基于未确知集合理论的搜索引擎性能评价模型

周文洁

（株洲职业技术学院，湖南 株洲 412001）

摘 要：应用未确知集合理论建立了一种对搜索引擎性能进行综合评价的数学模型，综合应用AHP法和熵值法给出了搜索引擎性能评价指标的分类权重，并通过实例分析表明了本方法的合理性和有效性。

关键词：搜索引擎；信息熵；指标分类权重

中图分类号：O223；G350 **文献标识码**：A **文章编号**：1673-9833(2008)04-0022-04

Performance Evaluation Model for Web Search Engine Based on Unascertained Set Theory

Zhou Wenjie

（Zhuzhou Professional Technology College，Zhuzhou Hunan 412001，China）

Abstract: A mathematical model of comprehensive evaluation for the performance of web search engine by means of the unascertained set theory is set up. It determines the performance evaluation index classification weight of web search engine after using the AHP method and entropy value method. Finally, an illustrative example is also given to show reasonability and effectiveness of the proposed method.

Key words: web search engine; information entropy; index classification weight

0 引言

1994年，雅虎的创建拉开了搜索引擎发展的序幕，经过10多年的发展，目前搜索引擎的数量庞大，并且在查询范围、检索功能、检索结果等方面走向多样化，这使得用户在选择搜索引擎时遇到了难题。因此，建立搜索引擎性能综合评价的数学模型，系统客观地对搜索引擎性能进行综合评价，具有较大的现实意义和应用前景。

文献[1]运用模糊综合评价方法、文献[2]运用优劣系数法、文献[3]运用灰色多层次方法对搜索引擎性能进行综合评价，均各有特点。但是这些方法的主观性较强，而且模糊综合评价方法不符合“非负有界性、可加性、归一性”的测度特性。本文在已有的研究基础上提出一种基于未确知集的搜索引擎性能综合评价模型，并综合应用AHP法和熵值法确定指标分类权重，比较符合客观实际。

1 未确知集合理论

1.1 未确知测度与未确知集合[4-6]

设U为论域，F是U上的性质空间，$\{F_1, F_2, F_3, \cdots, F_K\}$是$F$的一种划分，$E$为$F$上的$\sigma$代数。对任意$u \in U$，$A \in E$，我们想知道$u$具有性质$A$（或$u$处于状态$A$）的程度，即对$u$具有性质$A$的程度进行测量。作为测量结果的某种测度，显然要满足非负界为1、可加性、归一性等测量准则。

定义1[4,6] 设(F, E)是U上可测空间，若$\forall u \in U$，$A \in E$，存在映射μ，使$\mu_A(u)$满足：

$$0 \le \mu_A(u) \le 1, \quad (1)$$

$$\mu_F(u)=1, \quad (2)$$

$$\mu_{\bigcup_i A_i}(u) = \sum_i \mu_{A_i}(u), A_i \cap A_j = \phi (i \ne j), \quad (3)$$

则称$\mu_A(u)$为可测空间(F,E)上的未确知测度，称$(U,E,\mu_A(x))$为未确知测度空间。

收稿日期：2008-05-14

作者简介：周文洁（1967-），女，湖南湘乡人，株洲职业技术学院副教授，主要从事高等数学教育教学与研究.

定义 2[4,6]　设(F,E)是U上的可测空间，$\mu_A(u)$是u关于$A \in E$的未确知测度，则以$\mu_A(u)$为隶属函数确定了论域U上关于σ代数E的一个不确定性集合$\tilde{A}$，称$\tilde{A}$为U上的未确知子集。

1.2　未确知集合测度模型

设x_1，x_2，…，x_n表示n个待评价对象，记$X=\{x_1, x_2, \cdots, x_n\}$为论域；评价$x_i$（$x_i \in E$）有$m$项属性（指标）$I_1$，$I_2$，…，$I_m$，记$I=\{I_1, I_2, \cdots, I_m\}$，称之为属性空间；$x_{ij}$表示第$i$个评价对象$x_i$关于第$j$项指标$I_j$的观测值。对$x_{ij}$有$p$个评价等级$c_1$，$c_2$，…，$c_p$，记$C=\{c_1, c_2, \cdots, c_p\}$，称之为评价空间，若评价空间具有如下性质：

