互联网视频摘要信息自动抽取①

易荣锋[1] 朱六璋[2] 尹文科[1]

(1.中国科学技术大学 自动化系 安徽 合肥 230027; 2.安徽电力继远软件公司 安徽 合肥 230000)

摘 要: 提出一种识别视频播放页，并从中抽取视频摘要信息的方法，播放页的自动识别是通过三个判定要素的运用来实现，播放页内摘要信息的抽取是通过依次去除背景噪声、随机噪声、残留噪声来实现。有关实验结果表明，该方法具有较好的通用性。

关键词: 摘要信息抽取；互联网视频；网页净化

Automatic Extraction of Internet Video Summary

YI Rong-Feng[1], ZHU Liu-Zhang[2], YIN Wen-Ke[1]

(1.Department of Automation, University of Science and Technology of China , Hefei 230027, China;

2. Anhui Electric Power Jiyuan Company, Hefei 230000, China)

Abstract: This paper presents key techniques of acquining playing pages and extracting video summaries on playing pages. Automatic identification of playing pages makes the use of three determined elements. Extraction of a video summary in a playing page is done by removing background noises, random noises, and residual noises.

Keywords: video summary extraction; internet video; page purification

1 引言

近年来随着视频服务网站的蓬勃发展，对Internet上的视频信息进行数据集成已经成为互联网应用的迫切需求，而互联网视频摘要信息的自动抽取是该领域的关键技术之一。目前，各视频搜索网站所集成的视频摘要信息大都是从播放页链接的锚文本上获取，存在摘要信息过少的缺点。比如缺少视频的详细介绍、演员、导演，等等。通过对视频服务网站的观察可以发现，视频播放页的摘要信息远比播放页链接的锚文本丰富。因此从数据完整性的角度考虑，在视频播放页内抽取摘要信息是最优选择。通过通用网络爬虫获取视频播放页并在视频播放页内抽取摘要信息必须解决两个问题：播放页的自动识别和播放页上摘要信息的抽取。

在互联网早期，判断一个网页是否为视频播放页，可以通过查找其内有无object或者embed标签来确定。但随着网络应用技术的发展,该方法需面对以下问题:1)部分视频服务网站网页的源文件中不存在object或者embed标签；2)不能将视频播放页和音乐播放页区分开；3)大量的flash广告的应用导致难以区分播放器的播放内容是普通的视频文件还是广告。

针对这些问题，本文提出了一种识别视频播放页，并从中抽取视频摘要信息的方法（PI-SE）。其中视频播放页内摘要信息的抽取则是综合利用模板、网页结构、语义信息，通过逐步去除噪声信息，达到保留有用信息的目的。首先利用模板除去背景噪声，其次利用网页结构特点除去随机噪声，再次利用语义信息除去残留噪声。

① 基金项目:国家高技术研究发展计划(863)(2008AA01A318,2008AA01Z408)
收稿时间:2010-02-07;收到修改稿时间:2010-04-03

2 相关工作

2.1 播放页识别

目前学术届对于网页分类提出了若干方法，如文献[1]归纳的概率模型方法、关系学习方法、支持向量机方法，对于播放页的识别；文献[2] 将 SVM 方法应用到文本分类，并使用多项式核函数、RBF 核函数和 Sigmoid 核函数；文献[8]把 Bayes 网络用于文本分类。由于这些方法均没有充分利用视频播放页的相关特征，因此在播放页的识别准确率上以及识别效率上均不够理想。

2.2 网页去噪声

网页去噪声又称网页净化,是指将与当前网页主题无关的信息去除。文献[3]提出了一个去除网页中噪声内容的方法。该方法依据<table>标签构造网页的标签树，并依据<table>标签将一张网页规划为相互嵌套的内容块；而后，对于使用同一个模板生成的网页集，找出在该网页集中多次出现的内容，作为噪声内容。但该方法不适用于用非<table>标签来进行布局的网页。文献[4]提出一种基于可视化信息的去噪声方法,该方法利用页面中各元素的布局信息对页面进行划分，保留页面中间区域，而其它区域则认为是噪声。文献[5]提出一种提出了一种基于可视化信息的网页信息抽取方法，该方法必须先由人工训练出待抽取对象的特征描述。由于这两种方法都要预先提取页面中各元素的视觉信息，所以该方法存在速度慢的问题。文献[6] 提出 DSE (data-rich section extraction) 算法，该算法通过自顶向下比较两棵同模板的网页树，去除相同的子树，把剩余部分作为网页的主题内容。以上文献均没有对诸如相关视频和网友评论等的去除提出相应方法。

