RSS技术及其发展探讨

黄春贤,毛明志,钟 毅
(中山大学 信息科学与技术学院,广东 广州 510275)

摘 要:互联网的发展使得网络成为人们重要的信息来源,但传统的浏览方式存在一定不足。一种新的浏览技术RSS在近年来迅速发展,越来越多的Web站点为用户提供基于RSS的浏览方式。文中对RSS技术的各个方面做一个综述,分析了RSS的由来及发展状况,给出了RSS不同版本之间的对比;介绍了RSS技术的工作原理及其与传统的浏览方式的区别。对RSS的优点及应用领域作一个探讨,简单讨论了RSS的一些不足。作为一种新的网络浏览方式,RSS存在优点的同时存在一定的不足,但其将来必定会越来越完善。

关键词:RSS技术;信息浏览;RSS应用
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2007)05-0015-03

Study on RSS Technology and Development

HUANG Chun-xian, MAO Ming-zhi, ZHONG Yi
(Information Science & Technology College of Sun Yat-Sen University, Guangzhou 510275, China)

Abstract: As the development of Internet, people obtain information from network more and more. But traditional browse mode has some problems. Recently, RSS technology is developed quickly. More and more Web sites provide RSS browse for users. Analyses the development of RSS, and analyses the principle of RSS and differences between RSS and traditional browse. Then discusses the advantage and application of RSS. At last, discusses the shortages of RSS. As a new Browse mode of network, RSS has both advantages and shortages, but it will be perfect in future.

Key words: RSS technology; information browse; RSS application

0 引 言

随着互联网的飞速发展,网络成为人们重要的信息来源。但是现在的网络浏览方式也存在着一些问题。首先,不能为每个用户定制个性化阅读方案,每个人都必须面对同样的内容,不能由用户自主地选择感兴趣的信息类型;其次,现在网站的铺天盖地的广告、大量的图片、影音文件,会在浏览时减慢速度,影响用户的使用。

RSS技术的出现,为这些问题提供了一个很好的解决办法。RSS的定义为"Rich Site Summary(丰富站点摘要)"、"RDF Site Summary (RDF站点摘要,RDF是一种语义网技术)",还可以是"Really Simple Syndication(简易聚合)"[1]。这主要是因为该技术有不同的源头,不同的技术团体对其做出了不同的解释。实际上RSS是一种简单XML格式,用于为内容整合客户端提供选择性的、汇总过的Web内容。准确地说,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"[2],它能够用于共享各种各样的信息,包括新闻、简讯、Web站点更新、事件日历、软件更新、特色内容集合和电子商务等。

据不完全统计,美国提供RSS内容的网站数目从2001年9月的千余家激增至2004年9月的19.5万余家,三年中增长了近150倍。随着RSS内容数量的激增,RSS用户数也从2001年8月的10万用户增加到2004年8月的900万[3]。

文中从RSS技术的介绍出发,对RSS的发展、工作原理、优点、应用及存在的问题做一个探讨。

1 RSS的发展

RSS技术诞生于1999年的网景公司(Netscape)。当时网景公司定义了一套描述新闻频道的语言——RSS(RDF Site Summary或者Rich Site Summary),发布了一个0.9版本规范。目的是用来建立一个整合了各主要新闻站点内容的门户,但是0.9版本的RSS规范

收稿日期:2006-07-23
作者简介:黄春贤(1983-),男,广东龙川人,硕士研究生,研究方向为软件工程与CMM;毛明志,副教授,研究方向为软件质量、软件度量。

过于复杂，而一个简化的 RSS0.91 版本也随着 Netscape 公司对该项目的放弃而于 2000 年暂停[4]。

2000 年，RSS 技术标准的发展工作被 Dave Winer 的公司 User Land 所接手。通过 Dave Winer 的努力，RSS 升级后被众多的专业新闻站点所接受和支持。

2001 年，一个第三方、非商业组织根据 W3C 新一代的语义网技术 RDF 对 RSS 进行了重新定义，发布了 RSS 1.0 版，并把 RSS 定义为“RDF Site Summary”。

