Journal of Computer Applications
计算机应用,2012,32(8):2360 -2363,2368
ISSN 1001-9081
CODEN JYIIDU
2012-08-01
http://www.joca.cn

文章编号:1001 -9081(2012)08 -2360 -04
doi:10.3724/SP.J.1087.2012.02360

面向导航型网页关键词自动抽取的视觉模型与算法

彭　浩[1*],蔡美玲[1,2],陈继锋[1],刘　炽[3],余炳锐[1]
(1.湖南涉外经济学院 计算机科学与技术学院,长沙 410205;　2.中南大学 信息科学与工程学院,长沙 410083;
3.中国电力出版社 用电技术出版中心,北京 100005)
(*通信作者电子邮箱 penghaobox@yahoo.com.cn)

摘　要:导航型网页中往往包含了大量的噪声信息,为自动提取网页中的关键词带来了较大的困难。为此,提出一个新的网页表示模型 PIX-PAGE 和导航型网页关键词自动抽取算法 P-KEA。PIX-PAGE 模型利用提出的区域合并算法,将一张网页分割为适当粒度的区域;然后,依据人类视觉特点,对各区域进行视觉"奇异性"量化,同时利用奇异性传递规则进一步强化关键词相关区域的视觉"奇异性"。P-KEA 根据 PIX-PAGE 模型模型的视觉量化结果,能够较准确地找到视觉突出区域中的关键词。实验结果表明,与基于 DocView 模型的算法 DVM 相比,P-KEA 的准确率平均提高了 20.9%。

关键词:区域合并;视觉量化;网页表示模型;关键词自动抽取
中图分类号: TP391.4　　**文献标志码**:A

Visual representation model and automatic keywords extraction algorithm for hub Web pages

PENG Hao[1*], CAI Mei-ling[1,2], CHEN Ji-feng[1], LIU Zhi[3], YU Bing-rui[1]
(1. *College of Computer Science and Technology, Hunan International Economics University, Changsha Hunan 410205, China*;
2. *School of Information Science and Engineering, Central South University, Changsha Hunan 410083, China*;
3. *Power Technology Publishing Center, China Electric Power Press, Beijing 100005, China*)

Abstract: It is very hard to exactly extract keywords from hub Web pages because of its topic noise. To resolve this problem, a new sub Web page representation model and its automatic keywords extraction algorithm were proposed in this paper. At first, the new model segmented Web page into some blocks by using the block composition algorithm. Secondly, according to the visual recognition method of humanity, the new model computed the visual measurement of these blocks. At the same time, the transmission rule of visual measurement made blocks special where keywords were contained more specially. The automatic keywords extraction algorithm could exactly find these keywords in the most special blocks. The experimental results show that the proposed algorithm has bumped up by 20.9% on average in accuracy compared with keywords extraction algorithm based on DocView model.

Key words: block composition; visual characteristic measurement; Web page representation model; automatic keywords extraction

0　引言

聚焦爬虫[1]首先根据网页的位置预测主题相关性,再分析它的内容相关性,并反馈到预测模型中[2]。内容相关性的分析是基于主题表示来进行的[3]。关键词[4-5]集合是主要的主题表示形式之一[6]。因此,关键词的自动抽取成为聚焦爬虫研究的关键问题之一。针对主题型网页[7]或普通文本的关键词自动抽取,已经进行了较多的研究。

导航型网页中包含了大量的超链接,它的功能相当于书籍中的目录或者各章导读。然而,囿于商业利益的驱动,导航型网页中可能包含了大量的广告、友情链接、宣传等不相关的噪声信息,为自动提取网页中的关键词带来了较大的困难。对于主题型网页而言,它的关键词隐藏在内容文本之中,比如,频繁出现的词汇[8],或者中心度较高的词汇[9]。因为目标的不同,导致导航型网页的内容文本以超链接的锚文本为主,主题信息非常分散,从这些文本中提取的关键词很难应用于网页的主题描述。因此,导航型网页的关键词定义不能直接沿用主题型网页的方法。

