第七届华中地区大学生数学建模邀请赛

承 诺 书

我们仔细阅读了第七届华中地区大学生数学建模邀请赛的竞赛细则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们的参赛报名号为：10497156

参赛队员 （签名）：

队员 1：_____

队员 2：_____

队员 3：_____

武汉工业与应用数学学会

第七届华中地区大学生数学建模邀请赛组委会

第七届华中地区大学生数学建模邀请赛

编 号 专 用 页

选择的题号：B

参赛的编号：10497156

竞赛评阅编号：

第七届华中地区大学生数学建模邀请赛

题目：互联网搜索引擎的排名与设计

【摘 要】

互联网搜索引擎是用户在网络海量信息中检索到自己所需信息的有效途径，建立各搜索引擎的有效评价体系对用户和运营商均大有裨益，而设计个性化搜索引擎可以满足不同用户的不同需求，符合信息社会发展趋势。基于此，本文通过对搜索引擎相应基础知识的研究，数据挖掘和整合，结合层次分析法、灰色关联分析法建立评价模型；结合遗传算法，K-means 聚类分析建立个性化搜索引擎模型，并通过基于用户兴趣的空间向量模型对个性化引擎进行一般性推广。通过 matlab 编程求解，评价结果与所给排名进行对比验证评价模型，通过已知数据对聚类结果比照，做出对照点图验证模型的合理性。具体思想分析如下：

对于问题一，我们采用层次分析法和灰色关联分析法相结合的算法建立搜索引擎评价模型。层次分析法通过判断矩阵有效地将影响因素指标量化，相对客观地得到各评价指标所占的权重，也可以进一步对各搜索引擎进行评价，但仅用层次分析法无法利用客观真实的数据，从而会使评价结果有失偏颇。故我们结合灰色关联分析法，在层次分析法得到各评价指标权重的基础上，充分利用所挖掘整合出的数据。计算出各待评价引擎的灰色关联度，并据此进行排名，得到排名结果与题目所给排名进行比较。分析比较结果，并对各搜索引擎作出合理评价。在主流搜索引擎排名一致的前提下，部分排名有所差异，这与评价时选取指标和评价标准不同有关，故模型是比较成功的。对于实验中的数据，定量化的指标通过查阅相关文献和搜索引擎公布的官方数据，定性化的指标通过对相关概念查询建立相应的评价标准，成员通过亲自对搜索引擎进行相关实验检验，得到数据，具有相对客观性和真实性。

对于问题二，我们采用遗传算法改进的 K-means 聚类分析法建立科技论文文档搜索引擎模型。此模型的核心是对文档进行聚类分析，我们采用较为常用和经典的聚类方法 K-means 聚类方法，通过初始聚类种子不断迭代得到新的聚类种子，从而进行文档聚类。文档聚类前提是要对文档进行特征项提取，我们根据相关领域研究成果，利用 VSM 构造文档的空间向量，使其数据化便于聚类分析。对于该算法对孤立点敏感、初始聚类种子人为选取主观性大的缺点，采用遗传算法来改进，寻求最优解，得到聚类分析最优初始聚类种子。通过中国百科术语数据库中已分好类的文档与模型模拟出的聚类结果对比，分析对比点图，发现基本一致，故认为模型较为成功地对文档进行聚类，通过过滤非科技类文档输出，实现了个性化搜索科技论文文档的功能，模型较为完整和成功。

对于问题三，建立基于用户兴趣的空间向量的推广模型，该模型将用户兴趣，搜索历史等信息进行挖掘，建立了包含用户兴趣的文档个性化搜索引擎。我们建立用户兴趣的向量空间检索模型，自动提取用户兴趣的相关信息，结合问题二的聚类方法，过滤掉用户不感兴趣的信息，输出用户所需信息。便将各类文档根据用户兴趣输出，即实现了个性搜索引擎的一般化推广。

关键字：层次分析、灰色关联分析、遗传算法、K-means、个性化、搜索引擎

目录

一、问题重述

随着互联网的高速发展和普及，人们越来越依赖于互联网共享信息和获取信息。网络上的海量信息是我们巨大而宝贵的资源，而搜索引擎技术正是高效且方便地利用这些资源的有效手段。在搜索页面中，只要输入想搜索的内容的关键字，搜索引擎会立刻给出符合条件的链接。常见的搜索引擎有百度、谷歌、雅虎等，若要为给用户提供更好的服务，建立众多搜索引擎的客观综合评价体系是很有必要的。然而由于各大搜索引擎的通用性，仍然不能满足不同时期，不同背景，不同目的的需求，个性化搜索引擎技术便是针对这一问题而提出的。以此为背景分析解决以下问题：

1 建立数学模型，对现有互联网搜索引擎的优劣进行评价，给出排名前 5 名的搜索引擎，并将排名结果与[1]的排序结果进行比较。

2 建立数学模型，设计出具有个性化特色的互联网搜索引擎，并举例对搜索引擎进行验证。例如，如果是汉字成语搜索引擎，那么输入“张”时，可搜索出“张冠李戴”等成语；如果是中文人名搜索引擎，那么输入“张”时，可搜索出“张三丰”等人名；当然，也可以对某类图片、某类音乐等进行搜索。

3 将你设计的个性化搜索引擎进行一般性推广。

二、问题分析

2.1 问题一的分析

互联网搜索引擎为用户提供了获取网络海量信息的十分有效的手段，但当今搜索引擎数目种类繁多，建立搜索引擎的有效评价体系能对其进行综合评价，从而促使运营公司的技术对搜索引擎的问题解决和技术革新，为用户提供更为便利的服务。

问题 1 属于评价类数学问题，层次分析法是解决此类问题的常用方法，它可以有效分析评价指标，将定性化的指标因素定量化，从而计算出各指标的权重，可以对各搜索引擎进行初步的评价。但是，只简单通过层次分析法会带有一定的主观性，没有建立在客观可靠的数据基础之上，故我们采用灰色关联分析法和层次分析法结合建立模型。通过层次分析法确定各指标对应的权重，然后查阅资料文献建立各指标的评价标准，通过亲自验证各搜索引擎的功能，查阅各引擎的官方数据，并对其进行归一化处理，根据灰色关联度的相关原理，通过 matlab 编程计算得到各搜索引擎的灰色加权关联度，根据其大小对各搜索引擎进行排名和评价。

2.2 问题二的分析

如今搜索引擎种类纷繁复杂，而且普遍具有通用性，因此难以满足不同用户、不同背景、不同目的的需求，故设计出具有个性化特色的互联网搜索引擎是十分有必要的。

此问题可归结为数学中的聚类问题，是对相关索引结果进行聚类处理，并针对某项具体的分类为主，过滤掉其他类别的索引结果后进行排序输出，从而实现个性化的搜索。我们设计了一款索引科技论文文档的个性化搜索引擎，其核心内容是建立文档聚类的模型。文档聚类分析方面已有不少研究，我们采用较为常用和经典的聚类方法 K-means 聚类方法，此方法的前提是对文档的数学化处理，我们根据相关领域的研究成果，采用空间向量模型 VSM 提取文档的特征项，并对其赋予权值，使每个文本表示为特征项空间的一个点，一个文本集构成一个矩阵。可以通过相应的内积或余弦夹角表示两个文本之间的相似度。这样我们通过聚类种子迭代的方法就可以实现文档的聚类分析。K-means 算法具有较好的伸缩性和很高的效率，适合处理大文档集，只是该算法本身存在缺陷，对于文档中孤立的点十分敏感，少量的此类文档会对聚类结果产生很大影响。这样会使模型误差较大，为此我们结合遗传算法进行优化整合，遗传算法可以得到全局最优解，避免陷入局部极小点，而且利用可变长度的遗传算法进行初始点的选取，可以克服 K-means 算法类别人为确定的缺点。最后从已有明确分类的数据库中选取一定文档数进行聚类模拟对比，分析模型优劣。

2.3 问题三的分析

问题 2 中建立的个性化搜索引擎仅对科技论文类文档进行选择输出，具有很大的局限性，为了将模型推广到一般化，我们的模型需要加入基于用户兴趣和浏览历史的空间向量矩阵

三、模型假设

1. 假设题目所整合的数据真实可靠；
2. 假设除模型所列评价指标之外的因素对评价结果无影响；
3. 假设文档中不含有歧义词条，即一个文档只能隶属一种类别。

四、定义与符号说明

符号	含义
A	目标层判断矩阵
B_i	第二层第 i 个准则的判断矩阵
W_i	第 i 个对象的权重向量
λ	特征值
CI	一致性检验指数
CR	一致性比例
$\delta_i(k)$	第 i 个指标的第 k 个指标的关联系数
r_i	第 i 个对象灰色加权关联度
$D(t_1,t_2,...,t_n)$	文本特征向量，ti 表示各个特征项
$f(n)$	表示染色体的适应度

V	用户兴趣向量
δ	用户爱好程度

五、模型的建立与求解

5.1 问题一的分析与求解

5.1.1 数据的处理

在建模过程中，我们选取了日常生活中较为常见的十个搜索引擎为评价对象。对于评价指标，我们选取了四大类指标：索引数据库功能、检验功能、检验效果、用户体验，分别对应信息量、功能多样化、检索效率、用户评价四个相对完整的评价体系。为了得到可量化的数据，对模型中所需要的评价指标，我们通过搜索相关概念、查阅相关资料[1]，对标引广度、标引深度、过滤功能、基本检索功能实现、高级检索功能实现、交互程度、特色功能制定量化标准：采用返回搜索结果第一页前 20 个相关条目，针对指标进行数量判断；对于基本搜索和高级搜索功能实现、交互功能、特色功能则用被评价搜索引擎相关功能数量除以所有搜索引擎相关功能计算比率，得出评分如下表：

表 5.1.1 定性指标得分值及量化标准

评价指标	百度	谷歌中国	必应	搜搜	360搜索	量化标准
标引广度	8	12	4	8	5	前 20 条目中非纯网页的数量
标引深度	19	19	12	17	10	前 20 条目中的深度
基本检索功能	0.909	0.818	0.636	0.727	0.611	11 项基本搜索功能包含比率/%
高级检索功能	0.889	0.889	0.778	0.667	0.596	9 项高级搜索功能的包含比率/%
内容过滤	6	2	3	4	3	前 20 条目中不良网站和有害信息结果的数量
特色功能	0.791	0.837	0.279	0.349	0.766	13 项交互功能的实现比率/%

