基于信息量与信息熵的元搜索引擎排序算法研究

赖相旭[1] 韩立新[1,2] 曾晓勤[1] 王 敏[1] 吴胜利[3]
(河海大学计算机与信息学院 南京 210024)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]
(阿尔斯特大学计算机及信息科学学院 英国贝尔法斯特)[3]

摘 要 元搜索引擎集合了多个成员搜索引擎的结果,将结果进行一定的处理后再将处理后的结果返回给用户。其中对结果的重新排序直接影响到元搜索引擎的性能。基于通信领域上的信息量与信息熵提出一种计算结果相关度的算法——信息关联度IRD算法,再将算法进行特定的修正,并提出一种合并算法CombMul,将以上算法应用到元搜索引擎中,最终用MRR查准率来评价此方法。得到的MRR查准率数据表明,与广泛应用的Borda排序算法相比,IRD算法结果更为理想。

关键词 元搜索引擎,排序算法,信息关联度,IRD,信息量,信息熵,CombMul
中图法分类号 TP311.5 **文献标识码** A

Research of Ranking Algorithm Based on Information Quantity and Entropy in Meta Search Engine

LAI Xiang-xu[1] HAN Li-xin[1,2] ZENG Xiao-qin[1] WANG Min[1] WU Sheng-li[3]
(College of Computer and Information, Hohai University, Nanjing 210024, China)[1]
(State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]
(School of Computing and Mathematics University of Ulster, Northern Ireland, Belfast, UK)[3]

Abstract The meta search engine collects results from many search engines, using a certain way to treat the results and then returning back to the users. Reranking the results will directly affect the performance of meta search engine. This paper was based on information quantity and entropy which are used in communication field then presented a calculation algorithm——information related degree(IRD), after a particular amendment to the IRD algorithm, this paper also proposed a merging algorithm combMul. The above algorithms were applied to the meta search engine, and MRR precision was used to evaluate the algorithm . The MRR precision data show that IRD algorithm is even better compared with a widely used sorting algorithm——Borda.

Keywords Meta search mngine, Ranking algorithm, Information related degree, IRD, Information quantity, Entropy, CombMul

1 引言

元搜索引擎[1](Meta Search Engine, MSE)是一种建立在搜索引擎基础上、调用多个成员搜索引擎的搜索引擎,它获取所调用的成员搜索引擎的内容重新排序后,将结果返回给用户,其工作原理如图1所示。

从元搜索引擎的工作原理可以看出,各个元搜索引擎的差异主要由两个关键因素影响:成员搜索引擎的选取以及对结果的合成重排序。而其中最核心的技术在于对结果的合成重排序,结果排序的好坏将直接影响整个元搜索引擎的性能。因此,国内外研究元搜索引擎的重点也正是在于研究各个成员搜索引擎结果的合成重排序算法。

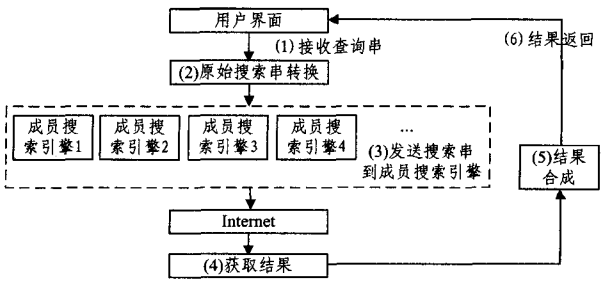

图1 元搜索工作原理图

以往对元搜索引擎排序算法的研究,从其发展历程来看,

到稿日期:2011-04-23 返修日期:2011-07-14 本文受国家自然科学基金项目(60673186和60971088),江苏省高校"青蓝工程"中青年学术带头人培养对象项目,教育部新世纪优秀人才支持计划(NCET-10-0327)资助。

赖相旭(1985—),男,硕士生,主要研究方向为模式识别、信息检索,E-mail:xiongyoo@163.com;**韩立新**(1967—),男,博士后,研究员,博士生导师,主要研究方向为信息检索、模式识别、数据挖掘;**曾晓勤**(1957—),男,博士,教授,博士生导师,主要研究方向为人工神经网络、模式识别;**王 敏**(1978—),女,博士,副教授,硕士生导师,主要研究方向为图像处理、智能计算和模式识别;**吴胜利** 男,博士后,讲师,主要研究方向为信息检索。

