

网页优化策略的模糊 C 均值(FCM)聚类算法研究^{*})

王玉龙 叶新铭 李秀华

(内蒙古大学计算机学院 呼和浩特 010021)

摘要 在对 Web 站点进行优化时,为了降低成本,往往需要在不改变硬件和网络配置的情况下提高网站的性能。此时,对构成网站的网页的修改就成为提高站点性能的主要途径。对网页的访问速度的测量已有很多成熟的方法,但是如何根据测试的结果指定合理的优化策略,却鲜有论述。本文使用 FCM 算法对测试结果和网站日志进行聚类分析,从而得到一个良好的优化策略。

关键词 Web, 优化, 模糊 C 均值(FCM), 聚类算法

Research on Fuzzy C-means Clustering Algorithm in Web Page Optimization Strategy

WANG Yu-Long YE Xin-Ming LI Xiu-Hua

(College of Computer Science, Inner Mongolia University, Huhhot 010021)

Abstract Web page optimization is an effective method to improve Web application performance. Before optimizing Web page we must know which one need optimization mostly. This paper puts forward a method which uses fuzzy C-means clustering algorithm to get the cluster centers of different Web page groups and a membership matrix in order to classify Web pages. With the help of the cluster center matrix and the membership matrix, we can pick out the Web pages whose optimization would lead to a giant improvement in performance and further provide advices on how to optimize those Web pages.

Keywords Web, Optimization, Fuzzy C-means (FCM), Clustering algorithm

1 引言

电子商务和电子政务的快速发展使得对 Web 应用的性能提出更高的要求。对于 Web 应用性能瓶颈的处理,也不再仅仅在应用开发阶段或集成测试时才进行。如何才能持续地提供高速的访问,成为电子商务/政务开发商和系统维护人员亟待解决的问题。

消除站点的性能瓶颈是提高站点性能的总策略。然而,具体施行方法的不同却会导致完全不同的结果。从投入回报率的角度出发,优化网页较之增加硬件资源和网络配置更合适,因此使用广泛。

然而,一个电子商务/政务网站可能包含成百上千个网页,如果盲目地进行优化,很可能事倍功半。实践证明,对 20% 的网页的优化,往往可以取得 80% 的性能提高。如何找到这 20% 的网页,是本文要解决的主要问题。本文根据网页的首次访问时间、二次访问时间、网页的访问频率和网页响应体大小等数据,使用 FCM 算法,对网页进行聚类分析,从而确定一个合理的优化策略。

2 基本概念

定义 1(聚类) 聚类就是将数据对象分组成为多个类或簇,使得在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大^[1]。

定义 2(模糊 C 均值聚类算法) 模糊 C 均值(FCM)聚类算法首先由 Dunn 于 1974 年提出,并由 Bezdek 于 1981 年改进^[2]。该算法能自动对数据对象进行分类并求出聚类中心和

每个数据点的隶属度,使得非相似性指标的目标函数达到最小,从而决定每个数据点的归属。

定义 3(网页首次访问时间) 从客户端第一次访问该网页所需时间,是访问该网页所发出的所有 HTTP 请求的响应时间的总和。

定义 4(网页二次访问时间) 在不清除浏览器缓存的情况下,从客户端再次访问该网页所需时间,是访问该网页所发出的所有 HTTP 请求的响应时间的总和。

定义 5(网页访问频率) 指定时间段内该网页被用户访问的次数。

定义 6(网页响应包大小) 对该网页的请求进行应答的所有 HTTP 响应包大小的总和,单位为字节。

对于哪些网页需要进行优化,不应由单一的指标决定。例如,对于访问速度慢但是访问频率极低的网页,优化就不是最迫切的。而对于首次访问速度慢,二次访问速度快的网页,也不一定需要优化。同时,最需要优化的网页,其各项指标需要满足哪些条件,也无法事先准确确定。因为没有有一个绝对的指标可以用来判断哪些网页的优化对网站性能的贡献度最大,所以我们采用模糊 C 均值聚类算法对所有被测网页进行聚类,从中推断出优化优先级高的网页,从而指导开发商以最小的劳动力代价获取最大的 Web 性能提高。

