互联网中 XML 网页的链接解析与信息采集

Link Analysis and Info – mining of Internet Resources Based on XML/XSL

杜义华　焦文彬　(中国科学院计算机网络信息中心　100864)

摘要:文章分析和介绍对互联网中 XML + XSL 网页资源链接解析和内容采集的方法,包括传统 HTML 中链接解析、XML 转换为 HTML 后链接解析、手工定制下 XML 链接解析和传统 HTML 信息采集、XML 信息抽取、XML 转换为 HTML 的信息采集等。

关键词:互联网信息采集　链接解析　XML 资源

互联网中有海量数据信息,目前网站页面多为 HTML 格式,由于 HTML 标记日益臃肿,文件结构缺乏条理,描述能力有限、有效数据提取复杂等已不再能满足网络上新的应用需求,作为 W3C 推荐的下一代网页发布语言,XML + XSL 方式是大势所趋,现已有一些网站如 37c 医学网、赛迪网等应用。但目前的各大搜索引擎 Spider 系统和互联网信息智能采集系统均为其于 HTML 格式的链接解析和内容提取,对 XML 检索无法支持或有很大局限性[1]。

1　网页链接解析

链接解析用于跟踪网站的新信息和进一步发现资源,即互联网上未知信息搜索[2]。

1.1　传统 HTML 中链接解析

传统 HTML 中标记定义明确,表示超链接用的标记有限。解析过程一般为取网页源文件中 href = 到 间、< area 到 > 块内的 href = 到 shape = 间,frame 到 > 块的 src = 与 > 间所有内容,然后剔除其中 < 与 > 间内容、单引号、双引号等干扰信息,对每块链接部分根据是否含 > 号可分出链接网址部分和链接标题部分,将链接网址部分与网页网址(URL)比较分析等进一步获取完整的 URL,链接文字部分若没有或不合法可进一步取它们源文件中 < title > 与 < title > 间内容。

1.2　XML 转换为 HTML 后链接解析

XML 使用 DTD 显示数据,使用 XSL 描述文档显示,XML 格式网页中各节点自行灵活定义,无法按传统 HTML 方式解析。

正如流览器在识别 XML + XSL 格式网页时先在客户端解析一样,我们也可以先利用 XSL 将 XML 转变成 HTML 语言再按传统 HTML 方式解析。方法为在获取 XML 源文件内容时,通过获取其中 XSL 文件地址,然后利用 XML 解析器(XMLDOM)将他们转结合转换为 HTML[3]。

1.3　手工定制下 XML 链接解析

通过转换为 HTML 语言后解析链接比较通用,适合全范围解析。由于相关超链接信息均存在 XML 文件的某类节点中,每次使用 XSL 转换会有性能上不必要开支,因此有时,特别是对某类网站信息定向跟踪时,为更高性能或仅为获取所需的部分链接,有必要采用手工定制的链接解析。

手工配置方法是先人为查看源 XML 或 XSL(流览器中查看源文件),找到超链接(包括文字、图片、附件)用节点名,添加在配置文件的 xmlhref 项中,系统解析时依此进行。同一个 XSL 文档对应的 XML 是同构的,故采用按每一个 XSL 文档指定所对应的 hreftext(链接用文字)和 hreflink(链接的网址)信息。

如对 http://www.37c.com.cn/的新闻频道网页中相关链接部分在配置 config.xml 中格式如下:

```
<xmlhref>
<xslsite>
<xslfile> http://www.37c.com.cn/info/info01/info01_detail.xsl <xslfile>
<hreftext> ritems/ item /itemtitle < hreftext >
<hreflink> ritems/ item / itemhref <hreflink>
```

</ xslsite > <xslsite > … </ xslsite >

</xmlhref >

其中 xslfile 用于指定配置有效范围，为便于系统实现，采用绝对网址格式，hreftext 和 hreflink 对应的节点采为标准 xpath 格式，考虑到每个 XML 文件根节点唯一，可以忽略根节点而交付程序自动判断实现。

1.4 完整链接解析流程

将以上几种方式结合起来，系统完整解析过程是：判断源网页格式，若为 HTML 则直接解析，若为 XML 格式，则从中找到 XSL 路径，检查 config. xml 中是否有相应 xslfile 的配置，如果有，直接通过配置的 hreftext 和 hreflink 解析，否则，采用通用的 XML 转换为 HTML 后解析链接。流程图如图 1 所示。

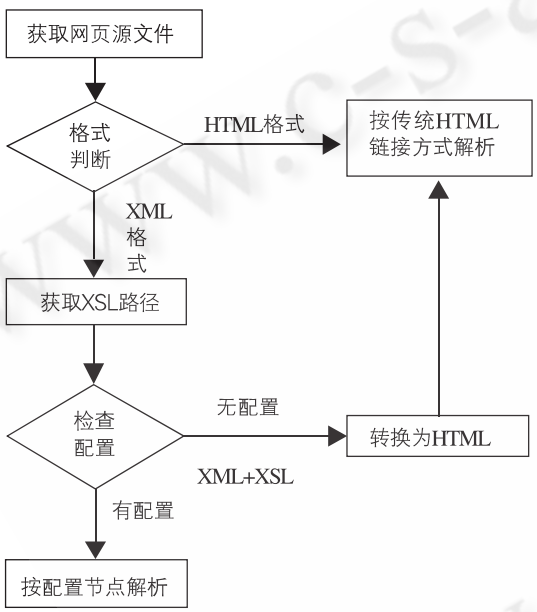

图 1

2 网页信息采集

网页信息采集指深入到站点和页面内部进行内容分析和分类整理，从网页中提取有效数据并按用户要求存储，如直接映射到指定数据库。

2.1 传统 HTML 信息采集

传统 HTML 中数据与格式语言混排，但很多网页采用动态发布技术实现或采用模板制作，有一定内在逻辑或规律。运用 html 分析技术，可以剥离出用户所需信息如标题、正文、作者等[4]。