交互程度	0.75	0.667	0.583	0.667	0.583	43 项特色功能实现比率/%

评价指标	搜狗	有道	雅虎中国	维基百科	宜搜	量化标准
标引广度	10	6	10	4	4	前 20 条目中非纯网页的数量
标引深度	10	6	10	4	4	前 20 条目中的深度
基本检索功能	0.875	0.694	0.674	0.598	0.610	11 项基本搜索功能包含比率/%
高级检索功能	0.832	0.741	0.633	0.599	0.623	9 项高级搜索功能的包含比率/%
内容过滤	3	6	2	5	4	前 20 条目中不良网站和有害信息结果的数量
特色功能	0.476	0.775	0.693	0.643	0.566	13 项交互功能的实现比率/%
交互程度	0.741	0.556	0.511	0.486	0.498	43 项特色功能实现比率/%

定量指标为标引数量、更新频率、查全率、查准率、死链数、查全率、查准率、检索时间，选择合适的评价角度，根据搜索引擎官方公布数据和查阅相关资料得到数据；对于用户界面体验，采用小范围内的问卷调查，得到十分制得分，总体数据如下表:

表 5.1.2 定量指标数值

评价指标	百度	谷歌中国	必应	搜搜	360 搜索	量化标准
标引数量	1	0.82	0.78	0.64	0.92	各搜索引擎数据库总量之比
更新频率	7	9	3	4	6	平均大规模更新次数/(次

						/(15d))
查准率	37	32	38	38	20	准确条目占搜索结果比率/%
查全率	0.37	0.45	0.23	0.28	0.20	搜索结果占总相关条目比率/%
死链接率	0.12	0.19	0.18	0.24	0.20	搜索结果无法显示比率/%
检索时间	0.001	0.002	0.0015	0.003	0.0025	搜索返回时间平均值/s
用户界面	6	9	8	6	7	问卷调查十分制得分

评价指标	搜狗	有道	雅虎中国	维基百科	宜搜	量化标准
标引数量	0.93	0.77	0.66	0.64	0.66	各搜索引擎数据库总量之比
更新频率	8	4	6	4	3	平均大规模更新次数/(次、(15d))
查准率	36	31	29	26	24	准确条目占搜索结果比率/%
查全率	0.40	0.21	0.23	0.19	0.17	搜索结果占总总相关条目比率/%
死链接率	0.18	0.17	0.21	0.21	0.22	搜索结果无法显示比率/%
检索时间	0.0015	0.0025	0.003	0.003	0.004	搜索返回时间平均值/s
用户界面	6	8	8	8	7	问卷调查十分制得分

对上述数据运用极值差法进行标准化无量纲处理，其中内容过滤、死链接率、检索时间是逆向指标，其余均为正向指标，得到处理结果如下表：

表 5.1.3 指标数据标准化

综合指标	评价指标	百度	谷歌中国	必应	搜搜	360 搜索
索引数据库功能	标引数量	1.0000	0.5000	0.3889	0.7778	0
	标引广度	0.5000	1.0000	0	0.5000	0.1250
	标引深度	1.0000	1.0000	0.2222	0.7778	0
	更新频率	0.6667	1.0000	0	0.1667	0.5000

检验功能	基本检索功能	1.0000	0.7074	0.1222	0.4148	0.0418
	高级检索功能	1.0000	1.0000	0.6212	0.2423	0
检验效果	查准率	0.9286	0.5714	1.0000	1.0000	0.2857
	查全率	0.7143	1.0000	0.2143	0.3929	0.1071
	内容过滤	0	1.0000	0.7500	0.5000	0.7500
	死链接率	1.0000	0.4167	0.5000	0	0.3333
	检索时间	1.0000	0.6667	0.8333	0.3333	0.5000
用户体验	用户界面	0	1.0000	0.6667	0	0.3333
	特色功能	0.9716	1.0000	0	0.1254	0.8728
	交互程度	1.0000	0.6856	0.3674	0.6856	0.3674

综合指标	评价指标	搜狗	有道	雅虎中国	维基百科	宜搜
索引数据库功能	标引数量	0.8056	0.3611	0,0556	0	0.0556
	标引广度	0.7500	0.2500	0.7500	0	0
	标引深度	0.7778	0.3333	0.1111	0	0.1111
	更新频率	0.8333	0.1667	0.5000	0.1667	0
检验功能	基本检索功能	0.8907	0.3087	0.2444	0	0.0386
	高级检索功能	0.8055	0.4949	0.1263	0.0102	0.0922
检验效果	查准率	0.7857	0.5000	0.3571	0.1429	0
	查全率	0.8214	0.1429	0.2143	0.0714	0
	内容过滤	0.7500	0	1.0000	0.2500	0.5000
	死链接率	0.5000	0.5833	0.2500	0.2500	0.1667
	检索时间	0.8333	0.5000	0.3333	0.3333	0
用户体验	用户界面	0	0.6667	0.6667	0.6667	0.3333
	特色功能	0.3530	0.8889	0.7419	0.6523	0.5143
	交互程度	0.9659	0.2652	0.0947	0	0.0455

5.1.2 模型分析

问题一属于评价问题，需要通过各个评价指标进行权重分析从而对搜索引擎进行评价，而层次分析法正是多层次多因素权重确定的有效工具。通过层次分析法构造判断矩阵导出排序权值，使定性的问题定量化，使结果更加客观。但仅用层次分析会使问题过于简单，而且带有很大的主观性。故将层次分析法和灰色关联分析法进行综合应用，可以吸收两种方法的优点，互补不足，而且提高了评价结果的客观性和准确性。

5.1.3 层次分析法分析及检验原理

1、构造判断矩阵

设比较 k 个因子$C=\{C_1,C_2,...,C_k\}$对某因素 B 的的影响大小，每次取两因子C_i,C_j最 B 影响，构造判断矩阵，判断矩阵权重（下层因素对上层因素影响相对重要程度）的计算方法采用几何平均法求权重，具体计算步骤如下：

①先计算$A=(C_{ij})_{k\times k}$中每行每列所有元素的几何平均值得到向量$M=[m_1,m_2,\ldots,m_k]^T$，

其中，$m_i=\sqrt[k]{\prod_{i=1}^{k}C_{ij}}(i,j=1,2,\ldots,k)$；

②对 M 作归一化处理，得到相对权重向量$W=[W_1,W_2,\ldots,W_k]^T$，其中，

$$W_i=m_i/\sum_{j=1}^{k}m_j(i,j=1,2,\ldots k)$$

2、层次权重单排序与一致性检验

判断矩阵 A 对应的最大特征值λ_{max}的特征向量 W ，经归一化后即为同一层次相应因素对于上一层次某因素相对重要性的排序权值，这一过程称为层次单排序；排序后引入完全判断一致性比例 CR 来衡量判断矩阵的一致性。具体计算步骤如下：

（1）计算判断矩阵 A 的特征值，公式为：

$$AW=\lambda_{\max}W \tag{1}$$

（2）计算完全一致性检验指数 CI，公式为：

$$CI=\frac{\lambda_{max}-k}{k-1} \tag{2}$$

（3）查表确定相应的平均随机一致性指标 RI

（4）计算一致性比例（CR），公式为：

$$CR=\frac{\mathrm{CI}}{\mathrm{RI}} \tag{3}$$

对于准则层各个因素，CR ＜ 0. 10 时，则认为判断矩阵的一致性是可以接受的，否则应对判断矩阵作适当修正。

3、层次权重总排序与一致性检验

确定某层所有因素指标对于总目标相对重要性的排序过程，成为层次总排序。从最高层到最低层逐层进行。设：

B 层 m 个因素B_1,B_2，…，B_m 对总目标 A 的排序为$b_1,b_2,\ldots$，b_m ；

C 层 n 个因素C_1,C_2，…，C_n 对上层 B 中因素为B_j的层次单排序为：

$C_{1j},C_{1j},\ldots,C_{nj}$ （j=1，2，…，m）

则 C 层第 j 个因素对总目标的权值为：$c_i=b_ic_{ij}$

设 C 层中与 B_j相关的因素的成对比较判断矩阵在单排序中经一致性检验，单排序一致性指标为 CI（j），相应的平均随机一致性指标为 R I（j），

则 C 层一致性比例为 CR，$CR=\dfrac{\sum_{j=1}^{m}CI(j)\,b_j}{\sum_{j=1}^{m}RI(j)b_j}$。对于准则层各个因素，CR ＜ 0. 10，则认为层次总排序结果具有较满意的一致性，并接受该分析结果。

5.1.4 数据灰色关联分析法分析及检验原理

（1） 确定比较对象（评价对象）和参考数列（评价标准）。设评价对像有 m 个，评价指标有 n 个，参考数列为 $x_0=\{x_0(\mathrm{k})\mid \mathrm{k}=1,2,\cdots\cdots,\mathrm{n}\}$，比较数列为 $x_i=\{x_i(\mathrm{k})\mid \mathrm{k}=1,2,\cdots,\mathrm{n}\},\mathrm{i}=1,2,\cdots,m$ 。

（2） 确定各指标值对应的权重。此题已用层次分析法确定各指标对应的权重 $w=[w_1,\cdots,w_n]$，其中 $w_k(\mathrm{k}=1,2,\cdots,\mathrm{n})$ 为第 k 个评价指标对应的权重。

（3） 计算灰色关联系数

$$\delta_i(\mathrm{k})=\frac{\min\limits_s\min\limits_t|x_0(\mathrm{t})-x_s(\mathrm{t})|+\rho\max\limits_s\max\limits_t|x_0(\mathrm{t})-x_s(\mathrm{t})|}{|x_0(k)-x_i(k)|+\rho\max\limits_s\max\limits_t|x_0(\mathrm{t})-x_s(\mathrm{t})|} \tag{4}$$

为比较数列 x_i 对参考数列 x_0 在第 k 个评价指标上的相关系数，其中 $\rho\in[0,1]$ 为分辨系数。其中，称 $\min\limits_s\min\limits_t|x_0(\mathrm{t})-x_s(\mathrm{t})|$、$\max\limits_s\max\limits_t|x_0(\mathrm{t})-x_s(\mathrm{t})|$ 分别为两级最小差及两级最大差。一般来讲，分辨系数 ρ 越大，分辨率越大；ρ 越小，分辨率越小。

