基于泛中文域名的网页关键词超链接功能探讨与实现

Introduction and Implementation of WebPage Content Keyword HyperLinks Based on Wildcard Domain Name

杜义华 （中国科学院计算机网络信息中心 管理信息服务中心 100864）

摘要：本文介绍一种基于泛中文域名的关键词超链接实现方法。主要是通过构建完整的关键词表、利用泛域名解析与虚拟中文域名的重定向技术和简单实用的添加链接算法，实现网页全文中所有专业术语、通用词均能点击和按关键词＋域名方式直接访问到相应站点、专题、网页或搜索页的充分互联，解决网页中关键词链接不全面、导向地址不准确、不便记忆或无法及时更新、链接方式不理想等技术问题。

关键词：泛域名 中文虚拟域名 关键词超链接

1 前言

超链接是互联网的重要特点，在频道栏目、标签（tags）、相关文章或热点推荐等线性结构导航与检索基础上，网页全文的关键词超链接能让各知识点多维网状互联，门户、行业或专门网站中全面完整的关键词链接标识有助于将信息立体式展示和为用户提供快捷的百科全书式阅读功能[1]。

网站中知识点和所涉及关键词可能很多，但由于信息整理量大、信息组织不能一步到位、一些关键词的导向页面不确定、超链接添加算法不完备等，目前只有少数网站的部分网页进行有部分关键词链接，如 http://news.sina.com.cn/c/2006-03-24/22048522691s.shtml 中部分词汇链至相关话题、人物专栏或搜索页（http://www.iask.com/n?k=），且链接点不全面、链接地址不便调整，尚没有网站系统能提供全面和专业的关键词标引服务。

泛域名技术能支持无限子域名，中文关键词作子域名能更直观简洁，多编码关键词参照表便于灵活扩展。引入和结合泛中文域名解析思路，设计构建全面关键词表和简洁添加超链接算法，能有效解决目前关键词链接中存在问题。开发实现通用网页关键词库管理平台和关键词链接添加插件，能推广适用于各类网站网页和信息发布平台。

2 关键词库构建

2.1 关键词定义

页面中关键词为直接从文章中抽取的自然语言（自由词），可能是规范术语、专用语或别名与简称，具有一词多义、多词一义和词义不清现象。关键词表可参考但不限于《汉语主题词表》、《医学主题词表》、《中医药主题词表》等公认主题词，可能根据业务需要还有大量机构、企业或人物名，所有词条均能对应到某主题或知识点。

以医疗保健类为例，关键词可包括中草药名、中成药名、方药名、西药名、疾病名、症状名、名医名院、食物名、与生活保健相关的各要素、机构组织、国家法规等。其中有别名现象如恶性肿瘤与癌症、胃十二指肠溃疡与消化性溃疡，有简称现象如中华人民共和国卫生部与卫生部、乙型肝炎与乙肝，此外，为保证语义完整和划词准确，一些惯用法词句即使没有对定专题介绍，也可采用上位主题词作关键词进行保护和参照，如儿茶酚胺与儿茶酚胺类、鼻炎与急性鼻炎、六味地黄与六味地黄丸、六味地黄口服液。

2.2 关键词导向地址

各关键词均对应到一个知识点，链接导向地址可以是一个网站地址、子站点或频道栏目专题首页、具体页面或相关搜索页面。如世界卫生组织可直接导向 WHO 网站、民族医药可链接至民族医药频道、禽流感可链接禽流感专题、非典防治方案可链接到方案的全

文页。

网站或课件中拥有大量知识信息素材，为加强对热点或知识点的展示，在按常规频道栏目或章节分类同时，可进一步挖掘内容间关系加工重组出大量专题。建设过程中，一些关键词对应知识点地址无法确定，或因信息或栏目专题的调整导致某些页面地址（URL）变化，常存在键词条设置不全、无法指向正确页面或已添加链接网页需要重新生成等问题。只有关键词本身是唯一的、确定的、不变的，因此我们采用泛域名的映射解析技术，将每一个关键词作为二级域名。

正如域名与 IP 地址关系一样，这种直接采用中文关键词的域名方式，相当于 URL 助记符，便于记忆，同时能保证链接稳定有效和导出设置灵活，当 URL 地址变化或指向需要调整时，只需修改对照表的相应记录。

2.3 关键词参照表

关键词表用于生成关键词词典文件和泛域名解析的重定向。其中关键词列具唯一索引，参照词用于解决多词一义现象。关键词、Big5 码、IDN 编码等列具有索引以提高解析速度。部分列数据冗余以避免嵌套查询或反复编码解码操作。在关键词表管理平台中，实现对 Big5 码、IDN 编码列和若有参照词时其链接地址列的自动维护。关键词表可同时具有优先级别、广告链接等属性。

