

# 基于链接分块的相关链接提取方法

王芳, 于浩, 谭红叶, 赵铁军

(哈尔滨工业大学 计算机学院 机器智能与翻译研究室, 哈尔滨 150001)

E-mail: fangwang@mtlab.hit.edu.cn

**摘要:**每个网页都包含了大量的超链接,其中既包含了相关链接,也包含了大量噪声链接。提出了一种基于链接分块的相关链接提取方法。首先,将网页按照 HTML 语言中<table>标签将网页分成许多的块,从块中提取链接,形成若干链接块;其次,根据相关链接的成块出现,相关链接文字与其所在网页标题含相同词等特征,应用规则与统计相结合的方法从所有链接块中提取相关链接块。相关链接提取方法测试结果,精确率在 85%以上,召回率在 70%左右,表明该方法很有效。

**关键词:**网页分块;链接块;相关链接提取

文章编号:1002-8331(2006)31-0110-04 文献标识码:A 中图分类号:TP391

## Relation Links Extracted Approach Based on Blocking Links

WANG Fang, YU Hao, TAN Hong-ye, ZHAO Tie-jun

(Machine Intelligence and Translation Laboratory, Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**Abstract:** There are lots of hyper links in a web page, including relation links and “noisy” links. A novel approach is proposed to extract relation links from page based on link block in this paper. The approach is composed of two steps. Firstly, a web page is partitioned into lots of blocks according to HTML tag <table> in a web page. Then links are extracted from blocks and lots of link blocks are gotten. Secondly, relation link block is obtained by using rules. For instance, relation link belongs to one block and their anchor text has common words with title of current page where relation link is located. The result of experiment shows that the method is effective, with above 85% precise rate and about 70% recall rate.

**Key words:** page segmentation; link block; relation link extraction

### 1 引言

网络给人们提供了一种很好的资源共享的方式,既包含大量的有用信息,也包含了大量无关冗余的“噪声”信息。近年来,网络信息资源爆炸式的增长,给人们获取自己想要的信息带来很大的难度,所以提供一种自动识别有用信息的手段显得尤为重要。

网页中标题及正文的内容称为网页的主题<sup>[1]</sup>。网页包含了大量的链接,与主题相关的链接被称为相关链接,该类链接一般是对网页主题信息的进一步说明或扩充,为人们提供更多有用信息;而与正文主题无关链接一般称为“噪声”链接,主要包含、导航链接、网站版权信息链接、服务链接、广告链接等,同一网站几乎所有网页都包含该类冗余信息<sup>[2,3]</sup>,为用户提供的信息较少,所以人们更关心相关链接所链接的内容。因此,有必要从大量链接信息中自动提出相关链接信息。自动提取的相关链接可直接用于网页的净化<sup>[4]</sup>,从含“噪声”的主题型网页中获取标题、正文及相关链接组成新的网页。另外,由于相关链接本身与网页主题相关性很大,所以自动提取相关链接方法可服务于特定主题爬行者<sup>[5]</sup>,还可以辅助网络社区发现与自组织(Self-Organization and Identification of Web Communities)<sup>[6]</sup>的研究。

同一网页从形式和内容上都包含若干块,且每块所含的信息重要性并不相同。网页许多块中包含的与主题无关的冗余信

息,常常会导致搜索及网络挖掘等研究结果的偏差。所以,基于网页分块的研究越来越多。目前主要的网页分块方法有以下几类:基于 DOM 的页面分割方法(DOM based segmentation)<sup>[4]</sup>,基于位置的页面分割(location-based segmentation)<sup>[8]</sup>和基于视觉的页面分隔(Vision-based Page Segmentation, VIPS)<sup>[9]</sup>。文献[2]中采用 VIPS 算法对网页分块后,采用机器学习的算法自动给出每块的重要性值。文献[3]利用 DOM Tree 方法对网页分块,利用块信息熵的不同来提取有用的信息块。本文研究发现相关链接在网页中成块出现的特点,并利用网页分块思路,此处与文献[2,3]相似,但与它们的不同点在于,本研究中主要针对块内链接进行分析,提取有用的链接信息——相关链接。