2002 年 9 月，UserLand 公司把 RSS 升级到了 2.0 版本，又将 RSS 再定义为“Really Simple Syndication”，并发布 RSS2.0。并交由哈佛大学法学院 Berkman 互联网和社会学中心进行维护[5]。

2003 年 2 月，Google 收购了美国大型的博客服务网站 www.blogger.com，使这个网站一直采用的一种近似于 RSS 的技术衍生版 Atom 迅速成为 RSS 领域标准之争的新的有力竞争对手。

至此，RSS 分化为 RSS0.9x/2.0 和 RSS1.0 两个阵营，RSS 迄今没有一个统一的标准，各种标准正在展开对话，表 1 是 RSS 各版本的比较。

表 1 RSS 各版本比较

Version	Owner	Pros	Status	Recommendation
0.90	Netscape		Obsoleted by 1.0	Don't use
0.91/0.92/0.93/0.94	User-Land	Drop dead simple	Obsoleted by 2.0	Use 2.0 instead
1.0	RSS-Dev Working Group	RDF-based, extensibility via modules, not controlled by a single vendor	Stable core, active module development	Use for RDF-based application
2.0	User-Land	Extensibility var modules, easy migration path from 0.9x branch	Stable core, active module development	Use for general-purpose, metadata-rich syndication

