基于用户查询意图识别的Web搜索优化模型

杨 艺 周 元
(重庆工商大学计算机科学与信息工程学院 重庆 400067)

摘 要 在对用户查询意图进行分析分类的基础上,提出了一种Web搜索优化模型。该模型通过识别用户查询意图来查询意图特征词和内容主题词的双重约束,再结合用户查询行为获得查询目标,既保证了用户查询意图的准确匹配,又自动过滤和屏蔽了不相关信息。与相关工作对比,其重点在于准确获取用户查询意图,提高用户满意度。实验结果表明,该模型在实现信息搜索准确性和用户对查询结果满意度方面比传统搜索方法有明显改善。

关键词 信息查询,意图识别,查询行为,优化模型
中图法分类号 TP391 **文献标识码** A

Web Retrieval Optimization Model Based on User's Query Intention Identification

YANG Yi ZHOU Yuan
(College of Computer Science and Information Engineering, Chongqing Technology & Business University, Chongqing 400067, China)

Abstract A Web retrieval optimization model was proposed based on the analysis and classification of user's query intention. It focuses on user's query intention identification, and can both ensure match user's query intention accurately and filter useless information automatically by means of characteristic word of query intention, content keyword and user's query behavior. Comparing with related work. This paper focused on user's query intention and user's satisfaction. Experimental results show that the model can improve significantly the accuracy in information retrieve and users' satisfaction compared with traditional methods.

Keywords Information retrieval, Intention identification, Query behaviors, Optimization model

大容量的数据仓库提供了海量的数据供互联网用户在不同的时间和地点访问。但是,无所不在的数据不等于信息的即时可用性。一般情况下,数据可以看成是封装的信息内容的外部表示。对存储数据的利用,实质上是查询并访问其中包含的信息,但是这样的信息检索活动并不是一件容易的事,当我们试图进行自动信息搜索时,对于给定的信息查询,查询意图必须正确识别,同时对于要检索的信息,其内容应该尽可能与查询意图匹配。本文通过一种Web搜索优化模型,根据用户的查询意图和查询内容对数据源Web网页进行匹配,从而获取并返回与用户查询意图最接近的搜索结果。

1 用户查询意图

向搜索引擎提交查询请求的用户均有其潜在的查询意图,获取用户意图对高效的查询是至关重要的。Broder等在文献[1]采用分类研究的方法将用户查询意图分为3类:

(1) 导航型(navigational):寻找某类特殊站点,这类站点能够为用户提供该站点上进一步的导航操作;

(2) 信息型(informational):寻找Web站点上某种以静态形式存在的信息,这是用户通常的一种查询;

(3) 事务型(transactional):寻找某类特殊的站点,这类站点的信息能够直接被用户下载或做进一步的在线操作,如购物、玩游戏等。

在信息搜索领域,关于用户查询意图的研究一直没有中断,Rose在文献[2]中将上述3类用户意图进行了更为细致的分类,将信息型意图细分为Directed, Advice等5个子意图,将事务型意图又细分为Download, Entertainment等4个子意图;文献[3,4]将上述3种意图作为类标签,研究准确的分类算法;文献[5]通过分析用户在搜索结果中的鼠标移动轨迹来推断查询意图属于导航型还是信息型。另外,还有一些用户意图分类的研究不是以上述3种类型为基础的。总之,关于用户查询意图的研究主要包括两个方面:一是使搜索引擎提供更好的交互功能,显式或隐式地获取用户意图;二是对用户意图尽可能准确地分类。

大多数用户在进行信息查询时,并不能十分准确地用查询语句表达自己的查询意图,因此,如何获取用户意图并最终满足该意图至关重要。基于此,根据查询意图分类对输入查询条件进行意图识别,大致确定用户意图类别。查询意图识别功能由一系列识别用户查询意图的句子和短语的规则表达式和查询词库组成。查询意图词库是用来存储与查询意图有关的特征词和特征词的组合库(例如:哪些,寻找,下载,是什么,如何等)。这些特征词是配合规则表达式来对查询短语或句子进行意图分类识别。根据文献[6,7]中对查询意图的6