采集过程是：用户通过分析指定网站或频道栏目下的网页元素，剖析网页源 HTML 代码与所需要数据项的对应关系，利用抽取过程编辑器定义和描述好 HTML 分析处理过程后，由内容替换抽取脚本的解释引擎依次执行和规整入库。其中脚本过程语言替换抽取过程实质为一些字符串处理操作组合，如简单替换命令、高级替换命令、抽取命令、赋值命令、规整命令[4]。

2.2 XML 信息抽取

与 HTML 不同，XML + XSL 方式中数据层跟展现层分开，数据结构清晰，信息的采集和整合相对容易。

用户可在分析某类 XML 源文件后，直接将相应节点与所需数据项对应即可。系统实现时，仍保留原对信息项进一步加工处理命令，需扩展一个命令 XML = (取 xml 中某节点数据)。如 www. 37c. com. cn 中新闻类批处理脚本(部分)为：

[操作内容]

新闻标题 XML = infotitle

新闻来源 XML = laiyuan

新闻来源 = 取 新闻来源 中的《 到 》之间的 全部内容

新闻时间 XML = date

新闻类别 XML = contenttype

新闻作者 XML = author

新闻主题词 XML = keyword

新闻内容 XML = content

新闻内容 = 将 新闻内容 中的 / > 替换为 >

新闻内容 = 将 新闻内容 中的 ^p 替换为

新闻内容 = 将 新闻内容 中的 <! -- 到 -- > 之间替换为

新闻内容 = 将 新闻内容 中的 <p 替换为 ^p <p

新闻内容 = 将 新闻内容 中的 </td > 替换为 ^p

新闻内容 = 将 新闻内容 中的 替换为 ^p

新闻内容 = 将 新闻内容 中的 替换为

新闻内容 = 将 新闻内容 中的 替换为

新闻内容 = 将新闻内容中的
 替换为 ^p

新闻内容 = 将新闻内容中的 <sub 替换为_sub

新闻内容 = 将新闻内容中的 </sub 替换为_/sub

新闻内容 = 将新闻内容中的 <sup 替换为_sup
新闻内容 = 将新闻内容中的 </sup 替换为_/sup
新闻内容 = 将 新闻内容 中的 替换为 _b_
新闻内容 = 将新闻内容中的 替换为_/b_
新闻内容 = 将新闻内容中的 <img 替换为 _img
新闻内容 = 将新闻内容中的 <到 >之间替换为
新闻内容 = 将新闻内容中的_img 替换为 <img
新闻内容 = 将 新闻内容 中的_b_替换为
新闻内容 = 将新闻内容中的_/b_替换为
新闻内容 = 将新闻内容中的_sub 替换为 <sub
新闻内容 = 将新闻内容中的_/sub 替换为 </sub
新闻内容 = 将新闻内容中的_sup 替换为 <sup
新闻内容 = 将新闻内容中的_/sup 替换为 </sup
规整 新闻内容

以上脚本功能包括有对新闻标题、来源等的获取，对新闻内容中换行、加粗、上下标格式的保留，图片的同步下载等，脚本由可视化编辑器定义后自动生成，在采集系统中自动加载和解释执行。其^p 为回车换行符，除 XML = 为新增的 XML 信息抽取命令外，其它同传统的 HTML 信息采集。

2.3 XML 转换为 HTML 信息采集

通过将 XML 信息抽取命令与传统 HTML 信息采集过程相结合，基本能满足网页信息采集需求。但由于 XSL 格式化功能强大而信息采集系统的逻辑处理部分相对简单，偶尔有少数信息隐含在 XSL 文件，如 XSL 文件中可能有当前位置信息、有一些经 XSL 筛选、排序或计算后的信息等。为准确和完整的获取所需信息，系统有时也可以将 XML 与其 XSL 结合转换为 HTML 后进行抽取。

3 结束语

以 XML 为基础的新一代 WWW 环境直接面对 Web 数据，仅基于传统 HTML 格式的链接解析和内容提取已无法满足应用需求。我们对前期开发和成熟应用的互联网信息采集系统改进，在解析模块部分引入对源文件格式判断、按配置处理和预转换为 HTML 功能，在抽取规整模块新增 XML = 命令和新调整解释引擎，保留原图形化配置、预览测试方便等特性，使得系统对 HTML 格式、XML 格式均能灵活高效的自动处理，并向下兼容。新升级后的系统已在中国科学院网站、百拇医药网等应用。

参考资料

1 Neel Sundaresan, Jeonghee Yi Mining, the Web for Relations, http://www9. org/w9cdrom/363/363. html.
2 Eric Ward , How Search Engines Use Link Analysis , http://searchenginewatch. com/searchday/article. php/2158431.
3 XSL 教程 http://www – 900. ibm. com/developer-Works/cn/xml/ccidnet/xslfund/index2. shtml.
4 杜义华、及俊川，通用互联网信息采集系统的设计与开发，计算机应用研究，2005. 1。