cdots, C_m\}$$

是与 S_2 对应的查询次数序列. 使用下式来统计 S_2 中前某个百分比(如 $x\%$)的查询词所对应的查询次数占总查询次数的比率 Y:

$$Y = \sum_{i=1}^{\lfloor m \cdot x/100 \rfloor} C_i \Big/ \sum_{j=1}^{m} C_j. \tag{1}$$

1) Cho Junghoo, Garcia-Molina Hector. Efficient Crawling Through URL Ordering. http://www-db.stanford.edu/~cho/crawler-paper/

计算结果如图 2(a)所示．可以看出，用户的查询词是非常集中的．例如，前 20% 的查询词的查询次数约占了总查询次数的 80%．对图 2(a)中的查询分布曲线进行函数拟合，得到其拟合函数，如图 2(b)所示．我们发现拟合函数具有幂函数的特征，其形式为

$$y = (-0.04103 + 1.01689x)^{0.1346}, \tag{2}$$

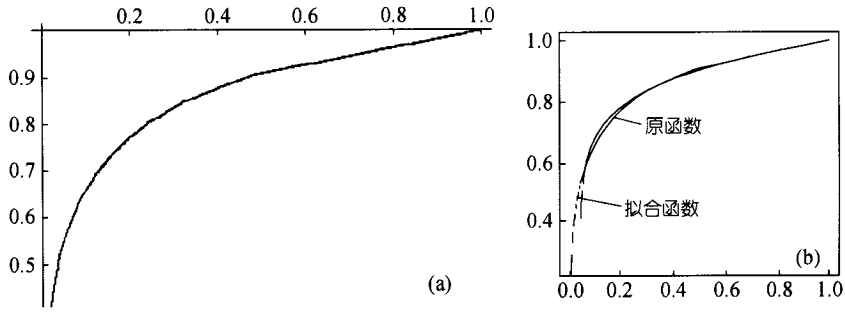

图 2　查询词的分布情况(a)和分布函数及其拟合函数(b)

这种幂函数($x^a(a<1)$)具有这样一个特征，即在 x 越接近 0 的地方，y 值增长越快，在 x 接近 1 的地方 y 的变化趋于平缓．这表明了查询词的分布具有很强的局部性：绝大多数用户查询的关键词落在了相对很小的一个集合上．

2.1.2　雷同查询词的统计

将序列 S_1 进行分组，每相邻 1000 项分为一组，并假设第 i 组的查询序列为 $A_i = \{q_{i1}, \cdots, q_{i1000}\}$，用 T_1 表示 A_1 中不同的查询项组成的集合，然后计算后面各组的查询项中有多少个查询项出现在 T_1 中，即对于 A_i，计算 Y_i 的值：

$$Y_i = \sum_{j=1}^{1000} c_{ij},\ C_{ij} = \begin{cases} 1, & q_{ij} \in T_1; \\ 0, & q_{ij} \notin T_1. \end{cases} \tag{3}$$

取不同的 i 值可以得到不同的 Y_i 值，其结果反映在图 3(a)中，其中横坐标表示组号，纵坐标表示该组查询项落在 A_1 中的个数．从图 3(a)可以看出，A_1 中的部分关键词或多或少地在其随后的多组(前 48 组)查询中也出现了，这表明用户的查询具有一定的稳定性．

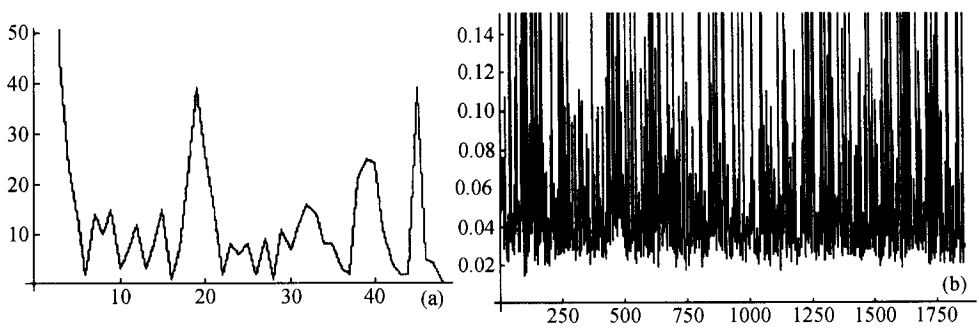

图 3　雷同查询词的统计(a)和相邻 1000 项查询词的频率的差平方和(b)

2.1.3 相邻 N 项查询词的偏差分析

对于相邻的两组 N 项查询项 A 和 B,假设 A 组中出现的不同的用户查询是($ab_1, ab_2, \cdots, ab_k, a_1, a_2, \cdots, a_n$),其中 ab_i 是 A 组和 B 组中所共有的,a_i 是 A 中出现但 B 中没有出现的查询. 而 B 组中出现的不同的用户查询是($ab_1, ab_2, \cdots, ab_k, b_1, \cdots, b_m$),其中 b_i 是 B 组中出现而 A 组中没有出现的.

