第31卷第3期 华中科技大学学报(自然科学版) Vol.31 No.3
2003年 3月 J. Huazhong Univ. of Sci. & Tech. (Nature Science Edition) Mar. 2003

元搜索引擎中检索结果排序的优化方法

文坤梅 卢正鼎 邓 曦 陈 莉

(华中科技大学计算机科学与技术学院)

摘要: 提出了一种新的基于概率模型的排序优化方法.利用贝叶斯规则,结合各组成系统平均执行性能的信息,推导出一种新的相关度计算公式,较好地解决了结果融合中相关度规范化和均衡化的问题.经实验验证,该方法对结果进行了最优化排序,其实际执行性能超出了现有的任何一个组成系统的性能.

关 键 词: 元搜索引擎; 概率模型; 结果优化排序; 排序融合

中图分类号: TP393.09; TP311.135 **文献标识码**: A **文章编号**: 1671-4512(2003)03-0049-03

在目前所存在的搜索引擎中[1],没有一个搜索引擎能够覆盖所有的WWW资源,大部分的搜索引擎都只能涉及到整个资源的一小部分.并且各类搜索引擎的信息来源差异较大,因此集成多个搜索引擎而产生的元搜索引擎具有比传统引擎覆盖面大,引擎效果更好且具有可扩展性等优点.其中对各个组成系统所返回的搜索结果进行排序是提高元搜索引擎效率的关键技术.

1 排序融合的关键技术

每一个成员搜索引擎都有自己的排序检索结果算法[2],根据用户所给定查询的相关度来排序文件.然而,这些方法千差万别,通常每一个算法都是某一搜索引擎提供者所特有的,并且算法不公开,这就使得融合以及排序来自不同数据源的数据结果变得非常复杂.

1.1 相关度的规范化

每一个成员搜索都有各自的尺度来衡量文件的相关度.例如,数据源 R_1 判断文件 f_1 对某一查询其相关度为0.1,而数据源 R_2 判断文件 f_2 对某一查询其相关度为1 000.如果想把来自 R_1 和 R_2 的结果融合成为单一的文件排序列表,那么 f_1 的相关度是高于或者低于 f_2 根本无法判断,因为不存在统一的标准.如何统一各个数据源的相关度,这就涉及到相关度的规范化问题.

1.2 相关度的均衡化

大多数搜索引擎的排序算法是不公开的,只有少数公开其算法.事实上,即使用同样的排序算法,在处理相关度问题时依然存在很大的困难.原因在于算法是基于不同的文件集合来排序文件.例如,R_1 是专门研究计算机科学的数据源,那么词语"数据结构"可能会出现在很多文件中,于是"数据结构"这个词语在 R_1 中将会有很低的相关度.而与此同时,如果数据源 R_2 和计算机科学完全不相关,并且 R_2 中出现过这个词语的文件很少,那么"数据结构"在数据源 R_2 中可能会有很高的相关度.

对包含"数据结构"这个词语的查询,R_1 可能会赋予文件较低的相关度,而 R_2 则会赋予较高的相关度.在同一个查询中,两个非常相似的文件 f_1 和 f_2,如果 f_1 在 R_1 中,而 f_2 在 R_2 中,却得到了不同的相关值.因此,即使数据源采用同样的排序算法,一个元搜索引擎仍然需要一些附加的信息用一种有效的办法来融合查询结果.

最好的解决办法就是综合考虑各个成员搜索引擎所给出的相关度,从而消除各个数据源本身带来的偏差.

2 检索结果排序的优化方法

在响应一个给定查询时,为了组合多个查询检索系统[3]所得到的文件排序列表,更好地解决上述两个问题,提出了一种新的概率模型.

2.1 一种元搜索引擎的概率模式

假定在响应给定查询时,元搜索引擎已经得

收稿日期: 2002-09-06.

作者简介: 文坤梅(1978-),女,硕士研究生;武汉,华中科技大学计算机科学与技术学院 (430074).

基金项目: 国家高性能计算基金资助项目 (99319).

到了各个成员搜索引擎的文件排序列表.同时也获取了一些简单的统计信息,包括关于组成系统的平均执行性能信息.这些信息都是以元数据形式给出的.基于这些信息,提出了一种概率模型并推导出优化的元搜索引擎策略.

