

基于 RSS 的个性化网络信息检索

闫晓妍

(河南城建学院图书馆, 河南 平顶山 467036)

摘 要: 个性化信息检索是面向用户检索的发展方向。RSS 因其高度准确性、动态聚合机制和服务的主动式、个性化等特点, 成为新型网络信息检索工具。将 RSS 技术引入信息检索中, 提出了基于 RSS 的个性化信息检索方式, 阐述了 RSS 个性化检索的实现机制、检索方式及策略, 分析了其优点, 并比较了与传统个性化检索方式的异同。

关键词: RSS; 个性化信息检索; 信息推送

中图分类号: G354

文献标识码: B

文章编号: 1002-1248 (2012) 01-0051-05

Personal Network Retrieval Based on RSS

YAN Xiao-yan

(Library, Henan University of Urban Construction, Pingdingshan 467036, China)

Abstract: The personal information retrieval is the development direction of facing user retrieval. Nowadays RSS becomes the most important new internet information resources searching tool because of its high precision, dynamic polymerization system and personalized and positive service features. This article applies RSS technology into information research, and puts forward the personalized searching based on RSS. It analyzes its principle, method as well as retrieval strategies, and makes a comparative analysis on personality retrieval based on RSS and the traditional personal search.

Keywords: Really Simply Syndication; Personal Retrieval; information push

个性化信息检索从本质上讲, 就是根据用户的兴趣、特点、个性化检索要求进行检索, 并得到高相关度的检索结果。其首要特点就是将“个性化”这一理念贯穿于整个检索过程, 摒弃了传统检索中检索结果千人一面的做法。1.0 时代, 个性化信息检索的实现方式一般基于如下两种技术: (1) 基于内容的个性化信息检索。根据用户设置的对内容的要求, 或者是用户曾经检索过的内容实现个性化。目前, 几乎所有的搜索引擎都支持用户设置检索要求, 但支持程度各不相同。(2) 基于协同过滤的个性化信息检索。根据用户的职业、兴趣爱好等找到与此用户有相似兴趣的其他用户, 然后将他们感兴趣的内容推荐给此用户。例如 Eurekster 利用特定网络社

区中用户使用检索结果的情况分析出用户更为关心的关键词和网站, 提供更富有个性化的检索结果^[1]。

1 Web2.0 环境下的个性化信息检索

Web2.0 产生了大量微内容, 网络信息检索不再是面对一个个的静态页面, 而是要链接到一个个不断更新的动态页面, Wiki、Blog 中聚集了许多有价值的思想、观点、知识和链接信息。在线新闻和博客是主要的频繁更新数据的数据源, 传统搜索引擎每日出动网络爬虫抓取网页信息的机制已经跟不上这些实时动态数据的更新速度。即使是 Google 也只能收录庞大网络信息中的一小部分, 对于时效性很强的信息更是显得力不从心。而最有效的方式就是

收稿日期: 2011-09-13

基金项目: 河南省高等学校青年骨干教师项目“Web2.0 环境下信息检索机制及其优化研究”研究成果之一。

作者简介: 闫晓妍 (1985-), 女, 情报学硕士, 毕业于郑州大学信息管理系, 数据库工程师, 现任职于河南城建学院图书馆。

使用 RSS 技术来解决信息聚合的问题。在 2.0 时代,出现了新的个性化信息检索方式,即基于 RSS 信息聚合技术实现的个性化信息检索。用户可以设置自己的 RSS 阅读器,只收集某些网站的某些类型的资料,按特定的时间间隔从提供信息的网站上更新资料,这样可以实现完全的个性化,将和用户不相关的信息真正的拒之门外。

2 基于 RSS 个性化信息检索的关键技术

RSS 是一种“轻量级、多用途、可扩展的元数据描述及联合推广格式”,用来描述 web 页面的内容摘要和访问地址。从本质上讲,RSS 不是内容,而是获取内容的一种渠道。RSS 技术主要包括两个重要的组件:RSS feed 和 RSS 阅读器。RSS 将网站看作一系列频道(Channels)的组合,各个频道又包含了一系列资源(items),因此通过对频道及所含资源的描述可实现对作为资源集合的网站的描述。