本文结合上述国内外相关研究,通过对许多代表性的中文网站(搜狐、新浪、网易、雅虎中文网站)中上百个网页的分析,提出了一种基于链接分块的相关链接提取方法。通过对不同网站多次测试得出,精确率在 85%以上,召回率在 70%左右,表明该方法非常有效。本文的相关链接提取方法主要采用链接分块的思想,辅以规则结合统计数据从中提取相关链接。通过用 HTML 中的<table>标签将网页分成若干块,从每块中提取链接形成链接块,使用相关链接文字与其所在页面标题包含相同词等特征应用规则,结合统计数据提取相关链接块。

本文后续内容安排如下,第 2 章详细介绍了相关链接提取

基金项目:富士通研发中心有限公司委托研究项目。

作者简介:王芳(1982-),硕士在读;于浩(1969-),副教授;谭红叶(1971-),博士;赵铁军(1964-),教授,博士生导师。

方法,其中包含了网页分块提取链接的方法及相关链接提取方法;第3章主要介绍了相关链接提取方法的测试结果;第4章结论中分析了相关链接提取的不足及针对不足提出的改进方案。

## 2 相关链接提取方法

本文仅针对主题型网页<sup>[9]</sup>,提出相关链接提取方法。首先对网页按照<table>标签分成若干块,从块中提取链接,形成若干链接块。其次,使用规则与统计相结合的方法,提取相关链接块。

### 2.1 链接分块提取

通常人们浏览网页的时候会发现,整个页面被分割成若干块,且同一块包含内容相似。如,相关链接、正文、公司版权链接、导航链接等信息都各成一块。通过对来自新浪、搜狐、网易、雅虎等多个门户网站的网页源文件分析,发现各大网站在对网页结构进行布局时基本上都用到了HTML语言中的容器标签<table></table>,将相似的内容放到同一个table结构中。利用上述网页构造的特点,本文使用<table></table>标签对网页线性分块,将夹在一对<table></table>中间的网页源文件看作一块,形成若干信息块,无需对<table>内的其他标签及内部嵌套<table>标签进一步分析。本文的相关链接提取方法,主要分析链接文本内容以判别相关链接,所以,网页分块时忽略了块间的层次及位置关系,但依然可以得到较好的结果。

按照网页中的容器标签<table>将网页分成若干块后,从网页分割后的若干信息块中提取链接形成许多链接块。链接提取中,一个链接对应网页HTML源文件中一个<a href=...>,从形如<a href="http://sports.163.com/"class="nav">体育</a>的HTML文件中提取链接。保存http://sports.163.com/作为链接URL,保存“体育”作为链接文本,最终得到许多链接块。

本研究过程中,从网上随机下载约10 000主题型网页,统计发现图片链接比例不足10%,所以在研究中,作者仅提取含有链接文字的链接,而未考虑图片链接。

### 2.2 相关链接提取

#### 2.2.1 相关链接特征及表示

##### 2.2.1.1 相关链接特征

通过对大量的、不同网站的、含有相关链接的、主题型网页分析发现,相关链接存在一些共同的特征。综合起来归纳为以下几点:

- (1)相关链接是成块出现的;
- (2)相关链接的链接文字的长度是有规律的,一般5~30字;
- (3)相关链接中链接文字一般不会出现某些词,如“首页”,“导航”等,本文中不出现在相关链接中的特殊词称为相关链接停用词;
- (4)相关链接的URL一般为站内链接,且格式较为规整,一般不包含“javascript”等特殊URL,这里将该类一般不出现在相关链接中的特殊URL称为相关链接停用URL;
- (5)相关链接的链接文字与其所在页面的标题一般具有相同的关键词;
- (6)某些网页包含相关链接开始标志,如“相关链接”、“相关主题”、“相关报道”等。