主要分 3 种：第一种是根据成员搜索引擎响应的先后顺序直接把结果返回给用户，即时间排序[2]。第二种是根据获取的结果在各成员搜索引擎中按位置排序，如 Borda 排序[3]、Round-robin 排序[8]。第三种是根据结果的文本信息给定权值，合并后将结果返回给用户，如动态分配相关分值法[4]。这 3 种排序算法有一定的科学性，但是也存在一定的弊端：第一种的响应速度是最快的，但对于结果的重叠率没有改善，排序结果不尽人意；第三种排序效果是最好的，但是复杂度较高；而第二种介于第一种和第三种之间。元搜索引擎主要在于提高最终排序结果的效果，在时间的响应上可以放低要求，因此本文在第三种排序算法的基础上，提出一种新的算法——信息关联度(IRD)。本算法是基于通信领域的信息量[5]和信息熵[5]而提出的，再加以第二种排序算法的辅助，对其修正后得到适用元搜索引擎的排序算法。

本文第 2 节介绍 IRD 算法的基本思路；第 3 节介绍 IRD 应用到元搜索引擎中所做的修正工作并提出合并算法；第 4 节介绍一种评价方法并利用实验数据来评价 IRD 算法；最后总结并提出下一步工作。

2 信息关联度 IRD 算法

2.1 信息量与信息熵

通信领域中用信息量表示传输信息的多少。信息是指消息中所含的有效内容，或者说是收信者预先不知道的内容。因此，信息量也可简单地理解为判断消息不确定性的大小。

信息检索中的相关度是衡量文本与查询串之间确定性大小的量。确定性与不确定是一组对应的关系，因此可以用信息量来衡量查询串与文本的相关度，其关系可概括为信息量与相关度成反比关系。信息量的计算公式为

$$I(x)=\log_a \frac{1}{P(x)}=-\log_a P(x) \tag{1}$$

式中，$P(x)$为 x 出现的概率。在信息检索中，完全可以用数值来确定字符出现的概率。只需要用待求字符串(用 *keyword* 表示)在文本(用 *text* 表示)中的比重来表示待求字符串的出现概率即可，即用搜索字符串长度与其在文本中出现的次数(用 *times*(*keyword*)表示)的乘积及其与文本总长度的比值来求得。

$$P(keyword)=\frac{len(keyword)\times times(keyword)}{len(text)} \tag{2}$$

然而，元搜索中对结果的排序会因为文本长度的不同而导致文本间的比较失去统一准线。

信息熵是每个符号所含的统计平均值，即平均信息量，计算公式为

$$\begin{aligned}H(x)&=P(x_1)[-\log_2 P(x_1)]+P(x_2)[-\log_2 P(x_2)]+\\&\quad\cdots+P(x_N)[-\log_2 P(x_N)]\\&=\sum_{i=1}^{N}P(x_i)I(x_i)=-\sum_{i=1}^{N}P(x_i)[\log_2 P(x_i)](b/\text{符号})\end{aligned} \tag{3}$$

信息熵计算的是每个符号的平均信息量，由此可以联想到将不同长度的文本的比较准线具体到每一个字符中，但是信息熵计算的是消息中所有字符的信息量。而在检索信息的相关度计算中只需要计算查询串或查询串相关的字符，因此将式(2)、式(3)改写成只计算相关字符的信息熵。假设文本中有与查询串或查询串相关的字符 M 个($M\leqslant N$)，则有