$C=\{c_1, c_2, \cdots, c_p\}$，$c_i \cap c_j = \phi (i \neq j)$，

$c_1 > c_2 > \cdots > c_p$（或$c_1 < c_2 < \cdots < c_p$），

即c_k比c_{k+1}强（或弱），则称c_1，c_2，…，c_p是评价空间C的一个有序分割类。

1.2.1　单指标未确知测度

即求出观测值x_{ij}属于各类c_k的等级测度μ_{ijk}。这里需要构造测度函数$\mu_{ij}(x)$，对每个评价等级c_k（$k=1, 2, 3, \cdots, p$）求出μ_{ijk}的值，从而得到对象x_i在单指标下的未确知测度评价矩阵：

$$\boldsymbol{B}_i=(\mu_{ijk})_{m\times p}=\begin{bmatrix} \mu_{i11} & \mu_{i12} & \cdots & \mu_{i1p} \\ \mu_{i21} & \mu_{i22} & \cdots & \mu_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{im1} & \mu_{im2} & \cdots & \mu_{imp} \end{bmatrix}, \quad i=1, 2, \cdots, n。 \tag{4}$$

1.2.2　指标分类权重的确定

评价指标权重确定的方法有多种，主要分为两类：一类是主观赋权法，如 Delph 法、AHP 法等；一类是客观赋权法，如主成分分析法、熵值法等。两类赋权法各有优缺点，这里将两种方法结合起来确定评价指标分类权重。

利用 AHP 法确定指标权重，设$\boldsymbol{W}'=(w_1', w_2', \cdots, w_m')$，然后利用熵值法确定评价指标的分类权重。由于 AHP 法没有利用决策矩阵的信息，而熵值法的最大特点是利用决策矩阵计算权重，故我们最后将由熵值法计算的指标分类权重与由 AHP 法计算的指标权重$\boldsymbol{W}'$进行综合，得到最终的指标分类权重，使确定的权重更为合理。

设对象x_i关于指标I_j的观测值x_{ij}使x_i处于c_1，c_2，…，c_p各个评价等级的未确知测度向量为：

$$\boldsymbol{\mu}_{ij}=(\mu_{ij1}, \mu_{ij2}, \cdots, \mu_{ijp}), \tag{5}$$

由此可知指标I_j对于对象x_i的分类做了多少贡献。

1）如果$\mu_{ij1}=\mu_{ij2}=\cdots=\mu_{ijp}=1/p$，说明指标$I_j$使$x_i$处于各个评价等级的程度相同，因而无法区分出$x_i$到底处于哪个评价等级，此时称指标$I_j$未对$x_i$的分类做出贡献。若用$\boldsymbol{\varepsilon}_{ij}$表示指标$I_j$关于$x_i$的分类权重，则$\boldsymbol{\varepsilon}_{ij}=0$；

2）如果p个μ_{ijk}中有一个$\mu_{ijk_0}=1$，其它的$p-1$个均为 0，则指标I_j使x_i处于c_{k_0}等级，此时称指标I_j对x_i的分类做了最大贡献。若用$\boldsymbol{\varepsilon}_{ij}$表示指标$I_j$关于$x_i$的分类权重，则$\boldsymbol{\varepsilon}_{ij}$这时取得最大值；

3）同理可以说明，$\boldsymbol{\mu}_{ij}$的p个分量取值越分散，$\boldsymbol{\varepsilon}_{ij}$越小；而取值越集中，$\boldsymbol{\varepsilon}_{ij}$越大。