##### 2.2.1.2 规则

链接块中每个链接的相关属性值用一个四元组( $Len, LT, LU, NUM$ )表示。

(1) $LEN$ 表示链接文本的长度(以Byte为单位);

(2) $LT$ 表示链接文本是否包含相关链接停用词,若包含取值为真,否则为假;

(3) $LU$ 表示链接URL是否包含相关链接停用URL,若包含取值为真,否则为假;

(4) $NUM$ 表示链接文本与其所在页面标题含有相同的词数。

另外,由于不同的网站相关链接开始标志并不相同,如用“相关专题”、“相关链接”、“相关文章”、“相关报道”和“相关话题”中的一个或几个;而还有些网站并不包含相关链接开始标志,所以为了提高本文相关链接提取算法的通用性,未考虑相关链接开始标志这一特征。实验表明,该算法仍然可以得到良好的效果。

针对上述特征,预先设定规则的方法进行相关链接提取,提出若干否定规则用于排除非相关链接块,定义主要谓词公式,用于半形式化描述规则。

LinkBlock( $x$ ): $x$  is link block (1)

Link( $y$ ): $y$  is link (2)

Attribute( $x, y$ ): $y$ 是 $x$ 的属性值 (3)

规则 1

For each Link( $y$ ) in LinkBlock( $x$ ),IF Attribute( $y, Len$ )<10 than count++;

IF the value of count greater than 2,Then  $\neg$  RelationLink( $x$ )

规则 2

For each Link( $y$ ) in LinkBlock( $x$ ),IF Attribute( $y, LT$ ) is TRUE, Then  $\neg$  RelationLink( $x$ );

规则 3

For each Link( $y$ ) in LinkBlock( $x$ ),IF Attribute( $y, LU$ ) is TRUE, Then  $\neg$  RelationLink( $x$ );

规则 4

For each Link( $y$ ) in LinkBlock( $x$ ),Sum+=Attribute( $y, NUM$ ), count++;

IF Sum/count Less than T,Then  $\neg$  RelationLink( $x$ );

上述规则中用到的相关链接停用词及相关链接停用URL词表,由手工采集不同网站相关信息获得。词表采集过程中作者发现,很多网站在建设过程中一般都应用模板,为了更快地生成网页及更好地进行网站维护,同时使得同一网站的大部分网页在整体风格上保持一致。同一网站几乎所有网页都包含导航链接、功能链接、公司版权等完全相同信息,而且一般使用的词都比较固定,这类词一般不会出现在相关链接文字中。如新浪网中的“首页”、“联系我们”、“发表评论”、“打印”。利用上述特点,大大简化了词表建立的工作量。然而,不同的网站在表达同一链接内容时使用具体的词又存在一定的差别,如用来发表评论的链接文本,在新浪网中对应链接文字“发表评论”,而搜狐网中则用“我来说两句”。网络文本内容的多样化,也给作者建立词表带来了很大的难度。所以,建立相关链接停用词表时应尽可能保持高的覆盖率,同时又要保证词表的精确率,以免将相关链接块错误地排除。

相关链接停用URL中包含少量的链接URL,“Javascript”,“mail to”等。

##### 2.2.2 相关链接提取方法

本文提出了一种通用的相关链接提取方法。对网页分块提取链接,形成若干链接块作为该算法的输入。通过计算链接块中每个链接的 $Len, LT, LU, NUM$ 值,应用上述规则,提取相关

链接。详细算法参见图 1。

```

1. Words of page title segmentation and initialize array Even
2. For Each Link Block
  { Sum=0; // Save the sum of the value of all links' NUM in current block
    Count=0; // Save the number of links in current block
    Second=False;
    For each link in current Block
      {
        Calculate the values of variables Len, LT, LU and NUM of link
        If (LT is True || LU is True || (Len < 10 && Second)) // Using rules 1, 2, 3
          Then { Sum=0; Count=1; Continue; }
        Else
          IF (Len < 10) Then Second=True;
          Else { Sum+=NUM; ++count; }
      }
    Save the value of Sum/count into Even [i];
  }
3. Select the max value from array Even and is as Even[Max]
4. IF Even[Max] > T Then It is Relation block // using fourth rule
   Else The page has no Relation block

