计算机科学2000Vol.27№.8

# 模糊分类与模糊匹配相结合的模糊检索

Fuzzy Retrieval Combined the Thinking of Fuzzy Type Parted with Fuzzy Match

李名标

（赣南师范学院数学与计算机系　赣州341000）

**Abstract** The works of research on the literature information retrieval in the information times is an important. In this paper, on the basis of thinking and theory of consistent retrieval by fuzzy type parted and inconsistent retrieval by fuzzy match, we give the method which combines consistent retrieval of fuzzy type parted with inconsistent retrieval of fuzzy match, as we have used fuzzy processing method of fuzzy type parted and fuzzy match. The results of retrieval are getting precise. Thus precision and recall of retrieval are also improved.

**Keywords** Fuzzy type parted, Fuzzy match, Consistent retrieval, Inconsistent retrieval, Fuzzy retrieval

## 1. 问题的提出

文献信息检索的过程，可以看成是一个用户“查询”和文献库之间“匹配”的过程。但是，当文献库中文献数量达到几百万甚至更多时，如果仍用用户的“查询”去逐一匹配文献库，将花费很长的时间，甚至使检索方法无法付诸实用。因此，将文献适当地分成若干子类，在子类中进行检索，能缩小检索范围，可加速匹配过程，将提高信息检索的效率。文[1]给出一种用模糊集理论来进行文献库模糊分类的方法，即模糊分类方法。与模糊聚类中的分类方法不同的是，这种模糊分类方法允许一个文献同时属于两类甚至几类，这对于文献分类是合理的，模糊分类完成后，在各子类中根据用户查询提问进行严格匹配，此即一致性检索。

一致性检索的不足在于，用户的查询提问在各子类中与文献严格匹配，其结果是，当用户提问与文献一致时，即文献库中有符合用户检索的文献，结果文献返回给用户；当用户提问与文献不一致时，用户只能得到“查无此文件”的结果信息。对于后一情况这种处理方式过于绝对化。有时，用户为检索某种文献，当库中没有与用户检索一致的文献时，为用户检出相近的文献对用户也有一定的参考意义。非一致性模糊检索的思想由此而产生。它在第二种情况下即文献库没有与用户严格匹配的文献时，按文献与用户提问的接近程度[2]，为用户提供一组参考文献。

## 2. 模糊分类与模糊匹配相结合的模糊检索

将模糊分类与模糊匹配的思想结合起来，就可得到一种适用于大型文献数据库的模糊检索方法。

### 2.1 模糊分类算法

文[1]中的模糊分类思想基于文献的特征描述，只适于小规模文献库的模糊分类。若根据国家标准的《中图法》和《科图法》来进行分类，并遵循从基础理论到应用技术、从低级到高级、从简单到复杂、从抽象到具体、从一般到特殊、从通用到专用、从宏观到微观、从定性到定量、从主要到其它等一些分类规则，我们就可得到一个规范的、适用于较大型文献库的一个模糊分类。文献的描述项可选择这样一些文献特征：文献名称、作者、出版者、主要的几个关键词、分类号、ISBN 编号等。文献就由这一组描述项来表征。设文献库 G 由 $X_1, X_2, \cdots, X_n$，n 个文献组成，每个文献 $X_k$ 由 m 个描述项 $d_1, d_2, d_3, \cdots, d_m$ 来表征。这样，每个文献都可用一个 m 维向量来表示：$X_k=(\varphi_{k1}, \varphi_{k2}, \cdots, \varphi_{km})$，式中 $\varphi_{kj}$ 由下式规定：

$$\varphi_{kj}=\begin{cases}1 & \text{表示 } X_k \text{ 中有描述项 } d_j\\ 0 & \text{表示 } X_k \text{ 中无描述项 } d_j\end{cases}$$

其中 $k=1,2,\cdots,n; j=1,2,\cdots,m$。于是，n 个文献和 m 个描述项的关系可用一个 $n\times m$ 阶矩阵 R 来表示。记文献库 $G=(X_1, X_2, \cdots, X_n)$，模糊分类算法为：

(1)不妨设 G 可分为 k 类：$G_1, G_2, \cdots, G_k$，每一类 $G_i$ 可看成论域：$G=(X_1, X_2, \cdots, X_n)$ 上的一个模糊子集，它们由隶属函数 $\mu_i(x)(i=1,2,\cdots,k)$ 来表征。