A 和 B 中的这些不同的查询项构成一个向量空间

$$(q_{ab1}, q_{ab2}, \cdots, q_{abk}, q_{a1}, \cdots, q_{an}, q_{b1}, \cdots, q_{bm}).$$

我们假设某查询词 q_i 在 A 中出现的次数为 F_{ai},对其规整化后作为其特征项 A_i,这样就得到了 A 组的特征向量

$$\boldsymbol{A} = (A_1, A_2, \cdots, A_k, A_{k+1}, \cdots, A_{k+n}, A_{k+n+1}, \cdots. A_{k+n+m}),$$

其中 $A_i = F_{ai}/1000$.

同样的方法可以得到 B 组查询项的特征向量

$$\boldsymbol{B} = (B_1, B_2, \cdots, B_k, B_{k+1}, \cdots, B_{k+n}, B_{k+n+1}, \cdots, B_{k+n+m}),$$

其中 $B_i = F_{bi}/1000$.

$$S = \sum_{i=1}^{k+n+m} (A_i - B_i)^2. \tag{4}$$

计算特征向量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的差平方和(如(4)式所示),结果如图3(b)所示. 可以看出,大部分的差平方和都是在0.02到0.06之间. 这一方面说明了每相邻 N 项之间的查询相差不是很大,另一方面说明了每相邻 N 项查询之间的差别很稳定. 即用户的查询不但在短时期内偏差不大,具有短期的相关性,而且这个偏差也比较稳定.

2.1.4 用户在输出结果中的翻页情况统计

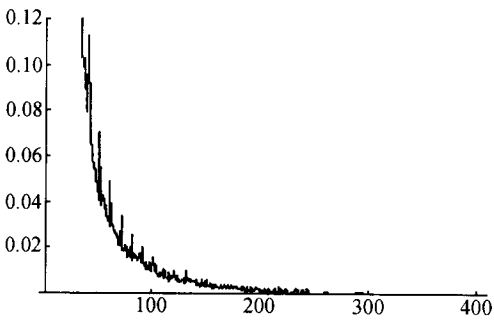

图4 用户翻页情况统计

我们用"天网"系统2000年4月份的查询日志来统计用户翻页情况和被点击URL的分布情况,其中该日志记录了近50万的用户点击情况(包括用户点击的URL及该URL所在输出结果中的页号). 假设"天网"系统中能够提供 n 个显示页面(在实际系统中 $n = 2000$,每个页面包含10个网页信息),用 $\{P_1, \cdots, P_n\}$ 来表示,它们对应的点击次数分别为 $C_1, \cdots, C_n$. 对第 i 个页面,我们根据(5)式计算其点击次数占总点击次数的百分比 Y_i. 得到的结果如图4和表1所示(横坐标是页号,纵坐标是该页面的被点击次数占总点击次数的比率).

$$Y_i = C_i / \sum_{j=1}^{n} C_j. \tag{5}$$

表1 用户在前5页的翻页统计

页号	1	2	3	4	5
百分比	47%	12.1%	7.4%	5.0%	3.7%

其中前面 5 页中 URL 点击次数占总点击次数的比例如表 1 所示，可以看出大部分的用户点击都落在前面几页中，如第一页的用户点击占总点击的 47%，而前面 5 页的点击占到了总点击的 75%以上．而图 4 表明用户很少浏览第 100 页以后的内容．这说明用户很少会在查询结果中翻很多页，用户一般就看看前面几页的内容而已．

2.1.5　用户点击 URL 的分布情况

基于 2000 年 4 月份“天网”系统的查询日志并使用 2.1.1 节中的方法得到用户点击 URL 分布的统计结果，如图 5(a)所示．其中横坐标是所选 URL 的数目占用户点击的 URL 总数的比率，纵坐标是所选 URL 的被点击数目占用户点击总数的比率．

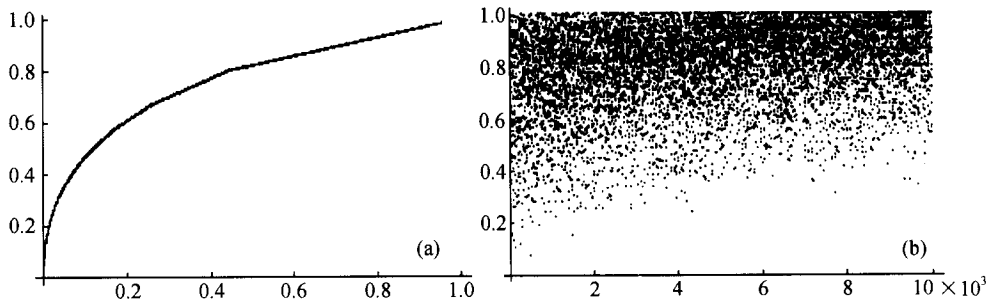

图 5　用户点击 URL 的分布情况(a)及考虑查询项与否的 URL 分布比较(b)