2.1 RSS feed

RSS feed 通常由 4 个主要元素构成: <channel>、<image>、<item> 和 <text input>, 其中, <channel> 元素是必需的, <item> 元素至少要出现一次。<text input> 和 <image> 元素是可选的,是否使用要视具体情况而定。<channel> 可以看作是 RSS 提要的标题信息,包括 <title> (标题)、<link> (链接,与该频道关联的 Web 站点的 URL)、<description> (描述,对频道的简要介绍)等子元素。许多频道子元素通常都是具有可选性的,如 <category> 用于描述元素的类别; <docs> 指向 RSS 提要文档的 URL; <lastBuildDate> 描述上次修改 channel 内容的时间等等。

项 <item> 通常是提要中最重要的部分,每个项都可以是关于某个 Blog、完整文档、电影评论、分类广告或者任何希望与频道连锁的内容的记录。频道中的其他元素可能不变,但项经常发生变化,可以有任意多个项,以前的规范限制为 15 个项,如果要保持向后兼容,仍须遵守这个上限。它包括一些常用的 item 元素:如 <title>、<link>、<description>、<author>、<category>、<pubDate>、<source> 等,分别描述了 item 的标题、URL、摘要、作者电子邮件、分类、发布时间、频道来源。

这些 <item> 可由网站管理员手工生成或由计算

机数据库系统自动生成,这一文档可以被其它站点收集并显示在其索引页面上,并且自动同步更新,以获取最新的信息。信息发布者将自己所生成的一个 RSS feed 发布到自己的网站或者将其提交给搜索引擎,每一个 RSS feed 都是 RSS 众多信息源的一部分。

2.2 RSS 阅读器

RSS 阅读器也叫信息聚合工具、RSS 客户端软件,是一个提供 RSS 数据的定制、聚合、检索、管理和显示等功能的客户端软件。目前,RSS 阅读器基本可以分为两类:一类是运行在计算机桌面上的应用程序,通过所订阅网站的新闻供应,可自动、定时地更新新闻标题。常见的有 RSS Reader、周博通、看天下、新浪点点通、以及网络蚂蚁等,开源社区也推出了很多优秀的阅读器。另一类,RSS 阅读器内嵌于已在计算机中运行的应用程序中,使用 IE 浏览器阅读 Blog 与新闻网站的 RSS 频道,或者不安装独立的 RSS 阅读软件而是为各电子邮件客户端和浏览器安装插件来增加 RSS 阅读功能。例如,NewsGator 内嵌在微软的 Outlook 中,所订阅的新闻标题位于 Outlook 的收件箱文件夹中,Pluck 内嵌在 Internet Explorer 浏览器中。

支持利用 RSS 服务的网站,通常会在其醒目的位置标注 **RSS** 或 **XML** 标志,用户可以从网站下载已经创建的 RSS feed 文件,除了常见的新闻网站之外,许多科技网站也采用这种方式发布最新消息。使用 RSS 阅读器,用户只需要点击希望阅读的 RSS feed 旁边的橙色小图标 **RSS** 或 **XML**,把浏览器窗口的 URL 拷贝粘贴到 RSS 阅读器,在 RSS 阅读器中加入该 RSS feed 所提供的信息频道,定制个性化信息内容。

3 基于 RSS 个性化信息检索的实现

3.1 实现机理

基于 RSS 技术的个性化搜索引擎充分支持用户相关性反馈机制,根据用户的兴趣爱好思维方式量身定做,主动地定期为用户查找信息。基于 RSS 的个性化信息检索通过任意关键词把相同兴趣的人集中在一起,一方面可通过强大的聚合功能为用户提供大量的相关信息资源,另一方面通过推送技术主动地为用户提供信息服务,充分利用网络资源为用

户创造价值。

基于 RSS 技术的个性化检索服务的“个性化”并非“个人化”，而是按照物以类聚、人以群分的原理，考虑用户所属的特征人群，参考与用户有相同爱好、兴趣及同行的选择，在支持普通搜索的同时，会提示与用户搜索内容相关关键词并且用户可以直接从搜索结果中订阅相关内容的 RSS^[3]。基于 RSS 的个性化信息检索的实现机理如图 1 所示。