设由测度μ_{ijk}所确定的信息熵为：

$$H(j)=-\sum_{k=1}^{p}\mu_{ijk}\cdot\log\mu_{ijk}, \tag{6}$$

令 $$\boldsymbol{v}_{ij}=1-\frac{1}{\log p}H(j)=1-\frac{1}{\log p}\sum_{k=1}^{p}\mu_{ijk}\cdot\log\mu_{ijk}, \tag{7}$$

令 $$\boldsymbol{\varepsilon}_{ij}=\frac{\boldsymbol{v}_{ij}}{\sum_{i=1}^{m}\boldsymbol{v}_{ij}}, \tag{8}$$

显然$0\le\boldsymbol{\varepsilon}_{ij}\le 1$，且$\sum_{j=1}^{m}\boldsymbol{\varepsilon}_{ij}=1$。

由信息熵的性质知：

1）当且仅当$\mu_{ij1}=\mu_{ij2}=\cdots=\mu_{ijp}=1/p$时，$\boldsymbol{v}_{ij}$取到最小值为 0；

2）当且仅当存在一个$\mu_{ijk_0}=1$，其它的$p-1$个均为 0 时，$\boldsymbol{v}_{ij}$取到最大值为 1；

3）μ_{ijk}的取值越分散时，$\boldsymbol{v}_{ij}$的取值越接近于 0；反之，μ_{ijk}的取值越集中，$\boldsymbol{v}_{ij}$的值越接近于 1。

由$\boldsymbol{v}_{ij}$的上述 3 条性质可知，式（8）定义的$\boldsymbol{\varepsilon}_{ij}$是指标$I_j$关于$x_i$的分类权重，称

$$\boldsymbol{\varepsilon}_i=(\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{im}) \tag{9}$$

为指标I_1，I_2，…，I_m关于x_i的分类权重向量。

将$\boldsymbol{\varepsilon}_i$与$\boldsymbol{W}'$进行综合，可计算出指标$I_1$，$I_2$，…，$I_m$关于$x_i$的最终的分类权重向量为：

$$\boldsymbol{W}_i=(w_{i1}, w_{i2}, \cdots, w_{im}), \tag{10}$$

其中 $$w_{ij}=\frac{w_j'\varepsilon_{ij}}{\sum_{j=1}^{m}w_j'\varepsilon_{ij}}。 \tag{11}$$

1.2.3　多指标综合未确知测度

由评价对象x_i的单指标测度评价矩阵$\boldsymbol{B}_i$和指标权重向量$\boldsymbol{W}_i$，得到x_i在m个指标I_1，I_2，…，I_m下的综合未确知测度向量为：

$$\boldsymbol{\mu}_i=\boldsymbol{W}_i\cdot\boldsymbol{B}_i=(\mu_{i1}, \mu_{i2}, \cdots, \mu_{ip}), \tag{12}$$

其中$\mu_{ik}=\sum_{j=1}^{m}w_{ij}\cdot\mu_{ijk}$（$i=1, 2, \cdots, n$；$k=1, 2, \cdots, p$）表示对象$x_i$属于评价等级$c_k$的综合未确知测度。

1.2.4　评价准则

由于评价等级是有序的，所以隶属度识别准则不适用，可以采用置信度识别准则。

置信度识别准则：设置信度为λ（$\lambda > 0.5$，根据问题背景和需要，通常取在 0.6 ~ 0.8 之间），如果

$c_1 > c_2 > \cdots > c_p$，令

$$k_0 = \min_k \left(k \left| \sum_{i=1}^{k} \mu_{ij} \geq \lambda, 1 \leq k \leq p \right. \right) , \quad (13)$$

则判定 x_i 属于第 k_0 个评价等级 c_{k_0}，且置信度为 λ。

令 $A=(p, p-1, \cdots, 2, 1)$，则 $q_i = \mu_i \cdot A$ 是评价对象 x_i 的分数，可按该分数对 x_1，x_2，$\cdots$，x_n 进行排序。