网页文本的关键词自动抽取以页面表示模型为分析基础。现有的表示模型有两类,一种把整个网页看作一些内容块的结构化表示,称之为结构模型,比如语义文本单元(Semantic Textual Unit,STU)[10]、基于文本语义的文本对象模型(Semantic Textual Unit-Document Object Model, STU_DOM)[11]、基于视觉的分块方法(VIsion-based Page Segmentation,VIPS)[12]和 DocView[7]等;另一种则直接把网页看作一个词汇网络,称为网络模型,比如文献[9]。不同的结构模型会把网页看作不同的语义结构。STU、STU-DOM 和 VIPS 模型都是为网页的信息抽取而提出的,因此,它们并没有充分注意到关键词的特征表示。DocView 是面向信息检索和关键词抽取而提出的,在保持网页原有的文本对象模型

收稿日期:2011-11-11;**修回日期**:2012-01-31。　**基金项目**:国家自然科学基金资助项目(60803024);湖南省自然科学基金资助项目(10JJ6092);湖南省大学生研究性学习和创新性实验计划项目(湘教通[2011]272 号,编号:393)。

作者简介:彭浩(1978 -),男,湖南长沙人,讲师,硕士,主要研究方向:Web 信息获取与处理、实时调度;　蔡美玲(1982 -),女,湖南长沙人,讲师,博士研究生,主要研究方向:网络计算、智能信息处理、图形图像信息处理;　陈继锋(1966 -),男,湖南浏阳人,教授,博士,主要研究方向:软件测试自动化;　刘炽(1979 -),男,湖南沅陵人,工程师,硕士研究生,主要研究方向:软件工程、嵌入式系统;　余炳锐(1988 -),男,湖南怀化人,主要研究方向:Web 信息获取与处理。

(Document Object Model,DOM)结构的同时,还注意到了关键词的字体视觉特征。不过,其视觉特征量化和权值传递规则存在较大的局限性。文献[9]则把一个网页看作由词汇构成的复杂网络,它对关键词的视觉特征计算与 DocView 一样。

我们注意到,一些书籍会将各章的主要内容概括性地描述出来并放在章首。导航型网页也借鉴了这种方法:利用一些概括性较强的小标题来引导读者浏览网页。因此,这些小标题恰好反映了导航型网页的主题。同时,这些小标题往往具有某种或多种视觉上的"奇异性"。基于这个发现,提出了一个新的 Web 页面视觉表示模型 PIX-PAGE(PIXel Web PAGE):首先,完整地抽象出了网页中的视觉特征;其次,通过独特的视觉量化方法和重要性传递规则,来表达并强化这些包含了主题关键词的视觉"奇异点"。借鉴了人类善于发现网页中"奇异点"的特点,PIX-PAGE 具有内在的抗干扰能力,是更一般的网页表示模型。还提出了区域合并算法。同时,在 PIX-PAGE 的基础上,针对导航型网页提出了一个新的关键词自动抽取算法 P-KEA(Pixel-page model based automatic Keywords Extraction Algorithm)。实验表明,算法 P-KEA 较大地提高了关键词自动获取的准确率

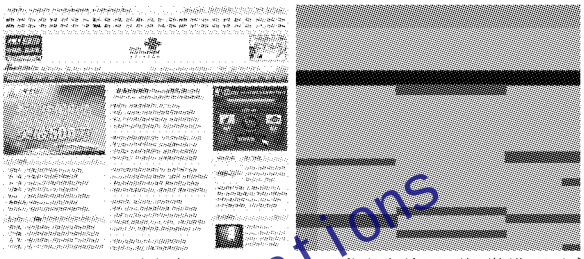
1　Web 网页的视觉表示模型 PIX-PAGE

为了让读者快速了解各章的内容,一些书籍会在各章的首页提供导读部分:用一些概括性的词汇描述本章的主要相关内容。导航型网页也借鉴了这种方法:利用各种分栏小标题来反映网页的主题分类特征。另一方面,网页作者在这些小标题上还施加了各种突出显示的视觉设计手段,使得读者一眼就能看出网页的主题是什么。因此,面向主题分类的网页关键词就隐藏在这些以分栏小标题为主的"关键区域"之中。基于这些发现,提出了一个新的视觉表示模型——PIX-PAGE来表达并强化这些视觉"奇异点"对于主题表达的重要性。