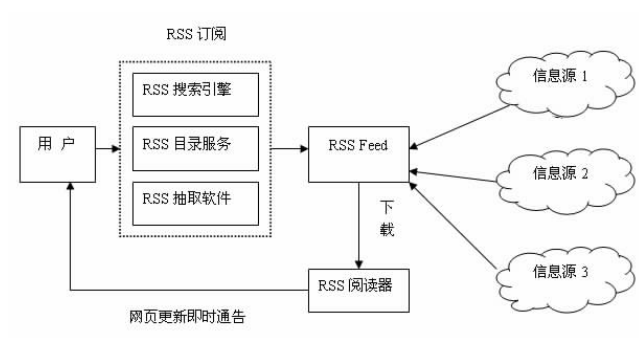


图 1 基于 RSS 个性化检索的实现机理

(1) RSS 订阅。RSS 订阅是基于 RSS 的个性化信息检索的重要步骤。用户可以通过 RSS 搜索引擎搜索网站中包含的 RSS Feed 文件；也可以利用 RSS 目录服务，一些网站将网页信息按照用户当前关注的热点或者大致主题分类（频道）进行信息聚合，并以 RSS 目录的形式呈现给用户；在浏览网页中发现感兴趣的信息源，即使没有提供 RSS 输出，也可以通过自动生成 RSS 的软件来定制。

(2) 网页更新即时通告。用户完成 RSS 订阅后，RSS 阅读器根据 RSS Feed 提供的地址定时获取相应的频道信息，一般是一小时或者用户定制的时间（RSS 信息检索方式要求网络带宽，太频繁会造成网络负担和资源浪费）。在 RSS 阅读器中可以浏览订阅信息的标题、摘要、链接等，定制信息源一旦有更新，就会向阅读器发出警报，从而实现一种主动的信息推送服务。

从信息检索的角度看，基于 RSS 的个性化信息检索主动地向用户推送信息，检索结果的反馈呈现实时性、连续性、动态性等特点。而传统网络信息检索是一种静态的、一次性的信息检索，一次检索返回给用户相应的一个结果，检索完成后用户想要获取相关的最新信息，需要重新检索。

3.2 检索方式

RSS 信息检索方式主要有以下几种：通过 RSS

搜索引擎搜索；利用 RSS 目录服务，通过分类来查找 RSS Feed；浏览过程中发现感兴趣的信息源，随时收录到自己的 RSS 阅读器中。

(1) 主题词 / 关键词订阅——RSS 搜索引擎。主题词 / 关键词订阅主要通过 RSS 搜索引擎，按照关键词、分类导航或标签服务，将众多的 RSS 信息资源进行集成，为用户提供主题词和关键词的查询方式。RSS 搜索引擎与其他搜索引擎原理相同，主要由用户界面、RSS feed 抓取器、本地索引数据库、查询预处理模块及控制模块组成。RSS feed 抓取器把一个博客网站的 URL 作为初始的种子。按照预先设置好的采集深度，对该 URL 所有的链接格式（“index.html”，“rss.html”，“rss_2.html”等等）分析搜索，来找出页面里的 RSS 种子，并把它们下载下来。通过控制模块控制 RSS 抓取模块的抓取过程，主要是对采集深度和采集的频度进行控制，尽可能在一定的硬件环境下，能够抓取更多更及时的 RSS feed。

国内专业的 RSS 搜索引擎不多，如 Feedss（<http://www.feedss.com>）、Feedsearch（<http://www.feedsearch.net>）和 Grassland（<http://www.cnblog.org>）等。中文 RSS 搜索引擎的不足之处是功能较简单且多数不支持 Tag 搜索。

(2) 网站信息分类检索——RSS 目录服务。在传统互联网搜索中，门户网站提供一种在线目录服务，也就是人为将互联网中的站点信息分门别类的放在一起，然后用户根据自己需要在这些分类中查找，如雅虎提供的分类目录。Web2.0 环境下，同样有许多网站提供 RSS 目录服务，如：Syndic8、News Is Free、The Feed Directory、看天下 RSS，有些不仅提供 RSS 目录服务，还提供 RSS 搜索引擎服务、在线 RSS 聚合服务。在自己提供的 RSS 订阅服务专题中，一般也提供自己站点的 RSS feed 目录，如：纽约时报、CNN 及国内新华网、人民网等新闻网站的 RSS 频道。

(3) 浏览订阅。对于习惯了 RSS 信息推送服务的用户，遇到自己感兴趣的传统资源网站、专题论坛不提供 RSS 输出确实感觉很遗憾。用户可以利用自动生成 RSS Feed 的工具（如 Feed43、Ponyfish 等）来实现 RSS 订阅。这种工具可以通过分析网页各项数据来自动生成一个可订阅的 RSS feed 地址^[5]。