从图 5(a)的统计结果可以看出，用户点击的 URL 实际上也是非常集中的，“天网”系统的数据库共维护了 100 多万个有效页面，但是被用户点击的 URL 只有 16 万多个，还不到总的有效页面的 1/6．而且在被点击的页面中常被用户点击的也是相当集中的，不到 1/3 的页面的点击次数占到了总点击次数的 2/3．这表明了用户点击 URL 也具有很强的局部性．

2.1.6　考虑与不考虑查询项时用户点击 URL 分布情况的对比分析

用户的每一次点击都是在某个查询的结果中进行的，因而我们还可以把用户的点击和相应的查询联系起来考察被点击 URL 的分布情况．其具体方法是：将点击的 URL 按其对应的查询词分类，统计每个查询词下各个 URL 点击的次数．

进而我们可以对这种方法的统计结果与 2.1.5 节的统计结果进行比较．在针对查询项的统计结果中，每个查询词 Q_i 下每个被点击的 URL 页面 U_j 都有一个点击次数 W_{ij}，在不考虑查询的 URL 统计中，该 URL 也有一个点击次数 W_j．在考虑查询项的点击次数的统计中，某个查询项 Q_i 下的 URL 点击次数形成一个向量

$$\boldsymbol{P}_i = (W_{i1}, W_{i2}, \cdots, W_{in}),$$

同时，这些 URL 在不考虑查询项时的 URL 点击次数也对应着一个向量

$$\boldsymbol{P} = (W_1, W_2, \cdots, W_n),$$

对于每个查询项，计算这两个向量的夹角余弦值

$$\cos(\boldsymbol{P}_i, \boldsymbol{P}) = \frac{\boldsymbol{P}_i \cdot \boldsymbol{P}}{|\boldsymbol{P}_i| \times |\boldsymbol{P}|}. \tag{6}$$

这个夹角的余弦值越接近 1，说明两个向量的夹角越小，角度越接近，两个向量中各个分量占的比率越接近，即两种统计中，各个 URL 的点击次数占的百分比越接近．我们对用户点

击次数最多的10000个词按上述方法做了比较,比较结果如图5(b)所示(横坐标是查询词的编号,纵坐标是与该查询词对应的按(6)式计算的余弦值). 由图5(b)可以看出,针对大部分的查询词计算出来的余弦值都是在0.8以上,这表明,在大部分的查询项下URL的点击频率和在所有查询项下URL的总点击频率基本上是一致的.

2.2 查询项分布过程的自相似性分析

2.1.3节的统计结果表明N项查询项之间的频率的差平方和在长时间内比较稳定,似乎具有自相似性的特征,我们进一步分析了查询日志,以验证用户查询是否具有自相似性.

自相似性直观上说就是一组序列在很长的时间范围内表现出结构上的相似性,其主要特点是长期依赖性[5]. 下面我们首先引入自相似性随机过程的定义[6,7].

定义1 设X是一个广义平稳随机过程,其均值为μ,方差为δ^2,自相关函数为$\rho(\tau)$. $\rho(\tau)$具有形式

$$\rho(\tau) \to \tau^{-\beta}L(\tau), \quad \tau \to \infty, \tag{7}$$

其中$L(\tau)$是一个在τ趋于无穷大时缓慢变化的函数,即$\lim\limits_{\tau\to\infty}\frac{L(\tau x)}{L(\tau)}=1$,对所有的$x>0$成立. 现将$X$分为大小为$m$,非交叠的子块(聚合过程),用每个子块的均值所组成的序列表示一个随机过程,即

$$X^{(m)}(t) = (X_{tm-m+1} + X_{tm-m+2} + \cdots + X_{tm})/m, \quad t \geqslant 1. \tag{8}$$

对每一个m,$X^{(m)}$都表示一个广义平稳随机过程,而$\rho^{(m)}(\tau)$表示$X^{(m)}$的自相关函数. 如果对所有的m,聚合过程$X^{(m)}$有着和X完全相同的自相关函数

$$\rho^{(m)}(\tau) = \rho(\tau), \ m \geqslant 1, \tag{9}$$

则称X为一个(严格二阶)自相似的随机过程,其自相似系数为$H=1-\beta/2$.

