K-means 算法在关键词优化中的应用①

林元国, 许振和, 范智勇

(莆田学院 现代教育技术中心, 莆田 351100)

摘 要: 关键词的分析和优化是搜索引擎优化中两个最繁重的环节. 首先利用 K-means 算法简化对关键词的分析, 并在此基础上提出基于关键词效能和价值率的网站优化策略, 最后给出应用实例. 该方法能快速提升网站关键词的排名并带来一定的访问量, 适用于各类企业网站.

关键词: 搜索引擎优化; K-means 算法; 关键词优化

Application of K-Means Algorithm in Keywords Optimization

LIN Yuan-Guo, XU Zhen-He, FAN Zhi-Yong

(Modern Educational Technology Center, Putian University, Putian 351100, China)

Abstract: Analysis and optimization of keywords are the two most onerous aspects in search engine optimization. This paper firstly simplifies the analysis of keywords by K-means algorithm, then puts forward the strategy of website optimization based on keyword effectiveness and value rate, finally gives out an application example. The method can quickly enhance ranking of site keywords and bring traffic to the website. It is applicable to all kinds of enterprise website.

Key words: search engine optimization; K-means algorithm; keywords optimization

搜索引擎已成为广大网民获取信息的一个重要工具. 搜索引擎优化(Search Engine Optimization, 简称 SEO) 是指采用相关技术对网站进行一系列优化,从而提高相应关键词在搜索引擎上的排名, 最终达到网站营销的目的. SEO 归根结底是关键词的优化. 在市场多元化以及各行业消费主体个性化需求的影响下, 涌现出大量新的关键词(特别是长尾关键词). 一方面, 这些数量庞大的关键词给网站运营者带来潜在的商机; 另一方面, 针对这些关键词的分析和优化, 也让大部分 SEO 工作人员承担巨大的工作量.

目前国内外对关键词优化的理论研究和技术应用比较多, 主要涉及关键词优化技巧[1-5]、关键词分析方法[6,7]和 SEO 策略[8,9]等方面. 但暂未提出一个有效的方法来简化关键词分析流程, 也没有一个完善的机制来管理关键词优化策略和进度. 而 K-means 算法[10-12]作为一种得到广泛使用的聚类算法, 其最大的优势就是容易快速实现对大型数据集的聚类, 因此该算法也适用于对大量关键词的聚类分析. 本文将以某中小企业网站为案例, 利用 K-means 聚类算法提高关键词分析的效率, 并提出一个科学可行的关键词优化策略.

1 K-means算法简介

K-means 算法是一种基于划分的聚类算法, 属于非监督学习方法. 它是一种已知类别数的聚类算法, 所生成的每个聚类内紧凑, 类间独立. K-means 算法被提出后, 在不同的学科领域得到广泛的研究和应用, 并延伸出许多不同的改进算法[13-15].

K-means 算法的基本思想是以数据集中的 k 个簇为中心进行聚类, 按照最邻近原则把数据集所有对象分到各个簇. 通过迭代过程, 逐次调整各簇中心的值,

① 基金项目:莆田市科技计划(2012G06)

收稿时间:2014-04-29;收到修改稿时间:2014-06-09

直至得到最好的聚类结果. 一般采用误差平方和函数作为收敛准则, 即最好的聚类结果就是误差平方和最小的 K 个簇. 误差平方和准则函数定义如下:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i} d(x_i,m_i) \quad (1)$$

其中, m_i 是簇 C_i 的聚类中心; $\sum_{x\in C_i} x_i$ 表示第 i 个簇中心位置, i=1, 2, …, k; $d(x_i,m_i)$ 表示数据对象 x_i 到 m_i 的欧式距离.

K-means 聚类算法描述[15]如下:

输入: 簇的数目 k 和包含 n 个对象的数据集 D.

输出: 平方误差总和最小条件下的 k 个簇.

步骤:

① 在数据集 D 中随机选取 k 个对象作为初始的簇中心, 并将数据集 D 中剩余的对象赋给最类似的簇;

② 判断平方误差 E 是否有明显变化, 若有, 转③, 若无, 转⑤;

③ 重新计算每个簇中对象的平均值;

④ 遍历数据集 D, 将 D 中所有对象重新赋给最类似的簇, 转②;

⑤ 算法结束.

K-means 聚类算法容易理解和实现, 适用于处理大数据集, 且其具有时间效率高等优点, 数十年以来, 该算法在国内外已被应用到包括金融数据分类、空间数据处理、生物学、考古以及图像检测分析等众多领域[16-19].