针对以上方法的不足，本文首先结合视频服务网站播放页的特点对各种噪声进行分类，然后对不同类型的噪声提出相应的去除方法。

3 播放页的自动识别

为了快速识别播放页，需要获取以下三个判定要素：

(1) 特征脚本 url：网页引入的某个外部脚本文件的 url，该脚本文件内含有播放器的特征 HTML 标签。

(2) 播放器父节点的 html 文本：对于播放器为动态生成的播放页，播放器节点的父节点具有独特的 id 属性或者 name 属性，因而可用来做为播放页判定的特征之一。

(3) 播放页模板集：对于所有不同类型的播放页的集合，各取其中一个播放页做为模板。

播放页模板集的获取步骤如下：

第一步：载入网页并运行其上的脚本(此步骤一般利用浏览器来完成)，判断有没有生成播放器的特征 HTML 标签。

第二步：分析网页中候选播放器对象的视觉特征，如大小、坐标等，以确定播放器的宽和高是否满足一定阀值，其右边界到页面的右边界的距离、上边界到页面上边界的距离、下边界到页面下边界的距离是否满足一定阀值。根据对视频服务网站的观察，以上四个阀值可分别设为：290，100,220,200,100,单位为像素。

第三步：利用基于树编辑距离的网页相似度对所获取的播放页集进行分类，从每一类中取一个播放页加入播放页模板集。

在获取以上三个要素之后，可采取如下步骤进行播放页快速判定：

第一步：网页是否引入了特征脚本 url 对应的外部脚本文件；

第二步：网页中的某段 html 文本是否严格匹配播放器父节点的 html 文本；

第三步：网页是否与某个播放页模板相似。

以下简要介绍一种基于编辑距离的网页相似性判断方法。

通过对大量网站的观察可以发现，同一网站的不同类型的页面(比如主页，视频列表页，播放面)，其源文件中的 head 部分一般是不会严格相同的，而对同一类型的页面，其源文件的 head 部分必定严格相同。因此，本文将网页相似性判定分为两个步骤：1)判断两个网页的源文件中 head 部分是否严格相同，是则进入步骤 2，否则判定为不相似；2)如果步骤 1 为是，则对两个网页的源文件应用改进的基于树编辑距离的

相似性判定方法，该方法首先将源文件转换成 DOM 树,然后前序遍历以<body>节点为顶点的子树，取出所遍历到的所有节点的节点名，得到一个节点名序列，再应用类似基于编辑距离的字符串相似性判定方法来进行相似性判定。

4 摘要信息抽取

4.1 相关定义

本文利用网页 DOM 结构、模板、语义信息，通过逐步去除噪声信息，达到保留有用信息的目的。为区分不同的噪声类型以便采用有针对性的去噪声方法，此处定义以下两种不同类型的噪声：

(1) 背景噪声：每个同一类型的网页中都存在、且结构和文本内容都完全相同的子树或者文本。如广告链接，站点导航，版权提示信息。以往的研究中所指的噪声系指此类噪声。如图 1 所示。

图 1 背景噪声

(2) 随机噪声：在同一类型的网页的 DOM 树中处于相同位置，但其文本内容并不完全相同,而又不属于当前播放视频的摘要信息。例如当前播放视频的相关视频、网友评论，等等。通过对播放页的观察可以发现，这类噪声的特点是它们由连续的且彼此结构相同的子树构成。如图 2 所示。

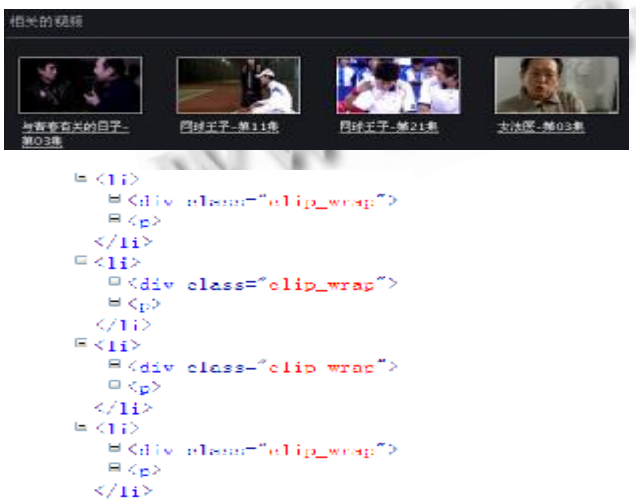

图 2 随机噪声(上边为视觉显示，下边为 DOM 结构(未全展开))