$$H(x)=\sum_{i=1}^{M}P(x_i)I(x_i)=\sum_{i=1}^{M}P(x_i)[-\log_2 P(x_i)] \tag{4}$$

由于式(4)并不是完全的信息熵公式，只涵盖了部分的字符，因此暂且称它为“部分信息熵”计算公式。

在分析部分信息熵与相关度关系之前，先分析概率 P 与部分信息熵中的独项 $P*\log_2(1/p)$之间的关系，见图 2。

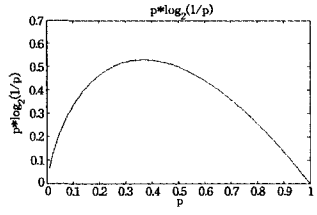

图 2 概率 P 与 $P*\log_2(1/p)$之间的关系曲线

如图 2 所示，函数 $P*\log_2(1/p)$并不是一个在 $P\in[0,1]$区间上单调的函数，因此部分信息熵 $H=\sum P*\log_2(1/p)$也不能判定为在 $P\in[0,1]$区间上是单调的函数。也就是说，信息熵在取值范围内具有不确定的单调性，故此信息熵不能延续信息量的物理意义来判断字符的相关性。用简单的话来说，一个字符的信息量越大，其相关性就越小。但是并不能说明部分信息熵越大其相关性越小。

2.2 信息关联度 IRD 算法

从上一节可以得出结论：提出的排序算法不仅要符合“信息熵”的“统计平均”的意义，而且要与“信息量”的实际应用意义相适应。由此提出信息关联度(Information Related Degree)算法，简称 IRD 算法，其计算公式如下：

$$H=\frac{I(keyword)}{len(text)} \tag{5}$$

式中，*keyword* 为最能代表搜索串的字符串，$I(keyword)$即为该字符串的信息量，*len*(*text*)为文本的长度。

2.3 代表性字符串的获取

我们知道，各个成员搜索引擎都已经有自己的分词技术，也就是说在返回给用户的信息中，都已经有分好词的字符串。因此，若再进行分词，不仅使得系统需要花费更大的内存去处理，而且需要加大时间和空间的复杂度，更重要的是这样的分词并不能反映出每个搜索引擎自身的特点。根据算法的特点，只要选取出现总字符(字符长度和出现次数的乘积)最多的那个词组。因此，分割基于最长字符匹配原则，但是需要注意的是，分割之后的关键词组不能全部权衡其信息量，不然同样会出现与上例所描述相悖的情况。要选取分割后的几个字符(串)组中有代表的一个，则该字符(串)应该具有更大的概率。又由于已经按照最长字符匹配原则，若分割后的词组出现相同的长度，根据式(18)可知道只需要选取其中字符(串)中出现次数最多、最频繁的那个字符(串)即可。

3 IRD 算法在元搜索引擎中的应用

元搜索引擎从成员搜索引擎中获取到的结果包含的信息有结果的标题、结果的摘要、结果的网址(不能被用户直接看见，以链接形式提供给用户，往往都是在标题中链接)、结果来源于哪个搜索引擎和结果在各自搜索引擎中原先的排序位置等信息。在应用 IRD 算法时，尽量结合这些信息来修正算法，这样能使得最终的结果更加准确。

3.1 无穷值的处理

观察式(5)，无穷值点有两种情况：(1) *len*(*text*)=0，出现

此类情况时，直接把 IRD 赋予 1。(2) $I(keyword)=\infty$ 时，根据式(2)，也即 $P(keyword)=\frac{len(keyword)\times times(keyword)}{len(text)}=0$ 知，排序的算法主要是根据数值的大小来排序，而结果集中有两个或两个以上的值是无穷大的话，那么它们之间就没有可比性。因此，需要对 IRD 式(5)作出修正：

$$H=\begin{cases}1, & len(text)=0\\ \frac{1}{len(text)}\times & \\ \left[-\log_2\frac{len(keyword)\times times(keyword)+1}{len(text)}\right], & \text{其他}\end{cases} \tag{6}$$

然而，当 $len(keyword)\times times(keyword)=len(text)$ 时，应用式(6)计算的 IRD 值会出现负数。为消除负数的影响，对此直接用 0 代替。

综合上述情况，对 IRD 式(6)作出修正如下：

$$H=\begin{cases}1, & len(text)=0\\ 0, & len(keyword)\times times(keyword)=len(text)\\ \frac{1}{len(text)}\times, & \\ \left[-\log_2\frac{len(keyword)\times times(keyword)+1}{len(text)},\right] & \text{其他}\end{cases} \tag{7}$$