2 基于未确知集的搜索引擎性能评价模型

2.1 建立搜索引擎性能评价指标体系

在综合国内外研究的基础上，我们认为要科学、合理、有效地评价搜索引擎性能，可以建立以下搜索引擎性能评价指标体系[1-3, 7-9]，详见表1。

表1 搜索引擎性能评价指标体系

Tab. 1 The evoluation index system of search engine function

一级指标	二级指标		细类
索引库	标引文件种类	I_1	FTP 文件；WWW 文件；Newsgroup 文件；Usenet 文件等
	标引深度及更新频率	I_2	全面 / 部分索引；考虑超文本的不同标记所表示的不同含义；收集页面中的超链接；更新频率快 / 慢
检索功能	基本检索	I_3	布尔检索；截词检索；邻近词检索；字段检索；区分大小写（英语）
	高级检索	I_4	加权检索；模糊检索；相关信息反馈检索；概念检索；自然语言检索；目录式浏览检索；多内码处理（中文）；多语种检索；多媒体检索
检索效果	查全率	I_5	查全率是指所检出的相关文献数量与文献库中所有相关文献总量之比
	查准率	I_6	查准率是指所检出的相关文献数量与所有检出的文献总量之比
	检索时间	I_7	从检索开始到显示检索结果的时间
亲和度	检索中与用户的交互情况	I_8	“个性化”查询界面；检索帮助信息；相关性排列；格式转换；交叉语言检索与翻译
	智能信息技术	I_9	信息过滤；信息挖掘；信息推送；学习功能

2.2 构建决策矩阵

设参加评价的搜索引擎为 x_i（i=1，2，$\cdots$，5），采取专家打分和统计实验相结合的方法，经过计算各指标的值得到决策矩阵[2]：

$$\boldsymbol{D}=(x_{ij})_{n\times m}=$$

$$\begin{array}{c} \quad I_1 \quad I_2 \quad I_3 \quad I_4 \quad\ I_5 \quad\ I_6 \quad\ I_7 \quad I_8 \quad I_9 \\ \begin{bmatrix} 4 & 3 & 2 & 6 & 0.38 & 0.61 & 0.12 & 3 & 3 \\ 2 & 2 & 5 & 7 & 0.89 & 0.78 & 0.34 & 4 & 2 \\ 3 & 4 & 4 & 6 & 0.74 & 0.83 & 0.28 & 3 & 1 \\ 4 & 1 & 3 & 8 & 0.63 & 0.44 & 0.09 & 2 & 2 \\ 3 & 3 & 5 & 5 & 0.91 & 0.90 & 0.50 & 3 & 4 \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} \end{array} 。$$

2.3 构造单指标测度函数，求出单指标测度评价矩阵

将搜索引擎性能的评价等级分为4级，即 c_1=优，c_2=良，c_3=好，c_4=差，显然 $c_1 > c_2 > c_3 > c_4$。由于各个指标没有统一的评价等级标准，我们通过下面的公式计算各等级的分界值（不妨设指标 I_j 是效益型的）：

指标 I_j 的等级分界值 $a_{jl}=I_j$ 的最小值 +

$$\frac{I_j\text{的最大值}-I_j\text{的最小值}}{\text{等级总数}-1}\times(l-1),(l=1,2,3,4),$$

从而得到指标分类标准矩阵：

$$\begin{array}{c} \quad I_1 \quad\ I_2 \quad\ I_3 \quad\ I_4 \\ \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \\ I_9 \end{matrix} \begin{bmatrix} a_1 & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ & & & \\ & & & \\ \vdots & \vdots & \vdots & \vdots \\ & & & \\ & & & \\ & & & \\ a_{91} & a_{92} & a_{93} & a_{94} \end{bmatrix} = \begin{bmatrix} 4 & 3.333\,3 & 2.666\,7 & 2 \\ 4 & 3 & 2 & 1 \\ 5 & 4 & 3 & 2 \\ 8 & 7 & 6 & 5 \\ 0.91 & 0.733\,3 & 0.556\,7 & 0.38 \\ 0.90 & 0.746\,6 & 0.593\,3 & 0.44 \\ 0.09 & 0.226\,7 & 0.363\,4 & 0.50 \\ 4 & 3.333\,3 & 2.666\,7 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \end{array}$$

