主题网页标签树邻接矩阵识别算法研究

宋 军 杨晓夫 李益才 王家伟

(重庆交通大学信息科学与工程学院 重庆 400074)

摘 要 随着Web编程技术的发展,同类主题网页可以采用不同的Html标签展示出视觉特征相同的网页信息,导致需要匹配Html标签名称的现有网页结构相似性算法无法准确识别同类主题网页。因此,提出一种主题网页标签树邻接矩阵识别算法,通过构造主题网页标签树邻接矩阵,并利用邻接矩阵的结构特征来计算网页之间的结构相似度以实现同类主题网页识别。实验结果表明,该算法的最佳性能达到查全率100%、查准率96%,平均性能达到查全率97%、查准率89%。

关键词 网页结构,Html标签,标签树邻接矩阵

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.063

Research on Recognition Algorithm for Subject Web Pages Based on Tag Tree Adjacency Matrix

SONG Jun YANG Xiao-fu LI Yi-cai WANG Jia-wei

(School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract With the development of Web program technology, the same type subject pages can show the same visual feature information of the Web page by using different HTML tags, resulting in existing Web structure similarity algorithm which measures the structure similarity of the Web page base on matching the HTML tag name information can't accurately recognize the same type subject pages. So, we proposed a recognition algorithm for the same type subject pages based on the tag tree adjacency matrix. This algorithm constructs Web page tag tree's adjacency matrix and recognizes the same type subject pages by computing the structure similarity between the Web pages through the tag tree adjacency matrix. The experimental results indicate that the optimal performance of the algorithm can reach 100% recall rate and 96% precision rate, and the average performance can reach 97% recall rate and 89% precision rate.

Keywords Web page structure, Html tag, Tag tree adjacency matrix

据英国《每日邮报》报道,全球互联网网站数量已经超过10亿且仍在持续增长。为提升公众查询票务信息、购票的便捷性,方便旅客出行,交通运输部在"十二五"期间实施了"公路水路交通出行信息服务系统"交通行业信息化重大工程,大力推动省域道路客运互联网售票系统建设。随着"互联网+"国家战略的实施,互联网企业也纷纷进入公路客运互联网售票领域,四川汽车票务网、重庆市公路客运售票网、12308汽车票务网、北京市公路客票网等国有大型客运企业和互联网企业建设的汽车票务网站纷纷上线。因此,研究如何从互联网海量的网页中快速、准确地获取汽车票务查询类主题网页具有重要的理论和应用价值。

同类主题网页通常具有相同的网页结构特征,利用网页结构特征来度量网页相似性从而实现对同类主题网页的识别成为当前的研究热点。SimRank页面相似度搜索算法[1]通过分析两个页面的链出、链入页面是否相同,判断两个页面是否相似,但无法高效准确地判断含有查询表单的汽车票务查询类主题网页。基于文档对象模型DOM的Web网页结构相似性算法[2]从根节点开始逐层比较文档的标签信息和内容信息,计算每层信息之间的差异,根据预设的阈值判断网页之间的相似性,无法有效地识别出标签不同但结构相同的相似性网页。文献[3]提出一种同时考虑文本、图片、音频、视频等多媒体内容的相似性度量方法,其虽然从不同角度提供了计算相似度的参考思路,但无法从结构上体现出网页之间的相似性。树编辑距离算法[4]通过计算修改、删除、增加网页标签树节点信息的操作步骤来度量网页结构相似性,算法复杂度高且不能很好地反映网页的层级结构。简单树匹配算法[5]通过计算两个网页标签树的最大匹配节点个数来判断网页的相似性,对树节点数量与顺序要求较严格,算法复杂度仍然较高。基于局部标签树匹配的网页聚类改进算法[6]利用标签树中模板节点与非模板节点的层次差异性,使用局部匹配完成网页结构的相似性计算。树路径匹配算法[7]通过两棵网页标签树各条路径上网页标签序列的匹配程度来判断网页结构的相似性。