到稿日期:2011-06-05 返修日期:2011-09-02 本文受重庆市教育委员会科学技术研究项目(KJ090728),重庆市科学技术委员会自然科学基金计划项目(cstc2011jjA90008)资助。

杨 艺(1971-),女,硕士,副教授,主要研究方向为信息处理及电子商务技术,E-mail:ycy88lj@sina.com。

种分类，再加入条件限制区分，总结出初步分类体系如下。

1.1 查询意图特征词和相关句子分类

类1：信息寻找意图。表示想寻找某些信息。所有出现与寻找动作有关的动词（寻找，查找，搜索，列出等），同时最后的中心词是属性或者属性值。例如我要找有关藏族人风俗习惯的相关资料、请列出有关鲁迅先生的所有著作。

类2：知识询问意图。表示想了解某些知识。例如世界最高的山是哪座山？亚洲都有哪些人种？导致感冒的原因都有哪些？烟台在哪里？最便宜的长虹彩电价格是多少？

类3：建议咨询意图。表示想获得某些建议。例如如何正确服用中药？怎样学习C语言？从重庆到桂林自驾游怎么走？

类4：资源下载意图。表示想下载各种资源。例如下载邓丽君的歌曲，风景图片下载，免费版杀毒软件下载。

类5：导航/URL意图。表示想获得某个网址。例如清华大学的网站，DELL公司网址。

类1－类3实际上是信息型查询意图的细分，都是根据查询关键字对静态信息进行查询，只是关注的角度不同；类4和类5分别属于事务型和导航型。具体的查询意图分类及对应的特征词如表1所列。

表1 查询意图分类及对应特征词

查询意图标识	查询意图类别	意图特征词
SI1	信息寻找	寻找，查找，搜索，列出……
SI2	知识询问	有哪些，在哪里，是哪个，是什么，是多少……
SI3	建议咨询	如何，怎样，怎么办……
SI4	资源下载	下载……
SI5	导航/URL	主页，网站，网址，地址，URL……

在对意图进行初步分类和分析对应特征词的基础上，再对查询意图识别进行建模。

1.2 查询意图识别模型

定义1 对用户查询 $Q(w_1,w_2,\cdots,w_j)$ 和意图集合 $I(i_1,i_2,\cdots,i_n)$，设 V_{Qi_k} 表示查询 Q 与意图 i_k 的吻合程度，V_{Qi_k} 越大，表明查询 $Q(w_1,w_2,\cdots,w_j)$ 越能反映意图 i_k。则可以为每一个查询 Q 定义一个意图吻合程度向量：

$$V_Q=\langle V_{Qi_1},V_{Qi_2},\cdots,V_{Qi_k}\rangle, i_k\in I \tag{1}$$

式中，V_Q 称为查询 Q 的意图模型，计算 V_Q 的过程称作对查询 Q 的意图建模。用户意图的判定最终建立在查询词与意图特征词的关系上，对查询意图的判定过程中，用 T_{i_k} 表示意图 i_k 的特征词集，在对查询 Q 进行分词处理和去重后，计算查询词与意图特征词的吻合程度，即对于一个查询 $Q(w_1,w_2,\cdots,w_j)$ 和特征词集 $T_{i_k}(t_1,t_2,\cdots,t_m)$，$k=1,2,\cdots,n$，有

$$V_{Qi_k}=\begin{Bmatrix}1,w_p\in T_{i_k}\\0,w_p\notin T_{i_k}\end{Bmatrix}, i_k\in I, p=1,2,\cdots,j \tag{2}$$

从而计算出该查询对各意图的吻合值 V_{Qi_k}。

由上述可知，在计算查询条件对意图的吻合值时，前提条件是在用户输入的查询语句中出现了显式意图特征词，其吻合值的精度受查询意图分类完备度和意图特征词集词汇准确度的影响。对于在查询语句中没有出现显式意图特征词的情况，比如用户输入的查询是“NBA赛事”，还需要进一步地收集用户查询意图数据来明确用户的真实查询意图。