2 基于K-means的关键词分析

2.1 关键词数据搜集

本文是针对谷歌搜索引擎进行相关的关键词优化研究, 因此关键词数据主要来源出自谷歌工具所搜集的和企业所需求的关键词. 首先以“触摸屏”为主题关键词, 利用谷歌关键词规划师工具(网址: https://adwords.google.com/ko/KeywordPlanner/Home)搜集到 231 个相关的关键词, 对应的数据项有本地每月搜索量(中国)、竞争程度和估算每次点击费用(CPC)等.

通过对企业和市场的调研分析, 筛选的关键词主要有“电容触摸屏”、“触摸屏厂家”、“触摸屏报价”、触摸屏技术原理以及触摸屏相关产品名称等; 同时去掉那些没有搜索量或脱离网站主题的, 最终得到 113 个待用的关键词. 最后利用谷歌搜索页面, 获得这些关键词对应的首页高质量网页数($Page_{1st}$)和搜索相关结果页面数(Intitle).

2.2 数据预处理

为了方便聚类, 对关键词的搜索相关结果页面数做一定的处理(Intitle/1000)而形成数据 Intitle'. 这样, 就得到一个 113*5 的矩阵 X(竞争程度、搜索量、Page1st、CPC、Intitle'), 其部分数据如下所示.

$$X=\begin{bmatrix} 0.41 & 2400 & 7 & 4.09 & 1820 \\ 0.38 & 1600 & 3 & 10.32 & 73 \\ 0.24 & 390 & 8 & 4.81 & 21.8 \\ 0.5 & 170 & 8 & 4.43 & 70.7 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

2.3 聚类结果分析

利用 MATLAB 7.1 软件和 K-means 算法, 将以上 113 个关键词数据聚为五类(即 k=5), 产生的 K-means 聚类结果如图 1 所示.

```
nr =

     8     2     4    80    19


    0.0003    0.2138    0.0061    0.0043    0.0311
    0.0004    2.0000    0.0050    0.0072    0.9465
    0.0003    0.0147    0.0075    0.0036    0.5122
    0.0002    0.0257    0.0042    0.0037    0.0157
    0.0003    0.0296    0.0046    0.0033    0.1007
```

图 1　关键词的 K-means 聚类结果

该聚类结果的数据形式亦可用表 1 来表达.

表 1　最终聚类结果

类名	竞争度	搜索量	$Page_{1st}$	CPC	Intitle'
类 1	0.0003	0.2138	0.0061	0.0043	0.0311
类 2	0.0004	2.0000	0.0050	0.0072	0.9465
类 3	0.0003	0.0147	0.0075	0.0036	0.5122
类 4	0.0002	0.0257	0.0042	0.0037	0.0157
类 5	0.0003	0.0296	0.0046	0.0033	0.1007

① 由表 1 得出一个结论, 即关键词的竞争程度和 Intitle 成一定的正比关系, 同时与聚类的关键词个数也成一定的正比例关系.

② 由表 1 可知, 类 4 的关键词数量最多, 竞争程度也最低. 类 2 关键词数量最少, 竞争程度也最高. 与

类 1、类 5 相比, 类 3 关键词对应的 $Page_{1st}$ 和 Intitle 都是最多的, 说明其竞争程度也更激烈, 但其搜索量反而是最少的, 故此类关键词效能(搜索量/竞争度)最差.

③ CPC 可以当作关键词商业价值的一个重要指标, 由表 1 可以看出, 类 2 关键词的商业价值最高. 类 4 关键词的 CPC 值排名第三, 而其竞争程度最低, 故该类关键词的价值率(CPC/竞争度)最高. 与类 3、类 5 相比, 类 1 关键词 CPC 值更高, 因此, 相对于前两者, 后者关键词的价值率也更高.

从分析结果来看, K-means 算法能高效可行地对关键词样本进行聚类, 实验数据容易理解, 便于分析不同聚类关键词之间的联系和区别.

同一样本对 k 的取值不同, 会产生不同的聚类结果. 本文对 k 取 3-7 不等的数值分别进行聚类实验, 通过比较发现, 当 k=5 时, 该关键词样本的聚类结果最佳. 因此, 在实验过程中, 应根据实际情况选择合适的 k 值, 以达到理想的聚类结果. 这样也为后期的关键词优化策略提供可靠的依据.

3 Keywords优化策略

从关键词特征的角度, 网站优化策略可分为四种, 分别为基于关键词竞争程度、基于关键词搜索量、基于关键词效能以及基于关键词价值率的优化策略. 所有网站的优化都宜用基于关键词效能的策略. 此外, 企业网站优化更应该注重关键词的价值率. 因此, 一个成功的企业网站优化策略必须兼顾关键词效能和关键词价值率.