（4） 计算灰色加权关联度。灰色加权关联度的计算公式为

$$r_i=\sum_{i=1}^{n}w_i\delta_i(\mathrm{k}) \tag{5}$$

其中 r_i 为第 i 个评价对象对理想对象的灰色加权关联度。

（5）模型分析。根据灰色加权关联度的大小，对各评价对象进行排序，可建立评价对象的关联序，关联度越大，其评价结果越好。

5.1.5 模型的建立与求解

1. 建立层次结构模型

通过查阅网络搜索引擎评价指标的论文，结合在近几年文献中出现频率较高的评价指标，以及指标是否能查询到可以量化的数据，该模型采用查全率、查准率等 14 个主要因素指标为依据建立层次模型，具体如下

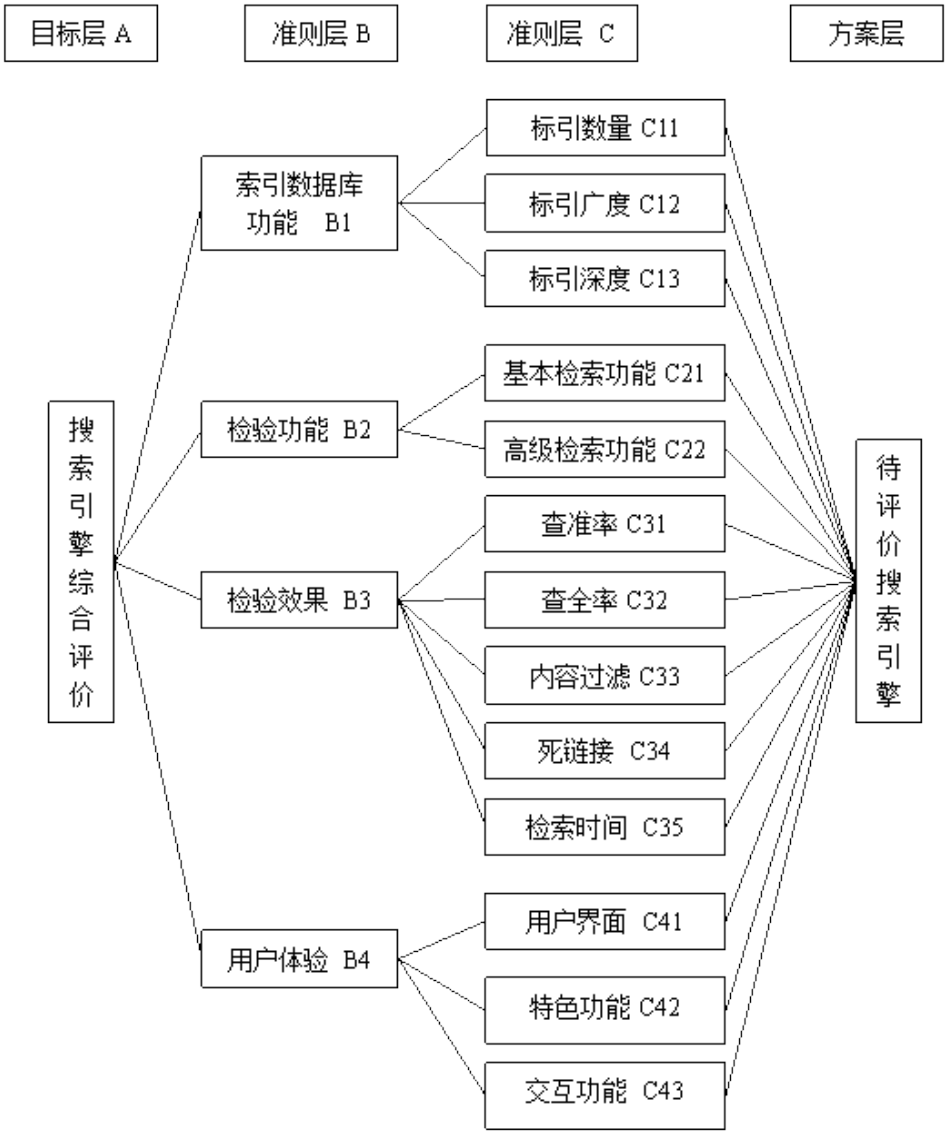

图 5.1.1　层次模型图

2．构造判断矩阵

根据相应文献查阅[2]及数据分析，结合领域专家意见，运用 Saaty 提出以 1- 9 为标度值的方法进行矩阵标度判断，经相应参考及数据分析，得到 B 层对 A 层指标相对重要性比较:

表 5.1.4 准则层 B 对目标层 A 的重要性比较

A-B	B1	B2	B3	B4
B1	1	2	1/3	3
B2	1/2	1	1/5	2
B3	3	5	1	7

B4	1/3	1/2	1/7	1

根据上表，得出判断矩阵：

A=$\begin{pmatrix} 1 & 2 & 1/3 & 7 \\ 1/2 & 1 & 1/5 & 2 \\ 3 & 5 & 1 & 7 \\ 1/3 & 1/2 & 1/7 & 1 \end{pmatrix}$ 通过 matlab 编程计算矩阵 A 的几何平均值

$m_i = \sqrt[k]{\prod_{i=1}^{k} c_{ij}}$ ，得 $M = [1.5833, 0.9250, 4.000, 1.976]^T$，对 M 进行归一处理得到权重向量 $W = [0.2179, 0.1228, 0.5872, 0.0723]^T$，根据 A W= λW 进行特征值的运算，得到λ = 4.0129利用公式 $CI = \frac{\lambda_{max} - k}{k-1}$ 计算一致性检验指数 CI = 0.0064,根据查表得到平均随机一致性指标 RI=0.90,利用 $CR = \frac{CI}{RI}$，求得 CR=0.0071<0.10，可知，判断矩阵的整体一致性是可接受的。同理可分别得到 B1、B2、B3、B4 的判断矩阵分别为：

B1=$\begin{pmatrix} 1 & 3 & 2 & 7 \\ 1/3 & 1 & 1/2 & 3 \\ 1/2 & 2 & 1 & 5 \\ 1/7 & 1/3 & 1/5 & 1 \end{pmatrix}$，　　B2=$\begin{pmatrix} 1 & 3 \\ 1/3 & 1 \end{pmatrix}$，

B3=$\begin{pmatrix} 1 & 3 & 5 & 4 & 3 \\ 1/3 & 1 & 5 & 4 & 3 \\ 1/5 & 1/5 & 1 & 1/3 & 1/4 \\ 1/4 & 1/4 & 3 & 1 & 1/2 \\ 1/3 & 1/3 & 4 & 2 & 1 \end{pmatrix}$，　　B4=$\begin{pmatrix} 1 & 1/3 & 1/5 \\ 3 & 1 & 1/3 \\ 5 & 3 & 1 \end{pmatrix}$

3．各层次单排序及一致性检验

通过 matlab 编程计算得到 A-B 层，B-C 层各比较矩阵的权重 w_i 、特征值 λ 随机一致性指标 CI 及一致性比率 CR 如下表

表 5.1.5 准则层 B 层对决策层 A 层权重

权重 w_i	特征值 λ	CI	CR
0.2179 0.1228 0.5872	4.0192	0.0064	0.7223

表 5.1.6 准则层 C 对准则层 B 的权重

P	权重 w_i	特征值 λ	CI	CR
索引数据库功能 B1	0.4905 0.1615 0.2878 0.0601	4.0192	0.0064	0.0071
检验功能 B2	0.0075 0.0025			
检验效果 B3	0.4373 0.2774 0.0488 0.0925 0.1440	5.2955	0.0739	0.0660
用户体验 B4	0.1047 0.2583 0.6370	3.0385	0.0193	0.00332

4．层次总排序及一致性检验

根据公式可得出各个判断指标相对于总目标，即在搜索引擎评价中所占的权重由大到小排列为：

表 5.1.7 各评价指标所占评价权重表

指标	查准率	查全率	标引数量	基本检索功能	死链接
权重	0.2568	0.1629	0.1069	0.0921	0.0846
指标	标引深度	检索时间	交互程度	标引广度	高级检索功能
权重	0.0627	0.0543	0.0406	0.0352	0.0307
指标	内容过滤	特色功能	更新频率	用户界面	
权重	0.0287	0.0186	0.0131	0.0076	

5．灰色关联法求解

此模型中比较对象为 10 个搜索引擎，参考数列为 14 个评价标准，为之前所确定的评价指标，各个参考指标对应的权重已通过层次分析法计算出来。

通过 matlab 计算灰色关联系数、灰色加权关联度、评价对象排名先后结果如下

Xishu=

$$\begin{bmatrix}
1.0000 & 0.5000 & 0.4500 & 0.6923 & 0.3333 & 0.7200 & 0.4390 & 0.3462 & 0.3333 & 0.3462 \\
0.5000 & 1.0000 & 0.3333 & 0.5000 & 0.3636 & 0.6667 & 0.4000 & 0.6667 & 0.3333 & 0.3333 \\
1.0000 & 1.0000 & 0.3913 & 0.6923 & 0.3333 & 0.6923 & 0.4286 & 0.3600 & 0.3333 & 0.3600 \\
0.6000 & 1.0000 & 0.3333 & 0.3750 & 0.5000 & 0.7500 & 0.3750 & 0.5000 & 0.3750 & 0.3333 \\
1.0000 & 0.6308 & 0.3629 & 0.4607 & 0.3429 & 0.8206 & 0.4197 & 0.3982 & 0.3333 & 0.3421 \\
1.0000 & 1.0000 & 0.5689 & 0.3976 & 0.3333 & 0.7199 & 0.4975 & 0.3640 & 0.3356 & 0.3552 \\
0.8750 & 0.5385 & 1.0000 & 1.0000 & 0.4118 & 0.7000 & 0.5000 & 0.4375 & 0.3684 & 0.3333 \\
0.6364 & 1.0000 & 0.3889 & 0.4516 & 0.3590 & 0.7368 & 0.3684 & 0.3889 & 0.3500 & 0.3333 \\
0.3333 & 1.0000 & 0.6667 & 0.5000 & 0.6667 & 0.6667 & 0.3333 & 1.0000 & 0.4000 & 0.5000 \\
1.0000 & 0.4615 & 0.5000 & 0.3333 & 0.4286 & 0.5000 & 0.5455 & 0.4000 & 0.4000 & 0.3750 \\
1.0000 & 0.6000 & 0.7500 & 0.4286 & 0.5000 & 0.7500 & 0.5000 & 0.4286 & 0.4286 & 0.3333 \\
0.3333 & 1.0000 & 0.6000 & 0.3333 & 0.4286 & 0.3333 & 0.6000 & 0.6000 & 0.6000 & 0.4286 \\
0.8585 & 1.0000 & 0.3333 & 0.3638 & 0.7971 & 0.4359 & 0.8182 & 0.6596 & 0.5899 & 0.5073 \\
1.0000 & 0.6140 & 0.4415 & 0.6140 & 0.4415 & 0.9362 & 0.4049 & 0.3558 & 0.3333 & 0.3438
\end{bmatrix}$$