由信息量的性质可得，$I\in[0,+\infty)$，则 $H\geqslant 0$。

3.2 文本来源的重要性

搜索引擎返回给用户的结果都会显示标题和摘要，而标题和摘要在网页中的位置有所不同。标题一般是所链接的网页的关键词，一般在 HTML 文本中的〈HEAD〉〈/HEAD〉标签内，表现形式通常有两种：第一，以〈meta name="Keywords"的属性来表现；第二，以〈title〉〈/title〉里面的内容来体现。这些都是属于概括该网页主题的语句。然而对于摘要，则是存在网页中的内容选段，表现形式也有两种：第一，在 HTML 文本中的〈HEAD〉〈/HEAD〉标签内以〈meta name="Description"的属性来表现；第二，存在于 HTML 文本中的〈BODY〉〈/BODY〉标签中。这些同属于对网页的摘要。用另一句话来说，就是标题的内容是整个网页内容的概括，而摘要是网页的选段，并不一定是网页的中心内容。因此，标题的内容显得比摘要的内容更加重要。需要对结果的标题和摘要赋以权值，才能区分其重要性。

由此提出以下公式：

$$H=\alpha H_{标题}+\beta H_{摘要} \tag{8}$$

其中令

$$\alpha+\beta=1 \tag{9}$$

在求 α 与 β 的值之前，首先设定标题和摘要的 IRD 临界值，也就是说各位置在取特定值的情况下，判定两种情况是一致的。由此设定以下值为概率的临界值：

$$\left(\frac{len(keyword)\times times(keyword)+1}{len(text)}\right)_{标题}=\frac{1}{3} \tag{10}$$

$$\left(\frac{len(keyword)\times times(keyword)+1}{len(text)}\right)_{摘要}=\frac{1}{2} \tag{11}$$

由题意可得：

$$\alpha H_{标题}=\beta H_{摘要} \tag{12}$$

假定标题与摘要的文本长度是一样的，将式(10)和式(11)代入上式，有

$$\alpha\log_2 2=\beta\log_2 3 \tag{13}$$

又由式(9)$\alpha+\beta=1$，可将上式变为

$$\alpha\log_2 2=(1-\alpha)\log_2 3 \tag{14}$$

解一元方程可得到

$$\alpha=0.6131 \tag{15}$$

$$\beta=1-\alpha=0.3869 \tag{16}$$

因此式(8)可确定为

$$H=0.6131H_{标题}+0.3869H_{摘要} \tag{17}$$

3.3 成员搜索引擎排序的影响

成员搜索引擎都有自己的排序算法，虽然会存在广告，但是其算法的重要性也不容忽视。因此，本文提出成员搜索引擎的排序算法对结果排序是存在影响的。为确定原来成员搜索引擎的影响带来的修正值，首先通过大量的搜索字符串统计式(17)的平均 IRD 的大小。本文以百度、Bing 和 Yahoo 3 个成员搜索引擎为例，表 1 为统计 100 次搜索串得到的各搜索引擎前 10 项结果的平均 IRD(数值保留到小数点后面 4 位)。

表 1 统计各个成员搜索引擎前 10 条结果的 IRD

位置	引擎		
	百度	Bing	Yahoo
1	0.0250	0.0487	0.0383
2	0.0338	0.0440	0.0637
3	0.0307	0.0276	0.0578
4	0.0320	0.0360	0.0361
5	0.0388	0.0455	0.0492
6	0.0377	0.0347	0.1323
7	0.0379	0.0448	0.0697
8	0.0329	0.0414	0.0568
9	0.0324	0.0293	0.0666
10	0.0339	0.0337	0.0923

如表 1 所列，应用式(17)得到的大多数结果只在百分位才有意义，只有极个别的数值会特别大或者特别小。因此，提出在原先的 IRD 结果的千分位加入修正值，修正的值为 $i/1000$，其中 i 为各个结果原先的排序位置，取值从 0 开始。对于大部分结果来说，加入修正值后对原先的结果会产生比较大的影响。但是对于原来少数数值特别大或者特别小的值来说，意义是不明显的。鉴于这少数部分已经有明显的确定意义(相关度的确定性已经可以确定)，因此加入修正值后对其实际意义不会产生很大的影响。