关键词	链接地址	参照词	Big5 码	IDN 编码
中华人民共和国卫生部	http://www.moh.gov.cn/		中华人民共和国卫生部	fiQ4Mp3EqsChE72E98Gko7CgbRkq7D
卫生部	http://www.moh.gov.cn/	中华人民共和国卫生部	卫生部	rlR479Ey7S
食疗	http://food.100md.com		食疗	pqYp66E
三七	/index/tcm/herb/0131/index.htm		三七	7gQL
田七	/index/tcm/herb/0131/index.htm	三七	田七	7gQx86G
高血压	/index/disease/k158/index.htm		高血压	omR993J8wL
高血压病	/index/disease/k158/index.htm	高血压	高血压病	omR890FqvM8qR
艾滋病防治条例	/html/law/20060215.htm		爱滋病防治条例	fsQx49CoyDlmCf8Nk0T4q5A
宠物	/index/health/topic/pet.htm		宠物	sbT234C
亚健康	/index/health/topic//yjkhtm		亚健康	‖Qp7Bl0Y
青春期	/index/health/ /qcq.htm		青春期	qiVrVx07I
…	…	…	…	…

3 泛中文域名解析

3.1 泛域名解析配置

泛域名解析是指将 *.域名解析到同一 IP，用于让域名支持无限子域名和防止用户错误输入导致的无法正常访问，目前常用于博客系统，但子域名均为英文字母和数字，中文子域名由于编码技术问题易造成无法正常访问，尚少见应用。配置方法是在 DNS 服务器的域名解析里面设置 *.a.com 的 A 记录或者 CName 记录指向某 IP 或者在域名转发里面设置 *.a.com 转发到 http://www.a.com，同时在此 IP 服务器上配置一个不指定主机头的 web 站点。

3.2 中文子域名编码转换

目前 Internet DNS 是 7 位 ASCII 编码环境，中文域名解析多以 PUNYCODE[2] 编码进行兼容转换。中文编码格式有国际标准（UTF－8）、国家标准（GB2312，GBK）和工业事实标准（BIG5），经过流览器提交编码后捕捉到的可能为国际化域名（IDN）或其它标准。如 CNNIC 的中文域名用户插件、TWNIC 的中文通、NETSCAPE7.1 以上版本、Mozilla browser－1.4 以上版本、Opera browser7.2 以上版本等支持和转换为 IDN 标准，Internet Explorer、3721 网络实名等仍采用 UNICODE 编码或 UTF－8 标准。

对于大量关键词若采用相应汉语拼音或英文作子

域名容易重复且不便记忆，直接采用中文关键词本身将很简洁直观，如 http://人参. 100md. com。采用 web 服务器上关键词多编码对照表和子域名捕捉处理程序，泛中文域名的实现可以不涉及更改流览者客户端设置或 DNS 服务器调整。针对当前对不同流览器或安装不同插件的流览器捕捉到的编码方式差异，我们采用在关键词表中预先设置好各类编码，即除简体中文外，同时对照有 BIG5 码（繁体中文）、IDN 或 punycode 编码、汉语拼音、英文（拉丁名）等字符串，用于兼顾多种标准和辅助解析，并实现对各类子域名如 http://当归. 100md. com、http:// 当归. 100md. com、http://danggui. 100md. com 的同时支持。

3.3 子域名重定向

DNS 泛域名解析将除明确定义（如 www）外的主机头均指向到 * 对应 IP 的 Web 站点。Web 站点中脚本程序（JSP、ASP、ASPX、PHP 等）捕捉取到子域名字符串，扫描关键词表，比较子域名字符串与表中各编码，定位到关键词；若有相应链接地址，重定向到相应地址（可使用框架页），若没有链接地址，通过参照词找到和重定向到参照的链接地址；若没有此关键词，提示没有开通此域名，或者重定向到某搜索页面。

4 超链接添加

4.1 实现思路

网页中添加超链接实质为字符串操作，如将邓铁涛 替换为 <a href = " http://邓铁涛. 100md. com/" target = "_blank" class = "kl" > 邓铁涛 ，但并不是将所有关键词直接替换为相应 HTML 代码这么简单，完善算法要求只对正文部分加链接、不能处理其中不应加链接字符、同一关键词同一页面的超链接不应多个和尽量避免破坏词句的断章取义，我们实现的流程和解决要点如下：

（1）定位取出正文内容。在制作和发布生成页面时采用 web 标准格式[3]，将内容（结构）与表达（样式）分离，正文内容用约定标签，网页中其它文字如当前位置、相关文章等在标签之外，以便处理时准确定位。

（2）保护一些文字和 HTML 代码。对于各类特殊用途的文字或字符，如 HTML 代码、已存在的链接文字、图片的 alt 文字、各 DOM 对象的 title 文字等进行保护处理，方法是按规则找到所有需保护字符存入数组后替换为识别用编号。

