支持Internet上个性化信息重组与发布的Web挖掘关键技术的研究*)

王大玲　胡明涵　于　戈　鲍玉斌
(东北大学信息科学与工程学院　沈阳110004)

摘　要　Internet上个性化信息的重组与发布是Web个性化技术的一个重要组成部分,这一领域目前存在的主要问题是:并非没有信息重组和发布的工具,而是缺乏能够使这类工具高效工作的支持技术。本文提出一种将流数据处理技术引入Web点击流、IP地址流及页面文本流挖掘和分析过程,研究基于Web数据流挖掘的用户行为和需求分析方法;将本体和领域知识引入Web内容挖掘过程,研究领域知识指导下的Web内容挖掘方法;将基于Web数据流挖掘的用户行为和需求分析与领域知识指导下的Web内容挖掘相结合,研究Internet上Web信息模式和Web用户模型及其相互关系的建立;将上述研究成果应用于实际,以期达到高效地支持Internet上满足用户个性化要求的信息重组与发布的目的。

关键词　个性化信息重组与发布,流数据挖掘,Web内容挖掘,领域本体知识

Study on Key Techniques of Web Mining for Personalized Information Reorganization and Distribution

WANG Da-Ling　HU Ming-Han　YU Ge　BAO Yu-Bin
(School of Information Science and Engineering, Northeastern University, Shenyang 110004)

Abstract　The personalized information reorganization and distribution is an important part on Web personalization. Now the major problem of this domain is not lacking the tools for information reorganization and distribution, but requirinig the supporting techniques for them work effectively. This paper proposes an approach including: Stream data processing technique is imported in analyzing Web click stream, IP stream and text steam for studying users' behavior and requirement based on stream data mining. Ontology and domain knowledge is introduced in Web content mining for guiding the mining process. Users' behavior and requirement analysis based on Web stream data mininig is combined with Web content mininig under the guidance of domain knowledge for creating Web information pattern, user model and their relation. The study results are used for supporting effectively personalized information reorganization and distribution of Internet according to users' requirement.

Keywords　Personalized information reorganization and distribution, Stram data mining, Web content mining, Domain and ontology knowledge

1　引言

目前,迅速增长的网上信息为我们的生活和工作提供了极大的方便,同时也使上网的用户越来越难以依靠自身的力量在Internet信息中获得自己所需要的内容,于是各种推荐系统、搜索引擎纷纷涌现。相对于Internet这样一个大的信息集合,推荐系统、搜索引擎提供的当然只是一个小小的"子集",但相对于用户需求的这个小的信息集合,这些系统和引擎提供的却是一个大得多的"超集",甚至这个"超集"还不完全包括用户需要的信息"子集"。用户面对这样一个大"超集"同样会因找不到自己所需要的内容而无所适从。就推荐和导航而言,一方面,网站提供的推荐和导航信息更多地是从网站的利益、而非用户的利益出发,因而用户难以从中得到自己所需要的信息;另一方面,即便网站考虑用户的需求,但不同的用户,其需求也是千差万别的,而许多网站提供的推荐则是千篇一律的形式和内容。就信息搜索而言,目前的很多网站提供的搜索界面虽然形式不尽相同,但其实质仍然是基于关键字、或在此关键字基础上扩充的新关键字的匹配[1],而这种匹配方法很难真正体现用户的需求。就结果的发布而言,用户无论是通过查询、还是浏览,所要求的可能仅仅是某些网页、甚至是网页中某个片断的信息,得到的却是大量相关、甚至不相关的页面和超链接,用户需要在这众多的页面和超链接中不断进行手工的或自动的"二次搜索",才能获得也许并非满意的结果。

上述一切问题,为网站的设计者提出了这样的要求:根据用户的个性化需求,为其重新组织相关的信息内容,并按其感兴趣的个性化形式予以发布或推荐,这就是Internet上个性化信息的重组与发布问题。目前的各种搜索引擎和推荐系统从广义上讲都是在进行信息的重组与发布工作,但其功能显然不能为用户所满意,因而问题的实质是:现在并非没有信息重组和发布的工具,而是缺乏能够使这类工具高效工作的支持技术。我们所述的"高效"并非搜索引擎和推荐系统本身的执行效率,而是其信息发布内容的高质量与形式的多样化,继而引发的是其支持技术的性能与效用,这涉及用户个性化需求的获取、针对该需求的个性化信息的搜索与推荐、针对该搜索或推荐结果的个性化信息展示、以及与之相关的理论与应用技术的研究。

2　相关技术

Web个性化推荐是上世纪末被提出的,目前主要采用两

*)国家自然科学基金资助项目(No. 60173051)。王大玲　博士,教授,主要从事Web挖掘与个性化技术的研究。胡明涵　讲师,在职博士生,主要从事自然语言处理技术的研究。于　戈　博士,教授,博士生导师,主要从事数据理论与技术的研究。鲍玉斌　博士,副教授,主要从事数据仓库技术的研究。