由此提出修正公式：

$$H'=H+\frac{i}{1000}=(0.6131H_{标题}+0.3869H_{摘要})+\frac{i}{1000} \tag{18}$$

式中，i 的取值范围是从 0 开始到总抓取数目减 1，即 $i=0,1,\cdots,n-1$。

可以验证，排名越后的结果，其加入的修正值就越大。因此，根据 IRD 的意义，其也就是越不相关的结果。而对于此，上式成立。

3.4 成员搜索引擎的信任度

对于不同的搜索引擎，搜索结果自然优劣不同。因此，就其信任度而言，也是需要进行区分的，即是需要进行加权的。

对于不同的搜索引擎，需要给予不同的加权系数，因此改

写式(18)为

$$H=\alpha_i(0.6131H_{标题}+0.3869H_{摘要}+\frac{k}{1000}) \tag{19}$$

式中，i 表示第 i 个搜索引擎，α_i 为第 i 个搜索引擎的加权系数，k 为该结果在第 i 个搜索引擎中排在第 k 个。且给定

$$\alpha_i<1 \tag{20}$$

现在需要确定加权值 α_i 在信任度不同的搜索引擎中的大小。

假如有搜索引擎 A 和 B，若对搜索引擎 A 的信任度要高于 B，设搜索引擎某一相同结果的 IRD 分别为

$$H_A=\alpha_A(0.6131H_{标题}+0.3869H_{摘要}+\frac{k}{1000}) \tag{21}$$

$$H_B=\alpha_B(0.6131H_{标题}+0.3869H_{摘要}+\frac{k}{1000}) \tag{22}$$

由于对搜索引擎 A 的信任度要高，因此对搜索引擎 A 搜索到的结果相关度要高。根据相关度与 IRD 的关系可以得到

$$H_A<H_B \tag{23}$$

假定 k 相同，结合上述 3 式可以得到

$$\alpha_A<\alpha_B \tag{24}$$

对于本文给定的 3 个成员搜索引擎，百度、雅虎中国和必应中国，若要确定各自优劣性，首先要根据式(18)观察其统计平均值，即将表 1 的各项数据分别加上 $i/1000$ 即可得到。

为更加直观、清楚地描述各自统计平均值的关系，将其变化采用趋势图描述，如图 3 所示。

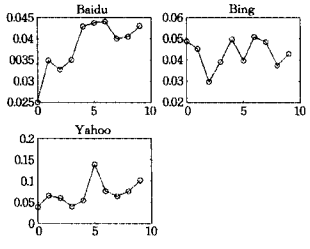

图 3　式(18)计算的统计平均数值趋势图

首先，以尊重各成员搜索引擎原先排序为前提，简单地说，若各成员搜索引擎的原先排序具有参考意义，那么用 IRD 计算后它们对应的值应该是按照升序排列的。如图 3 所示，Baidu 的曲线总体上是按照升序排列；Bing 的值相对平稳，区分度不大；而 Yahoo 的曲线在中部出现大的偏差，但是总的趋势也算符合要求。由此可以判断它们的信任度关系为百度>雅虎中国>必应中国。由式(24)，本文给出其各自的加权值分别为 0.9、0.8、0.7。

4　结果的合成

元搜索引擎对结果返回就是对结果的合成。合成方法非常多，如静态方法、CombSum、CombMNZ[9] 等。合成是将相同的结果进行合并后根据一定顺序返回给用户。结合修正后的 IRD 值的特点，现提出 CombMul (Combination of Multiplication)合成方法，公式如下：

$$H=\prod_{i=1}^{k}H_i \tag{25}$$

式中，k 是该结果重叠的次数，也即是所用的成员搜索引擎中含有 k 个该结果；H_i 为依据式(19)求得的第 i 个成员搜索引擎所求得的结果。

证明：要使式(25)成立，则需要证明出现次数越多的，排名应该越靠前，也即修正的 IRD 越小。

先来看看

$$H_i=\alpha_i(0.6131H_{标题}+0.3869H_{摘要}+\frac{j}{1000})$$

式中，$H_{标题}$ 和 $H_{摘要}$ 通过式(7)求得，即

$$H=\frac{1}{len(text)}\times[-\log_2\frac{len(keyword)\times times(keyword)+1}{len(text)}]$$