(2)确定隶属函数 $\mu_i(x)$，假定描述项 $d_j$ 在第 i 个模糊子类 $G_i$ 中出现的概率为 $P_{ij}$，则这些数据形成矩阵 $m_1$，该矩阵是一个 k 行 m 列的矩阵，即

$$m_1=\begin{Bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ P_{k1} & P_{k2} & \cdots & P_{km} \end{Bmatrix}$$

其中行表示类，列对应于描述项，$P_{ij}$由调查资料进行模糊统计得到。类 $G_i$ 的隶属函数由下式确定：

$$\mu_i(X_r)=\frac{\sum_{j=1}^{n}\varphi_{ij}P_{ij}}{\sum_{j=1}^{n}\varphi_{ij}}$$

其中 $i=1,2,\cdots,k$；$r=1,2,\cdots,n$

(3)求任意二个模糊子类两两交的隶属函数 $\mu_{ij}(Xr)$。把 $X_1,X_2,\cdots,X_n$ 隶属于 $G_i$ 的程度列成矩阵 $m_2$：

$$m_2=\begin{Bmatrix} \mu_1(x_1) & \mu_2(x_1) & \cdots & \mu_k(x_1) \\ \mu_1(x_2) & \mu_2(x_2) & \cdots & \mu_k(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \mu_1(x_n) & \mu_2(x_n) & \cdots & \mu_k(x_n) \end{Bmatrix}_{n\times k}$$

将 $G_i\cap G_j$ 的隶属函数用矩阵 $m_3$ 表示，则有：

$$m_3=\begin{Bmatrix} \mu_{12}(x_1) & \mu_{13}(x_1) & \cdots & \mu_{k-1\,k}(x_1) \\ \mu_{12}(x_2) & \mu_{13}(x_2) & \cdots & \mu_{k-1\,k}(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{12}(x_n) & \mu_{13}(x_n) & \cdots & \mu_{k-1\,k}(x_n) \end{Bmatrix}_{n\times\frac{k(k-1)}{2}}$$

其中，$\mu_{ij}(Xr)$是文献 Xr 在模糊集 $G_i\cap G_j$ 上的隶属函数，即 $\mu_{ij}(Xr)=\min\{\mu_i(Xr),\mu_j(Xr)\}$，其中 $i=1,2,\cdots,k-1$；$j=1,2,3,\cdots,k$；$r=1,2,\cdots,n$。

(4)给出模糊分类阈值 L，得到模糊分类子类。分类原则应使每个文献至少应分到一类中去，借助于 $m_3$，可确定阈值 L。阈值 L 介于0与1之间，L 取得越大，即分类精度越高，一种文献属于多个子类的可能性就越小；反之，L 取得越小，则分类越粗糙，一种文献同时属于多个子类的可能性越大。因此，L 的选择应适中。令 $L<\min\{\max\mu_{ij}(Xr)\}$，则可得到 λ 截集(普通集)：

$$G_i=\{X\mid\mu_i(x)\geqslant L\}(i=1,2,\cdots,k)$$

于是就得到文献库 $G=(X_1,X_2,\cdots,X_n)$的一个模糊分类 $G_1,G_2,G_3,\cdots,G_k$。这样对文献库 $G=(X_1,X_2,\cdots,X_n)$的匹配，就可转化成对 $G_i(i=1,2,\cdots,k)$的匹配，匹配范围小了，检索的效率得到了提高。

完成上述模糊分类有这样一些方法：完全的人工分类；机器自动分类；及人工与机器结合方式。机器自动分类需要计算机软件完成，该软件要有初步的切词、分词和语义理解功能，能自动地对文献进行特征标识抽取，再根据这些特征标识，利用上述算法由机器完成文献归类。机器自动分类涉及人工智能的诸多方面，技术难度大，其优点是可减少人工干预，提高分类速度，但分类的精度稍差。人工分类的优点是不言而喻的。由于人的智能因素和人在自然语言理解方面与机器的差别，人工分类是比较准确的。不足之处就是人工分类的效率稍低，当文献数量较大时尤显突出。Internet 上的 Yahoo! 目录式分类搜索引擎就是用人工方法完成分类的，Yahoo! 也正因如此，在众搜索引擎中以分类精细、准确而享有盛名。