2 RSS 原理

与传统的阅读相比，RSS 阅读方式有很大的不同，如图 1 所示。

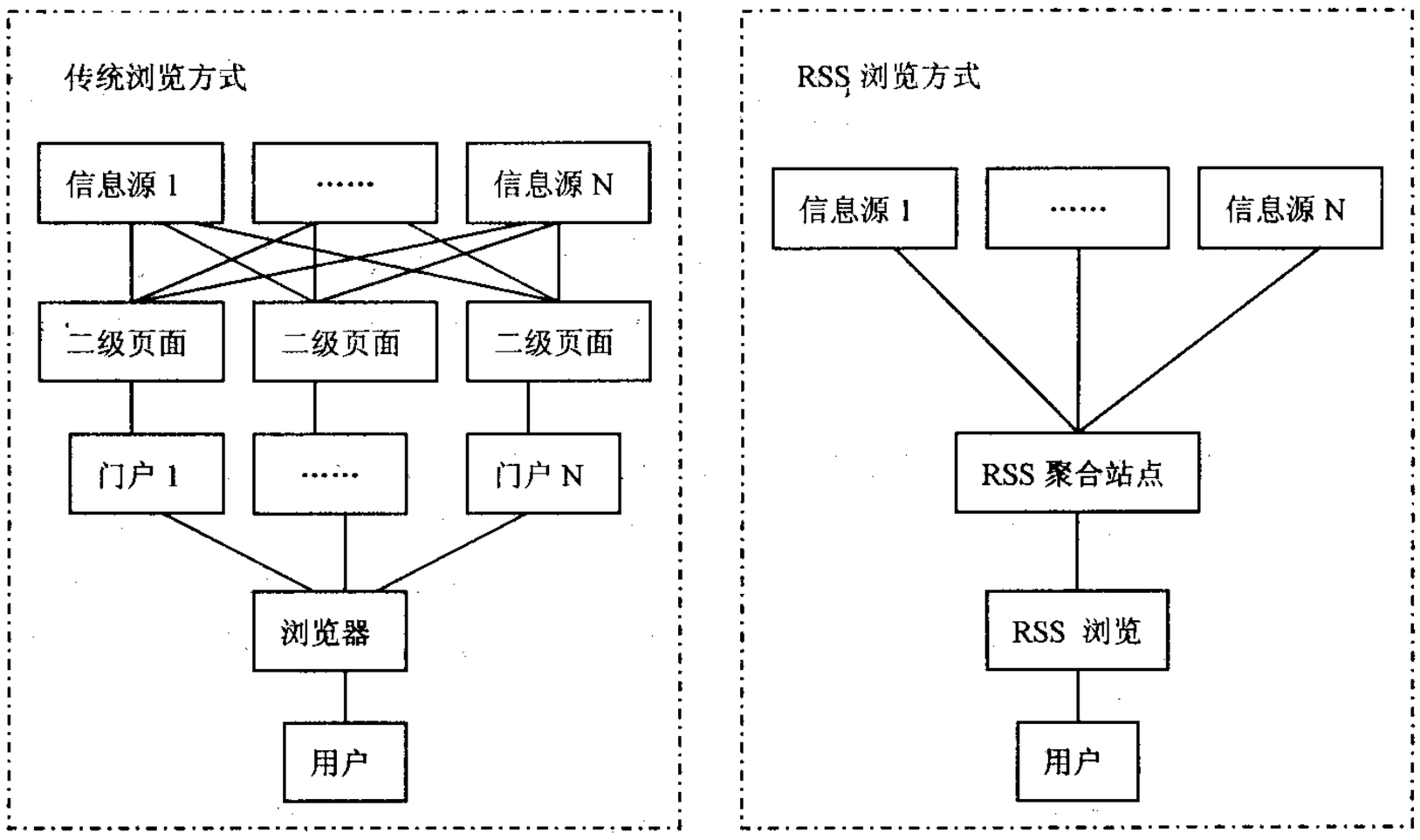

图 1 RSS 阅读方式与传统阅读方式的比较

RSS 是一种 XML 方言，用于连锁 Web 内容和元数据。RSS 文件由一个<channel>元素及其子元素组成。除了频道内容本身之外，<channel>还以项的形式包含表示频道元数据的元素，比如<title>、<link>和<description>。项通常是频道的主要部分，包含经常变化的内容。

频道一般有 3 个元素，提供关于频道本身的信息：

* <title>：频道或提要的名称。

* <link>：与该频道关联的 Web 站点或者站点区域的 URL。

<description>：简要介绍该频道是做什么的。

一个具体的 RSS 文档的形式及显示模式可能如下：

```
<? xml version="1.0" encoding="utf-8"? >
<rss version="2.0">
<channel>
<title>The Simplest Feed</title>
<link>http://www.link.com</link>
<description>A Simple RSS 2.0 Feed</description>
<item>
<title>A Sample</title>
<link>http://www.link.com/sample.htm</link>
<description>The Simple Sample</description>
</item>
</channel>
</rss>
```

从以上例子中可以看到 RSS 文档十分规范，可以清晰地获取各种信息供用户浏览。

3 RSS 优点

1)来源多样的个性化“聚合”。因为 RSS 基于 XML 格式，是一种被广泛采用的内容包装定义格式，因此信息源非常广泛，包括专业新闻站点电子商务站点、企业站点、甚至个人站点等。而在用户端，已经出现了很多商业的 RSS 阅读器，该阅读器能根据用户的喜好，有选择地聚合一些信息到该阅读器，从而用户既

能从多数据源进行阅读,又去掉了繁琐的信息检索过滤过程。

2)信息发布时效高、成本低。RSS 是一种信息聚合的技术,提供了一种更为方便、高效的互联网信息的发布和共享方式。当信息在服务器数据库中更新后,能很快地聚合到用户阅读器当中,极大地提高了信息的时效性和价值。此外,服务器端内容的 RSS 包装在技术实现上极为简单,而且是一次性的工作,使长期的信息发布边际成本几乎降为零,是传统的电子邮件、卫星传输、互联网浏览等发布方式所无法比拟的[6]。

3)无"垃圾"信息,便利的本地内容管理。RSS 用户端阅读器软件的特点是完全由用户根据自身喜好以"频道"的形式订阅值得信任的内容来源。RSS 阅读器软件完全屏蔽掉用户没有订阅的其他所有内容以及弹出广告、垃圾邮件等令人困扰的噪音内容。此外,对下载到阅读器本地的订阅 RSS 内容,用户可以进行离线阅读、存档保留、搜索排序、相关分类等多种管理操作,使阅读器软件不仅是一个"阅读"器,更是一个用户随身的"资料库"。

