目 录

张庆国

第六章 聚 类 检 索

6.1 问题的提出

在前面的章节中，我们讨论了基于倒排档的检索机制。倒排档检索方法有很多优点，但同时也存在着以下一些缺点：

① 需要一个辅助检索用的倒排档；

② 基本上是一种完全匹配方式，不能表示命中文献对用户在适用程度上的不同；

③ 不注意标引词的权值，对于同一篇文献的若干标引词往往是同等看待的，不注意它们重要程度的不同，也不注意词间关系；

④ 不能由一篇相关文献引导至另一篇相关文献；

⑤ 因为没有在机内存贮词间关系库，所以不易实现改善检索效果的自动化。

针对以上问题，国外近年来发展了一些聚类检索系统，它们在理论和技术上尚未十分成熟，因此又被称为实验性检索系统。这个名称也是相对传统的倒排档检索系统而言的。

我们将通过一些实际的检索系统来介绍这种检索机制，其中最主要的是SMART系统。

6.2 SMART系统

SMART系统与传统的基于倒排档的系统有以下几点主要的不同之处：

① 使用全自动的标引方式和检索提问构成方式；

② 集中与同一主题有关的文献，使之可以从一篇文献方便地导至另一篇；

② 通过处理文献与文献的匹配度，文献与提问的匹配度来完成检索，并且将输出文献按匹配度的降序排列；

( ) 通过先检索到的文献，自动改善检索提问式。

6.3.1 文献的向量表示和匹配度计算

设我们有文献——词矩阵

$$D=\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_N \end{array}\begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \cdots & T_M \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{pmatrix} \end{array} \qquad (6.1)$$

我们把其中的第i行当作一个M维向量，以代表文献D_i，即有

$$D_i=(a_{i1}, a_{i2}, \cdots, a_{in}) \qquad (6.2)$$

其中a_{ij}称为向量D_i的分量，表示标引词T_j对文献D_i的重要程度，或权

文献与文献之间可以计算其匹配度。匹配度的计算方法有很多，例如

①

$$Similar(D_i, D_j)=\sum_{t=1}^{M} a_{it}a_{jt} \qquad (6.3)$$

②

$$Similar(D_i, D_j)=\frac{\sum_{t=1}^{M} a_{it}a_{jt}}{\sum_{t=1}^{M} a_{it}^2 \cdot \sum_{t=1}^{N} a_{jt}^2} \qquad (6.4)$$

③

$$Similar(D_i, D_j)=\frac{\sum_{t=1}^{M} a_{it}a_{jt}}{\sum_{t=1}^{M} a_{it}^2+\sum_{t=1}^{M} a_{jt}^2-\sum_{t=1}^{M} a_{it}a_{jt}} \qquad (6.5)$$

其中式（6.4）称为余弦函数。是常用的一种计算匹配度值的公式。它的值域在0和1之间。

在一个有M个标引词的检索系统中。可以认为存在一个M维的文献空间。每篇文献可用这个文献空间中的一个向量来表示。即 $D_i=(a_{i1}, a_{i2}, \cdots, a_{iM})$ 其中各分量是该向量在各相应标轴上的坐标。

例如。下图是一个三维的文献空间。在这个文献空间中，三个文献向量

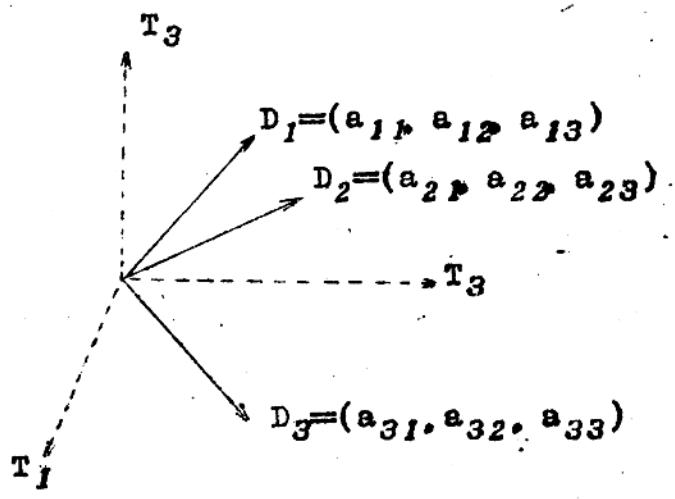


图 6.1 文献空间和文献向量

当文献被表示为文献空间中的向量之后。文献之间的相似程度（匹配度）还可以用其向量之间夹角的反比函数来度量。如果两篇文献完全相同。则其文献向量在文献空间中重合。夹角为0。而匹配度值为最大。