3 模糊 C 均值(Fuzzy C-Means, FCM)聚类算法的原理^[3]

FCM 算法基于求解下面的目标函数的最小值上:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, 1 \leq m < \infty \quad (1)$$

^{*} 基金项目:国家自然科学基金项目(60263002),内蒙古科技攻关项目(2002061002)。王玉龙 研究生。叶新铭 教授。李秀华 研究生。

其中 m 是任意一个大于 1 的实数, u_{ij} 是 x_i 在 j 簇中的隶属度, x_i 是 d 维被测量数据的第 i 个数据, c_j 是簇的 d 维中心, $\| * \|$ 是任何表达被测量数据与中心相似性的公式。

u_{ij} 还需满足如下条件:

$$\sum_{j=1}^C u_{ij} = 1, 1 \leq i \leq N$$

$$u_{ij} \in [0, 1], 1 \leq i \leq N, 1 \leq j \leq C$$

$$0 < \sum_{i=1}^N u_{ij} < N, 1 \leq j \leq C$$

模糊划分是通过用公式(2)(3)不断更新 u_{ij} 和 c_j , 而使上述目标函数不断优化得到的。

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

当 $\max_{ij} \{ |u_{ij}^{(k+1)} - u_{ij}^{(k)}| \} < \epsilon$ 时, 这个循环停止, 其中 $\epsilon (0 < \epsilon < 1)$ 是循环的终止条件, 而 k 是循环次数。这个过程收敛于 J_m 的一个局部最小值或鞍点。

算法由以下步骤构成:

1. 初始化 $U = [u_{ij}]$ 矩阵, $U^{(0)}$ 。
2. 在第 k 步, 使用 $U^{(k)}$ 计算中心向量 $C^{(k)} = [c_j]$ 。 c_j 由式(3)算得。
3. 更新 $U^{(k)}, U^{(k+1)}$ 。 u_{ij} 由式(2)算得。
4. 如果 $\|U^{(k+1)} - U^{(k)}\| < \epsilon$, 则停止循环, 否则执行第 2 步。

4 聚类有效性函数

聚类分析的最主要的缺陷是无论所给样本集的结构如何, 它总能将样本集进行分类。因此, 大多数聚类算法需要事先确定样本集的分类数。如果分类数选取得不合适, 则可能使划分的结果与样本集的真正结构不相符。关于样本集的最佳分类数问题, 属于聚类有效性问题。1995 年, Pal 与 Bezdek^[4] 比较了划分系数、Bezdek 的熵指标、Xie-Beni 指标、扩展的 Xie-Beni 指标及 Fukuyama-Sugeno 指标, 认为 Xie-Beni 指标具有较好的效果。Xie-Beni 指标的聚类有效性函数^[5]如下所示:

$$V_2^{\text{FCM}}(U; V; X) = \frac{\sum_{j=1}^n \sum_{i=1}^n u_{ij}^2 * d_{ij}^2}{n * (\min_{i \neq j} \{ \|v_i - v_j\|^2 \})} \quad (4)$$

$$\sigma_2(U, V, X) = \sum_{j=1}^n \sum_{i=1}^n u_{ij}^2 * d_{ij}^2 \quad (5)$$

$$\text{sep}(V) = \{ \min_{i \neq j} \{ \|v_i - v_j\|^2 \} \} \quad (6)$$

其中, n 为样本数, $V_2^{\text{FCM}}(U; V; X)$ 中的下标表示 FCM 算法中的加权指数为 2, d_{ij} 表示样本 i 与第 j 类聚类中心的距离。Xie-Beni 指标可以解释为 (U, V) 的总方差 σ (见(5)式) 与 V 的分异性指标 $\text{sep}(V)$ (见(6)式) 的比值。当 (U, V) 的总方差较小时, 则 u_{ij} 较大, 而 σ 较小。分类效果好的时候, 各类中心间的距离应该大, 即分异性指标 $\text{sep}(V)$ 比较大。由此, 当对应最佳类数 c^* 时, $V_2^{\text{FCM}}(U; V; X)$ 应该最小。