1.3 查询意图数据的获取

系统针对用户第一次输入的含糊查询条件，通过理解用户搜索过程中的查询意图的策略，给予用户查询意图动态匹配的搜索建议，来引导用户重新使用有效的查询语句。既明确了用户查询意图，又能够准确找到所需信息内容。在实际应用中，当用户输入查询条件时，搜索框的下方会同步构造一个从数据库返回的“建议”列表，来实时显示这些匹配查询意图的可选值，用户可根据需要选择或继续手动输入完善查询条件。这样可以在不中断用户当前操作的前提下，通过建立与用户的交互捕获用户查询的真实意图。

2 Web搜索优化模型

2.1 设计思路

当用户向搜索引擎提交一个查询语句时，系统先识别其查询意图，获得意图特征词，然后结合查询内容关键词，返回匹配用户查询意图和查询内容的初检网页集合。这种处理有助于在返回结果的TOP-N页面中尽可能少地出现与查询意图无关或者弱相关的信息。接下来，根据用户对初检结果的查询行为特征（如浏览时间长短或者下载操作）来判断Web页面的相关性，得到相关Web页面集，进而获得目标网页的URL。具体方法是：将用户对初检结果的点击浏览时间和下载操作作为判断Web页面相关性的两个参数，规定一个点击浏览时间阈值。如果某个Web页面被用户点击浏览的时间超过了这个阈值或者存在下载操作，则认为该Web页面是用户感兴趣的，是与用户查询意图目标相关的，应该提取出来组成目标URL集。其优化模型如图1所示。

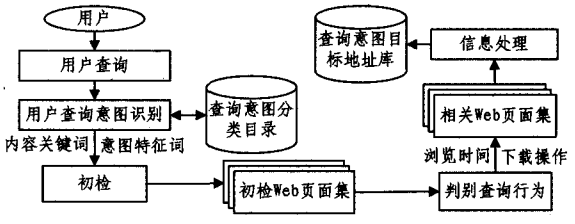

图1 Web搜索优化模型

2.2 相关定义

定义2 设 $Smodel(Web,query\text{-}item)=\langle Ikw,Ckw,P\rangle$，其中 $Smodel$ 为初检模型，Web为搜索引擎所有内容的集合，$query\text{-}item$ 为用户查询条件，Ikw 为查询意图特征词，Ckw 为查询内容关键词，P 为初检Web页面集。

定义3 设 $GModel(P,T,D,Ikw,Ckw)=\langle URL\rangle$，其中 $GModel$ 为二次检索模型，P 为初检Web页面集。T 为用户浏览页面时间，其中 T_i 为访问页面 P_i 的浏览时间，D 为用户对某个网页的下载（另存为）操作，其中 D_i 为用户对 P_i 的下载操作：如果 $D_i=0$，表示没有下载操作；而 $D_i=1$，表示有下载操作。设 θ 为浏览时间阈值，如果 $0<T_i\leqslant\theta$，则称 P_i 为经过页面；如果 $T_i>\theta$，则称 P_i 为兴趣页面；如果 $T_i>\theta$ 且 $D_i=1$，则称 P_i 为目标页面。URL为兴趣页面和目标页面的网页地址。

2.3 基于用户查询意图识别的搜索优化模型算法

Algorithm of Web Retrieval Optimization Model Based on User's Query Intention Identification(QII_AWROM)

Define:

Query-intentionURL //查询意图目标地址表结构

{

Urlid;//根据网页Url地址自动生成的唯一标识

T_i;//访问页面 Pi 的浏览时间
Ikw;//查询意图特征词
Ckw;//查询内容关键字
Url;//网页 Url 地址
};
Web,query-item,P,Ikw,Ckw;//见定义 2
T,D,θ;//见定义 3
Input:Web,query-item
Output:Query-intentionURL table sorted by users' browse time
Description:
(1) Do word processing for query-item;
(2) Identify users'query intention according to formula(2) in section 1.2;
(3) Go to search engine for obtaining P according to Ikw and Ckw obtained from step(2);
(4)For every p∈Top-N of P;
(5) if $T_i>\theta$ *or* $D_i=1$
(7) Put T_i,Ikw,Ckw,Url into Query-intentionURL;
(8) Urlid=+Urlid,i=+i;
(9) Return step(5)
(10) End if;
(11) End for;