(3) 残留噪声：残留噪声指去除背景噪声和随机噪声后所剩下的非摘要信息的文本或者子树。

本文中去噪声的基本过程如下：首先利用模板除去背景噪声，其次利用网页 DOM 结构特点去除随机噪声，最后利用语义信息除去残留噪声。

4.2 去除随机噪声

对于视频播放页而言，随机噪声的特点是它们由连续的且彼此结构相同的子树构成。只要结构相同的子树相继出现了 4 个以上，就认为这些子树为随机噪声。去除过程为：前序遍历待净化页面的 DOM 树，每到达一个节点就将该节点的子树与后继的三个兄弟节点所在的子树进行比较，如果它们的结构完全相同，则认为找到了解一组随机噪声。之后将处在该位置的所有具有此结构的子树全部去除。其中两棵子树结构完全相同的判定过程为：同时对两子树进行前序遍历，对于每次取到的两个节点(分别来自两子树中的其中一棵和另外一棵)，比较它们的节点名，如果相同则继续比较后面的节点对，否则认为两子树结构不相同;如果两棵子树的所有节点都对应相同，则认为两子树的结构完全相同。

4.3 去除残留噪声

对于视频播放页而言，残留噪声主要是某些用户评论。一般来说用户评论部分会在去除随机噪声时去除，但当用户评论数过少，这部分噪声就会被保留下来。本文使用一个启发式规则来去除这部分噪声。去除过程如下：取出去除了背景噪声和随机噪声后的待净化网页 DOM 的文本内容，然后去除符合下面模式的子字符串：以“关键字网友评论”或者“评论”开始，且其后部分含有关键字“说”或者“发表”或者“回复”或者“网友”。

5 实验

5.1 实验结果

使用本文提出的方法对 40 个视频服务网站进行抽取实验，并且对每个视频服务网站手工提取 100 个播放页和 100 个非播放页进行播放页识别准确率和召回率检验；实验结果如表 1。

表 1　实验结果

网址	发现模板数/正确识别个数/错误识别个数/抽取摘要条数
http://www.youku.com	
http://www.tudou.com	2 /100/5/10 万*
http://www.56.com	3/96/2/ 10 万*
http://v.sina.com	1/100/0/10 万*
http://6.cn	1/100/0/10 万*
http://www.ku6.com	2/100/5/10 万*
http://v.cnmo.com	2/100/0/11356
http://www.yoka.com/video/	1/98/0/793
http://v.9you.com	1/100/0/260
http://tv.hexun.com	2/100/0/560
http://v.zol.com.cn	1/100/0/15473
http://v.pconline.com.cn/	1/100/0/3417
http://www.mianfeishipinwang.cn/	2/100/3/3391
http://www.365shipin.com	1/100/0/77680
http://media.17173.com/	2/100/0/1648
http://live.csdn.net	1/100/0/989
http://www.tvtour.com.cn/	1/100/0/6275
http://tv.huzhai.com/	1/100/0/22230
http://www.che168.com/video/	2/100/0/24786
http://www.beijing2008.cn/video/	1/100/0/1573
http://www.openv.com/	2/98/3/5.4 万*
http://v.pcgames.com.cn/	1/100/0/5898
http://www.vodone.com/	2/94/4/7.2 万*
http://v.2008.163.com/	1/100/0/18776
http://video.mofcom.gov.cn/	1/100/0/1857
http://mv.2u.com.cn/	1/100/0/28549
http://v.hoopchina.com/	2/100/0/10691
http://www.laifu.org/shipin/	1/100/0/1677
http://www.joy.cn/	2/97/3/4.2 万*
http://www.karttv.cn/	1/100/0/1181
http://nv.qianlong.com/	1/100/0/4954
http://v.xgo.com.cn/	1/100/0/2567
http://v.chinadance.cn/	1/100/0/44537
http://v.eol.cn/	1/100/0/646
http://video.mylegist.com/	1/100/0/40
http://v.titan24.com	1/100/0/31842
http://www.seeju.com.cn	1/100/0/18586
http://www.ks52.com.cn/index.html	2/100/0/1155
http://v.jinghua.cn	1/100/7/1142
http://video.cnfol.com/	1/100/0/1208
http://www.openv.com/	2/98/3/5.4 万*
http://v.pcgames.com.cn/	1/100/0/5898
http://www.vodone.com/	2/94/4/7.2 万*
http://v.2008.163.com/	1/100/0/18776
http://video.mofcom.gov.cn/	1/100/0/1857
http://mv.2u.com.cn/	1/100/0/28549
http://v.hoopchina.com/	2/100/0/10691
http://www.laifu.org/shipin/	1/100/0/1677
http://www.joy.cn/	2/97/3/4.2 万*
http://www.karttv.cn/	1/100/0/1181
http://nv.qianlong.com/	1/100/0/4954
http://v.xgo.com.cn/	1/100/0/2567
http://v.chinadance.cn/	1/100/0/44537
http://v.eol.cn/	1/100/0/646
http://video.mylegist.com/	1/100/0/40
http://v.titan24.com	1/100/0/31842
http://www.seeju.com.cn	1/100/0/18586
http://www.ks52.com.cn/index.html	2/100/0/1155
http://v.jinghua.cn	1/100/7/1142