随着html 5、css 3等网页编程新技术的出现,在设计具有相同视觉特征的汽车票务查询网页时,可以采用不同的Html标签,通过级联样式编写出结构相同的网页,如图1所示。由于需要考虑节点的标签信息,现有网页结构相似性算法无法准确度量汽车票务查询类主题网页的相似度。因此,提出一种利用网页标签树邻接矩阵进行汽车票务查询类主题

到稿日期:2015-08-20 返修日期:2015-10-20 本文受国家自然科学基金(61573076)资助。

宋 军(1971-),男,博士,教授,主要研究方向为客运大数据分析、轨道交通通信与信号控制,E-mail:8335282@qq.com;杨晓夫(1990-),男,硕士生,主要研究方向为交通大数据分析、智能交通物联网;李益才(1970-),男,硕士,副教授,主要研究方向为数据挖掘与大数据处理、轨道交通通信系统;王家伟(1971-),男,硕士,副教授,主要研究方向为数据挖掘与大数据处理、软件工程。

网页识别的算法。

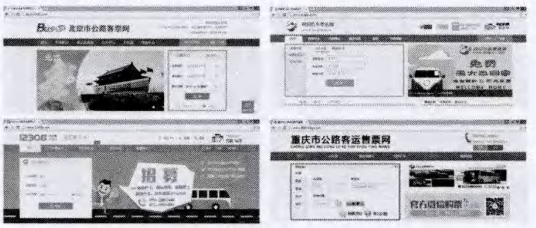


图1 汽车票务网查询页面

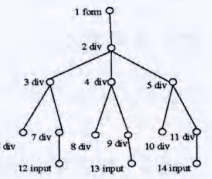
1 网页标签树

Html网页文本属于半结构化数据，由“头部”和“主体”两部分组成，“头部”提供网页信息和引用文档信息，“主体”提供网页具体内容，“头部”和“主体”由多种Html标签相互嵌套组成。北京市公路客票网、四川汽车票务网票务查询窗口的Html源代码片段如图2(a)、图3(a)所示。

Html网页的层次化结构可用文档对象模型(Document Object Model，DOM)表示。DOM定义了获取、修改、添加或删除Html元素的方法及所有Html元素的对象和属性，是页面上数据和结构的一个树形表示，可以转化为只表示网页结构特征的标签树。北京市公路客票网、四川汽车票务网票务查询窗口的标签树如图2(b)、图3(b)所示。

```
<form>
  <div class="cx-item">
    <div class="item">
      <div class="type">出发城市</div>
      <div class="div_input">
        <input id="departCity">
      </div>
    </div>
    <div class="item">
      <div class="type">目的城市</div>
      <div class="div_input">
        <input id="reachCity">
      </div>
    </div>
    <div class="item">
      <div class="type">乘车日期</div>
      <div class="div_input">
        <input id="departDate">
      </div>
    </div>
  </div>
</form>
```

(a)Html源代码

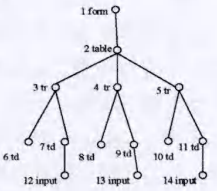


(b)标签树

图2 北京市公路客票网查询页面Html源代码片段及标签树

```
<form>
  <table width="100%" border="0">
    <tr>
      <td class="txt1">出发城市:</td>
      <td>
        <input id='bookStartCityName'>
      </td>
    </tr>
    <tr>
      <td class="txt1">达到城市:</td>
      <td>
        <input id="bookEndCityName">
      </td>
    </tr>
    <tr id="id_hidden_date">
      <td class="txt1">乘车时间:</td>
      <td colspan="2">
        <input name="searchDate">
      </td>
    </tr>
  </table>
</form>
```

