文章编号: 1673- 9469(2011) 01- 0106- 04

个性化搜索中隐私保护技术的探讨与研究

张 帆[1], 申艳光[2], 王 敏[2]
(1. 河北钢铁集团 邯宝公司, 河北 邯郸 056015; 2. 河北工程大学 信息与电气工程学院, 河北 邯郸 056038)

摘要: 从用户 个人信息的搜集、用户描述文件建立、搜索结果排序和系统评价四 个方面详细阐述了 个性化搜索 技术并比较分析了 不同技术的优缺点, 从防止未 经许可的数据访问和保护发 布数据的 安全两 个方面 介绍了隐私保护技术的研究状况, 并分析了目前 个性化搜索 中隐私保护技术的研究成果及其存 在的局限性, 最 后归纳出需要进 一步解决的问题。

关键词: 个性化搜索; 隐私保护; 用户描述文件; 搜索结果排序
中图分类号: TP309. 2 文献标识码: A

Discussion and research of privacy protection in personalized search

ZHANG Fan[1], SHEN Yan-guang[2], WANG Min[2]
(1. Hanbao Co., Ltd, Hebei Iron and Steel Group, Hebei Handan 056015, China; 2. School of Information and Electronic Engineering, Hebei University of Engineering, Hebei Handan 056038, China)

Abstract: For the threat of disclosure of user information and privacy caused by personalized search, privacy protection technology in personalized search becomes a recent research hotspot. First of all, the personalized search technology is expounded from four aspects including the collection of personal information, the establishment of user profile, search results sorting and system evaluation. And the advantages and disadvantages of different technologies are compared. Then, the privacy protection technology is discussed in detail from two aspects containing prevention of unauthorized data access and security of released data. In addition, the current research achievements and the limitations of privacy protection in personalized search are analyzed. At last, further problems which need to be solved are presented.

Key words: personalized search; privacy protection; rich user profile; search results sorting

随着互联网上信息数据爆炸式地增长, 互联网用户不再满足于人工分拣分类目录搜索的第一代搜索引擎和依靠超链接分析机器抓取技术的第二代搜索引擎提供的单一的搜索结果, 而希望得到和个人兴趣偏好更为相关的个性化搜索服务, 为了满足用户的这一需求, 应运而生了个性化搜索技术。

搜索引擎生存的关键是利用用户信息针对性地改善搜索服务质量, 提高用户搜索体验。为了向不同用户提供更为个性化、满意度更高的搜索结果, 个性化搜索时需要搜集、存储、挖掘和分析用户信息, 这不可避免地触及了个人隐私这一公众敏感神经。更为重要的是, 在搜索引擎所掌握的海量信息中, 除了个人隐私之外, 还有可能涉及到国家经济和政府机密信息, 威胁国家安全。随着个性化搜索技术的飞速发展, 隐私保护和搜索结果满意度之间急剧深化的矛盾已经成为了目前互联网技术研究亟待解决的问题。

1 个性化搜索技术

个性化搜索是以用户为中心的信息搜索技术, 它获取以多种形式表达的用户信息, 并综合利用这些用户信息, 提高搜索引擎的性能, 以满足不

收稿日期: 2010- 12- 02
基金项目: 国家自然科学基金(61075053); 河北省科技支撑计划项目(10220602D); 河北省教育厅科学研究计划项目(2009421)。
作者简介: 张帆(1980-), 女, 河北阜平人, 工程师, 从事计算机技术与应用。

同用户的个性化需求。目前绝大多数的研究主要集中在用户个人信息的搜集、用户描述文件建立、搜索结果排序和系统评价四个方面。

1.1 用户个人信息的搜集技术

用户个人信息的搜集技术主要包括显式、隐式和复合式三类方式。用户个人信息的显式搜集方式主要是请求用户的主动参与, 用户向搜索系统主动提供并描述其个性化需求的相关信息。用户个人信息的隐式搜集方式主要是搜集用户在操作过程中的行为。而复合式方式则结合了显式搜集和隐式搜集两种方式。

采用显式的用户个人信息搜集方式的系统有 SiteSeer 等[1]。用户个人信息显式搜集方式能使搜索系统获取准确的用户个人信息, 但需要用户花费多余的精力参与反馈, 降低了用户搜索体验。而 WebWatcher 等系统[2] 以及建立用户层级树的方法[3] 采用了隐式的用户个人信息搜集方式。用户个人信息隐式搜集方式避免用户在使用过程中被频繁要求做额外的操作, 但是存在搜集信息不准确, 不能准确反映用户意图的缺陷。考虑到显式和隐式这两种方法的优缺点, 不少系统采用对这两类用户信息搜集方式折中后的复合式信息搜集方法, 例如论文搜索系统 CiteSeer 等[4]。该复合式方法只要求用户在关键点上的主动参与, 为了保证最佳的用户体验, 在大多数时间上则采用隐式搜集的方式。