在SMART系统中。提问Q也被表成向量形式

$$Q=(q_1, q_2, \cdots, q_M) \qquad (6.6)$$

其中分量 q_j 表示检索词 T_j 对该提问的重要程度。或权。

需要指出的是，一般的布尔逻辑提问式可以转换成向量形式的提问式。例如

$$Q=(A+B)*(C*D)$$
$$=(A+B)*C*D$$
$$=(A*C*D)+(B*C*D)$$

于是这个提问式Q可用下个提问向量表示：

$$Q_1=(A, C, D)$$
$$Q_2=(B, C, D)$$

如果系统中只有A、B、C、D四个标引词，那么提问式Q_1和Q_2也可表成

$$Q_1=(1, 0, 1, 1)$$
$$Q_2=(0, 1, 1, 1)$$

提问向量与文献向量之间也可以计算匹配度。公式仍如式（6.3）、（6.4）、（6.5）所示。如需计算D_i与Q的匹配度，只要把公式中的a_{jt}换成q_t即可。常用的匹配度计算公式是余弦函数的式（6.4）。

用定量的方法计算文献与提问的匹配度之后，我们不再要求命中文献恰好包含全部检索词，我们这里只要求命中文献达到一定的与提问向量的匹配度，或者要求输出的文献满足一定的数量要求。例如，给定匹配度阈值为0.5，则所有与提问向量的匹配度值达到或超过0.5的文献都是命中的；或者，我们要求输出n篇命中文献，那么将系统中的文献分别与提问计算匹配度并将它们按匹配度值降序排列后，前面的n篇便是所需的。

6.2.2 聚类文件的生成和SMART系统的文档结构

计算了所有文献对之间的匹配度之后，我们可以构造一个聚类文件。这个聚类过程是一个从下到上的生成过程。

1、给定一个阈值。由这个阈值生成若干初始类目。在每一个类目中，任意两篇文献的匹配度都不小于这个阈值。设生成的初始类目为 C_1，C_2，…，C_Z。

如下图，每一个×表示一篇文献。它们的位置关系表示这些文献向量在文献空间中的分布。两个×相距越近，表示这两篇文献的相似程度越高（匹配度值高）。每个圆圈表示一个初始类目。在一个类目（圆圈）中，任意两个×之间的距离都不大于该圆的直径。显然，这里采用的是类似于“词团”结构的“文献团”结构。在一个文献团结构的类中，任一篇文献与该类中其他任一篇文献的匹配度都不小于指定阈值。

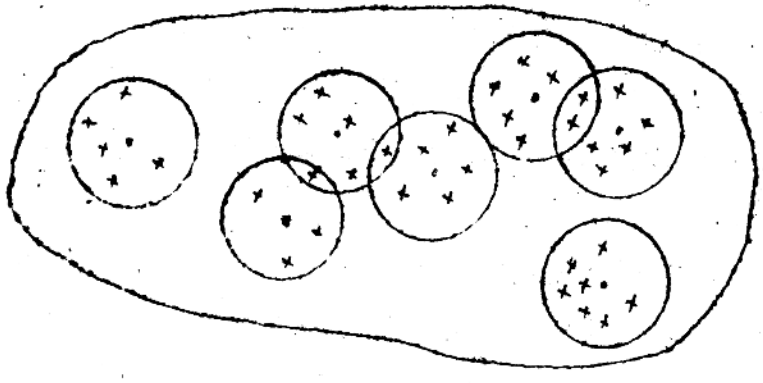


图 6.2　文献的聚类

可以想象，这若干个类目很可能不是互相独立的，即有的文献会同时出现在不止一个类目中。

在上图中，我们为了直观起见，仅给出了二维的文献空间。在一个M维的文献空间中，一类是一个M维的单位球域，其中任意两点之间的距离不超过二单位。

2．对于每一个初始类目C_v。给出一个类目向量$C_v=(T^C_1, T^C_2, \cdots, T^C_M)$。该类目向量各分量的值是各文献向量的相应分量的平均值。即