1.1　视觉模型 PIX-PAGE

人类要判断一个网页,特别是导航型网页的主题,比如,通过一个网站的首页来了解这个网站是否属于某类主题,往往就会从一些突出显示的区域(称之为视觉"奇异点")来判断,如页面头部的导航条,特别是网页中的各级分栏子标题。快速地找到这些视觉"奇异点"是人类的特长。人类能够很容易地从一张网页中找到它的关键词,正是由于关键词在网页中具有突出的视觉特征,比如,特别的背景、色彩、字体、大小和在页面中的位置等。而人类在快速定位到这些视觉"奇异点"时,往往忽略了"大众化"的视觉部分,而那些少数的、与众不同的区域被人类敏锐地捕捉到了。这至少说明了两个问题:1)如图 1(a)所示,少量的视觉"奇异点"区域与网页的主题分类特征具有很大的相关性,"大众化"区域却是这些主题的细化,因此,它们并没有概括性地表达主题,视觉特征为抽取网页中的关键词提供了非常重要的提示信息;2)如图 1(b)所示,人类往往会忽略那些在视觉上不太突出的区域,而重点表达并强化这些视觉"奇异点"区域。视觉表示新模型 PIX-PAGE 借鉴了这两个发现。需要说明的是,人类能够容易地识别出网页中的广告部分,因此,对于一般的网页读者来说,会将这些广告部分弱化为普通区域,故在图 1(b)的网页视觉模型中,"亮丽"的广告部分被表达为普通的灰色区域。

首先,PIX-PAGE 将一个网页近似地表达为一张平面图,它包含了多个不同的区域,一些区域较为"普通",而另一些区域则很"特别",对于读者来说,这些普通区域正是那些不需要特别关注的"大众化"区域,而特别的区域则对应于视觉"奇异点"。在 PIX-PAGE 的表示中,一张网页的视觉模型由若干个可视区域构成,这些可视化区域分为普通区域和特殊区域。PIX-PAGE 将网页的可视 DOM 叶子节点作为基本视觉子区域,这些基本区域还可以合并为更大的、级别更高的子区域。于是,PIX-PAGE 是一个包含多层嵌套结构的模型。在第一层中,每个子区域都是单个的基本子区域,而在更高的层次表示中,某些子区域肯定还包含了更多的基本子区域或组合子区域。因此,PIX-PAGE 可以形式化地表示为:PIX-PAGE = $\langle \phi, R, U \rangle$,其中,$U$ 是由所有基本子区域构成的集合,R 是子区域的合并规则,ϕ 是由同一层组合子区域构成的集合,即某个层次上的平面图。因此,$\phi = \langle \phi', R, U \rangle$,其中,$\phi'$ 是 ϕ 的下层平面图。



(a) 网页　　(b) 人类眼中的网页视觉模型示意

图 1　网页及人类眼中的视觉模型示意

(颜色越深表示越容易被注意到)

其次,基本子区域的合并规则是依据它们的 DOM 路径相似性来进行的,将某些"相邻"的区域进行合并。如何定义"相邻"关系?在图 1(b)中的第二块灰色区域对应于网页中的广告区域,它与下面的第三块区域在物理位置上是相邻的。但是,从主题分类的角度看,它们是完全不同的区域,不能合并,因此它们是不"相邻"的。对于图 1(b)中左下角部分的深灰色区域来说,它是由几个不同的子区域合并而成的,因此它们是"相邻"的。所以"相邻"关系不能简单地以物理位置来定义,应该以 DOM 路径的相似性来定义。这也正是网页作者在进行设计时的初衷——区域划分是用 HTML 标签来进行分割的。为了给出"相邻"关系的形式化定义,首先,定义两个区域 B_i 和 B_j 之间的距离为

$$dis(B_i, B_j) = |P_i - P_j| \tag{1}$$

其中 P_i 和 P_j 分别表示区域 B_i 和 B_j 在 DOM 树中到达共同祖先所经过的父节点数。比如,在同一个标签 < tr > 下的 < td > 和 < td > 所代表的区域之间的距离为 0。所谓"相邻"是指它们之间的距离小于某个阈值 δ,即 $dis(B_i, B_j) < \delta$,在 PIX-PAGE 中,只有相邻的区域才能合并为更高级的组合区域。

