

个性化元搜索结果整合算法的研究

李琴琴 汤小春 靳明星

(西北工业大学计算机学院,西安 710129)

摘要 元搜索是一种基于搜索引擎的搜索引擎,它将各个独立搜索引擎的结果经过融合呈现给用户,以此为用户提供更加全面的信息。但是要在数量庞大的搜索结果中快速地找到自己所需要的信息却不是一件容易的事。提出了一种基于用户兴趣的个性化元搜索引擎模型,通过提取用户个性化行为特征,建立用户兴趣库,并对搜索结果进行合理排序和整合,将用户最感兴趣的信息尽可能排在最前面,从而实现元搜索的个性化,提高用户搜索的查准率和效率。

关键词 元搜索 结果整合 用户兴趣 相关度

中图法分类号 TP391.3; **文献标志码** A

随着 Internet 的迅猛发展以及 Web 信息的飞速增长,网络已经成为人们获取信息的必要途径和重要手段,但是网络中的信息种类繁多、信息数目庞大、再加上新信息的不断出现以及网页的快速更新等特点,人们要找自己感兴趣或有用的信息,需要花费大量的宝贵时间。虽然各种各样的传统搜索引擎能够帮我们快速找到相关的主题,但是,传统搜索引擎搜索结果通常成百上千,有些结果跟用户的搜索关键字毫不相关,故无法满足人们对查全率、查准确率的要求,为了解决这些问题,元搜索引擎应用而生。

元搜索引擎是搜索之上的搜索引擎,它通过调用多个搜索引擎来实现搜索,并对搜索结果进行整合处理,能解决传统搜索引擎查询覆盖率低的问题^[1],但是元搜索引擎的各个成员搜索引擎索引数据库的覆盖范围、搜索算法和排序算法各不相同^[2],导致搜索出来的结果也不尽相同。目前,常见的元搜索引擎的结果整合方法有:

(1) 简单罗列式,只是将多个成员搜索引擎的搜索结果简单地罗列出来,没有考虑结果的相关度,这种方法的缺点显而易见;

(2) 基于相关度和位置的结果合成方法,这种方法的不足之处是仅仅依靠成员搜索引擎返回的有限描述信息来判断查询字符串和这个结果的相关性,没能结合用户兴趣和考虑成员搜索引擎的优先级问题^[3],导致所得的相关性信息局限性太大;

(3) 基于训练集的结果整合算法,这种方法在训练集计算耗时较大。

元搜索引擎综合了多个搜索引擎的搜索结果,提高了搜索的覆盖率,但是返回的结果往往数目庞大,并且很多结果与用户查询并不相关,这直接影响了用户检索的质量并增加了用户检索的代价。因此,本文提出一种基于用户兴趣的个性化元搜索引擎模型,系统通过对用户建立兴趣库,提取个性特征形成不同用户群,并对检索到的结果进行整合处理,返回给用户个性化的搜索结果。

1 基于用户兴趣的个性化元搜索

1.1 个性化搜索技术

个性化搜索技术是指根据不同用户的个性化行为采取不同的、有针对性的服务策略,提供符合用户个性化需求的服务内容。在本文中具体表现为针对不同的用户兴趣偏好,采用不同的目标站点,帮助用户更快,更准确地找到信息。

基于现有元搜索引擎结果整合算法存在的弊

端,本文针对查询结果整合问题,结合个性化搜索技术,提出了一种基于相关度和用户兴趣相结合的结果整合算法,通过用户反馈和对用户浏览日志进行挖掘,得到用户兴趣库,计算搜索结果与成员引擎的相关度、用户兴趣库中最常出现的关键词与搜索结果的相关度,然后考虑成员搜索引擎的权重、搜索结果的重复度等因素,最后计算搜索结果的权重值,对结果进行整合排序^[3]。

1.2 基于用户兴趣的个性化元搜索引擎模型

为了提高搜索引擎的查全率、查准确率,满足用户的个性化搜索,本文提出了基于用户兴趣的元搜索引擎,来实现元搜索引擎的个性化。基于用户兴趣的元搜索引擎主要的组成模块有:用户兴趣库模块、引擎调度模块、查询分发模块、结果处理模块、反馈模块,日志处理模块等(见图1)。