gsort =

0.8592 0.7146 0.7134 0.6312 0.6006 0.4550 0.4330 0.4050 0.3661 0.3494

ind =

1 6 2 4 3 7 8 5 9 10

（数字编号分别对应各搜索引擎，其所在位置先后代表排名先后）

故根据灰色关联分析法可得出结论，各搜索引擎的排名先后为：

百度>搜狗>谷歌中国>搜搜>必应>有道>雅虎>360 搜索>维基百科>宜搜

为了对各个搜索引擎进行评价，做出标准化后 14 个指标的折线图如下：

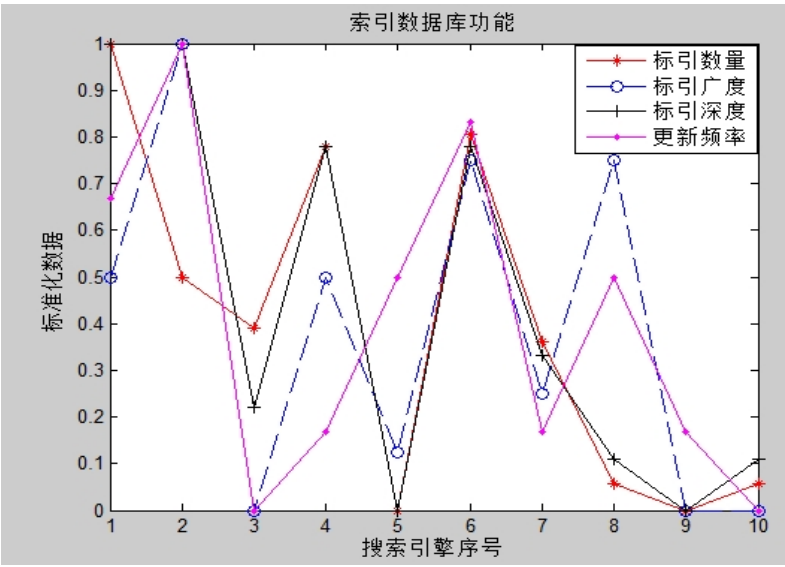

图 5.1.2 各搜索引擎基于数据库功能优劣对比图

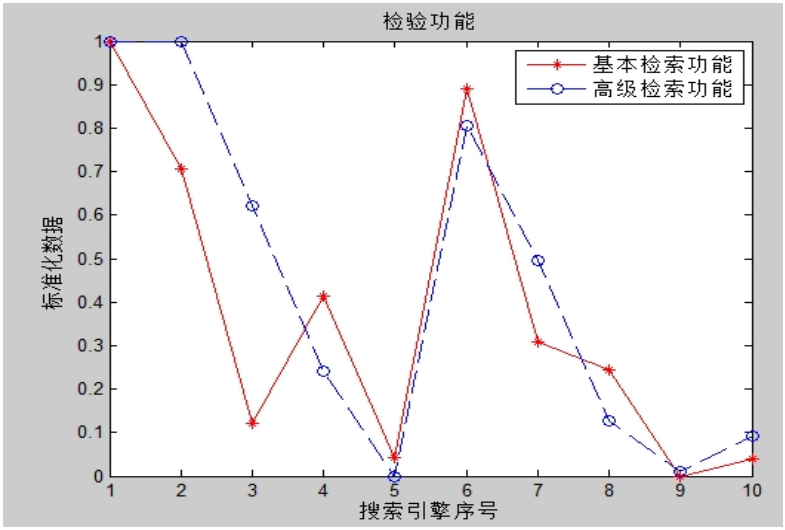

图 5.1.3 各搜索引擎基于检验功能优劣对比图

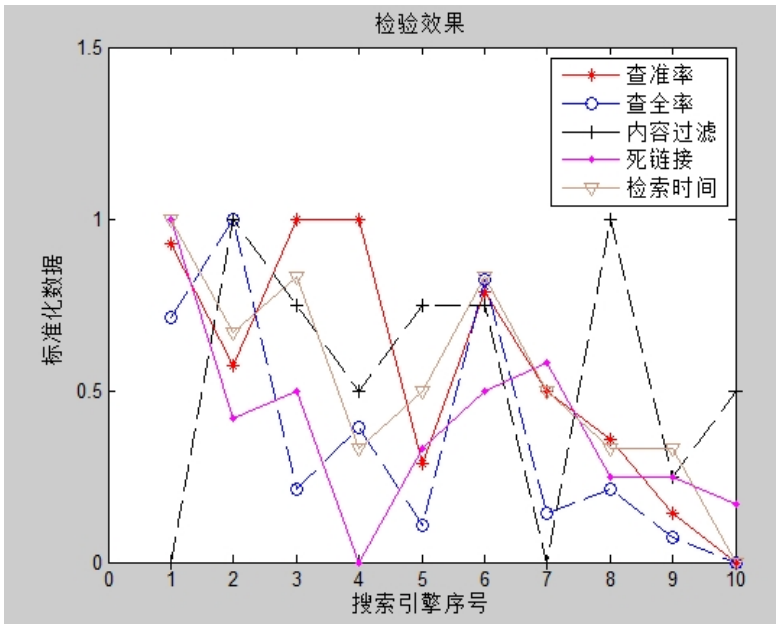

图 5.1.4 各搜索引擎基于检验效果优劣对比图

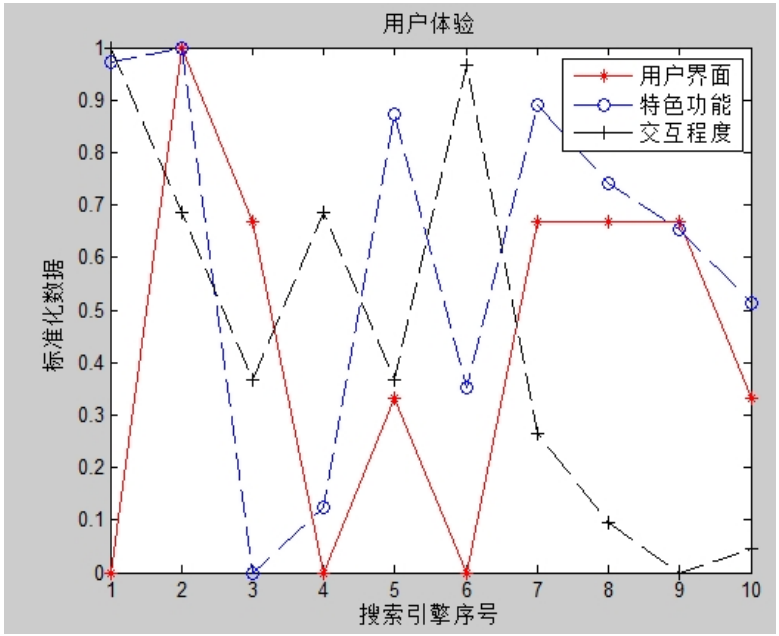

图 5.1.5 各搜索引擎基于用户体验优劣对比图

5.1.6 模型的评价和结果分析

通过各搜索引擎在 14 个评价指标优劣对比图来看，在占权重最大的大类指标——检验效果中，百度无论在查全率、查准率指标中都位居首位，这两个指标符合用户所关心的问题，对评价结果影响较大，故百度总排名靠前，而谷歌在更新频率、标引广度等指标高于百度，但我们通过图可以发现，在负向指标影响因素中，百度的过滤功能明显优于谷歌。故在搜索引擎发展过程中，除了要不断提高检索的准确性和全面性，也要注意提高引擎网站的数据库更新频率，以便减少死链接数等负面影响因素。另外要注意增强搜索引擎对不良信息的过滤以便增加用户的好感。此外，由于用户界面、特色功能指标所占权重较小，

各大搜索引擎更关注其通用性而或多或少忽视了个性化的搜索，这是各大搜索引擎的通病。

前五名的搜索引擎依次是百度、搜狗、谷歌中国、搜搜和必应搜索，此结果与题目中各搜索引擎排名进行对比发现：前两名一致，基本可以认为，目前在中国影响力较大的搜索引擎是百度和搜狗。由于新浪旗下爱问搜索引擎功能较为单一，不利于数据量化处理，统一比较，故在我们的模型中不予考虑。谷歌中国排名第三，与所给第五有些出入，这与我们获取的数据缺乏一定的时效性有关，且谷歌 2010 年宣布退出中国市场，这可能导致其排名下降，但不得不承认，谷歌仍然或曾经在中国搜索市场上占据很大一部分份额。搜搜排名第四，基本一致，另外，必应搜索入围前五，可见其影响力日益增大，分得搜索引擎市场的一杯羹。之后的有道、雅虎中国、360 搜索等的排名基本与我们使用频率大小一致，符合我们日常认知。由于对搜索引擎体系评价的指标选取不同，且评价标准也有差异，故不同的评价体系所评出的排名结果在主流搜索引擎排名一致的前提下，其他略有差异是很正常的，我们建立的模型所得出的排名结果与题目所给排名基本一致，因此，我们建立的模型是客观和成功的。

5.2 问题二的求解

5.2.1 数据处理

本模型采用“中国百科术语数据库”中的部分数据，随机选用选用百科数据库中的若干文档，数据如下：

表 5.2.1 原始实验数据表

类（卷）	哲学	科技	社会	政治	财经	教育
文档数	63	45	76	82	86	73
类（卷）	心理	军事	考古学	水利	机械	武术
文档数	61	54	66	38	76	42

由于文档集的预处理涉及到复杂的文档特征值抽取算法，需要对某一文档集的特征值作出大量的统计，最后得出数据。数据统计需要大量的人员在科学的统计方法指导下，合理有序的进行文档特征值这一数据集进行挖掘。目前国内并没有标准的文档分类数据集，限于已有的水平技术、参与论文人数和论文规定完成时间，故模型中遗传聚类算法的模拟借鉴相关文献[3]已有的统计数据。