验证一个随机序列的自相似性,直观上的办法是取不同的m值,看不同聚合过程的分布图形是否相似,是否仍符合一般自相似性序列的图形特征. 进而我们可以采用如下的数学方法从理论上验证一个随机序列是否是自相似性的[6]:

设$X=(X_1,X_2,\cdots,X_n)$为待验证的随机序列,μ是这个序列的均值,$S^2(n)$为这个序列的方差. 先按下式计算R/S统计值(rescaled adjusted range statistic):

$$R(n)/S(n) = [\max(0, W_1, \cdots, W_n) - \min(0, W_1, \cdots, W_n)]/S(n),$$
$$W_k = (X_1 + X_2 + \cdots + X_k) - k\mu, \quad k \geqslant 1, \tag{10}$$

然后计算$\lg(R(n)/S(n))$,对于不同的n,取横坐标为$\lg n$,纵坐标为$\lg(R(n)/S(n))$作图,各个点和原点的连线的斜率应该比较接近,如果用一条直线来拟合,这条直线的斜率如果在0.5到1之间,这个序列就满足自相似性,如果在0.7以上,这个序列就具有很强的自相似性了.

为从直观上观察查询分布图是否符合自相似性特征,先将用户的查询每500项分为一组,统计每一组中不同的查询项的个数,以组号为横坐标,不同的查询个数为纵坐标,得到图6(a). 然后调整组的大小为1000(即$m=2$),2000(即$m=4$),分别得到图6(b)和(c). 从这两幅图中可以看出查询序列聚合后的分布图形仍然保持了结构上的相似性.

接着,采用前面讲过的数学方法来严格地验证查询分布的自相似性. 基于"天网"日志中的数据,利用(10)式分别计算出$\lg(R(n)/S(n))$和$\lg n$,并以$\lg n$为横坐标,$\lg(R(n)/S(n))$为纵坐标得到函数图像如图6(d)所示. 可以看出几乎所有点都在斜率为0.58和0.82的直线之间.

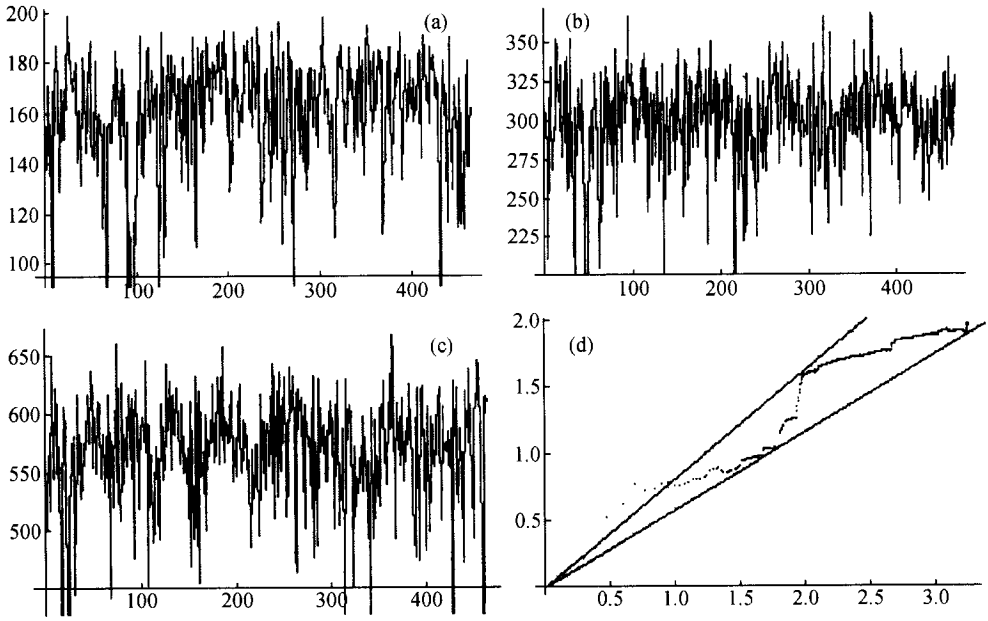

图 6　查询项分布及自相似性验证

a) 相邻 500 项中不同查询项的分布，(b) 相邻 1000 项中不同查询项的分布，(c) 相邻 2000 项中不同查询项的分布，(d) 查询项分布的自相似性验证

使用最小二乘法做直线拟合，求出拟合直线的斜率(Hurst 参数)是 0.67，即 Hurst 因子介于 0.5 和 1 之间，这样就验证了用户查询具备良好的自相似性. 2.1 节中的分析都是对“天网”某个时期的日志进行统计的，根据自相似性的特点，可以知道用户的查询是具有长期稳定性的，这样就可以将在前面的分析中得到的结果推广到“天网”长期以来的查询行为中，而且可以认为在今后相当长的时期，这些结果仍然有效.