根据函数 $V_2^{\text{FCM}}(U; V; X)$ 确定最佳类数 c^* 的步骤如下:

1. 给定 c 的范围是 $[2, \sqrt{n}]$ 。这是根据很多研究者的使用经验和一些理论依据给出的^[6]。
2. 计算当 $2 \leq c \leq \sqrt{n}$ 时每个整数 c 所对应的 V 值。

3. 比较各 V 的值, 取 V 最小时所对应的 c 值即为所求。

5 实验分析

5.1 网页样本聚类结果

本文以一个信息发布网站为实验对象。使用日志分析工具 Web Log Explorer Pro 2.51 对该网站 1 个月的日志数据进行分析, 得到每个样本网页的访问次数 (即以一个月计的访问频率)。以访问频率前 100 位的网页作为样本网页。使用 Web 性能测试工具 OpenSTA 对这 100 个网页进行测试, 得到每个网页的首次访问时间、二次访问时间、网页响应包大小等数据。

将以上数据进行整理, 就可以得到一个 100 行 5 列的矩阵。这是一个 4 维的高维数据空间 (第 1 列是网页的 URL, 我们认为对特征分析没有意义, 因此舍去)。共有 100 个网页记录, 网页的属性分别为网页访问频率、首次响应时间、二次响应时间、网页响应包大小, 如表 1 所示 (为了便于论述, 第一列用 ID 号代表 URL)。

表 1 网页样本数据

ID	访问频率	首次访问时间	二次访问时间	网页响应包大小
001	4925	6085	20	166773
002	2352	630	20	152956
003	1055	40	1	630
004	489	5217	20	34376
005	424	460	10	34201
...	...	...	...	...

本文使用 Matlab 7.0.1 中的矩阵运算判定聚类有效性。表 2 中是根据(4)式计算出的不同的 c 值对应的 $V_2^{\text{FCM}}(U; V; X)$ 值。

表 2 聚类有效性函数值

c	$V_2^{\text{FCM}}(U; V; X)$
2	0.5095
3	1.2561
4	1.6433
5	2.0123
6	2.0226
7	2.1078
8	2.7706
9	2.7138
10	2.7208

在 $2 \leq c \leq 10$ 的条件下, 当 $c=2$ 时聚类有效性函数 $V_2^{\text{FCM}}(U; V; X)$ 的取值最小 (为 0.5095), 所以可以分为 2 类。

使用 Matlab 7.0.1 中的 FCM 函数对表 1 中的数据进行聚类, 分类数取 2, 得到聚类中心矩阵为:

$$\begin{bmatrix} 1060 & 2280 & 20 & 145930 \\ 0100 & 1450 & 10 & 39770 \end{bmatrix}$$

聚类中心点在各个维的取值表征了该类的特征。在本例中, 聚类中心矩阵的第一行表示第一类网页的访问频率大约是 1060 次/月, 首次访问时间大约是 2280ms, 二次访问时间大约是 20ms, 网页的响应包大小在 145930 字节左右, 而第二类网页的访问频率、首次访问时间、二次访问时间和网页响应包大小的数据分别在 100 次/月、1450ms、10ms、39770 字节附近。因此网页样本可以分类如表 3 所示。

表3 网页聚类结果分析

类别	网页样本群体特征类型	访问频率	首次访问时间	二次访问时间	网页响应包大小
第一类	急需优化网页。该类网页被用户访问次数明显高,而每次访问时间较长,此类网页的更新对网站的性能提升影响很大。	高	较长	较长	较大
第二类	不需优化的网页。该类网页访问频率低而且访问时间较短,此类网页的更新对网站的性能提升影响很小。	低	较短	较短	较小

5.2 隶属度矩阵 U

U 为一个 2×100 的矩阵,表示了这 100 个网页样本分别属于 2 种类型的隶属度。由于样本数据较多,在此抽取 3 个样本进行分析。

网页 001: $U_1 = [0.9721 \ 0.0279]^T$

网页 003: $U_3 = [0.0678 \ 0.9322]^T$

网页 011: $U_{11} = [0.5684 \ 0.4316]^T$

可以看到,矩阵 U 每一列的和为 1。这是因为每个网页的各类隶属度之和为 1。取每一列的最大值,最大值在第几行,该网页就属于第几种类型。例如,网页 001,毫无疑问属于第 1 类。网页 003 也取最大值,因此它属于第 2 类。网页 011 虽然属于第 1 类,但是从它的隶属度中可以看出,它还具有第 2 类的特征。这说明该页的情况比较复杂,在选择是否优化该页时应该根据具体情况再进行判断。