5.2.2 模型分析

Web 以经成为人们获取信息的一个重要途径，由于 Web 信息的日益增长，人们不得不花大量的时间来搜索浏览自己所需要的信息。信息检索满足了人们的一定需要，但由于其通用性，仍然不能满足不同时期，不同背景，不同目的的需求。个性化搜索引擎技术就是针对这一问题而提出的，它为不同用户提供不同的服务，以满足不同的需求。为此我们设计一种基于遗传算法和 K-means 聚类的文档聚类算法和具备信息过滤功能的科技搜索引擎[4]，并主要对个性化搜索——科技论文文档搜索引擎的文档聚类建立了相应的数学模型。

对于大规模文档的聚类通常采用以 K-means 算法为代表的基于划分的聚类算法，K-means 算法具有良好的可伸缩性和很高的效率，适合处理大文档集。当结果簇密集并且各簇之间的区别明显时，采用此方法较好。该算法的缺点是

它要求用户必须事先给出要生成的簇的数目 k 值，不准确的 k 值会导致聚类质量的下降。此外，它对于“噪声”和孤立数据是敏感的，少量的该类数据能够对簇（类）平均值产生极大的影响。而遗传算法可以得到全局最优解，避免陷入局部极小点。故模型采用遗传算法和 K-means 算法结合，利用遗传算法思想对初始聚类中心进行优化选择，来代替 K-means 算法中随机找到初始点集的方法，克服了 K-means 算法对聚类中心十分敏感的缺陷；另外，利用可变长度的遗传算法进行初始点的选取，可以克服 K-means 算法类别人为确定的缺点。

5.2.3 模型相关原理及名词解释

1. 文本聚类

文本聚类是设计个性化搜索引擎的前提和基础，它是一个将文本集分组的全自动处理过程，是一种无监督的机器学期过程。簇是通过相关数据发现的一些组，簇类的文本和其它组相比更为相近。因此，文本聚类就是要找到这样一些簇的集合，簇之间的相似性最小而簇内部的相似性最大。其主要流程如下

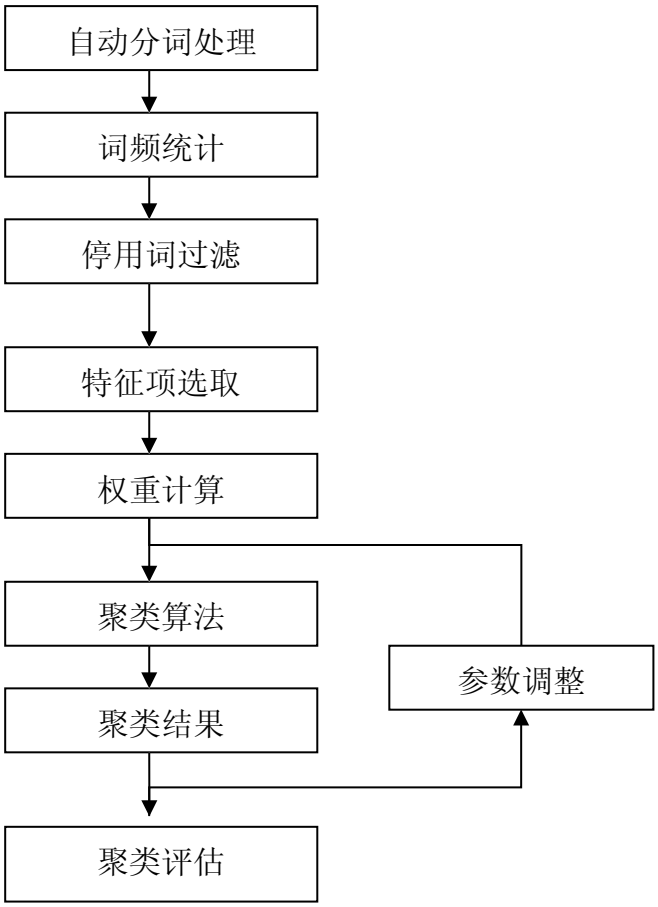

图 5.2.1 文本聚类的流程图

2. 特征项选取及文本的数学表示

[G.Salton,1988]提出向量空间模型 VSM，即使用向量表示文本，并成功应用到 SMART 系统中，它的核心概念描述如下：

项：

文本的内容被提取，由一些特征项来表示，这些特征项可以是字、词、句等语言单位，即文本表示 Document=D（t1，t2，…tn），其中，ti 表示各个特征项

权重：

对于特征项都赋予它一个权重 wi，用来表示第 i 个特征项在该文本中的重要程度。即：文本 Document=D（t1，w1,；t2，w2…tn，wn）权重大都是以特征项的频率为基础进行计算的，另外还可以加入其他信息来表示特征项的权重，比如特征项在文本中所在的位置等。

向量空间模型（VSM）：

根据上述两个概念，一个文本就形成了一个向量 Document=D（t1，w1,；t2，w2…tn，wn），其中 ti 表示该特征项所在的维度，wi 表示 ti 在该维度上的取值。即文本被表示成了特征项空间里的一个点；一个文本集就表示成一个矩阵，即特征项空间中的一个点集。

相似度：

相似度函数 Similar（D1，D2）表示两个文本之间的相似程度，在向量空间模型下，可以用相应的向量内积或者夹角余弦等方法来表示

简单遗传算法。

遗传算法的基本原理

遗传算法（Genetic Algorithms,GA）是一种基于自然选择原理和自然遗传机制的搜索算法，它是模拟自然界中的生命进化机制，在人工系统中实现特定目标的优化。遗传算法的实质是通过群体搜索技术，根据适者生存的原则逐代进化，最终得到最优解或准最优解。它必须做一下操作：初始群体的产生，求每一个体的适应度，根据使者生存的原则选择优良的个体，被选出的优良个体两两配对，通过随机交叉其染色体的基因并随机变异某些染色体的基因生成下一代群体，按此方法是群体逐代进化，直到满足进化终止条件。

5.2.4 遗传算法步骤

1.编码策略（相似度可变）

采用十进制编码，用随机数列 $\omega_1\omega_2\omega_3\omega_4...\omega_{n-1}\omega_n\omega_{n+1}$ 作为染色体，其中 $0 \le \omega \le 1$（i=2，3，・・・，n），ω_1=0；ω_2=1；每一个随机序列都和种群中的一个个体相对应

2.初始种群

先利用经典的近似算法——改良圈算法求得一个较好的初始种群。对于随机产生的初始圈：

$c = \pi_1...\pi_{u-1}\pi_u\pi_{u+1}...\pi_{v-1}\pi_v\pi_{v+1}...\pi_{n+1}$，记 $\Delta f = (d_{\pi_{u-1}\pi_v} + d_{\pi_u\pi_{v+1}}) - (d_{\pi_{u-1}\pi_u} + d_{\pi_v\pi_{v+1}})$，

若 $\Delta f < 0$ 以新路径修改旧路径，直到不能修改为止，就得到一个比较好的可行解。

直到产生 M 个可行解，并把这 M 个可行解转换成染色体编码。

3.相似度的定义及计算法

K 进制编码，码长为 L，种群数目为 N。令 c（1，n）表示中第 n 个个体编码的第 1 位，是 K 进制取值；s（看 k，l，n）表示种群中第 n 个个体的第 1 位的逻辑变量；m（k，1）表示种群中第 1 位为 k 的个体所占的百分比；θ（l）表示种群第 1 位的相似度。则种群相似度 θ 可由以下公式计算得出：

$$s(k,l,n)=\begin{cases}1, c(l,n)=k\\ 0, c(l,n)\neq k\end{cases}, k=0,\ldots,K-1; l=1,\ldots,L; n=1,\ldots,N$$

$$m(k,l)=\frac{1}{N}\sum_{N=1}^{N} s(k,l,n), k=0,\ldots,K-1; l=1,\ldots,L$$

$$\theta(l)=MAX(m(0,l),\ldots,m(K-1,l)), l=1,\ldots,L$$

$$\theta=\frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}}\theta(\mathrm{l}) \tag{6}$$

其中，相似度的值域为 $\frac{1}{K}\le\theta\le 10$ 。

4. 目标函数

目标函数为侦察所有目标的路径长度，适应度函数就去为目标函数。我们要求：

$$\min \mathrm{f}(\pi_1\pi_2\ldots\pi_n\pi_{n+1})=\sum_{\mathrm{i}=1}^{\mathrm{n}+1} d_{\pi_i\pi_{i+1}}$$

5. 交叉操作

交叉操作采用单点交叉。对于选定的两个父代个体 $f_1=\omega_1\omega_2\omega_3\omega_4\ldots\omega_{n-1}\omega_n\omega_{n+1}$，$f_2=\omega_1'\omega_2'\omega_3'\omega_4'\ldots\omega_{n-1}'\omega_n'\omega_{n+1}'$，随机地选取第 t 个基因处为交叉点，经过交叉运算后得到的子代个体为 S_1 和 S_2，S_1 的基因由 f_1 的前 t 个基因和 f_2 的后 n+1-t 个基因构成，S_2 的基因由 f_2 的前 t 个基因和 f_1 后 n+1-t 个基因构成。

交叉的方式有很多选择，应该尽可能选取好的交叉方式，保证子代能继承父代的优良特性。同时这里的交叉操作也蕴含了变异操作。

6. 变异操作

变异也是实现群体多样性的一种手段，同时也是全局寻右的保证。按照给定的变异率，对选定变异的个体，随机地取三个整数，满足 1<u<v<w<n+1，把 u,v 之间（包括 u 和 v）的基因段插到 w 后面。

7. 选择

采用确定性的选择策略，也就是在父代种群和子代种群中选择目标函数值最小的 M 个个体进化到下一代，这样可以保证父代的优良特性被保存下来。

5.2.5 K-means 聚类算法

已知 d 维空间 R^d，在 R^d 中定义一个评价函数 c：{X：X⊆S}->R^+,给每一个聚类一个量化的评价，输入 R^d 中的对象集合 S 和一个整数 k，要求输出 S 的一个划分：$s_1, s_2, \ldots, s_k$ 这个划分使 $\sum_{i-1}^{k} c(\mathrm{S}_i)$ 最小化。不同的评价函数将产生不同的聚类结果，最常用的评价函数定义如下

$$c(\mathrm{S}_i)=\sum_{r=1}^{|s_i|}\sum_{s=1}^{s_i|}(\mathrm{d}(x_i^r, x_s^i))^2$$