```

图 1 相关链接提取算法

对相关链接提取算法的几点说明：

(1) 该算法的输入

网页按块分割后，从中提取链接，形成的若干链接块。此处已经利用相关链接成块出现的这一特征。

(2) *Len*、*LT*、*LU*、*NUM* 值的计算

*Len* 表示链接文字的长度，由于本算法主要针对中文网页，链接文字为中文，一个中文要占 2 Byte，所以 *Len* 值为链接文字个数的 2 倍。*LT*、*LU* 的计算，将当前链接文字与链接 URL 分别与已建立的相关链接停用词表、相关链接停用 URL 词表进行匹配，如匹配则 *LT*、*LU* 为真，否则为假。*NUM* 值的计算，链接文字分词与该链接所在页面标题分词后的结果匹配，相同词的个数作为 *NUM* 的值。

### 3 实验结果

#### 3.1 实验测评标准

首先给出该算法的评估结果中准确率和召回率的定义：

准确率是从测试的页面中提取正确相关链接数目除以提取的所有的相关链接数目，表示为公式(4)：

$$\text{准确率 (precision)} = \frac{\text{提取正确的相关链接总数}}{\text{提取的相关链接总数}} \quad (4)$$

召回率是从测试的页面中提取正确相关链接数目除以测试页面中含有的所有相关链接总数，如公式(5)：

$$\text{召回率 (recall)} = \frac{\text{提取正确的相关链接总数}}{\text{页面中的相关链接总数}} \quad (5)$$

准确率和召回率反映了算法有效性的两个不同方面，两者应该综合考虑，不可偏废，因此，采用评估指标 *F* 测试值，表示如下：

$$F \text{ 测试值} = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \quad (6)$$

#### 3.2 相关链接域值确定实验及结果

第 2 章相关链接提取算法中用到了阈值，该值通过初步观察及多次实验确定。实验选取新浪 IT 领域 2 763 个网页。首先，对每个页面统计与页面标题含相同词均值最大的链接块的值。观察统计数据，初步确定了阈值为 0.3，在 0.3 附近取值进

一步实验，得到不同结果(表 1)。本文依据 *F* 值测度，最终确定了阈值为 0.25。由表 1 可以看出，阈值越大精确率越高，召回率越低，反之亦然。因此，可以根据不同需要调整阈值。

表 1 不同阈值得到的实验结果

| 程序提取链数 | 实际链接总数 | 错识别链接数 | 漏识别链接数 | 准确率     | 召回率    | <i>F</i> 值 | 域值   |
|--------|--------|--------|--------|---------|--------|------------|------|
| 11 504 | 15 671 | 1 061  | 5 228  | 90.777% | 66.64% | 0.768 5    | 0.35 |
| 11 948 | 15 671 | 1 140  | 4 863  | 90.458% | 68.97% | 0.782 6    | 0.32 |
| 12 079 | 15 671 | 1 163  | 4 755  | 90.372% | 69.66% | 0.786 7    | 0.30 |
| 12 118 | 15 671 | 1 165  | 4 718  | 90.386% | 69.89% | 0.788 3    | 0.29 |
| 12 989 | 15 671 | 1 622  | 4 304  | 87.512% | 72.53% | 0.793 2    | 0.25 |

#### 3.3 相关链接算法比较实验及结果

本文分别将文献[10]中开发的超链接分类系统原型 ClassifyLinks 中利用相关链接开始标志提取相关链接算法和本文提出的基于链接分块思想的相关链接通用提取算法对新浪和 Tom 两个不同网站的网页相关链接提取进行了对比测试，针对新浪网站带有明显的相关链接开始标志的 2 763 个网页两种方法的相关链接提取结果见表 2，针对 Tom 网站不带有明显的相关链接开始标志的 1 021 个网页两种方法的相关链接提取结果见表 3。