(a)Html源代码



(b)标签树

图3 四川汽车票务网查询页面Html源代码片段及标签树

比较图1中北京市公路客票网、四川汽车票务网、12308汽车票务网和重庆市公路客票售票网的票务查询网页可以发现，它们具有相同的界面视觉效果和相似界面结构的票务查询窗口。从图2、图3可以发现，北京市公路客票网、四川汽车票务网分别采用了不同的Html标签代码设计票务查询窗口，但两个票务查询窗口的标签树结构却相同。因此，可以通过分析比较票务查询窗口标签树结构的相似度判断它们是否同为汽车票务查询类主题网页。

2 树路径匹配算法

2.1 树路径

树路径是指从Html网页标签树的根节点到一个叶子节点所经历的所有Html节点标签信息组成的序列，表示为$P=(t_1,t_2,\cdots,t_n)$，其中$(t_1,t_2,\cdots,t_n)$表示该路径所经历的节点标签组成的序列。

北京市公路客票网查询窗口的标签树路径为$P_{P1}=(form,div,div,div)$，$P_{P2}=(form,div,div,div,input)$，$P_{P3}=(form,div,div,div)$，$P_{P4}=(form,div,div,div,input)$，$P_{P5}=(form,div,div,div)$，$P_{P6}=(form,div,div,div,input)$，其标签树路径集合为$h_P=(P_{p1},P_{p2},P_{P3},P_{P4},P_{P5},P_{P6})$。

四川汽车票务网查询窗口的标签树路径为$P_{Q1}=(form,table,tr,td)$，$P_{Q2}=(form,table,tr,td,input)$，$P_{Q3}=(form,table,tr,td)$，$P_{Q4}=(form,table,tr,td,input)$，$P_{Q5}=(form,table,tr,td)$，$P_{Q6}=(form,table,tr,td,input)$，其标签树路径集合$h_Q=(P_{Q1},P_{Q2},P_{Q3},P_{Q4},P_{Q5},P_{Q6})$。

2.2 树路径匹配主题网页识别

树路径匹配主题网页识别根据网页标签树最大匹配路径度量网页结构相似性。

定义1 树路径集合h_i中第m条路径P_{im}与树路径集合h_j中第n条路径P_{jn}的相似度为：

$$sim(P_{im},P_{jn})=\frac{clen(P_{im},P_{jn})}{\max(len(P_{im}),len(P_{jn}))} \tag{1}$$

其中，$clen(P_{im},P_{jn})$为树路径P_{im}与P_{jn}的最长公共标签序列长度，$len(P_{im})$为路径P_{im}的标签序列长度，$len(P_{jn})$为路径P_{jn}的标签序列长度。

定义2 树路径集合h_i中的路径P_{im}在树路径集合h_j的t条路径中的最大匹配树路径为树路径集合h_j中与路径P_{im}相似度最大的那一条路径，即：

$$h_jpath(P_{im})=\max(sim(P_{im},P_{j1}),sim(P_{im},P_{j2}),\cdots,sim(P_{im},P_{jt})) \tag{2}$$

定义3 树路径集合h_i与h_j对应的网页结构相似度为：

$$simhtml(h_i,h_j)=\frac{\sum_{m=1}^{pathnum(h_i)}h_jpath(P_{im})/pathnum(h_i)+\sum_{n=1}^{pathnum(h_j)}h_jpath(P_{jn})/pathnum(h_j)}{2} \tag{3}$$

其中，$pathnum(h_i)$为树路径集合h_i中的树路径总数，$pathnum(h_j)$为树路径集合h_j中的树路径总数。

以北京市公路客票网查询网页为模板，计算四川汽车票务网查询网页与其结构相似度的过程如下：

$P_{P1}=(form,div,div,div)$与$P_{Q2}=(form,table,tr,td,input)$的最长公共标签序列长度为1，$P_{P1}$的标签序列长度为4，$P_{Q2}$的标签序列长度为5。根据式(1)，$P_{P1}$与$P_{Q2}$相似度为：