$$T^C_k=\frac{1}{m}\sum_{i=1}^{m}T_{ik} \qquad (6.7)$$

其中m是该类目C_v中的文献篇数。

3．计算这些初始类目之间的两两匹配度。以作为“类对”的匹配度。

4．给定一个阈值。由这个阈值再生成若干高一级的类目。每一个类目中包含若干上一步生成的初始类目。而这些初始类目之间的两两匹配度都不小于指定的阈值。

5．重复上述步骤2．3．4．直至归并成为一个最高级类目。如下图所示：

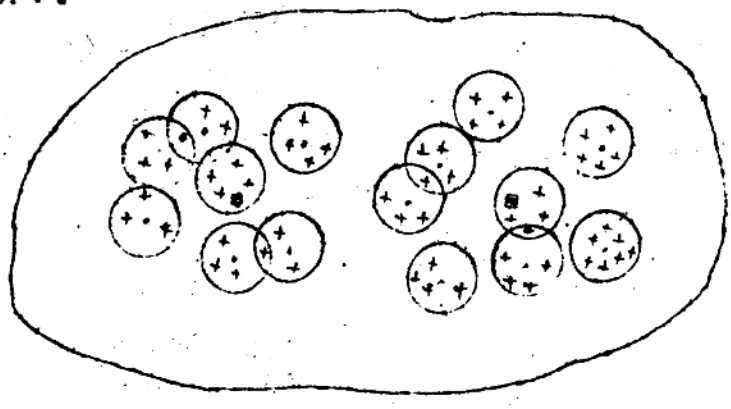


图6.3 文献聚类的层次

我们将这样生成的各级类目按从上到下的顺序分别称为一、二、三、…级类目。其中最后一级已是实际的文献。但由于文献和类目都被表示为向量。所以在处理手段上并无区别。

我们在这里采用的类目生成方法需要计算每一对文献之间的匹配度。计算量较大。另一种方法是先任意给出一些类目。然后计算

每一篇文献与每一个类目的匹配度，调整文献对类目的归属，逐渐逼近最佳的类目体系。这种方法的具体介绍从略。

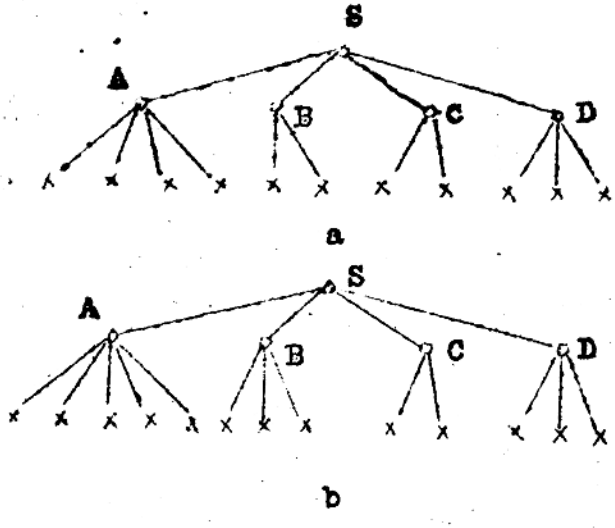
类目的生成并不是一次性的。当新的文献加入这个分类体系后，会给原类目体系带来以下两个影响：

1．由于初始类目的向量是由该类目中各文献向量所决定的，所以加入一个新的文献向量后，会改变初始类目的向量表示；又由于上一级类目的向量表示是由下一级类目的向量表示所决定的，所以后者的变化会影响前者。这种影响将一直波及到根。波及范围是分类树中从一个树叶（新加入的文献）到根的一条路径。