其中元数据包括:对任何一个查询,每一个成员搜索引擎所对应文件的相关度和不相关度,且这些都是未规范化的初始值.

给定 n 个检索系统返回的文件排序列表,$r_i(d)$被检索系统 i 赋值为文件d 的相关度(如果文件 d 没有被系统 i 检索,那么它的相关度就为∞).相关度是成员搜索引擎在评测文件 d 时提供给元搜索引擎的,评测建立在相关度规则之上[4].对给定的文件,假定:

$$Q_{re} = Q_r[r_1, r_2, \cdots, r_n]_{re};$$

$$Q_{ir} = Q_r[r_1, r_2, \cdots, r_n]_{ir},$$

式中,Q_{re}是给定的文件相关的概率值;Q_{ir}是给定的文件不相关的概率值.给定序列 $r_1, r_2, \cdots, r_n$,规定如果 $Q_{re} > Q_{ir}$,那么文件是相关的,反之则不相关.可以先计算出相关度的几率:

$$O_{re} = Q_{re}/Q_{ir}$$

然后根据这一价值尺度来排序.

应用贝叶斯规则,得到:

$$Q_{re} = Q_r[r_1, r_2, \cdots, r_n]_{re} \cdot Q_r[re]/Q_r[r_1, r_2, \cdots, r_n];$$

$$Q_{ir} = Q_r[r_1, r_2, \cdots, r_n]_{ir} \cdot Q_r[ir]/Q_r[r_1, r_2, \cdots, r_n].$$

$Q_r[r_1, r_2, \cdots, r_n]$ 这一项在实际中很难得到,以比率的形式估算,即

$$Q_{re} = Q_r[r_1, r_2, \cdots, r_n]_{rc} \cdot Q_r[re]/\{Q_r[r_1, r_2, \cdots, r_n]_{ir} Q_r[ir]\},$$

满足原始贝叶斯独立性假设,上式等同于

$$Q_{re} = Q_r[re] \prod_i Q_r[r_i \mid re]/\left\{Q_r[ir] \prod_i Q_r[r_i \mid ir]\right\}.$$

最后,由于仅仅考虑排序文件,可对其取对数,得到相关度的计算公式

$$\sum_{i=1}^{n} \log Q_r[r_i \mid re]/Q_r[r_i \mid ir],$$

式中 $Q_r[r_i \mid re]$是文件被成员系统 i 排列到水平 r_i 的相关概率值.同样,$Q_r[r_i \mid ir]$ 是一个文件可能被成员系统 i 排到水平 r_i 的不相关概率值.因此,对每一个成员搜索引擎,得到了文件的相关度,把所有成员搜索引擎的概率值比率的对数值相加,所得之和即为该文件的最终相关度.

相关度与非相关度的比率消除了相关度的规范化问题,屏蔽掉了各个成员搜索引擎中具体的相关度;另外各个成员搜索引擎的比率值相加这一点又综合考虑了各个搜索引擎所起的作用,实现了相关度的均衡化,从而客观地反映了文件的真实相关度.

2.2 方法评估与实验结果

利用基于概率模型的优化方法进行了实验,以目前比较通用的6种搜索引擎作为成员引擎集成了元搜索引擎 Mix,这五种成员搜索引擎分别是:新浪(sina)、网易(netease)、天网(pku)、雅虎(Yahoo)、搜狐(Sohu)以及 Google(对应 $i=1, 2, \cdots, 6$).Mix 在数据融合时采用了基于概率的检索结果优化排序方法,由于存在6个成员搜索引擎系统,因此 $i=6$,对任意文件 d,其相关度

$$\sum_{i=1}^{6} \log\{Q_r[r_i \mid re]/Q_r[r_i \mid ir]\} = \log\{Q_r[r_1 \mid re]/Q_r[r_1 \mid ir]\} + \log Q_r\{[r_2 \mid re]/Q_r[r_2 \mid ir]\} + \log\{Q_r[r_3 \mid re]/Q_r[r_3 \mid ir]\} + \log\{Q_r[r_4 \mid re]/Q_r[r_4 \mid ir]\} + \log\{Q_r[r_5 \mid re]/Q_r[r_5 \mid ir]\} + \log\{Q_r[r_6 \mid re]/Q_r[r_6 \mid ir]\}.$$

元搜索引擎 Mix 利用此相关度公式融合各个成员搜索引擎所返回的结果,不仅扩大了搜索的覆盖面,而且引擎效果更好,将用户真正想得到的信息赋予了较高的相关值.