$$sim(P_{P1},P_{Q2})=\frac{1}{\max(4,5)}=0.2$$

$P_{P1},P_{P2},P_{P3},P_{P4},P_{P5},P_{P6}$在树路径集合$h_Q$中的最大匹配路径由式(2)可得：

$$h_Q path(P_{P1}) = \max(sim(P_{P1}, P_{Q1}), sim(P_{P1}, P_{Q2}), sim(P_{P1}, P_{Q3}), sim(P_{P1}, P_{Q4}), sim(P_{P1}, P_{Q5}), sim(P_{P1}, P_{Q6}))$$
$$= \max(0.25, 0.2, 0.25, 0.2, 0.25, 0.2) = 0.25$$

$h_Q path(P_{P2}) = 0.2$，$h_Q path(P_{P3}) = 0.25$，$h_Q path(P_{P4}) = 0.2$，$h_Q path(P_{P5}) = 0.25$，$h_Q path(P_{P6}) = 0.2$

同理，路径 P_{Q1}，P_{Q2}，P_{Q3}，P_{Q4}，P_{Q5}，P_{Q6} 在树路径集合 h_P 中的最大匹配路径分别为：

$h_P path(P_{Q1}) = 0.25$，$h_P path(P_{Q2}) = 0.2$，$h_P path(P_{Q3}) = 0.25$，$h_P path(P_{Q4}) = 0.2$，$h_P path(P_{Q5}) = 0.25$，$h_P path(P_{Q6}) = 0.2$

根据式(3)，四川汽车票务网查询网页与北京市公路客票网查询网页的结构相似度为：

$$simhtml(h_P, h_Q) = \frac{\sum_{m=1}^{pathnum(h_P)} h_Q path(P_{Pm}) / pathnum(h_P) + \sum_{n=1}^{pathnum(h_Q)} h_P path(P_{Qn}) / pathnum(h_Q)}{2}$$
$$= [(\frac{0.25+0.2+0.25+0.2+0.25+0.2}{6}) + (\frac{0.25+0.2+0.25+0.2+0.25+0.2}{6})]/2 = 0.225$$

显然，根据树路径匹配算法所得出的两个汽车票务网查询网页相似度仅为 0.225，无法判定它们是同一类主题网页。

3 标签树邻接矩阵算法

3.1 标签树邻接矩阵

定义 4[8] G 为顶点集合 $V(G)=\{1,\cdots,n\}$ 和边集合 $E(G)=\{e_1,\cdots,e_m\}$ 构成的图，$A(G)$ 为 G 的 $n\times n$ 维邻接矩阵（A_{ij} 为 $A(G)$ 中的元素），$A(G)$ 的行与列元素由 $V(G)$ 进行索引。当 $i=j$ 时，$A_{ij}=0$；当 $i\neq j$ 时，若 V_i 与 V_j 不相邻，则 $A_{ij}=0$；若 V_i 与 V_j 相邻，则 $A_{ij}=1$。

通过宽度优先算法遍历各汽车票务网查询窗口标签树，构造出北京市公路客票网的标签树邻接矩阵 P：

$$P=\begin{pmatrix}
0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&1&1&1&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&1&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&1&1&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0
\end{pmatrix}$$

四川汽车票务网的标签树邻接矩阵 Q：

$$Q=\begin{pmatrix}
0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&1&1&1&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&1&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&1&1&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0
\end{pmatrix}$$

3.2 邻接矩阵主题网页识别

定义 5 模板网页的标签树邻接矩阵为 A，需识别的目标网页的标签树邻接矩阵为 B，目标网页与模板网页的结构相似度为：

$$sim(A,B) = \frac{n^2 - \sum_{i=1,j=1}^{i\leqslant n, j\leqslant n} |A_{ij} - B_{ij}|}{n^2} \tag{4}$$