对决策矩阵 $\boldsymbol{D}$ 中的数据构造单指标测度函数，不妨设 $a_{j1} < a_{j2} < a_{j3} < a_{j4}$，则：

当 $x_{ij} \leq a_{j1}$ 时，取 $\mu_{ij1}=1$，$\mu_{ij2}=\mu_{ij3}=\mu_{ij4}=0$；

当 $x_{ij} \geq a_{j4}$ 时，取 $\mu_{ij1}=\mu_{ij2}=\mu_{ij3}=0$，$\mu_{ij4}=1$；

当 $a_{jk} \leq x_{ij} \leq a_{j(k+1)}$ 时，取

$$\mu_{ijk}=\frac{\left|x_{ij}-a_{j(k+1)}\right|}{\left|a_{jk}-a_{j(k+1)}\right|}, \quad \mu_{ij(k+1)}=\frac{\left|x_{ij}-a_{jk}\right|}{\left|a_{j(k+1)}-a_{jk}\right|},$$

$\mu_{ijs}=0(s < k$ 或 $s > k+1)$。

根据单指标测度函数和指标分类标准矩阵，可以求得 x_1 的单指标测度评价矩阵为：

$\boldsymbol{B}_1=(\mu_{1jk})_{9\times 4}=$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.108\,9 & 0.891\,1 & 0 \\ 0.780\,5 & 0.219\,5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad (14)$$

x_i（i=2，3，4，5）的单指标测度评价矩阵略。

2.4　计算指标权重

由 AHP 法确定各指标的权重[1] 为 W'=(0.09，0.18，0.11，0.04, 0.15，0.26，0.08，0.06，0.03)，由熵值法，根据式（6~9）得到指标 I_1，I_2，…，I_m 关于 x_i（i=1，2，…，5）的分类权重分别为：

$$\begin{bmatrix} 0.127\,0 & 0.127\,0 & 0.127\,0 & 0.127\,0 & 0.127\,0 & 0.095\,5 & 0.078\,9 & 0.063\,5 & 0.127\,0 \\ 0.124\,4 & 0.124\,4 & 0.124\,4 & 0.124\,4 & 0.092\,7 & 0.077\,4 & 0.083\,3 & 0.124\,4 & 0.124\,4 \\ 0.074\,9 & 0.149\,8 & 0.149\,8 & 0.149\,8 & 0.132\,4 & 0.075\,3 & 0.043\,5 & 0.074\,9 & 0.149\,8 \\ 0.117\,5 & 0.117\,5 & 0.117\,5 & 0.117\,5 & 0.060\,0 & 0.117\,5 & 0.117\,5 & 0.117\,5 & 0.117\,5 \\ 0.062\,5 & 0.125\,0 & 0.125\,0 & 0.125\,0 & 0.125\,0 & 0.125\,0 & 0.125\,0 & 0.062\,5 & 0.125\,0 \end{bmatrix},$$

将两者综合，由式（10~11）得到指标 I_1，I_2，…，I_m 关于 x_i（i=1，2，…，5）最终分类权重分别为：

$$\begin{bmatrix} 0.102\,8 & 0.205\,7 & 0.125\,7 & 0.045\,7 & 0.171\,4 & 0.223\,4 & 0.056\,8 & 0.034\,3 & 0.034\,2 \\ 0.114\,0 & 0.213\,5 & 0.130\,4 & 0.047\,4 & 0.132\,6 & 0.191\,8 & 0.063\,5 & 0.071\,2 & 0.035\,6 \\ 0.062\,4 & 0.249\,5 & 0.152\,5 & 0.055\,4 & 0.183\,7 & 0.181\,1 & 0.032\,2 & 0.041\,6 & 0.041\,6 \\ 0.097\,1 & 0.194\,3 & 0.118\,7 & 0.043\,2 & 0.082\,7 & 0.280\,5 & 0.086\,3 & 0.064\,8 & 0.032\,4 \\ 0.048\,6 & 0.194\,6 & 0.118\,9 & 0.043\,3 & 0.162\,2 & 0.281\,1 & 0.086\,5 & 0.032\,4 & 0.032\,4 \end{bmatrix}。 \quad (15)$$