3 实验及分析

基于用户查询意图识别的 Web 搜索优化模型,本文实现了一个原型系统,相关算法及原型系统的实现是在目前通用搜索引擎基础上进行的。原型系统的初始界面像通用搜索引擎那样要求用户提交查询语句,搜索框的下方同步构造一个下拉列表,来实时显示匹配查询意图的可选特征词。当初始查询提交后,原型系统返回初始查询结果,用户在此点击感兴趣的页面,系统自动将相关信息存入 Query-intentionURL 表中并按浏览时间长短排序。

3.1 实验设计

1. 测试网页集的选取

利用 Google 的高级搜索功能,将搜索区域限制为 http://sina.com.cn, http://163.com, http://cn.yahoo.com,分别对新闻、体育、娱乐、财经、科技 5 个栏目进行检索,并选取各项目的前 200 个网页;其次从网址之家 http://www.hao123.com 选取了 100 个主页地址,组成测试网页集。

2. 实验方法

让 5 个查询用户分别针对上述 5 类查询意图设计查询条件,在原型系统中分别用 QII_AWROM 和传统方法进行查询。表 2 列出了其中一个用户的查询条件及所属查询意图。

表 2 查询条件及所属查询意图类别

查询语句	查询意图类别	意图特征词	内容关键词
14 届世游赛奖牌排行情况	信息寻找	查询	14、世游赛、奖牌
Ipad 的视频驱动是什么	知识询问	是什么	Ipad、视频驱动
怎样用手机编写微博	建议咨询	怎样	手机、微博
《传奇》歌曲	资源下载	下载	《传奇》、歌曲
清华大学今年招生信息	导航/URL	网址	清华大学、今年、招生

表 2 中第 1、4 和第 5 个查询语句中没有显式的查询意图特征词,需要用到文中 1.3 节描述的交互操作来获得。经过查询意图识别和内容关键词提取等操作,得到的意图特征词和内容关键词如表 2 第 3、4 列所示。

3.2 实验结果分析

根据各用户给出的查询条件,实验将传统查询的结果与采用 QII_AWROM 的结果进行对比,θ 取值为 60s。评价针对初检返回的 TOP-20 个结果的准确率 $P@20$=(兴趣页面数+目标页面数)/20 以及用户对查询结果的满意度 S 做比较[8]。

5 个用户给出的 5 类查询条件在初检返回的 TOP-20 个结果中的查询准确率 $P@20$(平均值)如图 2 所示。其中,采用 QII_AWROM 获得的结果标识为 With model,用传统查询方法返回的结果标识为 Without model。从图 2 可知,采用 QII_AWROM 的搜索结果准确率高于传统方法,即在初检时采用查询意图特征词和内容关键词的双重约束,再结合用户查询行为能实现优化查询。

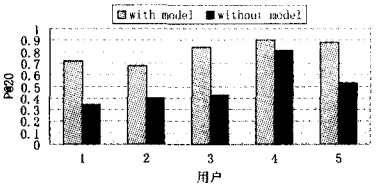

图 2 准确率 P@20 比较图

目前几乎没有搜索工具设计用户对查询结果进行评价的反馈机制,在此原型系统中设计了用户对该查询结果的满意度评价,分为 10 个等级,图 3 是用户对两种查询方法给出的评价。