再次,如何量化这些基本区域的视觉"奇异性"。从图 1(a)可以看出,网页作者在视觉"奇异点"上使用了不同的背景图片或背景颜色,不同的字号、字体,不同的字体颜色,还有特殊的位置等。因此,这些方面将是刻画一块区域的视觉"奇异性"的重要因素。如何量化每一个因素的视觉"奇异性"?我们注意到,视觉"奇异性"也就是它在整张网页中的"特殊性",或者"稀少性"。如果用"稀少性"来刻画区域的视觉"奇异性",则某个因素的"奇异性"可定义为:页面总面积/具有当前属性的区域面积。也可以定义为:1/具有当前属性的区域面积。因此,PIX-PAGE 将某个区域 i 的各视觉要素的视觉"奇异性"定义如下:

1)背景颜色的"奇异性"(将具有背景图片的区域直接定

义为某个较大的常数)：

$$B_i^{bg} = \frac{1}{\text{整张网页中具有当前背景颜色的区域的面积总和}} \quad (2)$$

2)字号的“奇异性”：

$$B_i^{fs} = \frac{1}{\text{整张网页中具有当前字号的区域的面积总和}} \quad (3)$$

3)字体的“奇异性”：

$$B_i^{fn} = \frac{1}{\text{整张网页中具有当前字体的区域的面积总和}} \quad (4)$$

4)重量的“奇异性”：

$$B_i^{fw} = \frac{1}{\text{整张网页中具有当前字体重量的区域的面积总和}} \quad (5)$$

5)字体颜色的“奇异性”：

$$B_i^{fg} = \frac{1}{\text{整张网页中具有当前字体颜色的区域的面积总和}} \quad (6)$$

从上述定义可以看出，PIX-PAGE 比 DocView 模型更具一般性。因为，DocView 中的“重要标签”只是 PIX-PAGE 中的字体重量这一个视觉要素，其他的要素没有考虑到。在导航型网页中，关键区域的视觉“奇异性”包含了更加丰富的设计手段。

区域视觉“奇异性”量化的另一个问题是，如何来综合这些因素的视觉“奇异性”而最终得到一个区域的视觉“奇异性”。研究发现，在这些因素中，并非每个因素对区域的整体视觉“奇异性”的贡献是同等重要的。因此，将这些因素进行加权求和，得到区域的整体视觉“奇异性”。具体讲，第 i 块区域的视觉“奇异性” B_i 就是

$$B_i = \alpha_1 B_i^{bg} + \alpha_2 B_i^{fg} + \alpha_3 B_i^{fn} + \alpha_4 B_i^{fs} + \alpha_5 B_i^{fw} \quad (7)$$

其中加权系数 α 可以根据具体的情况取不同的形式，比如，$\alpha_i = a^{k_i}(1 \leqslant k_i \leqslant 5, a = 2,4,8,\cdots)$，可以通过设置适当的加权系数来过滤掉一些友情链接和广告宣传等主题噪声。这些类型的主题噪声一般最多只是通过某种字体颜色来引起读者的注意，可以弱化前景的“奇异性”的贡献度，即减小 α_2 的值。

1.2 重要性传递规则

如前所述，假设：导航型网页中与主题分类相关的关键词主要分布于视觉“奇异点”之中。因此，PIX-PAGE 就直接用视觉“奇异性”来代表相应区域的主题相关的重要性。视觉“奇异性”大于某个阈值的区域称为关键区域，否则称之为普通区域。在 PIX-PAGE 的重要性传递规则中，遵循的是“强者愈强、弱者愈弱”的基本思想。在 PIX-PAGE 的每一层表示中，不同的区域之间的重要性不相互传递，仅在子区域与父区域之间进行传递。如图 2 所示，0 号和 1 号区域是一个关键区域，2 号区域则是一个普通区域。不过，它们可以合并为一个更高层次的组合区域。对于区域的重要性传递规则有两个问题需要解决：1)组合区域的重要性如何定义；2)组合区域与各个子区域之间的重要性如何传递。

如图 2 所示，当把三个区域合并为更高层次的组合区域时，根据重要性传递规则的基本思想，组合区域的重要性应该包含其中各个关键区域的重要性。另一方面，组合区域与各个子区域之间的重要性传递规则应该加强关键子区域的重要性，弱化普通子区域的重要性。因此，PIX-PAGE 的传递规则如下：