表 2 针对 2 763 个新浪网页(带有明显的相关链接开始标志)的测试结果

|            | 提取链接数  | 实际链接总数 | 错识别链接数 | 漏识别链接数 | 准确率     | 召回率    | <i>F</i> 值 |
|------------|--------|--------|--------|--------|---------|--------|------------|
| 利用相关链接开始标志 | 13 228 | 15 671 | 753    | 3 196  | 94.307% | 79.61% | 0.863      |
| 链接分块思想提取算法 | 12 079 | 15 671 | 1 163  | 4 755  | 90.372% | 69.66% | 0.787      |

表 3 针对 1 021 个 Tom 网页(不带有明显的相关链接开始标志)的测试结果

|            | 提取链接数 | 实际链接总数 | 错识别链接数 | 漏识别链接数 | 准确率     | 召回率     | <i>F</i> 值 |
|------------|-------|--------|--------|--------|---------|---------|------------|
| 利用相关链接开始标志 | 3 612 | 7 344  | 626    | 4 358  | 82.669% | 40.659% | 0.545      |
| 链接分块思想提取算法 | 6 034 | 7 344  | 850    | 2 160  | 85.913% | 70.588% | 0.775      |

并且，又选取了 Sohu、163、Tom 等不同网站 3 477 个页面对基于页面链接分块思想的相关链接通用提取算法进行了通用性测试，其中 Sohu 网页 1 215 个，Tom 网页 1 040 个，163 网页 1 222 个，测试结果见表 4：

表 4 针对 Sohu、163、Tom 等不同网站 3 477 个页面的测试结果

| 程序提取链数 | 实际链接总数 | 错识别链接数 | 漏识别链接数 | 准确率     | 召回率     | <i>F</i> 值 | 域值   |
|--------|--------|--------|--------|---------|---------|------------|------|
| 13 163 | 17 104 | 1 709  | 5 650  | 87.017% | 66.967% | 0.757      | 0.25 |

从测试结果可以看出，利用相关链接开始标志的相关链接提取算法在提取类似新浪网网页中带有明显的相关链接开始标志的相关链接时，准确率和召回率都很高，但是在提取类似 Tom 网网页中不带有明显的相关链接开始标志的相关链接时召回率明显下降，说明很多相关链接被漏识别，那是因为很多网页中不带有相关链接开始标志，所以这种依赖相关链接开始标志的相关链接提取方法通用性比较差，只适合某些特定的网站页面中的相关链接提取。而基于页面链接分块思想的相关链接通用提取算法在开放测试中准确率始终在 85% 以上，召回率

保持在 70%左右,不受其它条件限制,通用性比较强。

#### 4 结语

上文中实验结果表明,本文提出的算法的召回率还不够高。主要原因有以下几点:

(1)本文中相关链接提取算法仅考虑了相同词而未考虑近义词,是造成算法召回率下降的主要原因。

(2)按块提取链接中用<table>对网页分块,由于 HTML 语法的灵活性,并非所有网页都利用<table>进行网页结构划分,所以可能导致网页分块结果的偏差。

(3)有些相关链接,本身与网页主题并不相关,可能是近期热点话题推荐,或出于某种商业目的而被加入,是一种人为加入。

(4)本文算法统计相同词数时,使用的分词程序处理新词能力较弱,而链接文字包含大量新词,在一定程度上影响本文算法的精度。

针对上述不足,后续主要针对(1)(2)进行改进。对(1)的改进,考虑到相关链接文字与网页标题相似度的计算,主要使用 HOWNET 或同义词词林,从语义上计算二者的相似度,以提高精度。针对(2)的改进,采用不同的方法对网页分块后,提取相关链接,与本文结果比较实验。目前,已经将该算法应用于网页净化,准备进一步实验将其用于其它研究领域。

(收稿日期:2006年1月)

#### 参考文献:

[1] SHIH L K, KARGER D R. Using URLs and table layout for web classification tasks[C]//Proceedings International WWW Conference.