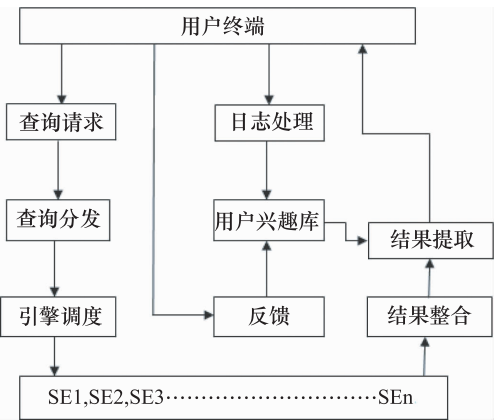


图1 基于用户兴趣的个性化元搜索引擎模型

用户兴趣库模块中存放了用户感兴趣的关键词,出现次数以及创建时间,更新时间等,当进行结果排序时,将用户兴趣库中的信息与查询结果进行匹配,从而将用户感兴趣或最需要的信息尽可能地排在最前面,本文将在下一部分介绍用户兴趣模型的建立及更新。

引擎调度模块是指研究元搜索引擎如何为用户选择数量合适并贴近用户查询要求的成员搜索引擎。常用的调度技术有好几种,本文采用用户手动选择技术即根据用户的喜好来选择成员搜索引擎。在实验中,选择的成员搜索引擎有 Google、Yahoo、Bing 等。

查询分发模块是指根据用户输入的关键词,将查询请求转化为各成员引擎能接受的格式,并提交给引擎调度模型选中的成员引擎。

结果处理模块有结果提取和结果整合两部分。结果提取是将各成员搜索引擎的搜索结果保存下来,结果整合是指将结果提取中保存的结果按照一定的规则进行排序,把用户最需要的信息排在最前面,其中,每条结果对应一个搜索引擎和用户兴趣的关键词。

反馈模块是指及时更新用户兴趣库的内容,以便满足用户不停变化的需求,在本文中通过两种方式进行用户反馈,一方面是通过用户的主观判断,修改用户自身的兴趣;另一方面是针对不同关键词之间的、关键词与成员搜索引擎之间的相关度来更新用户兴趣库。

日志处理模块存放的是用户浏览网页的信息数据,包括用户的访问时间、用户的 IP 地址、输入的查询串等,是用户兴趣库的数据来源,日志处理模块的存在解决了一般元搜索无法获取成员搜索引擎的网页相关度与数据库中词条等统计信息这一不足。

2 用户兴趣库的建立

本文采用显式获取和隐式获取相结合的方式进行用户兴趣库的建立,这样做的好处是减少了单纯显示搜集方法中的搜集步骤,也在大多数时间上避免用户在使用中感到不便^[4]。用户在使用系统时需要注册的,注册成功后,用户的基本信息就保存在系统的数据库中,比如用户的 id,可以根据用户 id 来提取用户感兴趣的查询,从而建立用户兴趣库。最终,将用户最需要、最感兴趣的搜索结果呈现给用户。

2.1 显式获取

系统中提供一种界面,在这个界面中用户可以自愿、显式地输入自己感兴趣的内容即关键词,然后在数据库中根据用户注册 id 记录用户的输入,包括关键词,创建时间,更新时间,出现次数等。

2.2 隐式获取

用户在长期的上网过程中会浏览很多的页面,其中不乏一些有价值的页面。可以在浏览器的记录中找到他们,对其进行收集,其集合表示为 M 。在时间段 T 内,用户浏览的网页按 M 中的数据来表示为 $p = \{(p_1, t_1, r_1), (p_2, t_2, r_2), \dots, (p_n, t_n, r_n)\}^{[5]}$ 。网页的兴趣度计算公示为 $a = t_i r_i / \sum t_j r_j$, 其中 t_j 为用户在时间 T 内浏览网页的所有时间之和, r_j 内网页出现的次数,基于此可以定义用户对网页的兴趣度,其中 a 为网页在时间段 T 的兴趣度, $1/\lg(\text{length})$ 说明网页长度对兴趣度的大小起负作用。

根据上面得到的网页兴趣度值,可以对用户浏览的网页进行过滤,给网页的兴趣度 R_i 设定一个阈值。只有 R_i 值大于这个阈值的网页才会把网页信息放入到日志处理模块中。然后对用户信息需求、偏好进行甄别、归纳、总结,分析用户的兴趣,并将信息进行整理、组织,从中分析出用户的信息偏好,再根据用户的这些信息建立用户兴趣库^[6]。其中包括关键词、出现次数、创建时间、更新时间等。当进行结果整合时,将用户兴趣库中的信息与用户查询条件进行匹配,从而将更加符合用户需求的信息排在最前边。