规则 1　将组合区域 B_i 的重要性 w_i 定义为其包含的所有普通子区域 b_{ik} 的重要性的最大值与各关键子区域 b_{ij} 的重要性 w_{ij} 之和，即 $w_i = \max_k\{w_{ik}\} + \sum_j w_{ij}$。

规则 2　关键子区域 b_{ij} 从父区域 B_i 获得的重要性增量 $\Delta w_{ij} = \frac{w_{ij}}{w_i} \times w_{ij}$。

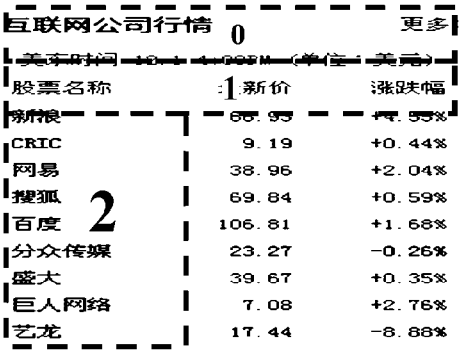


图 2　实例示意图

以图 2 为例，假设子区域 0、1 和 2 的初始重要性分别为 $w_0 = 5, w_1 = 2, w_2 = 1$。那么，经过一次传递后，$w_0 = 5(1 + 5/(5+2+1)) = 8.125$，$w_1 = 2(1 + 2/(5+2+1)) = 2.5$。它们的重要性初始差值为 3，而经过一次传递后，它们的重要性差值变为 5.625，大大地增加了。这说明 PIX-PAGE 的传递规则确实反映了“强者愈强，弱者愈弱”的基本思想。由规则 1 可知，完全由普通区域合并而成的更高层的组合普通区域，它的重要性改变会很小。根据规则 2 可知，一个关键子区域的重要性越大，它从父区域获得的重要性增量就会越大，因此，它的重要性得到了进一步的加强；反之，重要性越弱的子区域获得的重要性增量就会越小，从而，使得它的重要性相对地就被减弱了。而且，普通子区域根本就没有从父区域处得到重要性增量，因此，它的重要性被进一步地相对弱化了。

1.3 区域合并算法

在 PIX-PAGE 中，合并操作并不会无限地递归进行，否则，最后将整个网页合并为一个组合区域，对关键词自动抽取来讲就没有任何意义了。因此，在给定一个路径相似性阈值 δ 的前提下，利用式(1)计算出各个基本区域之间的路径相似性，然后将所有相邻的基本区域进行合并操作，最后，得到一个最高层的视觉表示模型。关键词自动抽取算法就在最高层的表示模型上进行。具体的合并算法如下：

输入：基本区域集合 $U = \{b_0, b_1, b_2, \cdots, b_{N-1}\}$，即 PIX-PAGE 的第 1 层表示模型；相邻关系阈值 δ；

输出：PIX-PAGE 的顶层表示模型，即组合区域集合 $\phi = \{B_0, B_1, \cdots, B_M\}$。

```
1)  for i = 0 to N - 1 do                          /* 计算距离 */
2)    for j = i to N - 1 do
3)      根据式(1)计算 b_i 和 b_j 的距离 s_ij
4)    end for
5)  end for
6)  while (| U | > 0) {                            /* 合并操作 */
7)    初始化一个空的组合区域 B_k
8)    从 U 中取出一个基本区域 b_i
9)    将 b_i 加入 B_k 中，且从 U 中删除它
10)   for U 中每一个基本区域 b_j do
11)     if (s_ij < δ) then
12)       将 b_j 加入 B_k 中
13)       从 U 中删除 b_j
14)     end if
```

15)　end for

16) 将 B_k 加入到 ϕ 中}

2　基于 PIX-PAGE 的关键词自动抽取算法

DocView 仅仅使用了部分的视觉信息来强化一些特殊区域,而 PIX-PAGE 则根据人类定位关键区域的基本原理,包含了更加完整而全面的视觉特征,以强化关键区域,是更为一般的视觉表示模型。