1.2 用户描述文件的结构

获取和组织用户个人信息形成用户描述文件, 该文件表达了用户的兴趣偏好, 在搜索过程中将准确的用户信息提供给搜索引擎, 返回给用户较好的搜索结果, 用户描述文件的结构分为树型和非树型两种。树型结构的用户描述文件都是基于目前网上最大的人工编制的分类检索系统—开放式分类目录搜索系统 ODP (Open Directory Project) 产生的, 它继承了 ODP 高度覆盖性和准确性以及消除二义性的优点, 例如: 基于 ODP 本体论概念的用户描述文件等[5], 但是 ODP 本身具有缺乏自由度和扩展性的缺陷。在非树型结构的用户描述文件方面, 有胖模式非结构化的用户描述文件等[6]。非树型结构的用户描述文件缺乏层次结构, 用户不能自治信息开放程度。

1.3 搜索结果排序方法

排序方法直接影响着个性化搜索结果。最初的研究根据网页本身的属性, 提出 PageRank 的概念[7], 对互联网上的页面进行评分, 在搜索时将得分较高的网页排在搜索结果列表的前面返回给用户。该方法没有利用任何用户信息, 因此无法提供更贴切用户自身、满意度更高的搜索结果。

随后展开的研究建立在通用的搜索结果基础上, 结合用户描述文件, 在客户端或服务器端进行搜索结果的重排序[2,6,8], 将贴切用户的个性化搜索结果排列在搜索结果列表中比较靠前的位置, 让用户更方便的找到自己满意的信息。在客户端重排序的方法受限于传输带宽, 致使排序准确度受限; 在服务器端重排序的方法可以得到较准确的搜索结果, 但加大了服务器的负载, 并且存在泄露用户隐私信息的威胁。

1.4 系统评价方法

目前对个性化搜索系统的评价一般都需要人工参与, 用户人工标注各个查询结果的正确性, 综合这些人工标注结果来评测个性化搜索系统的性能。主要常用三种方法: 准确率评价方法[9]、用户打分评测机制[10] 和 DCG 评测算法[11]。

准确率评价方法: 参与评测的用户标注每次查询返回的前 N 个结果的正确性, 系统利用每次查询前 N 个结果中标注为正确的结果所占比例作为评价指标来评价系统的性能, 评价指标的值越高则说明系统的性能越好。该方法的计算公式简单, 减少了参与评测用户的工作量, 容易实现, 但是在无指导的情况下用户标注时的随意性较大。

用户打分评测机制: 每个用户根据搜索结果与自己所需信息的符合程度对每次查询返回的前 N 个结果打分, 将所有用户对搜索结果打分的平均值作为系统性能的评价指标。该方法将用户对结果的评价划分为很多不同的等级, 给出将查询结果标注为某一等级的详细依据, 在一定程度上指导用户的评价行为, 更加规范用户的标准行为。

DCG 评测算法: 把 DCG (Discounted Cumulative Gain) 公式融入到对查询结果人工打分的方式中, 结合用户对搜索结果的打分和结果的排序位置, 将得出的计算值作为系统性能的评测指标。该方法结合用户的使用习惯, 对系统做出更加符合实际情况的整体评价。

综上,目前个性化搜索的研究重点在于如何提供更智能的搜索方式、个性化的搜索结果和高效的搜索能力,较少研究用户的隐私保护技术问题。

2 隐私保护技术

根据隐私泄露方式,目前针对隐私保护的研究主要集中在两个方面:用访问控制和加密手段防止未经许可的数据访问和用泛化数据的手段保护发布数据的安全。

2.1 防止未经许可的数据访问技术

主要有预防和检测两类手段。预防手段主要包括定义、执行和限制用户访问敏感信息和数据的访问控制技术和加密技术两种安全手段。检测手段主要包括审计和入侵检测两种数据安全技术。审计用于对数据访问、修改的事后审查。入侵检测对内、外攻击和误操作提供一种积极主动的实时保护,在系统受到危害之前拦截相应入侵,主要有基于基因算法的方法[12]等。

2.2 数据发布中隐私保护技术

目前最常用的隐私保护数据发布技术主要有k- 匿名[13]、l- 多样。k- 匿名使得每个元组的质量指标值都与其他 k- 1 个元组的质量指标值相同,从而保护了数据所有者的隐私。对 l- 多样性进行扩展,提出 t- 近似的概念[14],即每个匿名组中的敏感属性分布具有与整体的敏感属性近似的分布。

综上,虽然传统的数据安全技术和新兴的隐私保护数据发布技术能够对数据中的敏感信息和隐私信息起到较好的保护作用,但是,绝大部分传统的隐私保护技术无法直接应用于个性化搜索引擎中的隐私保护。

3 个性化搜索中隐私保护技术

目前针对个性化搜索中隐私保护技术研究的很少,主要提出建立用户信息层级树的方法,允许用户控制个人隐私的开放程度,但该层级树的建立缺乏一个统一的标准,不利于挖掘群体性的信息,同时也增加了用户的负担,其次分析了个性化搜索的隐私保护中可能遇到的种种问题,在此基础上提出了隐私保护的四个等级[8],但其等级的划分缺乏大量的理论依据,还需要进一步考证其准确性和实用性。

