一种基于页面聚类和排序算法的多元搜索引擎改进方案

张 泳 （浙江大学城市学院 310003）
吕 净 （浙江省能源集团财务有限责任公司 310003）

摘要：由于网络上信息数量庞大，多元搜索引擎可能会产生一个相当大的结果集，本文借鉴了 Web 挖掘中聚类算法 FCMA 和网页排序算法 HITS 的技术和思想，改进了多元搜索引擎的结构，以提高系统的查询效率。

关键词：多元搜索引擎 WEB 挖掘 页面聚类 网页排序

1 引言

多元搜索引擎是建立在搜索引擎技术基础之上的一种信息搜索系统。它本身不需标引和搜索网页，而是将查询请求提交给它所要调用的后台搜索引擎，由搜索引擎做真正的查询工作，多元搜索引擎再从各搜索引擎的查询结果中去除重复的查询结果并加以整合，最后由统一的用户接口提交结果。

但是搜索引擎用户一般不关心搜索引擎是怎样工作的，他们只对查询的结果感兴趣。从用户的角度来看，当前的多元搜索引擎主要存在以下两个不足：

（1）查询结果中无关信息过多，大多查询动辄返回成百上千甚至上万条信息；

（2）查询结果的显示顺序比较混乱，搜索引擎在对结果进行排序时没有区分信息类型（如文献、新闻、广告等）。

多元搜索引擎可以一次让多个搜索引擎并发查询，扩大了查询的覆盖面，从而产生一个非常庞大的结果集，那么如何精减结果以及如何将最重要的结果首先返回给用户就显得十分重要。

一个好的搜索引擎不仅能按照用户查询要求快速准确地把结果返回给用户，而且还能去伪存真，把貌似符合查询要求，实际离用户查询要求相去甚远的信息过滤掉。进行文档相关性评价，并最终按与用户查询相关程度来筛选出查询结果是搜索引擎的重要一环。另外，作为一个多元搜索引擎，如何能够将获取的信息按照相关度进行排序也是非常复杂的问题，因为不同搜索引擎在本身查询结果排序过程中采用的算法相差很大，甚至有一些未知的算法，而多元搜索引擎必须整合这些使用不同排序算法产生的结果，并以统一的结果形式返回给用户。这些都是在研究多元搜索引擎中遇到的难点，也是能否成功实现一个多元搜索引擎的关键。

2 多元搜索引擎改进方案

多元搜索引擎是建立在传统搜索引擎技术基础之上的一种信息检索系统。它利用下层多个搜索引擎提供的服务向上提供统一的检索服务。由于网络上信息数量非常庞大，多元搜索引擎可能会产生一个相当大的结果集，那么如何精减结果以及如何将最重要的结果首先返回给用户就显得十分重要。

Web 挖掘属于数据挖掘的分支，它属于知识发现的范围，而搜索引擎则以信息的检索为目的，它属于信息发现的范围，就其对于信息的开发层次而言，Web 挖掘要高于搜索引擎。

Web 挖掘技术虽然与搜索引擎存在着很大的区别，但是二者的关系十分密切。首先，二者的研究对象很相似，都把 Web 文档的处理作为主要内容。其次，二者的技术手段相互补充。搜索引擎来源于信息检索技术，经过几十年的研究，信息检索技术已经在文档内容表示、索引模型、匹配模型等方面发展的相当成熟。这些技术实际上就构成了 Web 挖掘的底层技术，Web 挖掘正是由于直接借鉴信息检索技术的经验，才使得它能够在更高层次上对 Web 资源进行更深一步的挖掘。反过来由于 Web 信息的不断膨胀以及人们对 Web 信息资源利用要求的不断提高，就要求搜索引擎借鉴 Web 挖掘中的技术，借鉴 Web 挖掘中的思想，使搜索引擎更加适应网络环境下对信息检索的需要。