2.3　用户行为的分布特征对搜索引擎系统设计的启示

2.3.1　用户行为启示

用户查询分布的局部性表明了使用查询 cache 的可行性：使用容量很小的 cache 就能命中大部分的用户查询，即可以用较小的空间取得较大的 cache 命中率.

用户雷同查询项的统计分析表明用户查询有一定的稳定性，说明了被缓存的查询信息具有长期有效性. 而相邻 N 项查询项的查询频率偏差很小且非常稳定，从另一个角度说明了查询 cache 的可行性：cache 替换过程不会因为用户查询短期内的变化而产生颠簸现象.

输出结果中翻页情况的统计分析表明用户通常只浏览前几页的内容，表明对输出结果进行排序是非常重要的. 越来越多的搜索引擎系统利用用户行为的反馈信息来提高查询结果排序的合理性. 它们通常根据一个 URL 被点击的次数来计算其权值. 但是如果把记录 URL 被点击次数的计数器放在磁盘上，会引起大量的磁盘 I/O，严重影响系统性能. 用户点击 URL 的局部性启发我们可以使用热点击 cache，这样用户点击过的 URL 及其点击次数和权值的修改都可以在一块较小的内存中完成.

然而如果在计算某个 URL 的被点击次数时没有具体到某个查询项，根据这样计算出来的 URL 权值进行输出结果的排序可能是不合理的. 如果对于每个用户查询项都维护一个 URL

列表,空间开销又会很大. 对查询项考虑与否的 URL 点击分布的统计分析表明,大部分查询项下的 URL 点击频率和所有用户点击过的 URL 点击频率是大致相同的,这样我们在实现热点击 Cache 的时候,就没有必要再去记录查询项的信息,只需要记录每个 URL 本身的信息即可,实现热点击 Cache 的空间开销和复杂程度就可以大大降低了.

另外,用户查询分布的自相似性表明其自相关函数是以双曲函数衰减的,即其具备长期依赖性. 它一方面说明了查询分布的局部性特征是长期有效的,为引入查询 cache 提供了理论基础,另一方面,类似于人们在发现了网络交通的自相似性后,利用自相似序列来测试 Web 服务器的性能,查询分布的自相似性对于设计和评价一个大型搜索引擎系统也具有重要的实用价值.

2.3.2 cache 替换策略研究

前面的统计分析表明了查询 cache 和热点击 cache 的可行性,下面以"天网"查询日志作为输入来对几种 cache 替换策略进行比较. 评测的替换策略包括FIFO, LRU 和 LFU 3 种,其中 LFU 是带衰减的 LFU,即每次发生替换时用某个衰减因子去衰减原来的查询次数并累加新的查询次数. 对于 FIFO 和 LRU,我们主要考察在不同的 cache 大小下的命中率,而对于 LFU,还考察不同衰减因子对 cache 命中率的影响.

图 7(a)示出了 FIFO, LRU 和 LFU 3 种策略下的 cache 命中率,其中对于 LFU,其衰减因子为 0.998. 从图中可以看出,当 cache 的大小到达一定程度时,cache 的命中率可以很高,例如当 cache 的大小为 500 时,这 3 种策略的 cache 命中率都达到了 60%以上. 同时,可以看出 LRU 和 LFU 替换策略下的 cache 命中率要比 FIFO 好得多,而在 0.998 的衰减因子下,LFU 与 LRU 的命中率相差不大,把图 7(a)中的部分结果局部放大,得到图 7(b),可以看出, LFU 比 LRU 略好一些,但效果不是很明显.

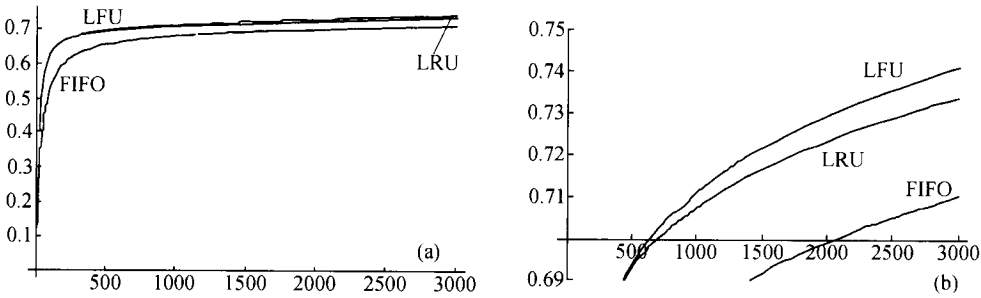

图 7 FIFO, LRU 和带衰减的 LFU 的 cache 命中率比较(a)和 3 种 cache 策略的局部比较(b)

实际上,LFU 在不同的 cache 大小下,取得最佳命中率的衰减因子都不同,如果在不同的 cache 大小下调整衰减因子,LFU 可以得到比 LRU 更好的命中率,表 2 显示了一些调整了衰减因子后的 LFU 的命中率,可以看出它要比 LRU 的效果好一些.