为叙述方便，将上式改写成

$$H=\frac{1}{y}\times[-\log_2\frac{x+1}{y}] \tag{26}$$

式中，x 和 y 都为整数，且 $x\geqslant 0$，$y\geqslant 1$。现在需要求得式(26)的最大值。在上式中，固定 y 值，则函数是随着 x 的增大而减小的。因此，当 $x=0$ 时具有最大的值。代入上式得

$$H=\frac{1}{y}\times(-\log_2\frac{1}{y})=\frac{1}{y}\times\log_2 y \tag{27}$$

对上式求导，得

$$H'=-\frac{\log_2 y}{y^2}+\frac{\ln 2}{y^2} \tag{28}$$

由导数的定义以及推论可得，若存在 y 值使得 $H'=0$，则该点就是式(26)的最大值点。

令 $H'=0$，则有

$\log_2 y=\ln 2$

$y=2^{\ln 2}$

将 $y=2^{\ln 2}$ 代入式(27)，得到

$H=0.4287$

也就是说，点 $(0,2^{\ln 2})$ 使得式(26)值最大，该最大值为 0.4287。结合 3.1 节可得

$0\leqslant H<1$

因此有

$0\leqslant 0.6131H_{标题}+0.3869H_{摘要}<1$

则

$$0\leqslant 0.6131H_{标题}+0.3869H_{摘要}+\frac{j}{1000}<1$$

又由式(20)所示，$\alpha_i<1$，则

$$0\leqslant\alpha_i(0.6131H_{标题}+0.3869H_{摘要}+\frac{j}{1000})<1$$

所以式(26)必然会随着 k 值的变大而变得越来越小。也就是说，同一结果在出现次数越多时，所得到的 IRD 会越小，而相关度会更加大。因此得证式(26)的可行性。

5　算法评价

评价搜索引擎排序算法优劣的方法有多种，本文采用 MRR 查准率[6] 的评价方法。MRR 查准率是结合了查准率与排序后的位置来给予评定的。MRR 查准率评定结果与排序算法的优劣成正比关系。

成员搜索引擎的选用如前面所述，将本文算法与较广泛使用的 Borda 算法做比较，选取 50 个查询串对用户做模拟查询。对于这样的查询串抓取成员搜索引擎的前两页数据，计算各个查询串的 MRR 查准率，最终统计这 50 个值的平均 MRR 查准率，结果如表 2 所列。

表 2　统计 50 个 MRR 查准率平均值

排序算法	Borda	IRD
平均 MRR 查准率	1.513	2.432

(下转第 191 页)

力、舒适度等级、质量等级、安全等级。

如果单是结合专家权重直接对专家的主张值进行加权求和可以得到这样的结论：汽车的“耗油量”是客户关心的第一工程技术指标。产生这个结论的原因主要是专家3的权重和主张确定值相对较大，由此可以看出此方法给结论带来了不稳定和不准确性。

通过案例的计算、比较与分析可以看出，本文提出的研讨信息模型将不确定性研讨信息进行了结构化与量化，从中提取到了论点框架、有效论点组并构造了支持分配函数，应用信息融合的技术对信息进行综合，得到了论点的可信度，从而生成了结论。为了消除论据冲突，采用论据支持度重新分配与平均证据修正的方法，一定程度上增强了信息融合的可靠性。

结束语 群体研讨中，专家个体因为有限理性、知识背景与经验的不同，对问题的思考会存在局限性或者差异性，因而研讨信息呈现出不确定性的特点。本文针对此特点提出了自然属性与人工属性集成的研讨信息模型。该模型中的主张支持值将专家的主观性进行了量化，通过构造支持分配函数来表示专家对论点的支持程度。从不确定性研讨信息中提取论点框架和有效论点组，为信息的融合提供了数据结构的基础。在信息融合方面，采用论据支持度重新分配方法，并结合平均证据去修正原始支持度，一定程度上缩小了论据间的冲突，增强了信息融合的稳定性和可靠性，促进了共识结论的生成。后续将进一步研究引起信息不确定性的其他因素，扩展模型结构，提高模型的处理能力。