这样在浏览网页时,用户可以随时将感兴趣的网站收集到自己的 RSS 目录,解决了那些没有提供 RSS 信息发布渠道网站的 RSS 信息检索问题。

(4) 信息板块订阅。网站上的信息资源通常都是按照类别等形成不同的信息板块,许多用户可能只希望了解网站某一板块的最新信息,比如新闻、天气、股票和体育信息等。通过 RSS 定制相应信息板块,可以不断整合并更新来自网站内及第三方网站的内容。如百度的个性化首页 (<http://my.baidu.com>),允许用户对“我的热门浏览”、“新闻”、“天气”、“音乐推荐”等板块进行个性化订阅。

例如,中国知网 (<http://www.edu.cnki.net/>) 提供了两种订阅方式,一是订阅期刊,可以在 CNKI 收录的 7 300 种期刊中选择需要的期刊,第一时间获取期刊信息;二是按照关键词订阅资料,包括新闻、学术论文、学术动态、最新科研成果,用户可以根据自己需要进行个性化订阅。

3.3 检索策略

(1) RSS 目录要涵盖 RSS 信息源的各种对象,包括博客、专题资源、技术技巧、编程方法等各种类型。除此之外,还要考虑 RSS 信息源的价值分类,领域专家的博客中提供信息的权威性、真实性都要高于一般的个人博客、网页。

(2) 订阅领域专家的博客。把与个人学习研究相关各领域有影响的专家学者的博客收藏到对应目录分类中。以图书馆学为例,知名学者范并思教授自 2004 年 11 月开博至今,其博客空间“老槐也博客”的访问量已接近 56 万(根据其博客首页 2008 年 5 月 27 日统计数据显示),是业界公认的图书馆学专业博客的领袖,程焕文、吴建中等图书馆学专业的知名教授也纷纷开博,除了有影响的专家学者外,开设专业博客更多的是普通科研工作者和高校学生,这些博客中所探讨的大多是专业热点、行业动态、会议信息等,具有很强的时效性和学术性。博客已经成为科研工作的重要信息来源,而传统的检索方式无法充分挖掘其中的丰富资源。

(3) 收集感兴趣的传统资源网站、专题论坛的 RSS 地址聚合到对应目录分类中。许多学科都设有专业网站或专业知识交流论坛,活跃在这些网站或论坛上的都是某一学科或专业的研究人员或从业者,从这些论坛、网站可以获得最新的专业或行业发展

信息。以图书馆学为例,除了中国图书馆学会网站 (<http://www.lsc.org.cn/>) 等官方网站外,还有 E 线图情 (<http://www.chinalibs.net>) 等具有一定商业性的专业网站;而生物医药类专业网站“生物谷” (<http://www.bioon.com/>) 更是吸引了一大批生物医学方面的专家学者和从业人员。万方数据 (<http://www.wanfangdata.com.cn/>) 的数字化期刊、生物医学文献网络资源中的 PubMed 数据库系统 (<http://www.pubmedcentral.nih.gov/>)、生物谷等都提供 RSS 订阅服务^[6]。这些专业网站具备的学科性,一般只有熟悉本专业情况者才能直接浏览对应的网站,有些也可以通过图书馆的各类学科导航获取网址。

(4) 收集各种社会性网站的 RSS 地址,特别是 Digg 类网站,里面有大量基于 Tag 分类的 RSS 地址,这些文章都是经过许多用户推荐出来的,具有一定的价值^[7]。

4 结束语

个性化信息检索是未来面向用户信息检索的一个发展方向。传统的个性化信息检索包括基于内容的个性化信息检索和基于协同过滤的个性化信息检索,并且一般需要引入人工智能技术,研究用户检索行为,建立用户模型,检索效果对用户模型的依赖性较大,表 1 分析了基于 RSS 的个性化信息检索与传统个性化信息检索的区别。

基于 RSS 的个性化信息检索具有个性化、实时性、语义性的特点,允许用户定制个性化的检索界面(如 mybaidu、My yahoo),可实现信息的主动推送,信息过滤和信息聚合功能,为用户提供“一站式”信息服务。与现有的其他信息检索方式相比,RSS 强大的聚合功能和推送功能使用户可以获得的相关信息更多、更新、更快,有助于用户知识的及时更新,并在一定程度上,强化了个性化服务。RSS 在网络营销等领域也有着巨大的影响力,如 e-Bay 上的商品讯息和 RSS 联系在一起,通过 RSS 阅读器等工具实时了解感兴趣的产品讯息。

目前,国内提供 RSS 输出的网站有限,无法搜索到理想的大量信息,但不能否认这种新型的信息检索模式有着巨大的优势和发展潜力,为广大用户提供更方便、及时的个性化检索服务。