其中，S_i 为划分形成的簇，x_i^r, x_s^i 分别为 S_i 的第 r 个和第 s 个元素，|S_i|表示元素个数，$\mathrm{d}(x_i^r, x_s^i)$ 为 x_i^r, x_s^i 的距离

该算法不断计算 S_i 的中心 $\bar{x}^{i}$，也就是聚类 S_i 中对象的平均值，作为新的聚类种子。实际使用的评价函数为

$$c(\mathrm{S}_i)=\sum_{r=1}^{|\mathrm{S}_i|} d\left(\bar{x}^{i} \quad x_r^i\right) \tag{7}$$

其中，$\bar{x}^{i}$ 为 S_i 的中心，其他符号同上式。

K-means 算法具体过程如下：

（1） 按一定原则选择 c 个类别中心；

（2） 在第 K 次迭代中，对任意一个样本，求其到 c 个中心的距离，将该样本归到距离最短的中心所在的类；

（3） 利用均值等方法更新该类的中心值；

（4） 对于所有的 c 个聚类中心，如果利用（2）（3）的迭代更新后，值保持不变，则迭代结束，否则继续迭代。

通过遗传算法找到适合聚类分析的初始类中心集 $v^{(0)}$，其中 $V=\{v_1, v_2, \ldots v_c\}$。利用文本的数学表示方法，即 Web 文档可以表示为 $Document=D(t_1, w_1; t_2, w_2 \ldots, t_n, w_n)$ 形式。

5.2.6 模型建立与求解

1.为了得到相对可靠的数据且便于对模型验证，我们选择“中国百科术语数据库”中的部分数据进行建模，随机选用百科全书数据库中若干卷的若干术语条目，得到数据，选择 12 卷的 762 个术语条目作为种群大小，随机产生由字符构成的初始种群。

2.将所选取的 762 个文档向量编码，形成一条染色体:

$$D_1, D_2, D_3, \ldots, D_{762}$$

其中 $D_i=(t_1,w_1;t_2,w_2\cdots,t_n,w_n)$ ，即一条染色体是由 c 个文档组合而成，长度为 c×n，形式如下所示：

$$w_{11},w_{12},\cdots,w_{1n},\ w_{21},\ w_{22},\cdots w_{2n},\cdots w_{i1},\cdots,\ w_{ij},\cdots w_{in},\cdots w_{762,1},\ w_{762,2},\ w_{762n}$$

在上面的染色体中前 n 项表示文档 D_1，依次类推，c 个 n 维文档组成了一条长度为 c×n 的染色体，该编码采用实数编码方式。

3.根据公式（6）计算种群相似度，相似度大于阀值 θ_0 时，在每个染色体后增加 n 个代码，即增加一个文档向量，也就是增加了下一阶段文档聚类类别个数。

4.计算每个个体的适应度

适应度反映个体的生存能力，是遗传算法的驱动力。Web 文档聚类可用染色体长度作为适应度的衡量标准。过长的染色体会使聚类个数过大，即聚类类别过多，无法使相似的文档聚集在一起；过短的染色体会导致相似度过低的文档划分到同一类中，因此，用 f（n）表示染色体的适应度，则 f（n）的取值范围是[0，1]。计算染色体的适应度。当 f（n）=0 时，表示该染色体过长或过短；当 f（n）=1 时，表示该染色体具有很强的繁衍能力。得到繁衍能力高的染色体。

5. 根据相关原理将繁衍能力高的染色体进行选择、交叉和变异操作；

6.不断重复（3）--（5），直到找到适应度趋近于 1 的染色体，即全局最优解为止。

通过计算，得到最优染色体的形式如下所示：

$$w_{11},w_{12},\cdots,w_{1n},\ w_{21},\ w_{22},\cdots w_{2n},\cdots w_{i1},\cdots,\ w_{ij},\cdots w_{in},\cdots w_{c1},\ w_{c2},\cdots w_{cn}$$

7．其中第 1 至第 n 为编码代表第一个聚类中心点，以此类推，从而得到 c 个聚类中心点，作为 K—means 聚类的初始中心点，便于 Web 文档聚类分析。

8． 我们根据遗传算法过程操作得到的染色体，计算得到 12 个最优初始中心点，作为 K-means 算法的初始聚类种子。

9．根据此 12 个聚类种子，通过公式（6）计算每个文档与所给聚类种子的相似度，将所研究的文档对象重新赋给最相似的簇。

10 更新聚类种子，即通过公式（7）重新计算每个簇中对象的平均值，用作象均值点作为新的聚类种子。

11 重复步骤 9、10，直到各个簇不在发生变化为止。

通过 Matlab 程序得出聚类分析的模型图如下：

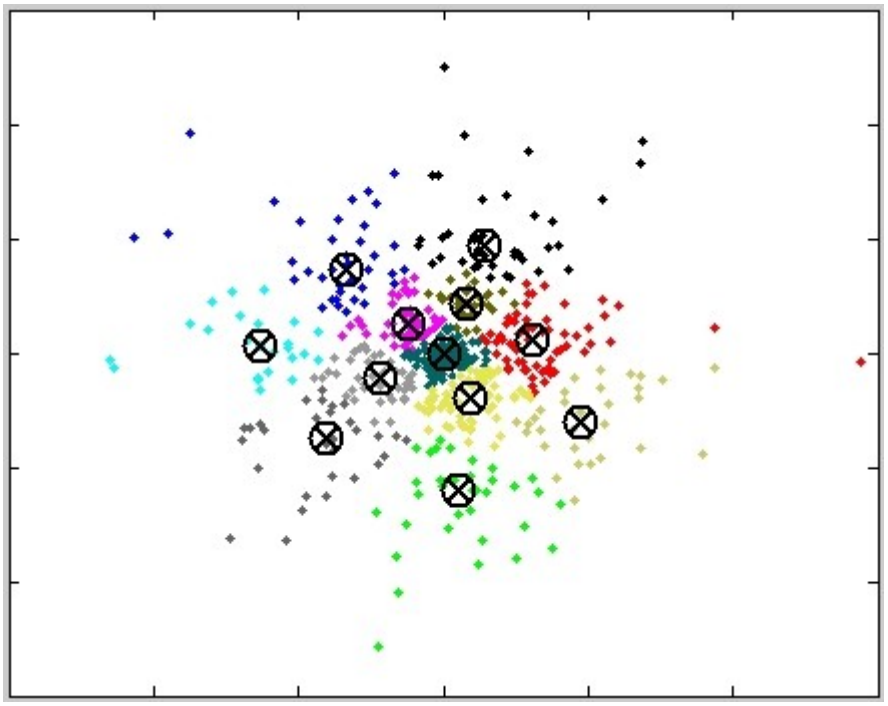

图 5.2.2 文档聚类分析结果图

建模过程中，762 篇文档得到 12 簇数据，即这些文档被划分为 12 个不同的种类，是建立在我们人为划分分类依据的。我们选用{科技，创新技术，科学实验，研究成果，科技网，硅谷，技术革新}等关键词为科技论文的评选标准，编码后与得到的 12 类分类结果进行相关性计算，相关性大的即可认为属于科技论文类文档，通过搜索引擎的排序功能将此类置前，从而实现检索科技论文文档类的搜索功能，即达到了我们设计个性化搜索引擎[5]的目的。

5.2.7 模型检验及分析

由于数据处理的中间过程较为繁琐，且依据数据来源和检验的可靠性，此模型的检验采用相对简易的检验方法，即通过查阅相关资料信息，通过 matlab 计算出基于我们所建立的模型对此 762 篇文档的聚类结果，并与数据库现有的分类结果对比，通过分析相关文档数的差异来对模型进行评价。具体如下：

表 5.2.2 聚类文档与分类文档数目比较表

类（卷）	哲学	科技	社会	政治	财经	教育
分类文档数	63	45	76	82	86	73
聚类文档数	61	45	70	88	85	69
类（卷）	心理	军事	考古学	水利	机械	武术
分类文档数	61	54	66	38	76	42
聚类文档数	78	61	74	28	68	35

：

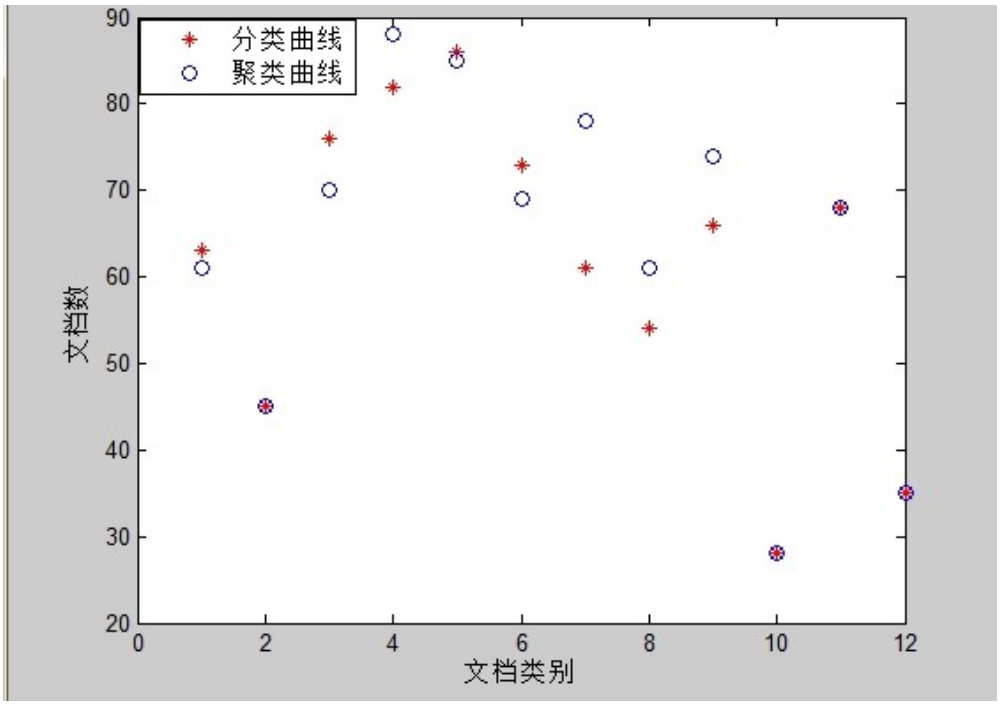

图 5.2.3 分类和聚类文档数对比点图

根据对比点图的结果可以看出：通过我们模型模拟出来的各聚类文档数与数据库中文档的分类数基本一致，没有出现“两极分化”的现象，即可以认为该模型能够有效地对文档进行聚类分析，误差在较小的范围内。进一步说明了基于遗传算法改进的 K-means 聚类算法有效解决了单独用 K-means 算法时对孤立数据敏感的弊端，具有很大的优越性。