3 个性化元搜索结果整合算法

个性化元搜索引擎中,由于各个成员搜索引擎对各自的排名算法不公开,这就导致了元搜索引擎在进行结果整合的时候,无法根据成员搜索引擎采用的算法排序方式来调整元搜索的排序方式,在成员搜索引擎提供的每个搜索结果中,应该为用户提供查询串在该结果中的权重,并且提供数据库中包含该查询串的文档数,从而方便元搜索引擎进行结果整合时使用^[7]。现在,元搜索引擎只有根据成员搜索引擎返回的结果定制相应的算法。以前的元搜索引擎在定制算法时,往往只考虑到了查询关键字与返回结果的相关程度,而忽略了用户的因素。事实上,每个用户在查看查询结果的时候,都会根据返回的简短的描述信息来判断这个搜索结果是

否包含自己想要的信息。因此在基于相关度和用户兴趣的元搜索引擎结果整合算法中,引入了搜索引擎的权重、网页的兴趣度^[8],以及用户的兴趣特征等概念。对于以查询串 q 来说,它和某个搜索结果 $result_i$ 的相关度定义如下:

$$Rank(q, result_i) = F(SEWeight, Position, Correlation, Repeats, Interest) \quad (1)$$

式(1)中 $SEWeight$ 的影响因子,即成员搜索引擎在本系统中的权重。 $Position$ 表示结果在成员搜索引擎结果集中的位置, $Correlation$ 表示查询字符串与结果相关联的程度, $Repeats$ 表示搜索出 $result_i$ 的搜索引擎的个数, $Internet$ 表示用户与该结果的匹配程度。下面分别对这几个影响因子进行讨论

3.1 搜索引擎的权重

各大成员搜索引擎各有特色,比如有的偏向于搜索结果多,有的偏向于搜索时间短。例如,百度侧重新闻搜索,Google 侧重学术搜索。搜索引擎所占的市场份额,反映了该搜索引擎的受欢迎程度。所以在本文中,基于搜索引擎所占的市场份额,计算出搜索引擎的权重,作为搜索结果排名的一个影响因子,假设搜索引擎 SE 所占的市场份额为 S ,则搜索引擎的权重可以定义如下:

$$SEWeightRank(SE_i) = 1 + s^2 \quad (2)$$

3.2 结果在成员搜索引擎中的位置

成员搜索引擎返回的搜索结果是按照搜索引擎自身的算法计算,且依据某种条件按降序排序后的一个有序结果集。排名越靠前的文档,与查询词的相关度越高。因此,将每个文档在结果集中的位置作为元搜索引擎结果整合排名的一个依据是很有道理的。假如对于某个查询词 q ,成员搜索时引擎返回了 m 个搜索结果,那么处于位置 k 的文档 $result$ 与查询词的位置相关度定义如下:

$$PosRank(q, SE_i, result_i) = \frac{m + k - 1}{m} (1 \leq k \leq m) \quad (3)$$

3.3 结与描述信息的相关度 $Correlation$

与查询词相关的描述信息片段主要有两部分:网页的标题和网页正文的摘要。标题一般是对整个网页信息的最精炼的概括,摘要则是对文档中出

现查询词的相关部分的提取。比如:如果查询字符串 q 出现在标题中,那么说明这个网页的重要行大于 q 出现在正文中的网页。同样地, q 出现在正文中的次数多的网页的重要行大于 q 出现次数少的网页的重要性。那么查询字符串 q 和返回结果 $result_i$ 的相关度可以表示如下(在本文中未考虑标题的影响)

$$\begin{aligned} & CorrectionRank(q, result_i) = \\ & length(result_i) \times \\ & \sum_1^{count(q, result_i)} \frac{Position(q, result_i)}{count(q, result_i)} \end{aligned} \quad (4)$$

式(4)中, $length(result_i)$ 表示结果的长度, $count(q, result_i)$ 表示查询词 q 在 $result_i$ 中出现的总次数, $Position(q, result_i)$ 表示 q 在 $result_i$ 中出现的位置。

3.4 结搜索结果的重复度 Repeat

搜索引擎由于其数据库的覆盖范围的差异,搜索算法和排序算法的不同,搜索的结果集不尽相同,但是也可能有重叠。如果一个文档能同时被多个搜索引擎检索出来,那么可以认为该文档与查询关键字的相关程度非常高,因而在元搜索排序中也应该排在最前边,因此将搜索结果集的重叠看作排序的影响因子之一,本文中,简化起见,我们将搜索结果 $result$ 的重复度定义为检索出 $result$ 的搜索引擎的个数占成员搜索引擎总数的百分比,定义如下

$$PepeatRank(q, result_i) = \frac{Search_count}{Total_count} \quad (5)$$

式(5)中, $Total_count$ 是成员搜索引擎的总数, $Search_count$ 是根据查询关键字搜索出的 $result$ 的搜索引擎的个数,由此可见, $PepeatRank$ 介于 $[0, 1]$ 之间。