2．类目中增加了新的文献或下位类后，可能会使该类目变得过大，因而影响检索。

对于前一个问题，解决办法是自加入文献的这个初始类目起，向上逐级重新计算其类目向量，直至树根。

对于后一个问题，解决办法是进行一些必要的类分裂工作，如下图所示：



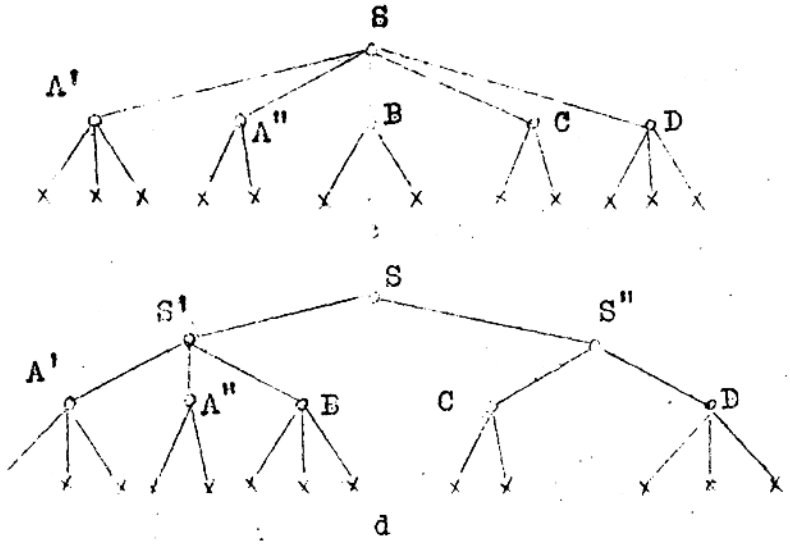


图 6.4 类的分裂

在SMART系统中，由于文献和提问都被表示成向量形式，所以增加新文献记录的工作机制与检索工作机制是基本相同的。下面给出它们的工作流程：

① 输入一个向量。输入向量可以是准备加入分类体系的文献向量，也可以是准备检索的提问向量。将所有二级类目排成一个待考察的类目队列。

② 在类目队列中选择一个结点，与输入向量计算匹配度。

③ 如果匹配度值小于指定阈值，则转②；否则转④。

④ 当前比较的结点是初始类目吗？如果是，则继续执行；否则转⑨。

⑤ 输入向量是文献向量还是提问向量？如是前者，则继续执行；否则转⑩。

⑥ 将输入的文献向量加入当前的初始类目，如加入后这一初始类目太大，则继续执行；否则转⑧。

⑦ 在准备分裂的结点路径中插入一个结点。

⑧ 新的文献记录已经加入所有应该加入的类了吗？如是，则转⑪，否则转②。

⑨ 将当前类目的所有儿子加进待比较的类的队列。转②。

⑩ 将当前初始类目的所有文献加进待比较的结点队列。检索出那些与提问有足够大匹配度值的文献。若已检出足够的文献记录，则停止。否则转②。

⑪ 出口。转类分裂程序。

本节最后给出SMART系统的文档结构：

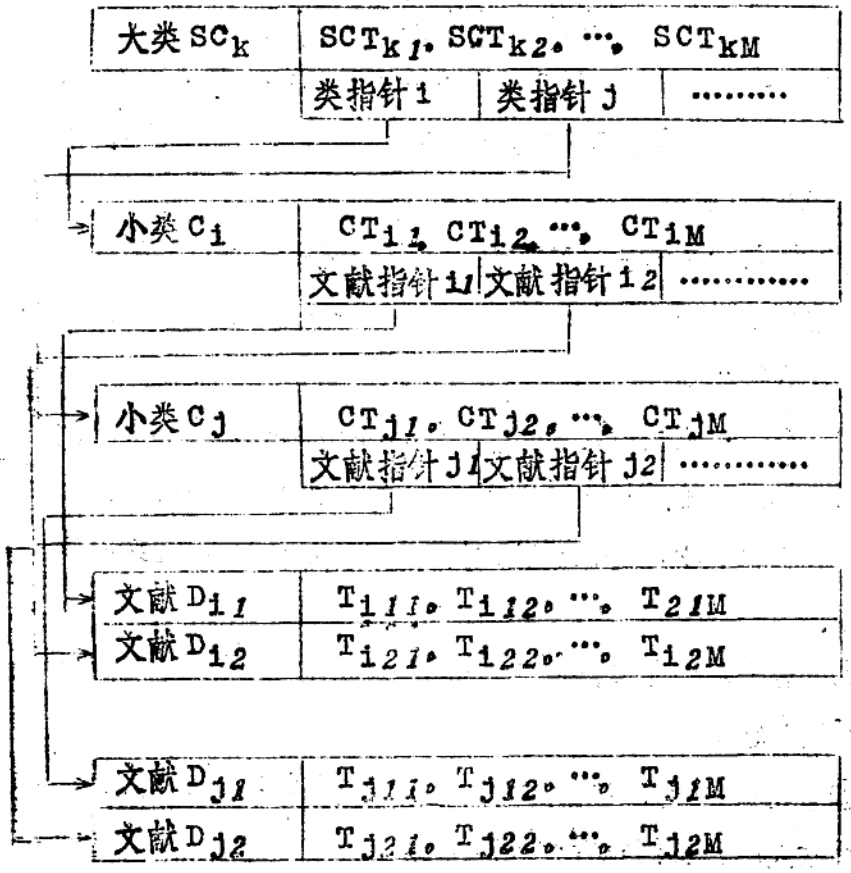


图 6.5 SMART系统的层次聚类文档结构

6.2.3 提问式的反馈调整

提问式的反馈调整是SMART系统的一个重要特点。因为SMART系统的检索机制并不是"完全匹配"的，而是通过匹配度值来表示输出文献与提问的满足程度。因此，我们可以并且应该尽量提高这种满足程度。

为讨论这种反馈调节机制，我们先看一种理想的然而也是不可能的情况：相关文献已知情况下的最优询问表达式。我们记这个最优询问式为Q_{opt}。