本文借鉴了 Web 挖掘中的技术和思想，将多元搜

索引擎加以改进。改进后的多元搜索引擎系统结构如图1所示。

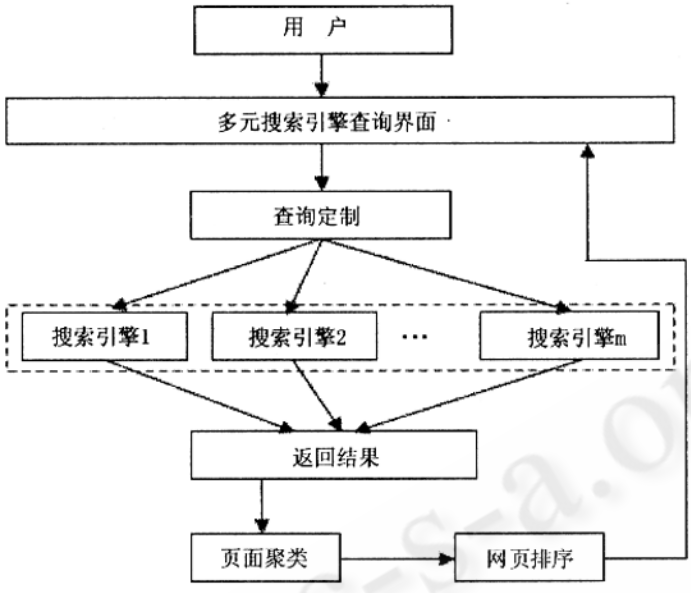

图 1　改进后的多元搜索引擎体系结构

改进后的系统通过页面聚类的算法从大量杂乱无章的查询结果中发现相似的网页，并根据这些页面的超链接结构进行排序，并把经过整合后的信息返回给用户，从而提高用户的查询效率。

3　页面聚类算法 FCMA

聚类(clustering)就是将物理或抽象的集合分组成为由类似的对象组成的多个类或簇(cluster)，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。通过聚类，我们能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系。

目前，在文献中可以找到大量的聚类算法。其中，FCM 是一种能自动对数据样本进行分类的方法，它通过优化模糊目标函数得到每个样本点对簇中心的隶属度，从而决定样本点的归属。而在实际生活中，Web 页面样本量巨大，无法有效地确定聚类数目，采用 FCM 算法对大样本聚类时将消耗大量的空间及时间，且有时会收敛到局部极小点上，FCM 算法的这一缺点限制了人们对 FCM 算法的使用。因此本文针对 FCM 算法上存在的不足，在算法结构上进行了具体的改进，利用对聚类的有效性函数的分析，提出了模糊 C 均值自适应算法 FCMA(Fuzzy C－means adaptation)，能够调整聚类数目 C，可以避免在聚类数目的选取上存在的主观性。

算法 FCMA 可以描述为：对象空间内的数据集 X 划分聚类。令 $X=\{x_1 x_2 \cdots x_n\}$ R^p 为模式空间 R^p 中的一个有限数据集；$\bar{x}_k$ $\{x_{k1} x_{k2} \cdots x_{kp}\} \in R^p$ 称为特征矢量或模式矢量，x_{kj} 为模式矢量 $\bar{x}_k$ 的第 j 个特征(属性)，对任意的整数 c，$2 \leqslant c \leqslant n$，$R^{c \times n}$ 用表示所有实的 $c \times n$ 阶矩阵集合。

目标函数 J_m 是各类中的样本到聚类中心的加权距离平方和大到最小，即为：

$$J_m=(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}^m,d_{ik}^2),1\leqslant m\leqslant\infty$$

要使目标函数 J_m 达到最小值的必要条件是：$u_{ik}=\dfrac{1}{\sum_{j=1}^{c}\left(\dfrac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}$，$v_i=\dfrac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}$

其约束条件为：

$$\sum_{i=1}^{c}u_{ik}=1,1\leqslant k\leqslant n \text{ 和 } u_{ik}\geqslant 0,1\leqslant i\leqslant c,1\leqslant k\leqslant n$$

这里，n 为总的样本数据数，c 为聚类中心数，U 为隶属函数矩阵，u_{ik} 是矩阵 U 的第 i 行第 k 列元素，代表第 k 个数据对第 i 个聚类中心的隶属度，$V=\{\bar{v}_1,\bar{v}_2 2,\cdots,\bar{v}_c\}$ 为聚类中心矩阵，$d_{ik}=\|\bar{x}_k-\bar{v}i\|$，采用的是欧氏公式，表示第 k 组数据对于第 i 个聚类中心的距离，$m(1,\infty)$ 是模糊指数，模糊指数越大，聚类的模糊程度就越大，Pal 等人从聚类有效性方面的实验研究得到 m 的最佳取值区间为[1.5, 2.5]，这里我们取 m=2。

下面给出聚类算法的自适应具体过程：

(1) 对于 q 个输入的系统，构造一组数据对：

$x_l=\{x_1,x_2,\cdots,x_k,\cdots,x_n \mid x_k=(x_{jk},y_k)\}$ 对应于 FCM 的第 j 个输入空间，x_k 是由 q 个输入和输出组成的第 k 组数据，n 是数据的总数，$j=1,\cdots,q$，设 $j=1,q=2$。

(2) 初始化聚类中心 $\bar{v}_i,i=1,\cdots,n$。

(3) 设 c=2，迭代次数 p=0，计算各个数据到聚类中心的距离 d_{ik}，计算隶属函数矩阵 $U^{(0)}=(u_{ik}^{(0)})$，u_{ik} 是矩阵 U 的第 i 行第 k 列元素，代表第 k 个数据对第 i 个聚类中心的隶属度。对于任一 k 列有 $\sum_{i=1}^{n}u_{ik}=1$，对于任一 i 行有 $0<\sum_{k=1}^{n}u_{ik}<n$。