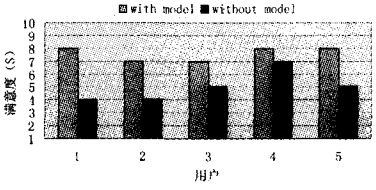

图 3 满意度(S)比较图

4 优化模型建模及实验方法讨论

本文提出的基于用户查询意图识别的 Web 搜索优化模型涉及查询意图、查询内容和用户查询行为,即从用户查询条件中识别出查询意图特征词并抽取内容关键词进行初检,再根据用户查询行为判定查询目标。实验结果表明,本文所提出的模型有效,但是也有一些问题需要提出来进行讨论。首先,提出的查询意图分类体系的完备性和类别间的独立不相关性依然值得考虑,意图特征词的筛选也需要进一步细化,用户的浏览时间阈值设置为 60s 是否合理也需要斟酌;其次,缺乏权威的大规模评测标准,且构造的测试集容量不够大,包含的查询个数通常在 10^2 数量级以上,这使得横向比较各种查询方法存在很大的困难,本实验是在初检结果 TOP-20 中来确定查询目标的准确率,其查全率也不能完全得到保证。另外,本文的满意度评价仅是靠用户单方面的打分确定的,到目前为止,还没有一套合理的定量的验证方法来说明在将查询意图正确分类的情况下,搜索引擎在用户满意度上究竟会如何提升以及有多大的提升[9,10]。

结束语 获取用户查询意图以及对用户查询意图分类是当前的热点研究问题。本文基于当前一种比较流行的查询意图分类方法(导航型、信息型、事务型),进一步细分出了 5 类查询意图及各类意图的特征词,从满足用户查询意图的观点

出发，提出了基于用户查询意图的 Web 搜索优化模型，实现了对搜索结果的优化。

为了更好地满足用户的查询意图，在较低时间代价下获得高质量的查询结果，进一步的研究工作包括：(1)探索新的用户查询意图，归纳出更准确的查询特征词，并基于此改进本文的优化模型；(2)进一步研究搜索内容关键词的抽取，并在抽取过程中引入领域知识；(3)探索和引入更多、更科学的查询结果评价方法[11]，对优化模型及查询结果进行更全面、客观的评价，并根据评价结果对模型进行改进；(4)对优化模型中设计的查询意图目标地址表的内容进行知识化处理以便查询复用。

参 考 文 献

[1] Broder A. A taxonomy of Web search[C]//SIGIR Forum. New York, N Y, USA: ACM Press, 2002: 3-10

[2] Rose D E, Levinson D. Understanding user goals in web search [C]//WWW '04: Proceedings of the 13the international conference on World Wide Web. New York, N Y, USA: ACM Press, 2004: 13-19

[3] Jansen B J, Booth D L, Spink A. Determining the user intent of Web search engine queries[C]// Williamson CL, Zurko ME, Patel-Schneider PF, et al., eds. Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2007: 1149-1150

[4] Ricardo A, Liliana C B, Cristina N. The intention behind Web queries[C]//Crestani F, Ferragina P, Sanderson M, eds. Proc. of the 13th Int'l Conf. on String Processing and Information Retrieval (SPIRE 2006). Berlin, Heidelberg: Springer-Verlag, 2006: 98-109

[5] Qi G, Eugene A. Exploring mouse movements for inferring query intent[C]//Myaeng SH, Oard DW, Sebastiani F, et al., eds. Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2008: 707-708

[6] 袁鼎荣，钟宁，张师超. 文本信息处理研究述评[J]. 计算机科学，2011，38(2)：9-13

[7] 吴晓辉，宋萍萍，张荣欣. 有无查询意图的分类与实现架构模型研究[J]. 情报科学，2009，27(12)：1829-1833

[8] 王大玲，于戈，鲍玉斌，等. 基于用户搜索意图的 Web 网页动态泛化[J]. 软件学报，2010，21(5)：1083-1097

[9] 罗长寿，康丽，刘国靖. 基于遗传算法的主题信息搜索系统研究[J]. 现代情报，2009，29(3)：176-181

[10] 凌波，周水庚，周傲英. P2P 信息检索系统的查询结果排序与合并策略[J]. 计算机学报，2007，30(3)：405-414

[11] Liu YQ, Fu Y P, Zhang M, et al. Automatic search engine performance evaluation with click-through data analysis[C]//Williamson CL, Zurko ME, Patel-Schneider PF, et al., eds. Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2007: 1133-1134