3.1 基于 keywords 效能的优化

根据综合能力(人力, 财力, 技术)以及竞争对手情况, 中小型企业网站应选择竞争程度较低的关键词. 否则关键词效能再好, 但因其竞争度高而使网站排名无法和大企业抗争, 最终导致整个优化工作的失败. 即基于关键词效能的优化首先要考虑关键词的竞争程度. 其次, 关键词搜索量也是一个重要的优化因素. 如果没有一定的搜索量, 虽然网站排名很容易上去, 但不能带来一定的流量, 也就达不到优化目标.

由实验得出的聚类结果的关键词效能如表 2 所示. 类 2 的关键词效能最好, 但其竞争度也最高, 中小企业的网站应慎用此类关键词. 类 3 的关键词效能最差, 而其竞争度也不低, 故不宜使用这类关键词. 类 1、类 4 和类 5 的关键词效能都比较好, 同时它们的竞争度适中, 也有一定的搜索量, 因此所有企业网站都可以选用这些关键词来优化.

表 2 聚类的关键词效能

类名	个数	占比	搜索量	关键词效能
类 1	8	7%	0.2138	713
类 2	2	1.8%	2.0000	5000
类 3	4	3.6%	0.0147	49
类 4	80	70.8%	0.0257	129
类 5	19	16.8%	0.0296	99

值得一提的是, 关键词效能具有时效性, 特别是那些具有季节性和节日性质的关键词, 往往会随着时间的变化导致其搜索量骤变, 相应的关键词效能会产生比较大的波动. 因此在基于关键词效能的优化过程中, 应当掌握好这类关键词的优化时机和进度.

3.2 基于 keywords 价值率的优化

关键词价值率和关键词的商业价值成正比, 而与关键词的竞争程度成反比. 根据分词原理, 竞争程度高的关键词宜放在首页, 中等竞争度的宜放在分类页和专题页, 竞争度低的宜放在内容页. 因此, 对于企业网站而言, 不能一味地讲究关键词的价值率, 还要重点策划不同竞争度的关键词在整个网站中的布局, 以达到最佳的优化效果. 较好的方法就是先按竞争程度对关键词进行归类, 再将每个分类按价值率高低对关键词进行排序, 最后把每个分类中价值率较高的一部分关键词分布在相应的页面来优化.

由聚类结果得出的关键词价值率如表 3 所示. 类 4 的关键词价值率最高, 此外其竞争度也最低, 因此这类关键词适合放在大量的内容页来优化. 类 2 的关键词价值率次之, 其竞争程度最高, 因而这类关键词适用于首页. 类 1 的关键词价值率也比较好, 其竞争度适中, 这类关键词应该放在分类页面和专题页面. 而类 3 和类 5 的关键词价值率比较低, 企业网站应慎用这些关键词.

表 3 聚类的关键词价值率

类名	个数	占比(%)	竞争程度	价值率
类 1	8	7	0.0003	14.3
类 2	2	1.8	0.0004	18
类 3	4	3.6	0.0003	12
类 4	80	70.8	0.0002	18.5
类 5	19	16.8	0.0003	11

在实际应用中，基于 keywords 效能的优化策略与基于 keywords 价值率的优化策略可能会发生冲突. 比如在本文中，以上两种优化策略对类 5 关键词有不同的采纳建议. 这就要求 SEO 决策者学会在两种关键词优化策略中权衡利弊，做出合理的选择.

不同的企业网站对以上两种关键词优化策略的运用侧重点也应有所不同. 具备强大 SEO 团队的企业可以选择竞争度激烈且关键词效能好的关键词；有产品优势的企业网站宜采用价值率较高的关键词来优化；而 SEO 技术薄弱又没有产品优势的企业则应该选用那些竞争度较低的长尾词来优化网站.

4　应用案例

以某触摸屏企业网站 hkace.net 为例，采用以上的关键词分析和优化方法，经过一个多月的网站优化工作，利用 CNZZ 站长统计工具，给出网站优化后的相关效果图，具体如下所示:

前后两个月的网站访问量比较如图 2 所示，在网站优化前的月访问 IP 只有 38 个，而在优化后的网站月访问 IP 增加到 99 个，增幅达到 160%.

图 2　前后两个月的网站访问流量对比

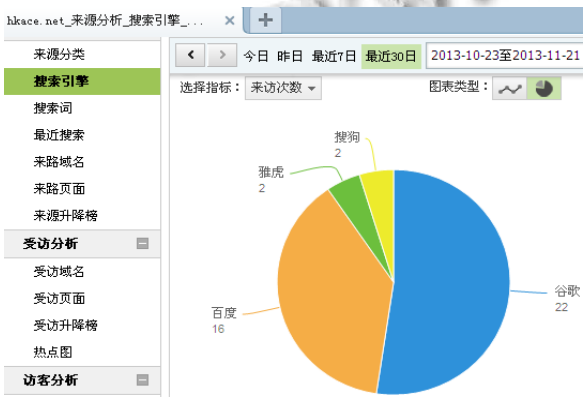

图 3　网站优化后的搜索引擎来源

查看搜索词(谷歌网页)