（3）依次扫描关键词库。在构建完整主题词库表和明确泛域名导出链接思路基础上，词库的词条采用按字符长度和优先级别排序，避开语义分析和断词处理的困难。

（4）对各关键词只对在正文中第一次出现者增加超链接，方法是当某关键词在文中首次出现时将其转换为链接 HTML 代码记录到中转数组，同时将原正文换为关键词前内容 + 对应数组的编号符 + 关键词后内容。当扫描所有词条后，再根据识别用编号从中转数组中还原成添加链接后的正文内容。

4.2 添加方式

关键词超链接可以在发布前手工添加、发布时自动添加、发布后在线添加。手工添加即编辑人员编排信息时人为设定、发布时添加指集成于发布系统[4]，在信息编审完毕生成静态页面后、更新到发布服务器前，由添加关键词超链接程序（模块）自动完成；在线添加为利用 HTML 包含的 js 脚本文件在用户流览网页时对正文部分分析和实现。

手工添加方式低效，不在讨论之列；发布时添加方式在服务器端完成，一次性添加，发布后对所有流览者有效，访问速度快，有利于搜索引擎收录，推荐使用，但注意当关键词表调整后需重新发布各页面；在线添加方式在客户端由 js 脚本实现，网页正文本身没超链接代码，用户可选择是否启用此功能，关键词表的调整能实时生效，但词条多时资源消耗较大，可适用于论坛或博客页面，相对较少采用。

4.3 部分代码（在线版 javascript 脚本）

```
//获取处理对象和正文内容
if ( typeof( theInfoContent) ! = " undefined" )
{ var obj = document. getElementById ( " theInfo-
Content" ) ; }
else { var obj = document. body ; }
var s = obj. innerHTML ;
//词条定义，按序排列，逗号分开
var strwords = " 中华人民共和国卫生部，艾滋病
防治条例，高血压病，高血压，青春期，卫生部，亚健康，
宠物，三七，田七，食疗" ;
var k = strwords. split ( " ," ) ;
//初始化中间数组
```

```
var mArray4Protect = new Array();
mArray4Protect[0] = "";
//保护特别标记为添加链接范围外内容,略
Special_ContentProtect();
//保护文中已有链接部分
Comm_ContentProtect("<A","</A>");
//保护各类 HTML 代码部分
Comm_ContentProtect("<",">");
//扫描词条,对首次出现的关键词按加链接后代码保护
for (var iLinks = 0; iLinks <= k.length; iLinks++) {LinkWords2ProtectedArray(k[iLinks],k[iLinks])}
//从中间数组依次还原
RestoreFromProtectedArray();
//回写到操作对象
obj.innerHTML = s;
//保护字符到中间数组,原相应内容换为编号标识
function Add2ProtectedArray(str4Protect){
  var iNext = mArray4Protect.length;
  mArray4Protect[iNext] = str4Protect;
  s = s.replace(str4Protect,"_" + iNext.toString() + "|")
}
//将出现关键词以超链接代码记录到中间数组
function LinkWords2ProtectedArray(strWord, strUrl){
var iPosition = s.indexOf(strWord)
if (iPosition == -1) {return ""}
  var iNext = mArray4Protect.length;
  mArray4Protect[iNext] = '<a href="http://' + strUrl + '.100md.com" target="_blank" class="bl">'+ strWord + '</a>';
  var strtemp ="_" + iNext.toString() + "|"
  s = s.substr(0,iPosition) + strtemp + s.substring(iPosition + strWord.length,s.length)
}
//扫描中间数组将各标识号还原
function RestoreFromProtectedArray(){
  for (var iarray = mArray4Pr-otect.length; iarray >=0; iarray--) {
    s = s.replace("_" + iarray.toString() + "|",mArray4Protect[iarray])
  }}
//将所有 * 与 * 间字符均保护
function Comm_ContentProtect(strStart, strEnd){
  var itimes =0; //防死循环
  while ((s.indexOf(strStart) != -1) && (s.indexOf(strEnd) != -1) && (itimes<5000)) {
  var strtemp = s.substring(s.indexOf(strStart),s.indexOf(strEnd) + strEnd.length);
    itimes++;
    Add2ProtectedArray(strtemp);
  }}
```

5 结束语

网页全文关键词超链接能充分发挥互联网链接优势,适用于各类网站。结合大量专题建设,作者尝试性提出和实现基于泛域名解析方式的超链接标识方法,并以医疗保键行业为例,整理专用或通用词条两万多,在百拇医药网各页面(如 http://www.100md.com/html/Dir/2003/09/18/96/613.htm)应用,半年来,网页支持各类流览器和插件,信息组建中调整扩展方便,信息互联效果较好。

参考资料

1 桑新民,当代信息技术在传统文化-教育基础中引发的革命,教育研究,1997.5. P17.

2 Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA), http://www.ietf.org/rfc/rfc3492.txt.

3 网页设计师:web 标准教程及推广, http://www.w3cn.org/.

4 杜义华、张亚,网站信息管理发布系统设计与应用,计算机系统应用,2005.1,P7.