参 考 文 献

[1] Kunz W, Rittel W J H. Issues as elements of information systems[R]. Berkeley, CA: University of California, 1970

[2] Hemant K B, Daniel J P, Daewon S. Progress in Web-based decision support technologies[J]. Decision Support Systems, 2007 (43): 1083-1095

[3] Ohbyung K, Kyoung Y K, Kun C L. MM-DSS: Integrating multimedia and decision-making knowledge in decision support systems[J]. Expert Systems with Applications, 2007(32): 441-457

[4] 李嘉，张朋柱. 群体研讨支持系统中研讨主题的自动可视化聚类研究[J]. 系统管理报，2009，18(3)：325-331

[5] Alonso S, Cabrerizo F J, Chiclana F, et al. A Web based consensus support system for group decision making problems and incomplete preferences [J]. Information Sciences, 2010(180): 4477-4495

[6] Alonso S, Viedma E H, Chiclana F, et al. Individual and social strategies to deal with ignorance situations in multi-person decision making[J]. International Journal of Information Technology and Decision Making, 2009, 8(2): 313-333

[7] Cabrerizo F J, Alonso S, Viedma E H. A consensus model for group decision making problems with unbalanced fuzzy linguistic information[J]. International Journal of Information Technology and Decision Making, 2009, 8(1): 109-131

[8] Chiclana F, Mata F, Martnez L, et al. Integration of a consistency control module within a consensus decision making model [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 16(1): 35-53

[9] Mata F, Mart L, Viedma E H. An adaptative consensus support model for group decision making problems in a multi-granular fuzzylinguistic context [J]. IEEE Transactions on Fuzzy Systems, 2009, 17(2): 279-290

[10] Chen Y J. Development of a method for ontology-based empirical knowledge representation and reasoning [J]. Decision Support Systems, 2010, 50(1): 1-20

[11] 李德华，熊才权. 一种研讨信息组织模型及其在研讨厅中的应用[J]. 计算机应用研究，2008，25(9)：2730-2733

[12] Toulmin S. The uses of argument[M]. New York: Cambridge Univemity, 1958: 1-10

(上接第156页)

从表2中可得

$MRR_{IRD} > MRR_{Borda}$

由MRR查准率的特性可以得出，IRD排序算法明显要优于Borda算法。

结束语 元搜索引擎的核心研究在于对从成员搜索引擎中获得的结果进行重新排序。以往的排序算法在计算上有一定的缺陷，在研究以往算法并集合多种算法后提出修正后的IRD算法，经实验结果表明，其较以前的排序算法更有优越性，能有效提高排序结果的效率：将用户期望的信息排在前面。下一步将结合历史值调整计算结果的相关度[7]。

参 考 文 献

[1] Selberg E, Etzioni O. Multi-service search and comparison using the MetaCrawler [C] // Proceedings of the 4th International World Wide Web Conference. December 1995

[2] Hsu D F, Taksa I. Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval [J]. Information Retrieval, 2005, 8(3): 449-480

[3] Shimotsuruma, Yamato S, Kanagawa K. Information Retrieval on the Web [J]. ACM Computing Surveys, 2000, 32(2): 144-171

[4] 曹林. 元搜索引擎排序技术研究 [D]. 南京：河海大学，2009

[5] 樊昌信，曹丽娜. 通信原理(第六版)[M]. 北京：国防工业出版社，2008

[6] Mizzaro S. Relevance: The whole history [J]. Journal of the American Society for Information Science, 1997, 48(9): 810-832

[7] Wu Sheng-li, Bi Ya-xin, Zeng Xiao-qin. Retrieval Result Presentation and Evaluation [J]. Lecture Notes in Computer Science, 2010, 6291: 125-136

[8] Martl′nez-Santiago, Ureña-López L, Martl′n-Valdivia M. A merging strategy proposal: The 2-step retrieval status value method [J]. Information Retrieval, 2006, 9(1): 71-93

[9] Wu Sheng-li. Assigning appropriate weights for the linear combination data fusion method in information retrieval [J]. Expert Systems with Applications, 2009, 36(2): 2997-3006