(4) 计算 c 个聚类中心 $\bar{v}_i, i=1,\cdots,n$。

(5) 重新计算隶属函数矩阵 U。

(6) 计算目标函数 $J^{(p)}$，如果 $|J^{(p)}-J^{(p-1)}|\leq\varepsilon$，表示收敛，则迭代结束；否则，$p=p+1$，转向第(4)步。

(7) 如果有效性函数 S 达到最小值。聚类过程结束，否则，聚类数 $c=c+1$，然后转向第(3)步，即取 S 随 c 的增加而成为最小点的值作为聚类数。

$$S=\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\delta_{ik}\frac{\left\|\bar{x}_k-\bar{v}_i\right\|}{u_{ik}^{m}V_{ik}}}{n\min\frac{\left\|\bar{v}_i-\bar{v}_k\right\|^2}{\partial_i+\partial_k}}$$

1992 年，Botafogo 等人定义了在超文本环境下的 index 节点(入度大于平均值)和 reference 节点(入度小于平均值)。1997 年，Carriere 和 Kazman 提出网页的一种排名由它的出度和入度之和决定，还是没有利用到万维网的有向特性。1998 年 Page 提出了 PageRank 算法，这类似于 Pinski－Narin 计算影响力权值的权值传播模型的网页排名方法。之后 Kleinberg 又提出了 HITS 算法，将网页分成 hub 网页和 authority 网页，通过迭代可以将最终得到排名最高的 hub 网页和 authority 网页。在目前的网页排序算法中，PageRank 和 HITS 算法是两种比较优秀的算法，下面就对这两种算法进行比较。

PageRank 和 HITS 的迭代算法都利用了特征向量作为理论基础和收敛性依据。这也是超链接环境下的这类算法的一个共同特点。

从两者的权值传播模型来看，PageRank 基于随机冲浪(random surfer)模型将网页权值直接从 authority 网页传递到 authority 网页；而 HITS 将 authority 网页的权值经过 hub 网页的传递进行传播。

从两者的处理对象来看，都是针对整个万维网上的网页的一个子集进行排序、筛选，没有一个搜索引擎能够将万维网上的网页全部搜索下来。但是，PageRank 的处理对象是一个搜索引擎上当前搜索下来的所有页面，一般在几千万个页面以上；而 HITS 的处理对象是搜索引擎针对具体查询主题所返回的结果，从几百个页面扩展到几千几万个页面。

从两者的具体应用来看，PageRank 应用于搜索引擎服务端，可以直接用于标题查询并获得较好的结果；若要用于全文本查询，需要与其他相似度判定标准进行复合，以针对具体查询形成最终排名。HITS 一般用于全文本搜索引擎的客户端，对于宽主题的搜索相当有效，可以用于自动编撰万维网分类目录；通过找到指向某网页的 Hub 网页并以此为根集，可以查找该网页的相关网页；也可用于搜索引擎的网页排序。对于窄主题的检索，HITS 现在的能力还很弱，因为根集太小，筛选的效果将不会很大。

综上所述，多元搜索引擎的返回信息经过聚类后选用 HITS 算法进行页面排序较好。

4 总结

本文针对多元搜索引擎返回的信息杂乱无序这一问题，借鉴了 Web 挖掘的技术和思想，对查询返回的信息在关联规则发现的基础上进行页面聚类，然后对聚类后的页面进行排序，把最符合要求和最相关的网页优先提供给用户。

多元搜索引擎技术在未来的发展中，将进一步体现对当前功能和特性的丰富与完善，同时也将结合人工智能、多媒体技术和协同技术等领域的研究结果，提出新的研究方向。

参考文献

1 Monika R. Henzinger. Hyperlink Analysis for the WEB. IEEE Internet Computing, 2001, (1):45－50.

2 Gerorage Meghabghab. Discovering authorities and hubs in different topological Web graph structures [J]. Information Processing and Management, 2002,38(1):111－140.

3 王涛、孙河山，WEB 挖掘技术在搜索引擎中的应用[J]，信息系统，第 25 卷 2002 年 4 月.

4 黄于蓝等，搜索引擎技术的新发展—多元搜索引擎系统[J]，计算机工程，Vol. 28，2001 Proceedings of the 10th IEEE International Conference and Workshop on the ECBS' 03.

5 姚全珠、张杰，基于数据挖掘的搜索引擎技术[J]，计算机应用研究，2006 年 11 期.

6 马晓玲、吴永和，对于搜索引擎优化(SEO)的研究[J]，情报杂志，2005 年 12 期.