其中，$|A_{ij}-B_{ij}|$ 表示两个矩阵相同位置上元素的异同，可以反映两个网页标签树的结构差异。

北京市公路客票网查询网页的标签树邻接矩阵 P 与四川汽车票务网查询网页的标签树邻接矩阵 Q 完全相同，即 $|P-Q|=0$。据式(4)，四川汽车票务网查询网页与北京市公路客票网查询网页的相似度值为：

$sim(P,Q)=1$

因此，可以判定它们属于同一类主题网页。

4 实验及性能分析

实验采用 JAVA 语言编程和 JAVA 开源工具包实现，步骤如下。

步骤 1 利用百度搜索引擎，分别以“道路客运网”、“公路客运联网售票”、“公路旅客运输网”、“汽车客运网”、“公路票务网”等 5 个不同关键词进行网页搜索。

步骤 2 将 5 个不同关键词的百度搜索前 50 个网页作为主题网页识别算法的实验样本，形成 5 个实验网页样本集合。

步骤 3 以北京市公路客票网主页作为模板网页，分别利用标签树邻接矩阵算法和树路径匹配算法计算 5 个实验网页样本集合中每个网页的相似度值，并根据相似度值进行汽车票务查询主题网页识别。

步骤 4 根据实验结果，对标签树邻接矩阵算法和树路径算法的主题网页识别结果进行性能分析与比较。

采用查准率、查全率两个指标分析并比较标签树邻接矩阵与树路径匹配两个主题网页识别算法的性能。

查准率为被正确识别出是或者不是同类主题网页的数量与全部待识别网页总量的百分比，是衡量主题网页识别算法精准度的性能指标。

查全率为被正确识别出的同类主题网页数量与全部待识别网页中同类主题网页总量的百分比，是衡量主题网页识别算法覆盖能力的性能指标。

4.1 单个关键词的算法性能分析

表 1 给出了以“道路客运网”为关键词时百度搜索引擎返

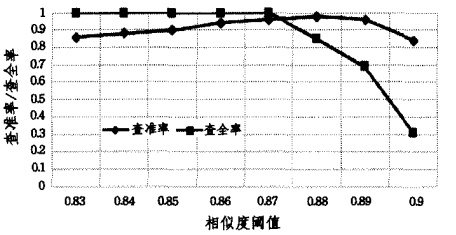
回的前 50 个网页和标签树邻接矩阵、树路径匹配算法计算的相似度值以及点击网页后的人工判断结果。