假设输入"肝炎"为关键字,经测试发现每个成员搜索引擎以及 Mix 系统都能响应查询,实验结果如表1所示,其中,c 为搜索效率;t 为搜索时间.

表1 各个组成搜索引擎与元搜索引擎的效率比较

搜索引擎	总页数	c/%	t/s
Sina	82 648	55.8	0.126
网易	77 000	52.0	0.162
天网	31 706	21.4	0.016
Yahoo	77 000	52.0	0.162
Sohu	82 648	55.8	0.126
Google	77 000	52.0	0.162
Mix	118 563	86.1	0.183

综上可见,由于采用了基于概率的检索结果排序方法,元搜索引擎的效率得以很大的提高,具体表现在搜索覆盖率的增大,同时响应时间并没有太大的变化,综合性能强于任何一个成员搜索引擎.此外,系统的强壮性也有很大的提高,当成员系统执行性能特别差时,元搜索引擎系统的执行性能也不会随之变得更差.

参　考　文　献

[1] 张晓辉，邵　华，常桂然. WWW上的信息发现与搜索引擎技术. 小型微型计算机系统，1998，19(6)：66～71

[2] Arasu A, Cho J, Hector G M, et. al. Searching the Web. ACM Transactions on Internet Technology, 2001, 1(1): 2～43

[3] 王继成，邹　涛. 基于Internet的信息资源发现技术与实现. 计算机研究与发展，1999，36(11)：1 369～1 374

[4] Lawrence S, Lee C. Context and page analysis for improved Web search. IEEE Internet Computing, 1998, 2(4): 38～46

An optimal for ranking results of web search in metasearch

Wen Kunmei　Lu Zhengding　Deng Xi　Chen Li

Abstract: The paper put forward a new optimal scheme based on a probability model. By using the Bayes rule and combining the information of the average performance of the component systems, a new relation degree formula was deduced and the problems on relational degree standardization and equilibrium were solved. The experiment results showed the new system provided the optimal ranking.

Key words: metasearch engine; probability model; results optimal ranking; ranking and fusing

Wen Kunmei　Postgraduate; College of Computer Sci. & Tech., Huahzong Univ. of Sci. & Tech., Wuhan 430074, China.

元搜索引擎中检索结果排序的优化方法

作者： 文坤梅， 卢正鼎， 邓曦， 陈莉
作者单位： 华中科技大学计算机科学与技术学院
刊名： 华中科技大学学报(自然科学版) ISTIC EI PKU
英文刊名： JOURNAL OF HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY(NATURE SCIENCE EDITION)
年，卷(期)： 2003,31(3)
被引用次数： 9次

参考文献(4条)

1. Lawrence s;Lee C Context and page analysis for improved Web search[外文期刊] 1998(04)

2. 王继成;邹涛 基于Internet的信息资源发现技术与实现[期刊论文]-计算机研究与发展 1999(11)

3. Arasu A;Cho J;HectorGM Searching the Web 2001(01)

4. 张晓辉;邵华;常桂然 WWW上的信息发现与搜索引擎技术 1998(06)

引证文献(9条)

1. 李建廷 基于模糊积分的元搜索引擎结果排序算法[期刊论文]-计算机仿真 2010(7)

2. 曹林. 韩立新. 吴胜利 元搜索引擎排序技术综述[期刊论文]-计算机应用研究 2009(2)

3. 丁一. 龚家才 基于半完全图在数据融合中的元搜索研究[期刊论文]-湖北师范学院学报（自然科学版） 2008(2)

4. 薛辉明 数字资源管理系统的设计与实施方案[期刊论文]-电脑知识与技术（学术交流） 2006(3)

5. 张辉. 隋佳 基于Z39.50的元搜索引擎优化策略[期刊论文]-中国图书馆学报 2006(2)

6. 施游. 刘宏 数字图书馆资源整合数据检索模型[期刊论文]-电脑知识与技术（技术论坛） 2005(11)

7. 刘竞 多元搜索引擎的研究与实现[学位论文]硕士 2005

8. 樊康新 基于服务器端的个性化元搜索引擎的研究与设计[学位论文]硕士 2005

9. 丁一 基于Web挖掘的个性化推荐服务研究[学位论文]硕士 2004