4)没有病毒。RSS 文件是一种比较简单的 AML 文本文件,并没有涉及任何可执行的文件格式。网站地址的解析也相对安全。

5)营销效果更加凸显。RSS 是用户主动订阅的,所以不容易产生排斥感;用户订阅的都是自己需要、感兴趣的内容,所以更容易转化成购买行为。

4　RSS 商业应用

1)新闻阅读。RSS 技术已经在新华社等新闻机构得到了有益的尝试,逐渐成熟走向商业化,并有望成为新闻出版业一项主流技术。Reuters. com 全球事业副总裁 Steven Schwartz 曾说:"随着网志持续成长,人们想要的内容阅读方式已经改变。我们希望随之调整,让使用者以他们想要的方式存取新闻。[7]"2004 年 8 月 9 日,新华网推出 RSS 聚合新闻,2004 年 9 月 9 日,Google 推出已带有 RSS 聚合新闻功能的简体中文新闻测试版,2004 年 12 月 9 日,百度新闻推出 RSS 新闻订阅,同时,新浪网推出"新浪点点通阅读器",提供 RSS 新闻定制、阅读、管理等一系列服务。自此以后,国内各大网站蜂拥而上,形成了一股影响遍及中国互联网业的 RSS 热潮。

2)RSS 在网络博客中得到广泛的应用。庞大的博客群体的存在,为 RSS 的应用提供了广阔的发展前景。由于 RSS 的提要是可以通过搜索引擎搜索的,因此很多博客都发表自己的网络日志的 RSS 提要,经由发表 RSS 摘要,博客们的网络日志的读者数量增加了,同时也便于让对同一主题感兴趣的人聚集在一起,成为一个个主题社区(Community)。读者还可以通过"回复"功能与文章的作者进行深层次的知识交流。

3)RSS 运用在电子商务中,比如 eBay,Amazon,或者是阿里巴巴,用户可根据自己感兴趣的商品进行定制,并且随时掌握最新标价等更新信息。一旦完成交易这个信息也就随之失效,这是商业上的运用,也是对 Web 的一种辅助。2005 年 3 月,在线超市亚马逊(Amazon. com)推荐数百个个性化的 RSS 提要,以供电子商务顾客使用,此举大大推动了 RSS 的发展[8]。

5　RSS 存在的问题

1)RSS 的商机问题。从读者角度出发,任何形式的在线广告都像垃圾一样在玷污着眼球。因为,相对于门户网站来说,RSS 的确是以一种很"单纯"的方式出现在用户面前的。但是作为内容提供商,他们却更加愿意从商业角度来看待 RSS 的价值。Weblogs Inc. 的创始人 Jason Calcanis 道出了内容提供商的心声:"如果用户不希望见到 RSS 广告,那么谁来为内容提供商付费呢? 如果没人付费,那么作者的收入哪里来呢?"百度副总裁刘建国 2005 年 9 月 2 日在互联网大会演讲表示,RSS 技术是一种好技术,但 RSS 的赢利模式还是看不清楚。而 RSS 要在中国互联网普及和发展,在很大程度上要依赖于拥有大量受众的商业网站,尤其是商业门户网站的大力推广。因此,只有在 RSS 上寻得新的赢利增长点,才能吸引各大商业网站的眼球,RSS 方有机会在中国互联网迎来大发展。

2)RSS 的标准之争。对于 RSS 本身而言,由于在发展过程中遗留下的分歧,RSS 0.9x/2.0 和 RSS1.0 各自为营,激烈竞争。而两者在互联网上均有广泛的应用,RSS 技术并没有一个统一的行业标准存在。在 2004 年 2 月初,Google 宣布其 Blogger 服务放弃 RSS 格式而选择了另一种技术 Atom,此举进一步加剧了这一技术的标准之争[9]。RSS 格式协议的主导者、哈佛大学研究员戴夫·温那在其 Blog 上发表言论,称愿意将 RSS 与其竞争者 Atom 合并为统一格式。然而此举并未得到 RSS 1.0 和 Atom 阵营的热情回应,RSS 技术标准至今还是个三国争雄的混战局面。