种技术：基于内容的过滤(Content-Based Filtering)和协作过滤(Collaborative Filtering)[2]。前者是利用信息资源与用户兴趣的相似性来过滤信息，后者是利用用户之间兴趣的相似性来过滤信息。由于基于内容的过滤和协作过滤技术各自的特点，目前，已有一些推荐系统探索将二者结合的推荐方式。

无论采用哪种技术，均需要尽可能准确地掌握用户的访问行为和浏览兴趣，即获得用户信息。目前，获得用户信息可采用显式收集(Explicit Collection)和隐式收集(Implicit Collection)两种收集方式[3]。前者是一种直接的信息收集方式，需要用户的反馈信息，后者是一种间接的信息收集方式，它不是从用户的反馈信息直接获得，而是从用户的上网数据中分析得到的。

Web 挖掘技术是实现隐式数据收集的一种有效的手段，更是支持基于内容过滤和协作过滤技术的有力工具。通过 Web 挖掘，可以进行页面内容的相似性分析，支持基于内容的过滤技术；可以进行用户访问模式的相似性分析，支持协作过滤技术。此外，通过 Web 挖掘的具体方法，可以获得应用于推荐的相关信息：页面关联分析可以获得用户一同浏览的页面的相关情况，页面访问量分析可以获得用户对网站中各网页的访问情况，用户访问模式聚类可以获得用户按访问兴趣的分组情况。总之，通过 Web 挖掘，我们能够在准确获得用户信息的基础上，为用户提供个性化推荐服务。

目前，将 Web 挖掘技术的研究应用于 Web 个性化服务中，已出现许多 Web 挖掘的原型系统和实际应用系统。其中一些典型的系统或产品包括：美国 Minnesota 大学和 Depaul 大学开发的 WebSIFT(推荐系统)[4]，德国柏林 Humboldt 大学研制的 WUM(序列挖掘系统)[5]，IBM T. J. Watson 研究中心开发的 IRA(电子商务推荐分析系统)[6]，美国 Net Perceotion 公司开发的 Net percerptions(浏览建议系统)[7]，香港科技大学开发的 FACT(查询处理系统)[8]等。这些系统和产品均不同程度地采用了 Web 挖掘和分析技术作为支持工具，因而在查询和搜索、推荐与发布等处理过程中取得了一定的成果。特别值得一提的是：著名的搜索引擎 Google 采用一些分布的自动搜索软件(Crawler)在网上搜索得到相关的 URL 列表后，将所对应的网页通过 Web 内容挖掘获得扩充的关键词，通过 Web 结构挖掘获得超链接信息，再利用这些扩充的关键词和超链接信息进行新的搜索。Google 正是利用了 Web 结构和内容挖掘技术，使其搜索技术在各种搜索引擎中更胜一筹[9]。

3 问题研究

目前的推荐系统大多是针对性较强的具体系统，而且主要根据用户的访问行为进行分析，其中用户模型的建立或者采用传统的统计学方法，或者采用的 Web 挖掘技术并没有真正考虑语义的内涵或没能及时反映用户信息需求的变化，因而其准确性受到影响。而目前的查询搜索引擎无论其查询的内容还是返回的形式均不能为用户所满意更是不争的事实。对此我们认为，除了应用技术本身的因素外，缺乏强有力的理论支持是一个主要原因，具体包括：

(1)信息收集过程不能很好地获取用户的访问需求。目前的显式收集方式和用户查询条件输入形式难以为用户提供表现其真正浏览兴趣的手段和方法，隐式收集方式由于缺少对用户要求的语义内涵的分析，因而不能从深层次获取准确的用户访问信息；

(2)用户行为分析不准确，更难以及时反映用户兴趣的变化。虽然 Web 挖掘技术从总体上为用户需求分析提供了理论支持，但由于缺少优秀的、能够很好地应用于用户行为分析的挖掘算法，导致 Web 挖掘技术在用户个性化信息的获取和分析中没有充分发挥其作用。特别是，有些用户的兴趣是长期而稳定的，有些用户的兴趣是随着时间和形势的发展而变化的，而目前所采用的分组数据挖掘方法、静态数据挖掘方法或者不能及时反映用户兴趣的变化，或者将对用户行为分析的前后结果割裂开来；

(3)推荐和搜索的结果离用户要求相去甚远。基于用户访问行为的推荐技术的研究虽然取得了一定的成果，但由于分组、静态数据挖掘算法本身的缺陷，使其推荐的准确性仍有待于改进。而基于内容的推荐及信息搜索目前所采用的关键字及其扩充的匹配方式主要来源于词频度的统计，缺少用户需求和推荐信息之间的语义及相关领域知识的分析，因而在很多情况下反而增大了搜索结果的冗余，也无法实现真正高质量的内容推荐。