搜索词	来访次数	占比
电容式触摸屏厂家	3	13.64%
电容触摸屏厂家	2	9.09%
3.5寸电容触摸屏	2	9.09%
触摸屏厂家	2	9.09%
电容触摸屏报价	2	9.09%
电容屏原理	2	9.09%
8寸电容触摸屏	1	4.55%
42寸多点触摸屏	1	4.55%
19寸电容触摸屏	1	4.55%

图 4　来自谷歌搜索的部分关键词

网站优化后一个月内的搜索引擎来源如图 3 所示，由于本网站是针对谷歌进行优化，因此来自谷歌搜索引擎的关键词访问来源占较大的比例，百度次之，同时也有少量搜索关键词来自雅虎和搜狗. 这也表明优化后的网站对于大多数的搜索引擎都比较友好.

2013 年 11 月份来自谷歌搜索的部分关键词如图 4 所示. 该网站首页、目录页、专题页以及内容页的关键词都有一定的排名. 大多数关键词排名已经进入前 50，一些长尾关键词排名甚至已在前 10.

综上可知，K-means 算法可以精简关键词的分析流程，进而减少整个网站优化的工作量. 基于关键词效能和价值率的网站优化策略是比较成功的，一方面能帮助网站在短时间内快速提升其关键词的排名，另一方面也可为企业网站带来一定的流量和询盘，从而达到理想的网站优化目标.

5　结语

在日新月异的网络营销时代，如何快速提高网站关键词排名并实现经济创收是一个严峻的挑战. 实践表明，基于 K-means 算法的关键词分析方法省时省力，能给 SEO 工作者带来诸多便利；而基于关键词效能和关键词价值率的 SEO 策略，则为众多企业提供一个系统高效的网站优化方案. 当然，SEO 本身是个周而复始的过程，本文提出的这些方法策略只是简明有效的，并不代表关键词优化工作可以因此而一劳永逸.

参考文献

1 范彦忠.SEO 技术研究.计算机应用与软件,2010,27(1):160–164.

2 Killoran JB. How to use search engine optimization techniques to increase website visibility. IEEE Trans. on

Professional Communication, 2013, 56(1): 50–66.
3 Beel J, Gipp B, Wilde E. Academic search engine optimization. Journal of Scholarly Publishing, 2010, 41(2): 176–190.
4 Moreno L, Martinez P. Overlapping factors in search engine optimization and web accessibility. Online Information Review, 2013, 37(4): 564–580.
5 陈经优.SEO 技术在电子商务网站中的应用.软件导刊,2012,11(8):166–167.
6 林元国,许振和.基于长尾关键词的 SEO 策略.计算机系统应用,2014,23(1):210–213.
7 刘文云,袁兆勇.面向搜索引擎的关键词优化统计分析.情报杂志,2013,32(1):77–80.
8 唐卫东,刘存后.基于关键词效能的搜索引擎优化策略分析.现代情报,2011,31(10):36–41.
9 李万辉,林瑞明.基于 SEO 的网站 IA 策略研究.图书馆学研究,2010,(19):44–49.
10 Roy DK, Sharma LK. Genetic K–means clustering algorithm for mixed numeric and categorical data sets. International Journal of Artificial Intelligence & Applications, 2010, 1(2): 23–28.
11 Laszlo M, Mukherjee S. A genetic algorithm that exchanges neighboring centers for K-means clustering. Pattern Recognition Letters, 2007, 28(16): 2359–2366.
12 Zalik KR. An efficient K-means clustering algorithm. Pattern Recognition Letters, 2008, 29(9): 1385–1391.
13 王千,王成,冯振元,叶金凤.K-means 聚类算法研究综述.电子设计工程,2012,20(7):21–24.
14 张文明,吴江,袁小蛟.基于密度和最近邻的 K-means 文本聚类算法.计算机应用,2010,30(7):1933–1935.
15 李虎,胡建龙,朱勇,马春波.基于密度聚类分析的入侵检测方法研究.计算机与数字工程,2013,41(2):254–323.
16 伍育红.浅议聚类分析方法.计算机科学,2012,39(6):325–327.
17 易云飞,包宗藩,罗世杰,张志平.改进 k-means 算法在高校计算机专业创新人才培养中的应用研究.软件导刊,2013,12(6):37–39.
18 左国才,周荣华,黎自强.改进k-means算法在电信CRM 客户分类中的应用.计算机系统应用,2012,21(11):153–186.
19 秦礼琦.聚类分析在优化排水管网监测点中的应用. 贵州大学学报(自然科学版),2013,30(2):123–134.