3)RSS 的安全性问题。

随着 RSS 应用的兴起,必然会引起黑客的兴趣,而安全问题必然会成为 RSS 必须面对的一个问题。对 RSS 安全性方面的问题主要包括:网络钓鱼攻击,目前最常见的网络钓鱼的方式就是假冒知名公司的名义发送电子邮件,RSS 可能成为该方式的下一个目标。

(下转第 124 页)

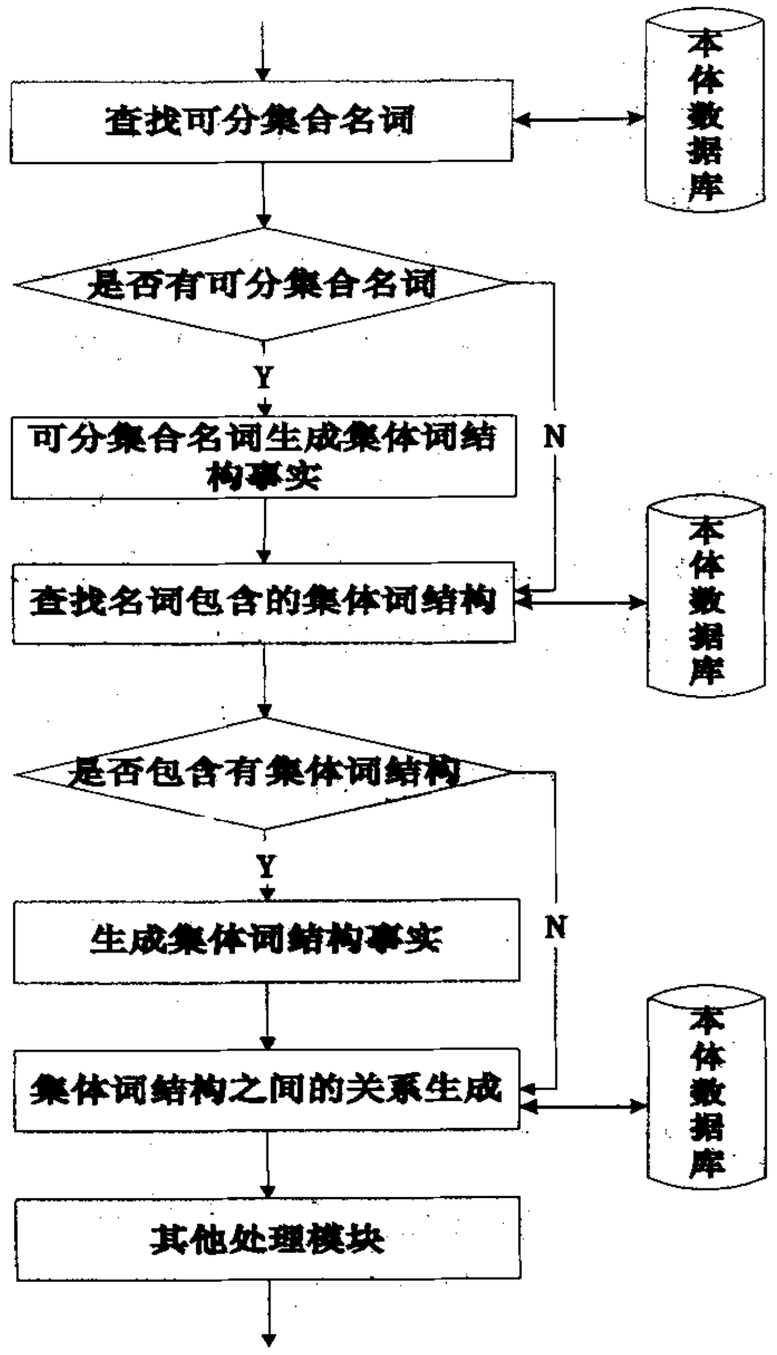

图 5 篇章分析部分集体词结构处理

集体词结构关系确定部分是理解集体词的关键。集体词结构之间的关系是集体词结构实际应用价值的主要体现。系统通过集体词结构差异查找模块完成结构间关系的确定。集体词结构差异查找模块包括结构的中心名词之间的关系比较、结构的元素数量比较、结构包含的所有元素的比较和集体标识词的位置四个方面,用来判断确定集体词结构关系。