表 2 调整后的 LFU 与 LRU 命中率的比较

	cache 大小					
	100	300	500	1000	2000	3000
LRU 命中率	0.629381	0.680018	0.692691	0.707481	0.723485	0.733972
LFU 命中率	0.629934	0.680690	0.694037	0.711096	0.729509	0.741040

综合以上结论，LRU 和 LFU 的 cache 命中率要明显好于 FIFO，而如果选取好的衰减因子，LFU 可以得到比 LRU 稍微好些的效果．考虑到实现的复杂性，LRU 和 FIFO 相对简单，而 LFU 在发生替换时要进行衰减，须遍历 cache，其替换时间要远远大于 LRU 和 FIFO．所以，综合考虑这几种替换策略，LRU 是最好的选择．

3　基于用户行为的海量网页信息的分布特征分析

3.1　背景

2000 年 4 月上旬，“天网”系统搜集了 1000000 国内网页，这些网页立刻作为新的数据对外提供服务．在随后的 14 天时间里，有 141779 篇网页通过“天网”的引导被用户访问，总访问次数为 400641．哪些网页被访问，访问了多少次都通过日志文件被记录了下来．下面将基于这些用户信息来考察网页信息的分布特征及其与网页重要度之间的关系．

本文对网页重要度的度量规则定义为：用户访问越多的网页越重要．需要指出的是，用户点击 URL 的行为是受“天网”输出页面中结果排序的影响的（例如 75% 左右的用户点击落在前 5 个输出页面中）．输出页面中某个 URL 的权值是使用文档向量空间模型和 tf * idf 算法计算出来的[1]，它反映了该 URL 和用户查询项之间的相关程度，这种排序本身就有其合理性．而且用户在点击一个 URL 之前，通常先浏览该网页的摘要等信息，只有对网页内容感兴趣才去点击它．所以这种受相关度排序影响的用户行为能够很好地反映网页的重要程度．下面首先定义几个网页参数，然后再具体考察其与网页重要度之间的关系．

定义 2　网页 P 的入度 $H(P)$ 是指整个网络中指向网页 P 的超链数目．

定义 3　网页 P 的镜像度 $C(P)$ 是指整个网络中 P 的镜像网页个数．

定义 4　域名深度是指域名中包含的子域的个数，目录深度是指域名中所包含目录的层数．这里将域名深度和目录深度统称为目录深度 $D(P)$．

3.2　海量网页信息的分布特征分析

将“天网”维护的 100 万网页按照被用户访问的次数按降序排序，设该 URL 序列为 U_1，U_2，…，$U_{1000000}$，其对应的用户点击次数依次为 V_1，V_2，…，V_{999999}，$V_{1000000}$，它们对应的网页入度为 H_1，H_2，…，H_{999999}，$H_{1000000}$，网页镜像度为 C_1，C_2，…，C_{999999}，$C_{1000000}$，URL 目录深度是 D_1，D_2，…，D_{999999}，$D_{1000000}$．另外，增加一个参照序列，使它对每一个 URL 赋予同等重要度，即 S_1，S_2，…，S_{999999}，$S_{1000000}$，其中 $S_i = 1$．图 8(a) ~ (d) 分别示意了被用户访问的 14 万多网页按照被点击次数、网页入度、镜像度及目录深度进行排序后的分布情况．可以看出，通常被用户点击越多的 URL，其网页入度和镜像度也相对地越高，目录深度表现得则不是很明显．

对 100 万网页消除镜像后得到 868357 个有效网页，其中有 131906 个有效的被访问网页，然后分别计算出有效网页和有效的被访问网页的入度总和、镜像度总和以及目录深度总和，列于表 3 中．从表中的比例关系可以看出，被访问网页的入度、镜像度都大于平均数（15.19%），而目录深度略小于平均数．如果我们根据入度 $H(P)$、镜像度 $C(P)$ 和目录深度 $D(P)$ 来计算网页 P 的搜索权值 $W(P)$