(注：带*的网站为仅仅采样其中部分视频摘要信息的网站；本实验忽略了含有符合视觉特征的播放器但在网站的所有含播放器的页面中比率过低的模板)

根据对实验数据的统计及对各测试网站实际情况的对照，得出的各项参数如下：模板召回率为 95%，模板准确率为 98%，识别召回率为，100%，识别准确率为，99%，净化准确率为 90%。

其中各参数的定义如下：

模板召回率：提取的有效模板占所有测试网站播放页模板总数的比率；

模板准确率：提取的有效模板占提取的模板总数的比率；

识别召回率：正确识别出的播放页个数占播放页总数的比率；

识别准确率：正确识别出的播放页个数占正确识别和错误识别的播放页个数之和的比率；

净化准确率：噪声被完全去除的网站数占全部测试网站的比率。

5.2 实验对比

使用文献[2]和文献[8]所述的方法进行播放页识别实验，所得到的识别召回率、识别准确率分别为64%、57%和68%、62%。使用文献[3]和文献[4]所述的方法进行摘要信息抽取实验，所得到的净化准确率分别为41%和47%。

5.3 实验分析

由于本文提出的播放页识别方法充分利用了视频播放页的相关特征，因而实验结果较为理想。实验中错误识别为播放页的网页主要来自于大型视频网站的播客主页以及一些具有视频广告的页面，这类页面具备的各个特征非常类似于播放页，因而较难排除。另外，由于本文提出的摘要信息抽取方法针对视频播放页的随机噪声提出了相应的处理方法，因而相比其它方法能达到更高的净化准确率。

6 总结

本文充分利用视频播放页的相关特征，提出了播放页识别的三个判定要素及其获取方法，进而提出了一种自动识别播放页，并通过逐步去除各种类型的噪声信息来获取当前播放视频的摘要信息的方法(PI-SE)，该方法对互联网视频的数据集成将具有十分积极的意义。后续工作将主要围绕进一步提高模板提取率、模板正确率、识别准确率、净化准确率展开。

参考文献

1 孙建涛,沈抖,陆玉昌,石纯一.网页分类技术.清华大学学报(自然科学版), 2004 ,44(1):51－68.

2 Leopold E, Kindermann J. Text categorization with support vector machine: How to represent texts in input space .Machine Learning, 2002:423－444.

3 Koller D, Saham iM. Hierarchically classifying documents using very few words. Fisher D, ICML 97. SanFrancisco:Morgan Kaufmann, 1997:170－178.

4 Lin SH, Ho JM. Discovering informative content blocks from Web documents. SIGKDD, 2002:588－593.

5 荆涛,左万利.基于可视布局信息的网页噪音去除算法.华南理工大学学报(自然科学版), 2004,32(s1):84－87.

6 Jan YW, Tsay JJ, Wu BL. WISE: A Visual Tool for Automatic Extraction of Objects from World Wide Web. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). 2005:590－593.

7 Wang JY, Lochovsky FH. Data-rich Section Extraction from HTML pages. Proc. of the 3rd International Conference on Web Information Systems Engineering (WISE.02). IEEE Computer Society, 2002:313－322.