综上所述，结合目前流数据挖掘、语义分析及自然语言处理技术方面的一些研究动态，我们提出如下解决问题的方法：

(1) 探讨基于流数据挖掘技术的 Web 数据流的预处理方法、数据结构、挖掘算法和结果分析，同时将语义提取技术纳入对用户需求的分析中，以解决“用户行为分析不准确、难以及时反映用户兴趣变化”的问题；

(2)探讨将本体领域知识、语义分析技术应用于 Web 内容挖掘的算法和数据结构，研究本体知识、领域知识对应于 Web 内容挖掘的存储和获取技术，以解决“信息收集过程不能很好地获取用户的访问需求”的问题；

(3)研究语义与 Web 挖掘的关系，进而研究将上述 Web 数据流挖掘与 Web 内容挖掘相结合的分析方法，建立 Internet 上 Web 信息模式和访问用户模型及其相互关系，深入研究相关模型在 Web 个性化信息搜索、重组、发布和推荐技术中的应用及实现机制，以解决“推荐和搜索的结果离用户要求相去甚远”的问题。

4 系统模型与实现技术

4.1 系统模型

我们设计的支持 Internet 上个性化信息重组与发布的 Web 挖掘系统模型如图1所示。

系统中所采用的技术主要包括基于 Web 流数据挖掘的 Web 使用挖掘技术、领域和本体知识指导下的 Web 内容挖掘技术、以及在此基础上的个性化信息重组与发布技术。我们采用4.2节所述的方法予以实现。

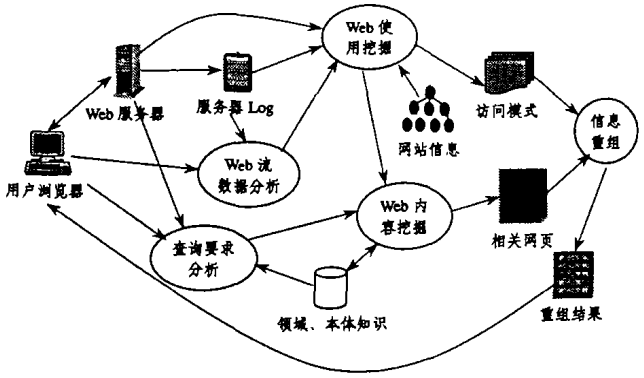

图1 支持 Internet 上个性化信息重组与发布的 Web 挖掘系统模型

4.2 实现技术

(1)流数据(Stream Data)是一种大量连续增长的、不断

变化的、有序的点组成的流，如传感器数据、电话记录等[10]，这种数据形式不同于传统的数据库数据，它需要特殊的方法来进行迅速的处理。Web 点击流、IP 地址流数据和页面文本流虽然具有流数据的性质，但对其处理结果的要求有别于传感器数据，前者对结果准确性要求甚于对响应速度的要求，后者则相反，前者被称为“离线流(Offline Stream)”，而后者被称为“在线流(Online Stream)”[11]。因此，我们重点研究离线流的性质及其相关的数据、规则的存储结构。具体地，分析现有流数据处理中相关理论和研究成果，分析 Web 数据流结构的特点，设计合适的存储结构。在此基础上，从流数据挖掘的思想出发，设计基于此结构的高效的 Web 使用挖掘算法；

(2)目前许多学者在自然语言处理及语义分析技术的研究中已取得了相当可观的成果，并在进一步开展该领域的研究。本体(Ontology)是描述概念及概念之间关系的概念模型，通过概念之间的关系来描述概念的语义[12]，领域知识(Domain Knowledge)是一个专门领域重要的问题或概念以及这些问题和概念之间的相互关系，本体领域知识构成了上述自然语言处理基础之上针对具体的、特定的领域所建立的语言要素的相关描述，目前这方面的研究、特别是针对 Web 内容挖掘方面的研究还比较薄弱。我们分析该领域现有的研究成果，并密切关注该领域新的发展趋势，以此作为我们开展 Web 内容挖掘的基础；

(3)以现有语言处理技术和本体领域知识表示技术及其可能的发展趋势为基础，进行本体和领域知识在 Web 挖掘中的表达、Web 挖掘与语义 Web 结合、领域知识在 Web 挖掘特别是内容挖掘中的应用研究，设计相关的数据结构和挖掘算法以及领域知识的获取机制及其在 Web 挖掘结果解释中的应用机制，研究针对本体领域知识和语义层次的数据结构和处理机制，补充、改进相关的本体领域知识，寻求语义 Web 与 Web 挖掘的有机结合；