### 2.2 模糊分类下的非一致性模糊匹配

在文献库的模糊分类前提下，可对某一选定的分类进行非一致性模糊匹配，形成基于文献特征性质的模糊匹配方法。算法可描述为：

(1)用户(检索者)选择子库类，不妨设 $G_{k0}=(X_{d1},X_{d2},X_{d3},\cdots,X_{dk0})$，文献特性为 $P=(P_1,P_2,\cdots,P_m)$，子库类 $G_{k0}$ 到特征集 P 的模糊关系矩阵 $R=(r_{ij})_{k_0\times m}$，这里 R 由模糊统计方法得到。

(2)用户根据特征集 P 提出检索条件。设用户的检索关键词记为 $B=\bigvee_{k=1}^{L}B_k$，其中 $B_k$(第 k 个检索分句)是特征集 P 上的模糊集合，$B_k$ 对于 P 的隶属函为 $B_k=(V_{1k},V_{2k},\cdots,V_{mk})$，$k=1,2,\cdots,L$，这里 $V_{ik}$表示 $B_k$ 对特征属性 $P_i$ 所要求占有的程度。

(3)确定优选阈值增量百分率 $S.0.5<s\leqslant 1$

(4)计算求出检索解向量。记矩阵 $Q=(V_{ik})$为 $m\times l$ 阶，Q 称为检索矩阵。又记

$$\sigma_{ik}=\frac{\sum_{j=1}^{m}\min(r_{ij},v_{jk})}{\sum_{j=1}^{m}\max(r_{ij},v_{jk})}$$

称为文献 $X_i$ 对检索提问的相近度，也就是 $X_i$ 与 $B_k$ 按定义所给的贴近度，它表示文献 $X_i$ 对检索提问 $B_k$ 的符合程度。

取二个特殊的阈值，设：

$M=\max\sigma_{ik}$　　$m=\min\sigma_{ik}$

其中 $1\leqslant i\leqslant n$，$1\leqslant k\leqslant l$。记 $\Psi=m+0.618(M-m)$；$\kappa=m+0.618(M-m)S$，Ψ 叫做优选阈值，κ 叫做扩选阈值，S 即为优选阈值增量百分率。将 $\sigma_{ik}$变换一下，记：

$$S_{ik}=\frac{\sigma_{ik}-m}{M-m}\cdot\frac{\sqrt{5}+1}{2}$$

$S_{ik}$的变化域为$[0,\frac{\sqrt{5}+1}{2}]$，可分为四种情况：

(a)$1\leqslant S_{ik}\leqslant\frac{\sqrt{5}+1}{2}$，亦即 $\Psi\leqslant\sigma_{ik}\leqslant M$；

(b)$S\leqslant S_{ik}<1$，亦即 $\kappa\leqslant\sigma_{ik}<\Psi$；

(c)$0.5\leqslant S_{ik}<S$，亦即 $m+\frac{\sqrt{5}-1}{4}(M-m)\leqslant\sigma_{ik}\leqslant\kappa$；

(d)$0\leqslant S_{ik}<S$，亦即 $m\leqslant\sigma_{ik}<m+\frac{\sqrt{5}-1}{4}$。

称 $X_i$ 是 $B_k$ 的优解，如果 $S_{ik}$属于情形(a)，当 $S_{ik}=1$(即 $\sigma_{ik}=M$)时，称 $X_i$ 是 $B_k$ 的最优解。当 $X_i$ 是 $B_k$ 的优

解时，采用记号 $t_{ik}=(X_i,S_{ik})=X_iS_{ik}$，此处 $X_iS_{ik}$ 是记号，不是乘积；当 $S_{ik}$ 属于情形(b)，称 X 是 $B_k$ 的扩解，此时采用记号 $t_{ik}=(X_i,S_{ik})=S_{ik}X$；当 $S_{ik}$ 属于情形(c)，称 $X_i$ 是 $B_k$ 的候解，此时采用记号 $t_{ik}=\underline{S_{ik}}X_i$；当 $S_{ik}$ 属于情形(d)，则采用记号 $t_{ik}=(X_i,S_{ik})=\theta$。

优解与扩解合称检索解，而候解不是检索解，其实际意义是作为参考备用的候补解。称 $T=(t_{ik})$ 为文献检索的预解矩阵。进一步得到文献检索的解矩阵，记为 $T'=(t_{ik}')$。