[12] 余肖生，司新霞. 基于聚类分析的元搜索引擎模型[J]. 重庆理工大学学报：自然科学版，2011，25(6)：69-72

(上接第 255 页)

CSO 更快达到最优的原因。

结束语 依据云的随风飘浮、降雨、形态的反复变化等自然现象构造了云搜索优化算法。为了保证算法收敛性，结合云团内部水滴信息的局部性，将其有序化为差商信息，提出了带差商信息的云搜索优化算法，并证明了其收敛性。benchmark 函数的数值实验展现了两算法优秀的寻优能力，特别是差商信息的加入大大提高了 DCSO 的收敛速度。

现有智能优化算法要么无法保证收敛到极值点，要么只能依概率收敛到全局最优点，收敛速度无法保证。而本文利用差商和梯度的关系证明了 DCSO 类似经典算法的收敛性，因而其与经典算法一样至少具有线性收敛速度。

参 考 文 献

[1] Li Yu-ying, Wen Qiao-yan, Li Li-xiang. Modified chaotic ant swarm to function optimization [J]. The Journal of China University of Posts and Telecommunications, 2009, 16(1): 58-63

[2] Holland J. Adaptation in Natural and Artificial Systems [M]. Ann Arbor, MI: Univ. of Michigan Press, 1975: 1-9

[3] Goldberg D E. Genetic Algorithms in Search, Optimization, and Machine Learning [M]. New York: Addison-Wesley, 1989

[4] Kennedy J, Eberhart R C. Particle swarm optimization [C]// Proceedings of the IEEE International Conference on Neural Networks. 1995, 4: 1942-1948

[5] 李太勇，吴江，朱波，等. 一种基于距离度量的自适应粒子群优化算法 [J]. 计算机科学，2010，37(10)：214-216

[6] Karaboga D, Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm [J]. Journal of Global Optimization, 2007, 39: 459-471

[7] Chen Ting-yu, Chi Tzu-ming. On the improvements of the particle swarm optimization algorithm [J]. Advances in Engineering Software, 2010, 41: 229-239

[8] Rudolph G. Convergence Analysis of Canonical Genetic Algorithms [J]. IEEE Transactions on Neural Networks, 1994, 5 (1): 96-101

[9] Zheng Yong-ling, Ma Long-hua, Zhang Li-yan, et al. On The Convergence Anlysis and Parameter Selection in Particle Swarm Optimization [C] // Proceedings of the Second International Conference on Machine Learning and Cybernetics. Xi'an, China, IEEE, 2003: 1082-1087

[10] Trelea I C. The Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection [J]. Information Processing Letters, 2003, 85: 317-325

[11] Jiang M, Luo Y P, Yang S Y. Stochastic Convergence Analysis and Parameter Selection of the Standard Particle Swarm Optimization Algorithm [J]. Information Processing Letters, 2007, 102: 8-16

[12] Zeng Jian-chao, Jie Jing, Cui Zhi-hua. Particle swarm optimization [M]. Beijing: Science Press, 2004

[13] 张光卫，何锐，刘禹，等. 基于云模型的进化算法[J]. 计算机学报，2008，31(7)：1082-1091

[14] 戴朝华，朱云芳，陈维荣，等. 云遗传算法及其应用[J]. 电子学报，2007，35(7)：1419-1424

[15] 张光卫，康建初，李鹤松，等. 基于云模型的全局最优化算法[J]. 北京航空航天大学学报，2007，33(4)：486-490

[16] 刘禹，李德毅，张光卫，等. 云模型雾化特性及在进化算法中的应用[J]. 电子学报，2009，37(8)：1651-1658

[17] 赵小平. 差商最速下降法及其收敛性 [J]. 华东化工学院学报，1992，18(6)：807-812

[18] 孙文瑜，徐成贤，朱德通. 最优化方法 [M]. 北京：高等教育出版社，2004：94-114