5.3 问题三的求解

5.3．1 模型分析

个性化信息服务越来越成为信息检索领域中研究的热点，然而现今的搜索引擎往往具有通用性，排序机制又没有考虑到用户的偏好，因而很难满足不同背景、不同目的的用户需求。

我们在问题二中建立的模型已经设计出了个性化搜索引擎，不过只是针对科技论文文档的搜索，若要考虑将个性化搜索引擎进行一般性推广，则需要将用户兴趣，搜索历史等信息进行挖掘，建立了包含用户兴趣的文档个性化搜索引擎。我们建立用户兴趣的向量空间检索模型，自动提取用户兴趣的相关信息，结合问题二中对文档的聚类分析，过滤掉不符合用户信息的类别，从而实现了基于用户兴趣对不同文档进行检索输出的功能，即实现了个性化搜索引擎的一般推广。

5.3．2 模型的相关原理

1．用户兴趣模型的表达

为有效的表达用户兴趣偏好，有必要为每个用户建立一个用户兴趣模型，即用户描述文件，它能有效的表达用户的特征与用户的标准。目前为止该模型的表达没有统一标准。本题利用加权兴趣树模型表达用户的兴趣爱好，它融合了向量空间模型和信息过滤模型，有效的表达用户的领域偏好。

2.用户模型结构

用户模型结构由文档预处理、文档分类、兴趣生成以及兴趣更新四部分组成。

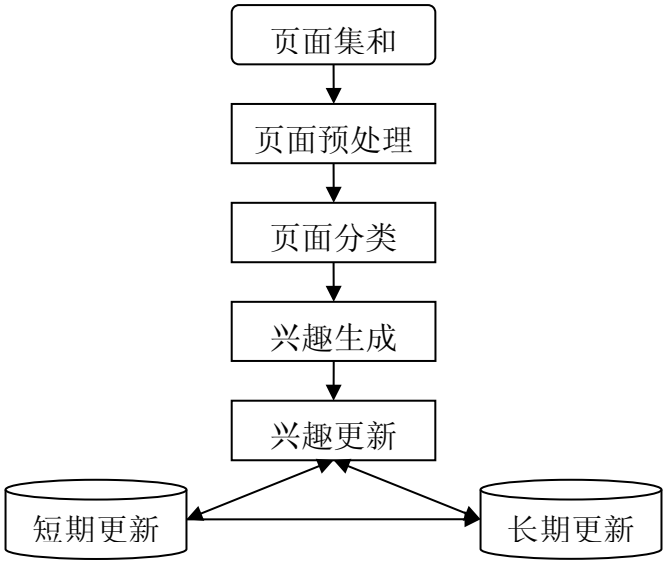

图 5.3.1　用户模型结构图

其中页面预处理、页面分类两个问题已经在问题二得到解决。用户模型主要目的就是兴趣生成。

用户兴趣类向量表示方法

Step 1：统计用户兴趣模型中的文档数；

Step 2：提取文档特征集 K={k1,k2,・・・,km}；

Step 3：统计特征词 k_i 在文档中出现的次数 n_i；

Step 4：计算各特征词的权值

3.用户兴趣特征挖掘与提取[6]

用户兴趣的挖掘与提取是建立用户兴趣偏好模型的第一步，高效并且有效的用户兴趣模型有助于提高整个系统的效率和提高用户描述文件对用户兴趣描述的准确度。

用户兴趣向量[7]的挖掘分为两部分：

（1）用户信息和自身访问历史记录的挖掘；

（2）同一时期不同用户访问记录的挖掘。

数学模型表示为：

$$V = aV_1 + (1-a)V_2 \tag{8}$$

其中，V_1 代表用户自身信息训练的兴趣向量，V_2 代表由其他用户信息训练的兴趣向量。a 和（1-a）是依据向量，V_1，V_2 在挖掘用户兴趣向量是所起作用的大小规定的影响因子。

4.用户注册信息和自身访问历史记录的挖掘。

为用户在数据库中建立一个用户行为表，用来记录每一个用户最近一段时间访问过的网页及访问时的相关数据，在进行用户信息注册时为该用户建立相应记录，其内容包括：用户 ID、访问的网页名称、该次访问查询关键词、该网页在本地硬盘的位置、网页的 URL、点击次数、用户访问时间长度、最后一次

访问时间、网页长度。该表按照最后一次访问时间倒序排列。当表内容过多时，替换最久没访问的网页，也就是最后一次访问时间最远的记录。

将用户行为表中存在的记录对应的网页向量提取出来，点击次数（n_i），网页长度（l_i）和访问时间长度（t_i）从某些方面反映了用户对网页的重视程度，因此，将点击次数、网页长度、访问时间长度作为参考参数。这样用户自身信息训练的兴趣向量V_1可以用数学方法表示为：

$$V_1 = \sum_{i=0}^{k} v_i \times p_i \times t_i = \sum_{i=0}^{k} v_i \times \frac{n_i}{\sum_{i=0}^{k} n} \times \frac{t_i}{l_i} \tag{9}$$

其中，k 代表此时用户行为表中当前用户对应的网页数；v_i代表每个网页的特征向量；n_i代表每个网页的点击次数，$\sum_{i=0}^{k} n_i$代表所有网页总点击次数的值；t_i代表每个网页用户访问时间长度，l_i代表每个网页长度。

5. 同一时期不同用户访问记录的挖掘。

用户查询具有局部性，尤其是背景相近的用户，在同一时期检索的内容都具有共性。因此，对同一时期不同用户访问记录进行挖掘，也是挖掘用户兴趣向量的一个重要方面。

首先，查询用户日志，提取出最近一段时间内使用过搜索引擎的所有不同用户，然后访问用户信息表，提取用户的相关信息，包括：职业、专业、爱好。将当前用户的这些信息与最近一段时间内访问网页的其他用户同类信息进行比较，找到其他用户与当前用户的相似程度β_j。由其他用户信息训练的兴趣向量可以表示为如下形式：

$$V_2 = \sum_{j=0}^{m} u_j \times \beta_j \tag{10}$$

其中，β_j为第 j 篇网页的访问用户与当前用户的相似程度；m 是找到的用户行为表中除当前用户外，所有其他用户对应的网页数量；u_j为第 j 篇网页的特征向量。

其中，$\beta_j = ra_j + (1-r)b_j$，r 和（1-r）是依据$a_j$ 、b_j ，在挖掘不同用户兴趣向量时所起作用的大小规定的影响因子。

a_j、b_j定义方法如下：

由于爱好的度量值b_j是一个 0␣ 1 之间的小数，因此为了使数据影响相同，职业、专业分成几大类，不同类之间规定相似度，选择同类的内容相似度都为 1，选择不同类的内容直接看他们的类相似度。

爱好度量值b_j，它是根据规定的类别对每一类提取出相同个数的特征确定的。用户甲的爱好向量为D_1；用户乙的爱好向量为D_2。用户甲和乙的爱好相似程度为：

$$Sim(D_1,D_2)=\frac{\sum_{i=1}^{8}d_{1i}d_{2i}}{\sqrt{\sum_{i=1}^{8}d_{1i}^{\ 2}}\sqrt{\sum_{i=1}^{8}d_{2i}^{\ 2}}} \tag{11}$$

于是，对于用户兴趣向量可以最终表示为：

$$V=aV_1+(1-a)V_2=a\sum_{i=0}^{k}v_i\times\frac{n_i}{\sum_{i=0}^{k}n_i}\times\frac{t_i}{l_i}+(1-a)\sum_{j=0}^{m}u_j\times(ra_j+(1-r)b_j) \tag{12}$$

5.3.3 模型的建立和分析求解

基于问题二已建模型的拓展性文档个性化搜索引擎加入用户兴趣的向量空间检索模型，利用用户兴趣模型优化查询，使之更贴近用户语义，从而提高搜索引擎的查准率与个性化程度。

向量空间模型

向量空间模型是文档表示的常用方法，表达文档和用户兴趣比较直接的做法是利用文档特征。用户兴趣是多方面的，可以根据其浏览过的文档选取合适的主题词来表达用户兴趣。根据上文方法可以得到用户兴趣向量 V。

令 di 表示文档向量，D={d1,d2,・・・，dn}^{T},n 是基于问题二已建好的模型聚类好的文档类数[8]；Q={q1,q2,・・・，qn}表示查询向量。

结合通过数据挖掘得到的用户兴趣向量，过滤掉相关性小的文档类，对输出结果进行重新排序。最终爱好程度计算方法如下：

$$\partial=\frac{C\bullet(\sum_{j=1}^{m}W_{V_j}\bullet W_{V_{ij}})}{\sqrt{\sum_{j=1}^{m}(W_{V_j})^2\bullet\sum_{j=1}^{m}(W_{V_{ij}})^2}} \tag{13}$$

式中 V 是用户兴趣向量，D_i是文档类，C 是类别权重，W_D表示兴趣爱好权重，W_{Di}表示文档类向量权重。该值越大文档接近用户的可能性就越大。

要进行推广的文档个性化搜索引擎，就是在问题二设计的科技论文文档个性化搜索引擎基础上上加入用户兴趣模型。最后的理想结果是通过挖掘用户访问的历史记录获取用户兴趣。面对杂乱无章的文档首先进行特征表示，然后依照问题二的文档聚类模型对文档进行聚类分析。通过文档类与用户兴趣的相关性，选择性地选取文档。输出与挖掘的用户兴趣相关性大的文档类别，过滤掉或者通过排后相关性小的文档类，来分辨杂乱无章的信息，满足不同用户不同需求。

5.3．4 模型检验及分析

该模型不仅考虑了文档与检索关键字的相关性，同时考虑了用户的爱好程度，以此来对用户兴趣进行预测。并利用了用户兴趣向量对用户检索结果进行过滤，从而是用户搜索得到的结果能满足用户的需要。作为一般性推广时，用户兴趣模型可以加入长期兴趣和短期兴趣实时监测用户兴趣变化，不断更新用户兴趣。