(4)在挖掘算法设计方面，分析现有 Web 挖掘和流数据挖掘算法的优势和不足，从离线流性质、Web 数据流、Web 内容特别是 Web 文本数据的特点出发，结合统计学理论、机器学习理论和方法，从更广义的角度设计 Web 挖掘算法，并在应用领域知识进行挖掘结果优化方面进行深入研究；

(5) 分析现有的推荐和信息搜索方法，考虑基于流数据处理的 Web 使用挖掘结果和在领域知识及语义 Web 指导下的 Web 内容挖掘结果更加有机的结合，设计并实现支持个性化信息重组及发布机制的高效算法；

(6)目前的 Web 挖掘方法中将用户、网页、文本等不同类型对象分别进行挖掘和建立模型。而实际上，用户访问网页，网页由若干文本组成，文本又是用户访问网页的实体所在，显然，它们之间有着内在的必然联系，这种联系也必将导致这三种不同对象的模型之间的相互影响。因此，我们在设计挖掘算法时考虑揭示它们之间的关系及相互影响，在用户、推荐页面或信息搜索结果与网站相关内容之间建立起相互关联，进一步提高信息发布的质量。

综合上述研究内容，以下是我们解决的关键问题：

(1)为及时准确地获取 Web 数据流并进行处理、保证足够的挖掘结果准确性和尽可能快的响应速度，相关的数据结构及基于此结构的数据预处理和挖掘算法的研究是我们解决的关键问题之一；

(2)目前文本挖掘、领域本体知识表示以及自然语言处理技术已经取得了一定的成果，应用它们并通过 Web 挖掘改进它们，在现有文本挖掘和语义 Web 的基础上达到新的突破，是我们解决的关键问题之二；

(3)在 Web 信息模式和用户模型建立的基础上，将它们有机地结合起来，寻求对个性化信息重组和发布技术有利的支持，是我们解决的关键问题之三。

结语 Web 个性化技术本身已经不是一个很新的研究课题，但由于其所涉及的新的内涵、广泛的应用前景以及目前在相关理论支持和应用研究中存在的问题，使其相关支持理论和应用技术的研究仍是一个热门话题，并具有更加深入而广阔的拓展空间。

纵观本文的系统模型和研究内容，我们认为其创新之处主要包括如下几方面：

(1) 基于流数据处理技术的 Web 数据流挖掘及用户访问行为分析方法，包括相关的数据结构及基于此类结构的挖掘算法；

(2)领域本体知识指导下的 Web 内容挖掘方法，包括相关的领域知识存储结构和获取机制、Web 内容挖掘算法、以及在此基础上建立的语义 Web 结构；

(3)从寻求用户、网页、内容之间关系的目的出发所进行的模式、模型及其相互关系的建立机制；

(4) 基于流数据处理技术的 Web 使用挖掘结果与领域知识指导下的 Web 内容挖掘结果的有机结合，以及在 Internet 上个性化信息重组与发布中的支持作用的研究。

参 考 文 献

1 Agichtein E, Ipeirotis P, Gravano L. Modeling Query-Based Access to Text Databases. In: WebDB2003, 2003. 87～92

2 McDonald D W. Ubiquitous Recommendation Systems. IEEE Computer, 2003, 36(10): 111～112

3 Aggarwal C, Yu P. Data Mining Techniques for Personalization [J]. In: IEEE Data Engineering Bulletin, 2000, 23(1): 4～9

4 Mobasher B, Dai H, et al. Effective Personalization Based on Association Rule Discovery from Web Usage Data. In: Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), Atlanta, Georgia, USA, Nov. 2001

5 Humboldt University Berlin. WUM: A Web Utilization Miner. http://wum.wiwi.hu-berlin.de/. 2001

6 Yu P. Data Mining and Personalization Technologies [J]. In: Database System for Advanced Applications (DASFAA99), Taiwan, 1999. 6～13

7 海脉经济资讯. 几个 Web 数据挖掘方向的公司介绍[J]. 网络经济周刊. http://www.hermes.orm.on, 2001(62)

8 Diao Y, Lu H, Chen S, et al. Toward Learning Based Web Query Processing. In: Proc. of 26th VLDB2000. Cairo, Egypt, 2000. 317～328

9 Google. http://www.google.com

10 Charikar M, Chen K, Farach-Colton M. Finding frequent items in data streams. Theor. Comput. Sci., 2004, 312(1): 3～15

11 Manku G, Motwani R. Approximate Frequency Counts over Data Streams. In: Proc. of 28th VLDB2002, Hong Kong, 2002. 246～357

12 Chen X, Zhou X, Schert R, et al. Using an Interest Ontology for Improved Support in Rule Mining. In: DaWak 2003, LNCS 2737, 2003. 320～329