4 结束语

个性化搜索系统已被广为开发,但其研究的重点还处于如何提高搜索结果质量和用户体验的环节。虽然传统的数据安全技术与新兴的隐私保护数据挖掘和发布技术能够对敏感数据起到较好的保护作用,但目前绝大部分的隐私保护技术无法直接应用于个性化搜索中的隐私保护。针对个人信息搜集、传输、使用、存储和挖掘的保护方法研究方面还处于起步阶段和缺乏对个性化搜索中隐私保护整体框架研究的情况,将来有必要对此开展研究来解决个性化搜索中的隐私保护问题,推动个性化搜索引擎的健康发展。

参考文献:

[1] JAMES RUCKER, MARCOS J, POLANCO. Personalized navigation for the Web [J]. Communications of the ACM, 1997, 40(3): 73 - 76.

[2] DUNJA MLADENIC. Machine learning for better Web browsing [A]. Proc. of AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces[C]. American: Association for Artificial Inteuigence, 2000. 82- 84.

[3] YABO XU, BENYU ZHANG, ZHENG CHEN, et al. Privacy- enhancing personalized Web search [J]. In: Proc. Of WWW2007, May 8- 12, 2007: 591- 600.

[4] BOLLACKER KURT D, LAWRENCE STEVE, GILES C LEE. Discovering relevant scientific literature on the Web [J]. Intelligent Systems and Their Applications, 2000, 15 (2): 42- 47.

[5] AHU SIEG, BAMSHAD MOBASHER, ROBIN BURKE. Web search personalization with ontological user profiles [A]. Proc. of CIKM' 07, November 6- 8[C]. Lisboa, Portugal: Acm, 2007. 525- 534.

[6] JAIME TEEVAN, SUSAN T. DUMAIS, ERIC HORVITZ. Personalizing search via automated analysis of interests and activities [A]. Proc. of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 05), August 15- 19[C]. Salvador, Brazil: Acm, 2005. 449- 456.

[7] LARRY PAGE, SERGEY BRIN, MOTWANI R, et al. The Pagerank citation ranking: bringing order to the Web [R]. Technical Report, Stanford University, 1998.

[8] XUEHUA SHEN, BIN TAN, CHENGXIANG ZHAi. Privacy protection in personalized search [R]. SIGIR Forum, June

2007: 4- 17.

[9] FERRAGINA P, GULLI A. A personalized search engine based on Web snippet hierarchical clustering[A]. International World Wide Web Conference[C]. Chiba, Japan: Acm, 2005. 801- 810.

[10] PA CHIRITA, W NEJDL, R PAIU, C KOHLSCH TTER. Using ODP metadata to personalize search[A]. Proceedings of the 28th annual international ACM SIGIR[C]. Salvador, Brazil: Acm, 2005. 178- 185.

[11] J TEEVAN, ST DUMAIS, E HORVITZ. Personalizing search via automated analysis of interests and activities [A]. Proceedings of the 28th Annual International ACM SIGIR[C]. Salvador, Brazil: Acm, 2005. 449- 456.

[12] 苏璞睿, 李德全, 冯登国. 基于基因规划的主机异常入侵检测模型[J]. 软件学报, 2003, 14(6): 1120- 1126.

[13] TOCHUKWU IWUCHUKWU, JEFFREY F, NAUGHTON. K - Anonymization as spatial indexing: Toward scalable and incremental anonymization [A]. Proc. of VLDB 2007[C]. Vienna, Austria: Acm, 2007. 746- 757.

[14] NINGHUI LI, TIANCHENG LI, Suresh venkatasubramanian t- closeness: Privacy beyond K- anonymity and l- diversity [A]. Data Engineering[C]. Istanbul, 2007. 106- 115.

(责任编辑　刘存英)

(上接第 99 页)

参考文献:

[1] BARREL J. Criteria for the recognition of ancient delta deposits [J]. Geological Society of American Bulletin, 1912, 23: 377- 446.

[2] 李宏伟, 邓宏文, 肖乾华, 等. 热欧地区残留可容纳空间分布与储集砂体预测[J]. 石油学报, 2002, 23(4): 29 - 32.

[3] 樊太亮, 吕延仓, 丁明华, 等. 层序地层体制中的陆相地层发育规律[J]. 地学前缘, 2000, 7(4): 315- 321.

[4] 樊太亮, 吴贤顺. 从古地貌谈层序格架中储层的发育规律[J]. 地球学报, 2002, 23(3): 259- 262.

[5] MARTINSEN R S. Depositional remnants, part 1: Common components of the stratigraphic record with important implications for hydrocarbon exploration and production [J]. AAPG Bulletin, 2003(87): 1869- 1882.

(责任编辑　刘存英)