3.5 用户兴趣特征 Interest

在用户兴趣库中选择出现次数最多的 10 个关键词,用成员搜索引擎返回的结果去匹配这 10 个关键词,计算返回结果和 $keywords$ 的匹配程度,作为排名的一个依据。搜索结果和 $keywords$ 的匹配程度越高,说明该搜索结果越符合用户的需求,在结果排名中就应当越靠前。用户兴趣库中出现次数最多的 10 个 $keywords$ 和搜索结果 $result_i$ 之间的匹配程度定义如下

$$\begin{aligned} & InterestRank(keyword_1, keyword_2, \dots, keyword_{10}) = \\ & \sum_{j=1}^{10} \left(\frac{count(keyword_j, result_i)}{length(result_i)} \right) length(keyword_j) \times \\ & Title(keyword_j, result_i) \end{aligned} \quad (6)$$

式(6)中 $count(keyword_j, result_i)$ 表示 $keyword_j$ 在 $result_i$ 中出现的次数, $length(result_i)$ 表示搜索结果 $result_i$ 的长度, $length(keyword_j)$ 为 $keyword_j$ 的长度, $Title(keyword_j, result_i)$ 表示 $keyword_j$ 是否出现在标题中。

3.6 用搜索结果总排名计算

上面分别讨论了元搜索引擎进行结果整合时考虑的几个影响因素,以及影响值的计算方式。因此可得出,对于一个查询串 q ,检索结果 $result_i$ 从搜索引擎 SE 取得的权值计算公式如下:

$$\begin{aligned} Rank(q, SE_i, result_i) = & SEWeight(SE_i) (PosRank \times \\ & (q, SE_i, result_i) + Correlation- \\ & Rank(q, SE_i, result_i) + InterestRank(keyword_1, \dots, keyword_{10})) \end{aligned} \quad (7)$$

从以上可以得出对于查询串 q 的搜索结果 $result_i$ 在元搜索结果整合排名中的计算公式如下

$$\begin{aligned} Rank(q, result_i) = & Rank(q, SE_i, result_i) + Pepeat- \\ & Rank(q, result_i) \end{aligned} \quad (8)$$

下面给出基于用户兴趣和相关度相结合的结果整合算法:

输入:由结果提取模块得到的各搜索引擎返回的 n 条结果。

输出:按结果权重由大到小排列好的 m ($m \leq n$) 条结果。

Begin:

(1) 初始化结果列表 AllResults,

初始化最终结果列表 LastResults;

(2) ForAllResults 中的每一个 Result;

(3) ifResult 中包含查询词、用户兴趣库中的关键词,则计算结果在 AllResults 中的重复次数,删掉其它重复项;

(4) elseif 继续执行(2);

(5) 根据公示(7)计算 Result 的权重值;

(6) LastResults.add(Result);将符合条件的结果放入最终结果列表中;

(7) EndFor;

(8) 对 LastResults 根据每个结果的权重值进行排序;

(9) 返回排序后的结果。

4 实验结果

本文提出了元搜索基于用户兴趣的结果整合算法,建立了基于用户兴趣的元搜索引擎系统,通过大量的实验挖掘出了用户感兴趣的关键词,建立了用户兴趣库,随着实验次数的不断增加,搜索引擎针对不同用户的兴趣词搜索能力的差异逐渐地体现出来,通过实验绘出了 Google,Yahoo,Bing 三个搜索引擎与用户兴趣词比如 Cure 的相关度变化曲线(见图2)。

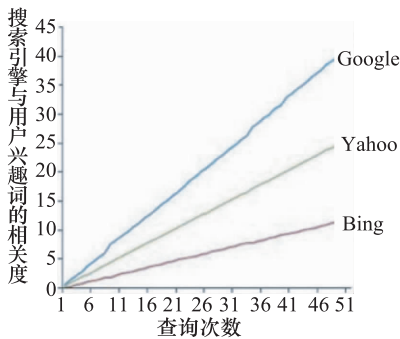


图2 用户兴趣词 Cure 的相关度的比较

通过程序实验上述提出的算法,并与 Google、Yahoo、Bing 等搜索引擎进行平均相关度比较,结果如表1所示,从实验结果可以看出,该算法的平均相关度明显好于单个搜索引擎,充分证明了算法的可行性。注明:表中 MyTest 为文中提出的算法。