设文献数据库中共有N篇文献。对于某一用户提问，系统中有相关文献R篇，不相关文献N—R篇。记这个相关文献群为D_R，不相关文献群为D_{N-R}。考虑一个检索词（标引词）T_k，在相关文献群D_R中，标引词T_k的平均权值为

$$\frac{1}{R}\sum_{i\in D_R} T_{ik} \qquad (6.8)$$

其中$i\in D_R$表示这里的文献D_i属于D_R。

类似地，在不相关文献群D_{N-R}中，标引词T_k的平均权值为

$$\frac{1}{N-R}\sum_{i\in D_{N-R}} T_{ik} \qquad (6.9)$$

可以理解，在最优询问式Q_{opt}中，检索词T_k的权值应规定为

$$(Q_{opt})_k = C\left(\frac{1}{R}\sum_{i\in D_R} - \frac{1}{N-R}\sum_{i\in D_{N-R}} T_{ik}\right) \qquad (6.10)$$

其中C是一个常数。

实际上，无论是相关文献群D_R还是不相关文献群D_{N-R}，在检索前都是不知道的，因为否则也就不需要检索了。并且，一般也

不可能用几何的方法在文献空间中划分出两个互相独立的区域。其中一个恰好包含全部相关文献。另一个恰好包含全部不相关文献。聚类文件虽然在一定程度上聚集了学科相近。主题相关的文献。但也并非绝对如此。与同一检索课题相关的文献分布与不相关的文献分布仍然会有各种复杂的情况。如下图所示：

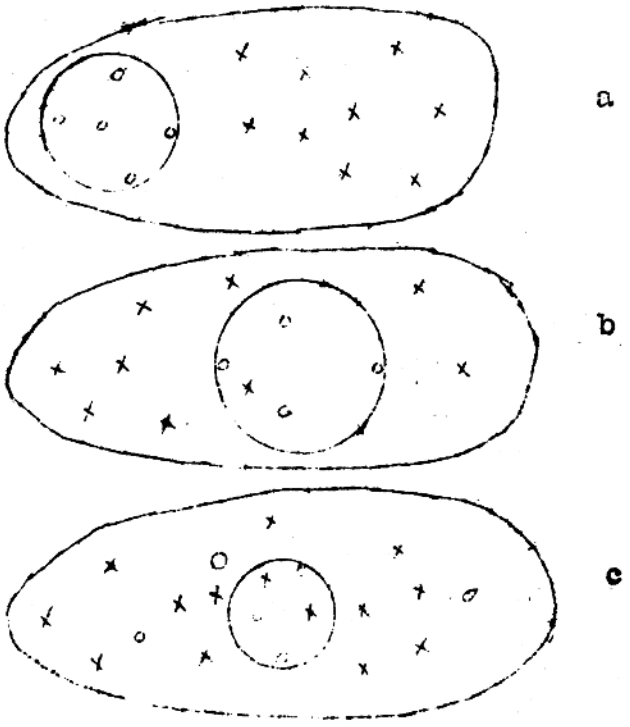


图 6. 6 相关文献与不相关文献的分布

上图中。o表示相关文献。x表示不相关文献。容易看出。a的情况最好。我们可以画出一个圆。其中包含了全部相关文献。而没有包含不相关文献。c的情况最差。我们无论在什么地方画圆和画多大的圆。都不可避免地要遗漏一些相关文献和包含进一些不相关文献。提问向量实际上是文献空间中的一个点（提问向量的端点）。输出时用户指定的阈值实际上是一个半径。这两个值共同确