(上接 100 页)

高,并且随着损失因子  $\lambda(a_2, c_1)$  取值的增大,准确率得到进一步的提高。由此可见,最小风险贝叶斯方法减少了由正常邮件被误判为垃圾邮件所带来的损失。

同时,随着损失因子的增大,查全率相对有所下降。查全率与准确率之间存在一种相辅相成的关系。从用户的角度来说,对准确率的要求通常比较高,而对于有几封垃圾邮件没有被过滤掉的接受度相对较宽。因此,在实际应用中,可以根据实际情况对  $\lambda(a_2, c_1)$  的取值进行调整。

由图 5 仍可知, HM 模型的准确率增长幅度略低于 MM 模型,而查全率明显都优于 MM 模型。因此,综合考虑查全率与准确率两个方面,作者认为基于 HM 模型的最小风险贝叶斯算法在邮件分类问题上得到较为优秀的效果。

#### 5 结束语

本文应用基于混合模型的最小风险贝叶斯方法构造邮件分类器,并根据中文邮件的特点进行分词、特征提取等预处理工作。通过在不同结构的中文邮件数据集上进行实验的结果表明,此种方法在整体上提高了邮件分类器的性能。目前,该分类模型在实际的邮件服务上应用,取得了较好的效果。

New York, USA:[s.n.], 2004.

[2] SONG Rui-hua, LIU Hai-feng, WEN Ji-rong, et al. Learning block importance models for web pages[C]//Proceedings International WWW Conference. New York, USA:[s.n.], 2004.

[3] LIN S-H, HO J-M. Discovering informative content blocks from web documents[C]//The Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(SIGKDD'02), 2002

[4] CHEN J, ZHOU B, SHI J, et al. Function-based object model towards website adaptation[C]//Proceedings of the 10th World Wide Web conference(WWW10). Budapest, Hungary:[s.n.], 2001.

[5] 张志刚,陈静,李晓明.一种 HTML 网页净化方法[J].情报学报,2004, 23(4)

[6] CHAKRABARTI S, BERG M van D, DOMC B. Focused crawling: a new approach to topic-specific Web resource discovery[C]//Proceedings of the Eighth International World Wide Web Conference. Toronto, Canada:[s.n.], 1999.

[7] FLAKE G W, LAWRENCE S, GILES C L, et al. Self-organization and identification of web communities[J]. Computer, 2002, 35(3): 66-71.

[8] KOVACEVIC M, DILIGENTI M, GORI M, et al. Recognition of common areas in a Web page using visual information: a possible application in a page classification[C]//The Proceedings of 2002 IEEE International Conference on Data Mining(ICDM'02). Maebashi City, Japan:[s.n.], 2002.

[9] CAI D, YU S, WEN J-R, et al. VIPS: a vision based page segmentation algorithm, Microsoft Technical Report, MSR-TR-2003-79[R]. 2003:210.

[10] 王华兵. Web 页面链接信息抽取与分类的研究[D]. 哈尔滨工业大学, 2004.

(收稿日期:2006年4月)

#### 参考文献:

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2004.

[2] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997:412-420.

[3] SAHAMI M, DUMAIS S, HECKEMAN D, et al. A Bayesian approach to filtering Junk E-mail[C]//Proceedings of AAAI-98 Workshop on Learning for Text Categorization, 1998.

[4] RENNIE J D M, SHIH L L, TEEVAN J, et al. Tackling the poor assumption of Naïve Bayes text classifiers[C]//Proceedings of ICML-03. Washington DC, USA:[s.n.], 2003:616-23.

[5] LIN Y P, CHEN Z P, YANG X L, et al. Mail filtering based on the risk minimization bayesian algorithm[C]//Proceedings Industrial Systems and Engineering III, 6th World Multi-conference on Systemics, Cybernetics and Informatics(SCI 2002), 2002:282-285.

[6] MITCHELL T M. 机器学习[M]. 曾华军, 张银奎, 译. 北京: 机械工业出版社, 2003.

[7] Sendmail.org[Z/OL]. <http://www.sendmail.org>.