利用 PIX-PAGE 表示模型进行网页的关键词自动抽取,首先需要将网页的 HTML 代码转换为 DOM 表示模型,然后从中提取可视叶子节点,作为 PIX-PAGE 的基本区域。这就是 PIX-PAGE 的第一层平面图表示。根据式(2)~(7)量化第一层表示模型的各个基本区域的视觉“奇异性”。五个视觉要素是存在某种视觉优先级的,要根据优先级来合理设置它们的加权系数 α_i。另外,对于字体重量视觉要素的“奇异性”是根据 CSS 规则的 font-weight 属性值,以及加粗标签,比如 , <hn>(n = 1,2,…,6)等来计算的。根据各个基本区域的视觉“奇异性”,即它的主题重要性来判定它是否是关键区域。

其次,就是合并基本区域;接着,根据重要性传递规则,计算各个组合区域的重要性,以及各个子区域从父区域处获得的重要性增量。

Web 页面作者可能没有丰富的视觉设计知识,因而,可能没有在视觉信息上提供足够的主题相关性。因此,不能只考虑视觉“奇异性”,还得加上一些其他的因素。一般来讲,与主题相关的关键词会被反复提及。加入词频可以将那些视觉上不太突出的区域的主题相关性得到合理的加强。因此,在 P-KEA 看来,一个关键词的主题相关性不仅包括它在 PIX-PAGE 表示模型中的视觉“奇异性”,还包括它在文本中出现的频率。

设网页中的总词汇数为 N,记关键词 i 出现在 PIX-PAGE 中的 m 个基本区域中,在基本区域 j 处的词频为 f_{ij},区域 j 的重要性为 w_j,那么,它的主题相关性即为

$$t_i = \sum_{j=1}^{m} \frac{f_{ij}}{N} \times w_j \tag{8}$$

算法的最后一步要计算出各个词汇的主题相关性,并进行降序排列,选出前面的若干个词汇作为网页的关键词集合。算法的具体描述如下:

输入:HTML 网页 p;

输出:关键词集合。

1)将网页 p 解析为 DOM 树;

2)构造 PIX-PAGE 模型;

3)提取出所有可视叶子节点,作为 PIX-PAGE 的基本区域;

4)根据式(2)~(7)计算各个基本区域的视觉“奇异性”,即重要性;

5)根据图 3 的区域合并算法,得到 PIX-PAGE 的顶层表示模型;

6)在 PIX-PAGE 的顶层表示模型上,根据奇异性传递规则;

7)计算各个组合区域的奇异性;

8)计算各个基本区域的奇异性增量,并更新它的奇异性;

9)根据式(8)计算所有词汇的主题相关性;

10)将词汇进行降序排列,取出前面的若干个作为网页的关键词集合返回。

3　实验结果与分析

导航类型网页(Hub Web Pages, HUB)包含了大量的主题噪声。为了检验在 PIX-PAGE 模型基础上提出的关键词自动抽取算法 P-KEA 的性能,用 Java 实现了该算法。在互联网上选取了新闻、经济、技术、社会、体育、娱乐、综合门户和电子商务,这样 8 类典型的导航型网页进行实验,并与 DocView 模型进行了性能比较。实验数据集均取自相关领域的权威网站,各个测试数据集的网页数量不等,且选取深度最大为 3(首页为第一层,链出一次就将深度值加 1,即最多为第三层)。DVM 表示基于 DocView 的关键词自动抽取算法,P-KEA表示提出的基于 PIX-PAGE 表示模型的关键词抽取算法。表 1 给出了两个算法的测试数据集及实验结果,其中的准确率定义如下:

$$\text{准确率} = \frac{\text{算法返回的正确的关键词个数}}{\text{算法返回的关键词总数}} \times 100\% \tag{9}$$

尽管某些网页中给出了关键词集合,但可能因为搜索引擎优化而并不准确。所以,在实验中,各网页的标准关键词集合是人工给出的。如表 1 所示,基于 DocView 模型的算法 DVM 的平均准确率为 52.9%,而基于视觉模型 PIX-PAGE 提出的关键词自动抽取算法 P-KEA 则为 73.8%,平均提高了 20.9%。在实验中,算法在每类网页上返回的关键词的数目正好等于人工给出的标准关键词集合的数目,因此,准确率与召回率及 $F1$ 均相等。