表 1 “道路客运网”关键词的实验结果

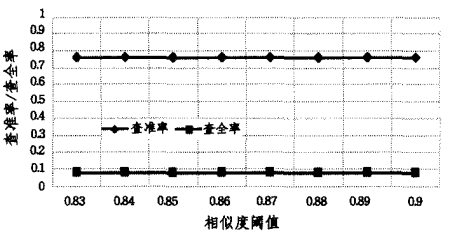
序号	百度搜索的网页	邻接矩阵	树路径	人工判断
1	http://jl.bus365.com	1.00	1.00	是
2	http://keyun.96520.com	0.92	0.42	是
3	http://www.hn96520.com	0.91	0.63	是
4	http://maipiao.96900.com.cn	0.91	0.08	是
5	http://www.xaglkp.com	0.90	0.52	是
6	http://www.12308.com	0.89	0.39	是
7	http://www.12308.com	0.89	0.39	是
8	http://www.12308.com	0.89	0.39	是
9	http://www.zgdlys.cn	0.89	0.23	否
10	http://www.968980.cn	0.89	0.22	是
11	http://www.jslw.gov.cn/busSearch.do	0.88	0.42	是
12	http://www.sdytjy.cn	0.88	0.13	是
13	http://www.chinarta.com/News/kTransit	0.87	0.15	否
14	http://www.jslw.gov.cn/index.do	0.87	0.10	是
15	http://sp.sxtrans.net	0.87	0.04	是
16	http://www.sqk.com.cn	0.86	0.23	否
17	http://www.zgkyxx.com	0.85	0.75	否
18	http://www.etest8.com/dlys	0.85	0.61	否
19	http://zjyz.zjt.gov.cn/cfcms	0.84	0.13	否
20	http://www.chinarta.com	0.83	0.13	否
21	http://www.gzsygj.gov.cn/default.aspx	0.82	0.33	否
22	http://www.ctis.cn/html	0.00	0.00	否
23	http://www.glchx.com/buyphelp	0.00	0.00	否
24	http://www.glchx.com/buyphelp	0.00	0.00	否
25	http://www.hnkyz.com	0.00	0.00	否
26	http://www.dlky.dl.gov.cn	0.00	0.00	否
27	http://www.jxyz.gov.cn/Index.shtml	0.00	0.00	否
28	http://www.crta.org.cn	0.00	0.00	否
29	http://www.gxyz.gov.cn	0.00	0.00	否
30	http://www.xbus.cn/Class/highway	0.00	0.00	否
31	http://www.jt.sh.cn/export/bmcx/sjky	0.00	0.00	否
32	http://www.hzyg.gov.cn	0.00	0.00	否
33	http://www.crtm.cn	0.00	0.00	否
34	http://dlys.gdcd.gov.cn	0.00	0.00	否
35	http://www.daoluyunshu.com	0.00	0.00	否
36	http://www.gs-ys.com/home/index	0.00	0.00	否
37	http://www.hrbkyz.com/public/AA/index.jsp	0.00	0.00	否
38	http://www.jxyz.gov.cn/Index.shtml	0.00	0.00	否
39	http://www.hljjt.gov.cn	0.00	0.00	否
40	http://www.tol.org.cn/jsyzcms/html/index.html	0.00	0.00	否
41	http://www.gs12328.com/passengerhelp	0.00	0.00	否
42	http://www.twwtn.com/Policy/60_282413.html	0.00	0.00	否
43	http://www.moc.gov.cn	0.00	0.00	否
44	http://news.sohu.com/20150610/n414740254.shtml	0.00	0.00	否
45	http://www.hbys.gov.cn	0.00	0.00	否
46	http://www.ygc.czs.gov.cn/ygc	0.00	0.00	否
47	http://www.ycygj.cn	0.00	0.00	否
48	http://www.gxyz.gov.cn	0.00	0.00	否
49	http://www.cqtransit.com	0.00	0.00	否
50	http://zgzks.jxedt.com	0.00	0.00	否

根据表 1 中的实验结果，相似度阈值取 0.83～0.9 时，两种算法的查准率与查全率如图 4 所示。

实验结果表明，在百度搜索引擎返回的前 50 个网页中，有 13 个汽车票务网页。标签树邻接矩阵算法在相似度阈值取 0.83～0.87 时，识别出全部 13 个汽车票务网页，查全率保持在 100%；相似度阈值取 0.87 时，将 9、13 号两个网页误识别为汽车票务网页，查准率为 96%，达到最佳性能；相似度阈值取 0.88 时，虽然只将 9 号网页误识别为汽车票务网页，查准率达到 98%，但 14、15 号两个汽车票务网页被漏检，查全率迅速下降至 85%。由于只能识别与模板网页采用相同 Html 标签设计的网页，在百度搜索引擎提供的 50 个待识网页中，树路径匹配算法只识别出 1 个与北京市公路客票网模板网页结构相同的吉林省公路客票网网页，其查全率仅有 8%，查准率只有 76%。因此，标签树邻接矩阵算法识别相关主题网页的性能远高于树路径匹配算法。此外，相似度阈值超过 0.87 后，标签树邻接矩阵算法的查全率快速降低且查准率也开始回落，因此在实际应用中不应将相似度阈值参数设置得过高。