2.5　求出多指标综合测度评价矩阵

由单指标测度评价矩阵式（14）和指标分类权重式（15），根据式（12）可求得多指标综合测度评价矩阵为：

$$\begin{bmatrix} \mu \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} = \begin{bmatrix} 0.147\,1 & 0.293\,8 & 0.261\,9 & 0.297\,1 \\ 0.188\,8 & 0.115\,1 & 0.451\,7 & 0.244\,4 \\ 0.355\,0 & 0.483\,5 & 0.120\,0 & 0.041\,6 \\ 0.226\,6 & 0.153\,0 & 0.080\,8 & 0.539\,6 \\ 0.594\,6 & 0.235\,1 & 0.040\,5 & 0.129\,8 \end{bmatrix},$$

取置信度 λ =0.6，可以得到 5 个搜索引擎性能的评价等级为：x_5、x_3—良，x_1、x_2—中，x_4—差。

各搜索引擎所得分数为（q_1，q_2，q_3，q_4，q_5）=（2.291 0，2.248 3，3.151 8，2.066 6，3.294 5），故它们的排序为：$x_5 > x_3 > x_1 > x_2 > x_4$。与文献[1]对比，$x_3$ 与 x_5 交换了位置。而从直观来看，9 项指标中 x_5 有 4 项指标超过 x_3，且这 4 项指标值都是 5 个搜索引擎中最好的，另外 x_5 有 2 项指标与 x_3 相同，故本文结果更合理一些。

3　结语

本文将未确知集合理论应用于搜索引擎性能的综合评价，将 AHP 法和熵值法综合应用于求指标分类权重，把决策者的先验知识与决策矩阵的信息结合起来，使得到的指标分类权重更合理。同时未确知集合理论引入指标分类权重概念，使得系统的推理算法合理，得到的合成可信度具有可解释性。本文所建立的搜索引擎性能综合评价的未确知集合测度模型，计算简便，结果合理，既可划分等级，又可进行排序，是一种有效实用的搜索引擎性能评价方法。

参考文献：

[1] 刘正春，蒋福坤. 搜索引擎性能的模糊综合评判[J]. 数学的实践与认识，2004，34(7)：24-28.

[2] 刘正春. 搜索引擎综合评价模型研究[J]. 数学的实践与认识，2004，34(9)：7-14.

[3] 汪新凡. 搜索引擎性能的多层次灰色评价模型[J]. 情报科学，2006，24(12)：1845-1848.

[4] 刘开第，李万庆，庞彦军. 未确知集[J]. 数学的实践与认识，2006，36(10)：197-204.

[5] 王　瑜. 要地防空中目标未确知威胁的测度变权评价[J]. 系统工程理论与实践，2003(2)：111-115.

[6] 刘开第，曹庆奎，庞彦军. 基于未确知集合的故障诊断方法[J]. 自动化学报，2004，30(5)：747-756.

[7] 陶跃华. 因特网搜索引擎评价体系[J]. 计算机工程与科学，2001，23(3)：25-27.

[8] 宛　玲，杨秀丹，杜晓静. 试析中文搜索引擎的评价标准[J]. 情报科学，2000，18(1)：28-30.

[9] 凤元杰，王坚毅，刘正春. 搜索引擎主要性能评价指标体系研究[J]. 情报学报，2004，23(1)：63-68.

（责任编辑：罗立宇）