六、模型评价与推广

问题一建立了基于层次分析法和灰色关联分析法的搜索引擎评价体系。此模型有很大的优点和独到之处。体现在对数据的挖掘处理和两种算法的有效结合使用。数据选取过程中，定量化的指标通过查阅相关文献和搜索引擎公布的官方数据，定性化的指标通过对相关概念查询建立相应的评价标准，成员通过亲自对搜索引擎进行相关实验检验，得到数据，具有相对客观性和真实性。将层次分析法与灰色关联法结合起来，有效改善了层次分析过于主观的问题，而灰色关联法充分利用所得数据，使评价全面化，模型是比较成功的。

问题二建立了基于遗传算法改进的 K-means 聚类法的个性化引擎设计模型。此模型的可取之处在于用遗传算法对 K-means 聚类法进行了改进，有效解决了其本身的缺陷。利用 VSM 算法建立文档的空间向量，使得模型很完整，且算法很合理。只是这样会使得模型的复杂度增大，增加了工作量。另外，对模型验证时由于文档特征项提取过程过于繁琐，我们对已有文献数据进行模型验证，数据量较小，可能使模型的评价力度不足

对于模型推广，问题三给出了基于用户兴趣空间向量建立的个性化引擎推广模型，实现了科技论文文档向各类文档检索的推广，这是基于问题二文档聚类的基础上。而我们可以更一般化，改变聚类方法，实现对图像，音频、视频、游戏等聚类，结合此兴趣化模型，实现一般化个性搜索引擎的推广，具有很大的前景。

七、参考文献

【1】 吴婷,肖建华.基于 AHP 的搜索引擎评价方法研究[J].现代情报，2008(8)：起始页码-结束页码
【2】方曦，李娜，葛月华.基于 AHP 与 TOPSIS 方法的中文搜索引擎评价体系[N].科技导报，2012，30（14）；
【3】张兴华.搜索引擎技术及研究[J].现代情报，2004（4）；
【4】陆宏菊.基于遗传算法与模糊聚类的网络信息过滤信息的研究[C].山东：山东师范大学，2008；
【5】曾春，邢春晓，周立柱.基于内容过滤的个性化搜索算法[N].软件学报，2003，14（5）；
【6】王华.基于用户兴趣分析的个性化搜索引擎研究[C].北京：首都师范大学，2009；
【7】崔丽杰，刘伟.基于用户兴趣的个性化搜索引擎的设计[C].计算机与现代化，2008（7）；
【8】谭颖.文本挖掘中的聚类算法[C].吉林：吉林大学，2009；
【9】司守奎 孙玺菁.数学建模算法与应用[M].北京：国防工业大学出版社，2011；
【10】姜启源，谢金星，叶俊[M].北京：高等教育出版社，2003；

附：matlab 代码

灰色关联分析算法

```
clc,clear
A=[1 0.82 0.78 0.92 0.64 0.93 0.77 0.66 0.64 0.66
   8 12 4 8 5 10 6 10 4 4
   19 19 12 17 10 17 13 11 10 11
   7 9 3 4 6 8 4 6 4 3
   0.909 0.818 0.636 0.727 0.611 0.875 0.694 0.674 0.598
0.610
   0.889 0.889 0.778 0.667 0.596 0.832 0.741 0.633 0.599
0.623
   37 32 38 38 28 35 31 29 26 24
   0.37 0.45 0.23 0.28 0.20 0.40 0.21 0.23 0.19 0.17
   6 2 3 4 3 3 6 2 5 4
   0.12 0.19 0.18 0.24 0.20 0.18 0.17 0.21 0.21 0.22
   0.001 0.002 0.0015 0.003 0.0025 0.0015 0.0025 0.003
0.003 0.004
   6 9 8 6 7 6 8 8 8 7
   0.791 0.837 0.279 0.349 0.766 0.476 0.775 0.693 0.643
0.566
   0.75 0.667 0.583 0.667 0.583 0.741 0.556 0.511 0.486
0.498];     %n 个指标（行），m 个对象（列）
for i=[1:8 12:14]           %正向指标
    A(i,:)=(A(i,:)-min(A(i,:)))/(max(A(i,:))-min(A(i,:)));
end
for i=9:11                  %逆向指标
    A(i,:)=(max(A(i,:))-A(i,:))/(max(A(i,:))-min(A(i,:)));
end
A
[m,n]=size(A);
canko=max(A')'              %参考序列的取值
t=repmat(canko,[1,n])-A; %求参考序列与每一个序列的差
mmin=min(min(t));           %计算最小差
mmax=max(max(t));           %计算最大差
rho=0.5;                    %分辨系数
xishu=(mmin+rho*mmax)./(t+rho*mmax)    %计算灰色关联系数
%guanliandu=mean(xishu)                  %取等权重，计算关联
度
W=[0.1069 0.0352 0.0627 0.0131 0.0921 0.0307 0.2568 0.1629
0.0287 0.0543 0.0846 0.0076 0.0186 0.0460];%各指标的权重

guanliandu=W*xishu;
```

```
[gsort,ind]=sort(guanliandu,'descend')%对关联度从大到小排序
```

层次分析算法

```
clc,clear
display('请输入判断矩阵 A(n 阶)');                %在屏幕上显示
A=input('A=');                                  %从屏幕上接收到判断矩阵 A
n=length(A);                                    %A 的长度，A 是方矩阵

X=ones(n,100);
Y=ones(n,100);
M=zeros(1,100);
M(1)=max(X(:,1));                               %X 第一列的最大元素赋值给向量 M 的第一个元素
Y(:,1)=X(:,1);                                  %X 的第一列赋值给 Y 的第一列
X(:,2)=A*Y(:,1);
M(2)=max(X(:,2));
Y(:,2)=X(:,2)/M(2);
p=0.0001; i=2; k=abs(M(2)-M(1));                %初始化 p i k
while k>p
    i=i+1;
    X(:,i)=A*Y(:,i-1);
    M(i)=max(X(:,i));
    Y(:,i)=X(:,i)/M(i);
    k=abs(M(i)-M(i-1));
end
disp('特征向量');
disp(Y(:,i));
a=sum(Y(:,i));
w=Y(:,i)/a;
t=M(i);
display('权向量');display(w);
display('最大特征值');display(t);

%一致性检验
CI=(t-n)/(n-1);                                 %一致性指标
RI=[0 0 0.58 0.90 1.12 1.24 1.32 1.41 1.45 1.49 1.51 1.54 1.56 1.58 1.59];%计算的标准，随机性指标
CR=CI/RI(n);                                    %计算一致性
if CR<0.1
    disp('矩阵的一致性可以接受');
```

```
    disp('CI=');disp(CI);
    disp('CR');disp(CR);
else
    disp('矩阵的一致性不可接受！');
end
```

遗传聚类算法

```
function [BESTX,BESTY,ALLX,ALLY]=GAFCM(K,N,Pm,LB,UB,D,c,m)
M=length(LB);%决策变量的个数
%种群初始化，每一列是一个样本
farm=zeros(M,N);
for i=1:M
 x=unifrnd(LB(i),UB(i),1,N);
 farm(i,:)=x;
end
%输出变量初始化
ALLX=cell(K,1);%细胞结构，每一个元素是M×N矩阵，记录每一代的个体
ALLY=zeros(K,N);%K×N矩阵，记录每一代评价函数值
BESTX=cell(K,1);%细胞结构，每一个元素是M×1向量，记录每一代的最优个体
BESTY=zeros(K,1);%K×1矩阵，记录每一代的最优个体的评价函数值
k=1;%迭代计数器初始化
%% 第二步：迭代过程
while k<=K
%% 以下是交叉过程
 newfarm=zeros(M,2*N);
 Ser=randperm(N);%两两随机配对的配对表
 A=farm(:,Ser(1));
 B=farm(:,Ser(2));
 P0=unidrnd(M-1);
 a=[A(1:P0,:);B((P0+1):end,:)];%产生子代a
 b=[B(1:P0,:);A((P0+1):end,:)];%产生子代b
 newfarm(:,2*N-1)=a;%加入子代种群
 newfarm(:,2*N)=b;
 for i=1:(N-1)
  A=farm(:,Ser(i));
  B=farm(:,Ser(i+1));
  P0=unidrnd(M-1);
  a=[A(1:P0,:);B((P0+1):end,:)];
  b=[B(1:P0,:);A((P0+1):end,:)];
  newfarm(:,2*i-1)=a;
     newfarm(:,2*i)=b;
```

```
 end
 FARM=[farm,newfarm];
%% 选择复制
 SER=randperm(3*N);
 FITNESS=zeros(1,3*N);
 fitness=zeros(1,N);
 for i=1:(3*N)
  Beta=FARM(:,i);
  FITNESS(i)=FIT(Beta,D,c,m);
 end
 for i=1:N
  f1=FITNESS(SER(3*i-2));
  f2=FITNESS(SER(3*i-1));
  f3=FITNESS(SER(3*i));
  if f1<=f2&&f1<=f3
   farm(:,i)=FARM(:,SER(3*i-2));
   fitness(:,i)=FITNESS(:,SER(3*i-2));
  elseif f2<=f1&&f2<=f3
   farm(:,i)=FARM(:,SER(3*i-1));
   fitness(:,i)=FITNESS(:,SER(3*i-1));
  else
   farm(:,i)=FARM(:,SER(3*i));
   fitness(:,i)=FITNESS(:,SER(3*i));
  end
 end
 %% 记录最佳个体和收敛曲线
 X=farm;
 Y=fitness;
 ALLX{k}=X;
 ALLY(k,:)=Y;
 minY=min(Y);
 pos=find(Y==minY);
 BESTX{k}=X(:,pos(1));
 BESTY(k)=minY;
%% 变异
 for i=1:N
  if Pm>rand&&pos(1)~=i
   AA=farm(:,i);
   BB=GaussMutation(AA,LB,UB);
   farm(:,i)=BB;
  end
 end
 disp(k);
    k=k+1;
```

```
            end
            %% 绘图
            BESTY2=BESTY;
            BESTX2=BESTX;
            for k=1:K
                TempY=BESTY(1:k);
                minTempY=min(TempY);
             posY=find(TempY==minTempY);
             BESTY2(k)=minTempY;
             BESTX2{k}=BESTX{posY(1)};
            end
            BESTY=BESTY2;
            BESTX=BESTX2;
            plot(BESTY,'-
    ko','MarkerEdgeColor','k','MarkerFaceColor','k','MarkerSize',2)
            ylabel('函数值')
            xlabel('迭代次数')
grid on
```