(a)邻接矩阵相似度匹配识别算法



(b)树路径相似度匹配识别算法

图 4 “道路客运网”关键词的算法性能比较

4.2 算法综合性能分析与比较

表 2 给出了相似度阈值取 0.83～0.90 时，对百度搜索引擎根据 5 个不同关键词搜索出来的网页，标签树邻接矩阵算法、树路径匹配算法识别汽车票务查询类主题网页的查准率与查全率。表 3 给出了标签树邻接矩阵算法、树路径匹配算法针对 5 个不同关键词的平均查准率与查全率。

表 2 不同关键词实验数据汇总

相似度阈值	关键词	邻接矩阵		树路径	
		查准率	查全率	查准率	查全率
0.83	道路客运网	0.86	1	0.76	0.08
	公路客运联网售票	0.86	0.96	0.54	0
	公路旅客运输网	0.86	1	0.72	0.07
	汽车客运网	0.90	1	0.7	0.4
	公路票务网	0.90	0.96	0.64	0.33
0.84	道路客运网	0.88	1	0.76	0.08
	公路客运联网售票	0.86	0.96	0.54	0
	公路旅客运输网	0.90	1	0.72	0.07
	汽车客运网	0.88	0.96	0.7	0.4
	公路票务网	0.92	0.96	0.64	0.33
0.85	道路客运网	0.9	1	0.76	0.08
	公路客运联网售票	0.86	0.96	0.54	0
	公路旅客运输网	0.90	1	0.72	0.07
	汽车客运网	0.88	0.96	0.7	0.4
	公路票务网	0.92	0.96	0.64	0.33
0.86	道路客运网	0.94	1	0.76	0.08
	公路客运联网售票	0.86	0.96	0.54	0
	公路旅客运输网	0.90	1	0.72	0.07
	汽车客运网	0.86	0.92	0.7	0.4
	公路票务网	0.88	0.89	0.64	0.33

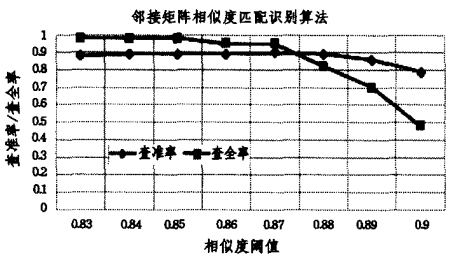
（续表）

相似度阈值	关键词	邻接矩阵		树路径	
		查准率	查全率	查准率	查全率
0.87	道路客运网	0.96	1	0.76	0.08
	公路客运联网售票	0.86	0.96	0.54	0
	公路旅客运输网	0.92	1	0.72	0.07
	汽车客运网	0.84	0.88	0.7	0.4
	公路票务网	0.90	0.89	0.64	0.33
0.88	道路客运网	0.98	0.85	0.76	0.08
	公路客运联网售票	0.80	0.78	0.54	0
	公路旅客运输网	0.88	0.79	0.72	0.07
	汽车客运网	0.88	0.84	0.7	0.4
	公路票务网	0.9	0.85	0.64	0.33
0.89	道路客运网	0.96	0.69	0.76	0.08
	公路客运联网售票	0.80	0.65	0.54	0
	公路旅客运输网	0.86	0.71	0.74	0.07
	汽车客运网	0.80	0.64	0.7	0.4
	公路票务网	0.90	0.81	0.64	0.33
0.90	道路客运网	0.84	0.31	0.76	0.08
	公路客运联网售票	0.68	0.35	0.54	0
	公路旅客运输网	0.84	0.57	0.74	0.07
	汽车客运网	0.78	0.56	0.7	0.4
	公路票务网	0.80	0.63	0.64	0.33