表1 元搜索与单个搜索引擎平均相关度比较

搜索引擎	MyTest	Yahoo	Google	Bing
平均相关度	0.77	0.54	0.68	0.63

5 结论

元搜索引擎的结果合成算法的好坏直接影响着整个元搜索引擎系统的效率。基于用户兴趣的元搜索引擎系统继承了元搜索引擎搜索范围广的特征,同时提出了结果整合算法,将相关度、用户兴趣和用户反馈考虑到结果整合和排序中,在一定程度上解决了搜索引擎结果庞大,用户缺乏耐心去查看所有搜索结果的问题,提高了查准率,使得最终显示给用户的结果更加地合理和真实。

参 考 文 献

1 原福永,梁顺攀. 元搜索引擎的现状与发展. 计算机工程与设计, 2005;26(12):3278—3280

2 徐宝文,张卫丰. 搜索引擎与信息获取技术. 北京:清华大学出版社,2003

3 徐 科,崔志明,郑冬冬. 元搜索引擎中基于用户兴趣的查询结果合成技术. 微电子学与计算机,2006;23(7):199—201

4 韩 娜,沈西挺,刘 岩. 基于用户兴趣的个性化搜索系统研究. 软件导刊,2010;9(1):38—39

5 Kaw _ aiLamandchiHOIJeung. RankAggregationforMetasearchEngines. InACM,2004

6 LiuKinglup, Yu C M, Wei yi N P. Astatisticalmethodforestimatingtheusefulnessoftextdatabases. IEEE Transactionson Knowledgeand Data Engineering,2002

7 张健沛,李连江,杨 静. 个性化搜索引擎排序算法的研究与改进. 第三届全国信息检索与内容安全学术会议;1994—2010: 516—520

8 陆安江,董旭晖. 个性化元搜索引擎模型的研究与设计. 计算机与现代化,2011;(1):139—141

(下转第 8368 页)

由表 2 判别结果可知,中宁地区饱和粉-砂土液化等级从轻微液化到严重液化均有分布,液化指数 1.14~26.5。其中,轻微液化孔数 6 个,中等液化孔数 4 个,严重液化孔数 2 个。

4 结论

随着中宁县城市建设的高速发展,特别是高层建筑的大量兴建,建筑物基础埋深增大及大面积地下广场的兴建,地基土中饱和粉-砂土的地震液化现

象对建筑物的影响非常明显。中宁县历史上是地震多发、强度较大的地区,地层岩性主要由第四系冲洪积粉土、粉细砂和卵砾石土组成,地下水位埋藏较浅,地震液化危害程度相对较高,进行工程建设时,应采取有效的抗液化措施。

参 考 文 献

1 谢定义. 土动力学. 北京:高等教育出版社, 2011

2 中华人民共和国国家标准编写组. GB 50011—2010. 建筑抗震设计规范. 北京:中国建筑工业出版社, 2010

3 陈国兴. 岩土地震工程学. 北京:科学出版社, 2007

Evaluation on Seismic Liquefaction of Saturated Silty-sandy Soil in Zhongning County

ZHU Sai-nan, CAO Guang-zhu

(College of Territorial Resources, Kunming University of Science and Technology, Kunming 650093, P. R. China)

[Abstract] Zhongning County located in the west section of Wei-ning seismic belt is seismically-active and intensive area in history. silty-sandy soil liquefaction phenomenon caused by earthquake led foundation failure, and commonly accompanied by large-scale ground subsidence, slip, ground fissures, sandboils and watersprouts phenomena are presented. In order to identify the extension and hazard of foundation soil seismic liquefaction of Zhongning County, geotechnical engineering investigation reports data selected from 12 sites within the county to evaluate seismic liquefaction are used.

[Key words] seismic liquefaction liquefaction grade standard penetration test Zhongning County

(上接第 8365 页)

Research of Result Conformity Algorithm based on Personalized Meta-search

LI Qin-qin, TAN GXiao-chun, JIN Ming-xing

(School of Computer, North-Western Polytechnical University, Xi'an 710072, P. R. China)

[Abstract] Meta-search, which is a search engine-based search engine, provides more comprehensive information for users by presenting the integrated results from individual search engines, but it's not an easy task to find the information required for the users from the huge number of search results quickly. A personalized meta-search engine model based on user interest is presented, which gives users the most interesting information at the top as far as possible by extracting the behavior of individual users, establishing user interest database, sorting and integrating search results reasonably, to achieve a personalized meta-search and to improve user search precision and efficiency.

[Key words] Meta search result conformity user interest relevance