表 1 基于 RSS 的个性化信息检索与传统个性化信息检索的比较

	基于 RSS 的个性化信息检索	传统个性化信息检索
检索工具	RSS 搜索引擎及 RSS 阅读器	搜索引擎
检索机理	RSS 定制	用户模型、用户定制
检索对象	支持 RSS 输出的网站	互联网中的网页
标引方式	RSS Feed (XML 格式的元数据), 包含<channel>、<item>、<title>、<link>、<description>、<pubData>等元素	全文索引
检准率	只返回用户定制的信息	返回结果中混杂大量不相关信息
数据更新	用户可定制更新频率, 可同步更新、实时性强	取决于搜索引擎的索引数据库的更新频率以及用户模型的更新频率
语义性	RSS 本身是一种 XML 文档, 包含丰富语义信息	关键词的机械匹配, 不能很好体现语义相关性
个性化	用户可定制个性化用户界面, 选择自己感兴趣的信息源, 定制更新频率、结果显示方式等	所有用户输入相同检索词, 返回相同结果
结果呈现	动态的、连续的、个性化	静态的、一次性的
信息过滤	个性化定制, 完全过滤掉广告、垃圾邮件等无关信息	协同过滤
信息聚合	可聚合不同网站的信息, 实现 RSS Feed “合烧”	可聚合网站内信息

参考文献:

- [1] 崔建海,等.Web 环境下的个性化信息检索技术[J].现代图书情报技术,2005,(9):45-49.
- [2] 黄继征.RSS 技术在图书馆信息推送服务中的应用[J].大学图书馆学刊,2006,(5):35-36,42.
- [3] 刘晶,杨万成.基于简易聚合技术对电子商务个性检索的应用[J].计算机工程与设计,2008,(1):173-174.
- [4] 封硕,等.基于 RSS 的分布式博客搜索引擎的研究[J].情报杂志,2007,(8):96-97.
- [5] RSS Feed 的应用以及周边[EB/OL].<http://www.showeb20.com/?p=457>. [2008-4-10].
- [6] 赵研科,赵革科.Web2.0 环境下专题文献检索途径分析[J].中国科学导报,2007,(8):8-10.
- [7] 许丽华.RSS 技术及其信息聚合策略研究[J].教育技术导刊,2007,(3):47-48.
- [8] 肖莉.论 RSS 聚合技术对竞争情报工作的影响[J].图书情报工作,2007,(4):128-131.

(上接第 47 页)

(2) 宣传与推广。希望在更多的学校有链接;希望更多的读者知道并使用;希望更多的馆和馆员加入到咨询服务中来。除了本科院校,面向数量更多的高职高专开展宣传推广活动。

(3) 计划建立安徽省光盘共享中心,便于各校使用;同时,在省数图会的门户网站上提供高校课程资源共享平台试用:<http://ahadl.org>。

(4) 完善省数图业务及管理运行制度,例如咨询时间控制、节假日咨询反馈处理方法、文献传递业务激励机制、宣传推广队伍建设等。

(5) 文献资源方面,根据第一年的资源经费情况,选择了一些具有联采可能性的数据库,方案已经收集,正在对比和论证过程中;二期经费如果足够,将会考虑更多资源。

6 结束语

安徽省高校数字图书馆在各位关心支持数字图

书馆事业的领导、老师的关心下成立了,但并不表示数字图书馆的事业已经圆满结束。各高校图书馆要结合本校的实际情况,合理统筹规划,在以实体馆藏文献资源建设的基础上,大力整合、开发和利用数字文献资源,在共建共享的道路上继续努力,实现安徽省高校图书馆的可持续发展。

参考文献:

- [1] 杨琼,刘会娟.安徽省高校图书馆数字化资源建设的策略研究[J].农业图书情报学刊,2011,(2):57.
- [2] 林雅萍.图书馆数字化管理[M].上海:上海辞书出版社,2008,(5):152
- [3] 胡雪琴.安徽:信息化为教育均衡发展鼓帆助力.[2006-04-11].http://www.eol.cn/20060411/3184583_1.shtml.
- [4] [6]赵国荣,赵永俊.浅析高校数字图书馆信息资源建设与共享[J].晋图学刊,2011,(2):33-35.
- [5] [7]彭少杰,岳国庆.安徽省高校图书馆电子文献资源共建共享探究[J].通化师范学院学报,2009,(8):60.