3 结 语

群体的处理在数学领域占有很大的比重,文中基于数学领域提出了集体词结构概念及其应用。而笔者已经将这种处理模式应用到了智能辅导系统中。经过反复的、大量的句子测试,推理过程符合人的思维习惯,达到了预期的目的。但是汉语表达形式千变万化,这些分类并不能完全涵盖整个语料库,还需要进一步细划。文中的下一步工作就是继续查找语料库,完善知识库并重点完善与集体词结构相关的处理。

参考文献:

[1] 朱德熙. 语法讲义[M]. 北京:商务印书馆,1984.

[2] 周建设. 中国逻辑语义论[M]. 长沙:岳麓出版社,1996:231-240.

[3] 余以胜,张玉峰. 基于本体论的知识库系统研究[J/OL]. 情报杂志,2003(7). http://e37.cnki.net/kns50/detail.aspx?QueryID=7&CurRec=1.

[4] Zhu Cheng, Wang Zhenjie. Ontology Mapping For Interaction in Agent Society[C]//Services Computing, 2004(SCC2004) Proceedings. 2004 IEEE International Conference. China: Shanghai Jiao Tong University, 2004:619-622.

[5] Allen J. 自然语言理解[M]. 第 2 版. 刘 群,张华平,骆卫华等译. 北京:电子工业出版社,2005.

(上接第 17 页)

恶意代码的二次攻击,RSS 阅读器通过数据库和网页保存用户订阅的信息,如果用户下载恶意的 Feed,那么以后每次用户查看该信息时,恶意代码就会被执行。恶意脚本攻击,为了丰富页面元素,允许方便地用 HTML 的元素来规定字体、颜色、图片显示等元素。很多 RSS 阅读器中自带的 Web 浏览功能,并没有完善地过滤 JavaScript 等脚本语言的功能,因此诸如页面跳转、弹出窗口、运行 ActiveX 控件、修改用户注册表等功能都很容易实现。

6 结 论

RSS 是近年来迅速普及的技术,文中从 RSS 的发展、工作原理、应用、优点、缺点等方面作了全面的阐述。与传统的阅读方式相比,RSS 有成本低、效率高、针对性强的优点,因此能广泛应用于新闻阅读、博客、电子商务网站等领域。但另一方面,RSS 存在一定的不足,如应用过于单一、版权问题、安全性问题等。

参考文献:

[1] Hammersley B. Content Syndication with RSS[M]. [s.l.]: O'Reilly, 2003.

[2] Bates C. XML in Theory and Practice[M]. [s.l.]: John Wiley & Sons, 2003.

[3] Hammersley B. Developing Feeds with RSS and Atom[M]. [s.l.]: O'Reilly, 2005.

[4] 曾宇昆,王清明,杨卫冬,等. XML 模式到关系范式的映射[J]. 计算机工程,2005,31(8):37-39.

[5] 江泽文. RSS:即将到来的互联网新革命[J]. 传媒观察,2005(9):46-47.

[6] 张德杰,高厚礼. RSS 技术及其电子商务应用分析[J]. 计算机与现代化,2005,19(11):82-84.

[7] 李子臣,王晓丽. 引擎竞争的两大焦点:RSS 技术和桌面搜索模式[J]. 中国信息导报,2004(10):54-56.

[8] 伍玉伟. RSS:网络信息"聚合"利器[J]. 图书情报论坛,2006(1):72-73.

[9] 魏 英. Internet 环境下自动新闻发布系统[J]. 计算机应用,2004,24(12):294-296.