$$W(P) = f(H(P),\ C(P),\ D(P)), \tag{11}$$

则 $W(P)$ 应当与 $H(P)$ 和 $C(P)$ 呈现某种正比关系，与 $D(P)$ 成反比关系．这表明网页入度和

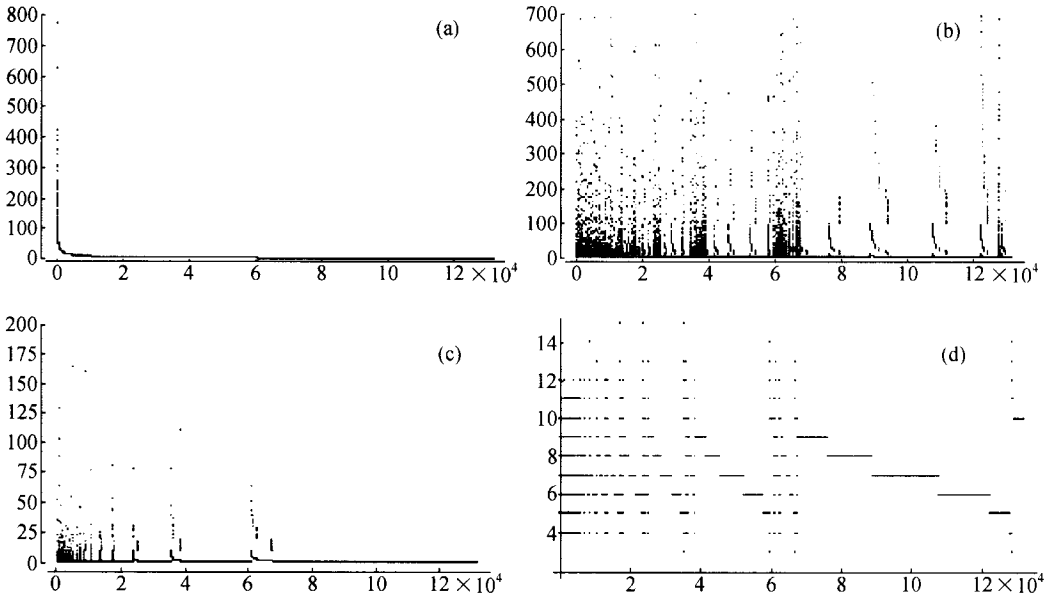

图 8 (a) 被访问次数, (b) 网页入度, (c) 镜像度, (d) 目录深度

镜像度越高,或目录深度越浅,网页重要度越大. 优先搜集短目录的网页还可以得到一个好处:避免了深度优先搜索,以搜集更多网站的重要网页.

表 3 各网页参数的分布

网页参数	总数	对被访问的部分求和	百分比
网页数	868357	131906	15.19%
入度	2876462	589868	20.51%
镜像数	999999	160896	16.09%
深度	6479796	947179	14.62%

把 URL 序列$\{U_i\}$对应的用户点击序列$\{V_i\}$、入度序列$\{H_i\}$、镜像度序列$\{C_i\}$、目录深度序列$\{D_i\}$以及参考序列$\{S_i\}$分别进行规整化,得到新的序列$\{V'_i\}$, $\{H'_i\}$, $\{C'_i\}$, $\{D'_i\}$和$\{S'_i\}$. 以序列$\{V'_i\}$为例,其规整化方法如下式所示:

$$V'_i = \frac{V_i}{\sum V_j}. \tag{12}$$

对于$\{H'_i\}$, $\{C'_i\}$和$\{S'_i\}$,我们分别求与$\{V'_i\}$的差平方和,得到

$$\text{mis\_}H = \sum (H'_i - V'_i)^2 = 2.604480 \times 10^{-4}, \tag{13}$$

$$\text{mis\_}C = \sum (C'_i - V'_i)^2 = 6.321719 \times 10^{-5}, \tag{14}$$

$$\text{mis\_}S = \sum (S'_i - V'_i)^2 = 6.298261 \times 10^{-5}. \tag{15}$$

可以看出,网页入度与访问次数的偏差最大,而镜像度与访问次数的偏差和参考序列与访问次数的偏差比较接近. 这一结果出乎我们的预料,尤其是我们认为网页入度最有可能与被访问次数相一致,以便被用作影响结果定序(result ranking)的因素. 这说明了在一般情况下,

URL 的入度与受用户查询相关度排序影响的用户点击行为呈现某种反比关系，在进行结果定序时不能简单地认为某个 URL 的入度越大，其检索权值就越高.

通过对实际数据的分析，发现国内有影响的网站的主页、技术文档和书籍的目录主页多获得了比较高的入度，而一般的网页入度都较低. 进一步分析发现，网站一般组织成类似树形的结构，如图 9 所示. 大多数载有文章的网页(图中的空心矩形)，它们极少被外站的网页所链接，在本站一般也只被一个网页所链接. 而主页(图中的椭圆)，它们既有超链指向站外，也被站外网页所链接. 图中的实心矩形是目录网页，它们被站外网页所指向，但自己的超链不向外指. 图 9 中的三角为一些专门向外指的网站，常被命名为“网络导航”，“友情链接”之类.