表 1　测试数据集及实验结果(网页类型均为 HUB)

数据集	文本数量	准确率/%	
		DVM	P-KEA
新闻	18	61	71
经济	7	55	75
技术	7	59	69
娱乐	23	51	78
社会	17	41	63
体育	13	53	79
综合门户	20	53	76
电子商务	19	50	79

DocView 模型注意到了 Web 网页的视觉特征对表达文本的主题具有重要的暗示作用。但是,一方面它没有全面、准确地量化节点的视觉强化特征,另一方面 DocView 模型在传播节点的影响因子时采用了全局传播的模型,这可能没有真实反映网页中不同位置和区域的节点之间的主题相关性关系,反而会增加噪声节点的权值。P-KEA 是在提出的 PIX-PAGE 模型基础上进行关键词的自动抽取。PIX-PAGE 根据人类在网页中,特别是导航型网页中定位并识别网页主题的基本原理,完整地表达并强化了那些被网页作者强调的内容块或词汇,使得与主题相关的关键词在网页表示模型中更加突出地表现为“奇异点”;另一方面,PIX-PAGE 模型本身的视觉“奇异性”量化方法和重要性传递规则通过“强者愈强、弱者更弱”的基本思想,能够有效抑制噪声的干扰。因此,基于 PIX-PAGE 模型的 P-KEA 的性能得到了较大的提高。但是,同时也应注意到,有些网页的主题关键词可能并没有使用足够的视觉特征来强调它;另一方面,可能有些热点词汇会被使用一些特殊的视觉强调手段。因此,可能会降低新模型 PIX-PAGE 和关键词自动抽取算法 P-KEA 的准确率。

(下转第 2368 页)

率 r_a、季度装备修复时间加权平均值和季度装备报废量 N_d 作为 DNEMP 的变量特征并建立了它们的数学模型，通过实例和比较验证了这些特征选择的合理性、模型构建的准确性以及所提方法的先进性，该方法可视为是对目前已有 SVM 算法研究成果在一定程度上的补充，也说明了支持向量机是一种实现 DNEMP 预测的有效工具。本文所提出的理论方法只适用于单目标主动性行为的行为载体特征确定，针对多目标的研究还有待进一步深入。

参考文献：

[1] 郝杰忠，杨建军，杨若鹏. 装备技术保障运筹分析[M]. 北京：国防工业出版社，2006：23－24.

[2] VAPNIK V N. The nature of statistical learning theory [M]. 2nd ed. New York: Springer-Verlag, 1999.

[3] CHALIMOURDA A, SCHÖLKOPF B, SMOLA A J. Experimentally optimal *V* in support vector regression for different noise and parameter settings [J]. Neural Networks, 2005, 18(2): 205－205.

[4] SUN JUN, FENG BIN, XU WENBO. Particle swarm optimization with particles having quantum behavior [C]// CEC 2004: Proceedings of the Congress on Evolutionary Computation. Piscataway: IEEE, 2004, 1: 325－331.

[5] SUYKENS J A K, van GESTEL T, de BRABANTER J, *et al*. Least squares support vector machines[M]. 3rd ed. Singapore: World Scientific Publishers, 2003.

[6] 李仁兵，李艾华，李亮，等. 支持向量机在导弹动力系统推力预测中的应用[J]. 系统仿真学报，2010，22(4)：934－937.

[7] CRISTIANINI N, SHAWE J. An introduction to support vector machines and other kernel -based learning methods[M]. 2nd ed. Cambridge, UK: Cambridge University Press, 2000.

[8] 白鹏，张喜斌，张斌. 支持向量机理论及工程应用实例[M]. 4 版. 西安：西安电子科技大学出版社，2008：79－81.

[9] 杨俊燕，张优云，朱永生. ε 不敏感损失函数支持向量机分类性能研究[J]. 西安交通大学学报，2007，41(11)：1315－1320.

[10] 于青，赵辉. 基于 GA 的 ε-支持向量机参数优化研究[J]. 计算机工程与应用，2008，44(15)：139－141.

[11] 孙祥，徐流美，吴清. Matlab7.0 基础教程[M]. 3 版. 北京：清华大学出版社，2005：37－39.