5.3 聚类结果分析

根据隶属度矩阵可以得到每一类中具体包含哪些网页。在本实验中,属于第一类的网页样本有:001、002、010、011、012、013、015、023、031、036、049、052、059、064、071、079,共 16 个网页,占样本空间的 16%。

到此,已经找到了应该进行网页优化的所有网页。但是,不同的数据可能得出不同的结果。如果在使用该算法对网页数据进行聚类时,得到需要优化的网页所占比例较大,还可以根据其隶属度对这些网页进行降序排序。然后根据实际需要,挑选出 20% 左右的网页进行优化。

此外,对于本实验数据,根据其中心矩阵还可以得到这样一个信息:第一类网页的响应包大小都比较大,而且首次响应时间和二次响应时间的差距很大。由此可以推断,导致这类网页访问速度慢的原因很可能是网页附带资源太多,而不是因为在 Web 服务器上生成该网页的计算时间过长。所以,对这 16 个网页进行优化时,重点应放在减少网页资源上。而对于其他类型网站的网页数据,可能能够得到不同性质的中心矩阵。例如,矩阵的某个行向量显示该类网页的点击率较高,首次响应时间和二次响应时间都较长,但是首次响应时间和

二次响应时间之间的差距很小,而且网页响应包的大小较小,则很可能意味着该类网页是一类需要大量后台计算或操作但是数据生成量较小的网页。对于这类网页的优化重点,就应该放在网页后台代码的优化上。

另外,对混合类型的网站(静态网页和动态网页都占有较大的比例)的网页进行聚类分析时,很可能得到 3 个或更多的分类。此时应该根据每一类所对应的中心矩阵的行向量的数据进行具体分析,从而指定合适的网页优化策略。

结论和未来工作 本文结合使用性能测试工具、日志分析工具,通过模糊 C 均值算法对目标网站的网页进行了分析,从而得到了对网页进行优化的良好策略。

在对网站进行性能优化时,开发人员总是希望投入最少的时间和精力而获得最好的结果。很多时候这是相互矛盾的。但是,在投入时间和精力一定的前提下,尽量获得更好的结果却是可行的。实践证明,80% 的性能提升往往来自 20% 网页的优化。然而,网页的多个属性都对性能产生影响。在事先不能确定每一个属性对性能影响的比重的情况下,无法通过单一指标的比较筛选出需要优化的网页,更无法给出优化时的重点。

本文通过模糊 C 均值算法,不仅筛选出需要优化的页面,同时对这类页面的特性进行了分析,以便指导对其进行优化的侧重点。

在下一步的工作中,我们将致力于实现一个自动化分析工具,使得日志的分析、网页响应时间和响应大小的获取、聚类的执行都在一个统一的工具中自动地进行。用户仅仅告知该工具网站的域名,并且允许该工具访问网站日志即可获得该网站的网页优化策略建议。

参考文献

- 1 An application of fuzzy clustering in group-positioning analysis [J]. Proc Natl Sci, Counc ROC(C), 2000, 10(2): 157~167
- 2 Michalopoulos M, Dounias G D, Thomaidis N T. Decision making using fuzzy C-means and inductive machine learning for managing bank branches performance [EB/OL]. <http://citeseer.nj.nec.com/458829.html>, 2002
- 3 Xie X, Beni G. A validity measure for fuzzy clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 814~847
- 4 Pal N R, Bezdek C. On cluster validity for the fuzzy c-Means model. IEEE Trans, 1995, 3: 370~379
- 5 Xie X L, Beni G. A validity measure for fuzzy clustering. IEEE Trans, 1991, 13(8): 841~847
- 6 于剑,程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学(E 辑), 2002, 32(2): 274~280