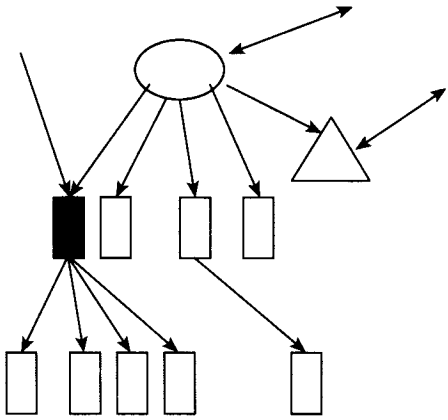

图 9　站内网页的树形结构

进而通过分析“天网”的用户查询日志发现，绝大多数用户查询是针对普通网页的，只有少量的是在查找一些站点主页，如北大、清华、中科院、新浪网和方舟等. 因而使用所有的网页入度来和被访问次数计算差平方和肯定是失败的. 在这一启发下，我们对 100 万网页中的 1 万多个主页重复求与被访问次数的差平方和，得到

$$\text{mis\_}H' = \sum (H'_i - V'_i)^2 = 7.16635 \times 10^{-4}, \tag{16}$$

$$\text{mis\_}C' = \sum (C'_i - V'_i)^2 = 6.93268 \times 10^{-3}, \tag{17}$$

$$\text{mis\_}S' = \sum (S'_i - V'_i)^2 = 7.17072 \times 10^{-3}. \tag{18}$$

从(16)～(18)式可以看出，网页入度成为与用户访问次数相关度最高的网页参数，镜像度与用户行为的相关度也较高. 综合(13)～(18)式的结果，得到如下启示：在搜索引擎提供服务时，应当将网站查询和一般网页的查询区分处理，一方面可以缩小输出结果范围，提高检索质量，另一方面可以为这两类查询采用不同的相关度定序算法. 如用户在进行网站查询时，除了根据查询项与网页的相关度计算该网页的基本权值外，还可根据其入度和镜像度计算附加权值，检索子系统综合这两个权值以进行结果定序. 而对于普通网页，计算附加权值时就可以不考虑网页入度.

4　总结

本文基于“天网”系统的日志数据，对用户行为的分布特征进行了统计分析，并使用用户行为信息考察了海量网页信息的部分网页参数，得到了如下的结论：

(ⅰ) 用户查询项和 URL 点击的分布具有强烈的局部性和稳定性，启示我们采用查询 cache 和热点击 cache 来缩短用户查询的响应时间，提高检索性能和并发查询支持能力；

(ⅱ) 对几种典型的 cache 替换策略的模拟测试结果表明，LRU 可以用比 LFU 更低的开销获得比 FIFO 更高的缓存命中率，因而是搜索引擎系统的优选方案；

(ⅲ) 用户查询过程的自相似性对于设计和评价搜索引擎系统将具有积极的指导意义；

(ⅳ) 本文对网页重要度的评价结论，启示我们可以利用入度、镜像度、目录深度等网页参

数来计算相应 URL 的搜集权值,以得到较好的搜索导向策略;

(V) URL 的入度、镜像度等参数与用户行为反馈后的相关度的方差分析,启发我们不能简单地利用网页入度来影响检索的附加权值,而应当区分网站主页和一般网页,并采用不同的权值计算方法,这将有利于改进检索质量.

目前,基于 LRU 替换策略的查询 cache 已经成功地应用于"天网"系统,并取得了很好的效果(近 70%的用户查询请求能够在 1 ms 内得到响应). 我们将进一步把 URL 的入度和镜像度作为影响计算网页权值的因素,重新设计网页的定序算法. 另外,今后我们还打算在更大规模信息量(如千万量级)的基础上重新验证本文的结论.

参 考 文 献

1 Liu J, Lei M, Wang J, et al. Digging for gold on the web: experience with the webgather. In: Proceedings of the 4th International Conference on High Performance Computing in the Asia-Pacific Region, Beijing, 2000. New York: IEEE Computer Society Press, 2000. 751 ~ 755

2 Salton G. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983

3 Church K W, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistic, 1990, 16(1): 22 ~ 29

4 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: Proceedings of 7th World Wide Web Conference. Amsterdam: Elsevier Science, 1998. 107 ~ 117

5 Crovella M, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes. In: Proceedings of 1996 ACM SIGMETRICS Conference, Philadelphia, PA, USA, May 1996. 160 ~ 169

6 Will E, Leland. On the self-similar nature of ethernet traffic (extended Version). IEEE/ACM Transactions on Networking, 1994, 2 (1): 1 ~ 15

7 赵晓芳, 刘 欣, 徐志伟. 网络交通自相似特性的分析及应用——具有单一登录点的机群网络服务器的性能评测. 计算机研究与发展,36(9): 1032 ~ 1038