[12] LIBSVM—A library for support vector machines[EB/OL]. [2010－08－20]. http://www.csie.ntu.edu.tw/～cjlin/

[13] IVAKHNENKO A G. Sorting methods for modeling and clustering (survey of the GMDH papers for the years 1983－1988) the present stage of GMDH development[J]. Soviet Journal of Automation and Information Sciences, 1988, 21(4): 1－13.

[14] KIM D, SEO S-J, PARK G-T. Hybrid GMDH-type modeling for nonlinear systems: Synergism to intelligent identification [J]. Advances in Engineering Software, 2009, 40: 1087－1094.

[15] 朱帮助，魏一鸣. 基于 GMDH-PSO-LSSVM 的国际碳市场价格预测[J]. 系统工程理论与实践，2011，31(12)：2264－2271.

[16] 邹昊飞，夏国平，杨方廷. 基于两阶段优化算法的神经网络预测模型[J]. 管理科学学报，2006，9(5)：28－35.

[17] ENKE F, MÜLLER J-A. Self-organizing data mining [J]. Systems Analysis Modeling Simulation, 2003, 43(2): 231－240.

[18] HUANG C-L, WANG C-J. A GA-based feature selection and parameters optimization for support vector machines [J]. Expert Systems with Applications, 2006, 31(2): 231－240.

（上接第 2363 页）

4 结语

网页的关键词自动抽取是聚焦爬虫研究中非常重要的问题。主题型网页的关键词自动抽取已经有了较多的研究，但是，导航型网页的关键词自动抽取问题研究还不多，仍是一个技术难点。我们注意到，网页作者往往会利用各种视觉强化手段，把一些与主题相关的词汇突出地显示出来。借鉴人类善于发现图像中“奇异点”的原理，将 Web 网页表示为由普通区域和关键区域构成的平面图，提出了一个新的视觉表示模型 PIX-PAGE。通过有效的视觉量化方法，准确地表达并强化了与网页主题关键词相关的特殊区域的视觉“奇异性”。PIX-PAGE 的视觉量化方法和遵循“强者愈强、弱者愈弱”的重要性传递规则，有效地提升了模型的抗主题噪声的能力。因此，在此基础上提出的关键词自动抽取算法 P-KEA 的性能得到了较大的提高。

参考文献：

[1] CHAKRABARTI S, van den BERG M, DOM B. Focused crawling: a new approach to topic-specific Web resource discovery [J]. Computer Networks, 1999, 31(11－16): 1623－1640.

[2] CHAU M, CHEN H. Incorporating Web analysis into neural networks: an example in Hopfield net searching [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2007, 37(3): 352－358.

[3] 周立柱，林玲. 聚焦爬虫技术研究综述[J]. 计算机应用，2005，25(9)：1965－1969.

[4] WU XIAOYUAN, BOLIVAR A. Keyword extraction for contextual advertisement [C]// WWW'08: Proceedings of the 17th International Conference on World Wide Web. New York: ACM, 2008: 1195－1196.

[5] 刘远超，王晓龙，刘秉权，等. 信息检索中的聚类分析技术[J]. 电子与信息学报，2006，28(4)：606－609.

[6] 陈竹敏. 面向垂直搜索引擎的主题爬行技术研究[D]. 济南，山东大学，2008.

[7] 李晓明，闫宏飞，王继民. 搜索引擎——原理、技术与系统[M]. 北京：科学出版社，2006：98－103.

[8] 韩客松，王永成，滕伟. Web 页面中文文本主题的自动提取研究[J]. 情报学报，2001，20(2)：217－222.

[9] 任克强，赵光甫，张国萍. 基于带权语言网络的网页关键词抽取[J]. 计算机工程与应用，2008，44(8)：155－157.

[10] BUYUKKOKTEN O, GARCM-MOLINA H, PAEPCKE A. Seeing the whole in parts: text summarization for Web browsing on handheld devises [C]// WWW'01: Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001: 652－662.

[11] 王琦，唐世渭，杨冬青，等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展，2004，41(10)：1786－1791.

[12] CAI DENG, YU SHIPENG, WEN JI-RONG, *et al*. VIPS: a vision-based page segmentation algorithm, MSR-TR-2003-79 [R]. Redmond: Microsoft Research Corporation, 2003.