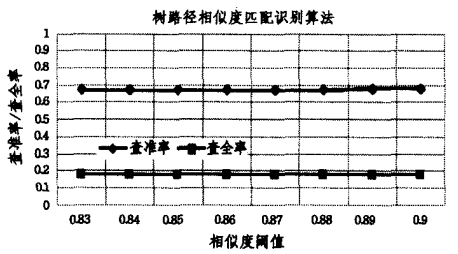
表3　不同阈值下的平均查准率与查全率

相似度阈值	邻接矩阵(平均)		树路径(平均)	
	查准率	查全率	查准率	查全率
0.83	0.88	0.98	0.67	0.18
0.84	0.89	0.98	0.67	0.18
0.85	0.89	0.98	0.67	0.18
0.86	0.89	0.95	0.67	0.18
0.87	0.90	0.95	0.67	0.18
0.88	0.89	0.82	0.67	0.18
0.89	0.86	0.7	0.68	0.18
0.90	0.79	0.48	0.68	0.18

根据表3中的实验结果，相似度阈值取0.83～0.9时，标签树邻接矩阵算法、树路径匹配算法针对5个不同主题关键词的平均查准率与查全率如图5所示。



(a)邻接矩阵相似度匹配识别算法



(b)树路径相似度匹配识别算法

图5　算法综合性能比较

实验结果表明，标签树邻接矩阵算法在相似度阈值取0.83～0.87时，平均查全率保持在97%，平均查准率达到89%，且相似度阈值取0.85时，达到查全率98%、查准率89%的最佳性能；相似度阈值超过0.87后，查全率快速降低且查准率也开始下降。树路径匹配算法针对不同主题关键词的平均查全率仅为18%，查准率只有67%。

结束语　汽车票务查询类主题网页标签树邻接矩阵识别算法能有效地度量使用不同Html标签表征相同视觉结构特征的网页，弥补了现有算法在这方面的不足。由于网页的多样性和复杂性，该算法仍然存在少量的错误判断。后续工作将进一步优化算法，提高算法查准率，并将其应用于相关Web垂直搜索引擎开发。

参考文献

[1] Lin Zhen-jiang, Lyu M R, King I. PageSim: A novel linkbased measure of Web page similarity[C]// Proc of the 15th WWW Conf. Los Alamitos: IEEE Computer Society Press, 2006: 1019-1020

[2] Kang Chun-ying. DOM based Web Page to Determine the Structure of the Similarity Algorithm[C]// The Workshop on INtelligent Information Technology Applications IEEE. 2009: 245-248

[3] Shi Peng, Ding Lian-hong, Liu Bing-wu. Similarity Computation of Web Pages[C]// IEEE International Symposium on Knowledge Acqustion and Modeling Workshop, 2008. Kam Workshop, 2008: 777-780

[4] Zhang Rui-xue. Research & Application of Web Similiarity Based on DOM Tree[D]. Dalian: Dalian University of Technology, 2011(in Chinese)
张瑞学. 基于DOM树的网页相似度研究与应用[D]. 大连：大连理工大学，2011

[5] He Xin, Xie Zhi-peng. Measurement of Web page structure similarity based on simple tree matching algorithm[J]. The Research and Development of Computer, 2007, 44(Suppl): 1-6(in Chinese)
何昕，谢志鹏. 基于简单树匹配算法的Web页面结构相似性度量[J]. 计算机研究与发展，2007，44(Suppl.)：1-6

[6] Li Rui, Zeng Jun-yu, Zhou Si-wang. Structural Similarity Measurement of Web Pages Based on Simple Tree Matching[J]. The Development of Computer, 2010, 30(3): 818-820(in Chinese)
李睿，曾俊瑀，周四望. 基于局部标签树匹配的改进网页聚类算法[J]. 计算机应用，2010，30(3)：818-820

[7] Liao Hao-wei, Yang Yan. An Improved Web Structure Similarity Based on Matching Algorithm of Tree Paths[J]. Journal of Jilin University, 2012, 50(6): 1200-1202(in Chinese)
廖浩伟，杨燕. 一种改进的基于树路径匹配的网页结构相似度算法[J]. 吉林大学学报，2012，50(6)：1200-1202

[8] Bapat R B. 图与矩阵[M]. 吴少川，译. 哈尔滨：哈尔滨工业大学出版社，2014