(5)解向量返回给用户(检索者)。根据解矩阵 $T'$，将各列元素数字相加(θ的数字算作0)，设 $t_{ik}'$ 对应的数字为 $\alpha_{ik}$。令：

$$W_i=\sum_{k=1}^{L}\alpha_{ik}\quad i=1,2,\cdots,n$$

再取：$\xi=2S\times0.618$，则 B 对 $G_{k0}$ 的检索总向量为：

$$(X_1\frac{W_1}{\xi},X_2\frac{W_2}{\xi},\cdots,X_{k0}\frac{W_{k0}}{\xi})$$

按 $W_k$ 的大小顺序重排，便得到 B 对 $G_{k0}$ 的有序总检索向量：

$$(X_{r1}\frac{W_{r1}}{\xi},X_{r2}\frac{W_{r2}}{\xi},\cdots,X_{rk0}\frac{W_{rk0}}{\xi})。$$

其中 $W_{r1}\geqslant W_{r2}\geqslant\cdots\cdots\geqslant W_{rk0}$。这种检索解不限于是(但包括)传统意义下的一致检索解，它也可以是一种给检索者(用户)有灵活使用余地的相近检索解。

这里描述的算法，同文[1]、[2]中描述的算法不同。本算法的特点是，先进行模糊分类，目的是缩小文献集的规模；然后，检索只需在子类的小范围内进行，子类中的检索是一种非一致性模糊检索，既能提供一致性求解结果，又能提供非一致性参考求解结果。本算法能有效地提高检索的速度、查全率与查准率，随着文献库数量的加大，这种有效性将更加显现出来。

## 3. 算法的可扩充性与可维护性问题

文献更新是经常(每周或每月)要进行的。文献库更新了，分类怎样进行？是文献库整个重新分类？还是只需新增的文献依据一定策略加进已有的分类中？显然，后者的系统开销要小得多。模糊分类，依据抽取的文献分类号和若干模糊分类号来进行，它的可扩展性和可维护性是较好的。

具体做法是，对于新增文献 $X_{n+1},\cdots,X_{n+s}$，建立分类号和若干模糊分类号 type 和 ftype_1，ftype_2，ftype_3(比如说只取3个模糊分类号)，根据 type 的值，决定 $X_{n+1}\in G_{v0}$，或根据 ftype_1的值，决定 $X_{n+1}\in G_{u0}$，或根据 ftype_2的值，决定 $X_{n+1}\in G_{v0}$，或根据 ftype_3的值，决定 $X_{n+1}\in G_{w0}$。

对于其它新增文献 $X_{n+2},\cdots,X_{n+s}$，依照执行。这种按学科分类号进行模糊分类的方法，其扩展性是很好的。这种模糊扩展分类，适合用人工方法完成。

## 4. 模糊分类与模糊匹配在 Web 检索中的应用

目前 Internet 中的中文搜索引擎主要有两种检索方式，即全文检索和目录式分类结构方式。全文检索基于关键词的严格匹配，返回给用户的是含有关键词的引擎索引库中的文献全文；目录式分类结构也是基于关键词的严格匹配，但返回给用户的是含有关键词的引擎索引库中的目录分类。全文检索结果虽全但过于繁杂，目录分类结构结果分门别类清晰但不完全，不能满足用户更进一步的检索需求。而且，两者都不具有更多的模糊处理功能，可将模糊分类与模糊匹配的思想结合并用于 Web 检索中。首先，用户输入检索关键字，严格匹配，检出与关键字匹配的所有站点目录，这就是第一步的分类。然后，用户选择分类，再在这个分类下，按第一步已输入的关键词进行全文检索亦即在分类中进行全文检索，就避免了单独的全文检索和分类目录检索的一些缺点，同时又保留了全检索和目录分类检索各自的优点。

## 参 考 文 献


1 楼世博，等. 模糊数学. 科学出版社，1985
2 汪培庄. 模糊集合论及其应用. 上海科技出版社，1983
3 汪培庄，等 模糊系统理论与模糊计算机. 科学出版社. 1996
4 常桂然，等. Web 信息检索服务系统与搜索引擎. 计算机科学. 1998，25(5)
5 汤兆魁. 数学模型在情报检索系统中的应用. 现代图书情报技术，1994(1)
6 王克明. 情报检索系统最佳效果评估方法浅析. 现代图书情报技术，1994(1)
7 Gudivada V N. Information retrieval on the World Wide Web. IEEE Internet Computing，May 1997