

# 基于用户辅助估计的相关网页搜索聚类

何拥军, 龚发根

(广东科学技术职业学院 计算机工程技术学院, 广东 珠海 519090)

**摘要:**随着 web 上的信息急剧增长, 如何有效地从 web 上获得高质量的信息已经成为当今热门研究主题之一。在信息检索、数据挖掘、人工智能等领域, 如何提高搜索信息结果的相似度, 以提高搜索信息的质量, 是众多研究的主要思考方法。文中在链接分析的基础上, 基于 SAHN 分级聚类算法提出了以用户辅助估计进行相关网页的聚类搜索方法, 与普通的聚类方法相比, 实验通过比较三种常用的相似性聚类方法在提高搜索结果中的应用, 发现结合用户辅助估计方法可以更好地提高搜索结果的满意度, 达到更好的搜索效果。

**关键词:**信息检索; 链接分析; 相似性

**中图分类号:**TP311

**文献标识码:**A

**文章编号:**1673-629X(2011)07-0112-04

## Clustering of Related Pages Based User-Assisted Estimation

HE Yong-jun, GONG Fa-gen

(Guangdong Institute of Science and Technology, Zhuhai 519090, China)

**Abstract:** With the rapid growth of the information on the web, how to effectively obtain high quality information has become one of today's hot topic in information retrieval, data mining, artificial intelligence, search for information on how to improve the results of similarity search for information in order to improve the quality of many of the major ways of thinking. In the link analysis based on hierarchical clustering algorithm based on SAHN proposed to estimate the associated auxiliary web users clustering search method, and the common clustering methods, experiments by comparing the similarity of the three commonly used clustering methods to improve the search results in the application of estimation methods that can be combined with user assistance to improve the search results to better satisfaction, to achieve better search results.

**Key words:** information retrieval; links analysis; similarity

## 0 引言

如何能更好地利用好 web 上有用信息, 通常都借助于搜索引擎基于关键字来检索符合某一搜索主题的相关页面, 搜索的结果页面往往在内容和主题上具有相似性, 这里的相似性往往是衡量搜索结果质量好坏的主要评价指标。还有一些情况就是用户一开始已经知道关于某一主题的一些页面, 需要通过这些页面来发现更多的在内容和主题上相关的页面, 最常用的就是基于链接分析的一些常用方法, 如文献计量方法中公共文献应用, 二分匹配图法, 基于链接分析的聚类等<sup>[1,2]</sup>, 这些方法也是用来评价科学文献之间主题相似性的最主要方法。

在科学文献搜索中, 往往是通过相互引用的参考

文献来搜索更多主题内容相似的文章, 该思想应用与 web 页面搜索的时候, 也就体现在页面和页面之间的链接性。相互链接的两个页面之间往往具有一定的相似性, 或者在主题上比较接近。基于链接分析的方法是把整个 web 上的所有页面之间的链接状态看成是一个图结构, 可以连通, 也可以不连通。每个页面看成是图的节点, 页面之间的链接看成是图的边, 在应用中往往是从一个临近的 web 链接图中按照一定的算法规则提取出一个 web 子图。这个子图中相互链接的页面往往具有一定的相似性, 这些具有相关性的页面往往就是用户需要搜索的结果页面, 因此链接分析很容易归结为对图的分析。链接分析的应用往往就是要如何从整个网络的 web 链接图当中抽取一些连通的链接子图, 这些链接子图所包含的边页面就是用户需要得到的信息。

在公共文献引用和二分匹配方法等链接分析当中, 往往首先都是在局部范围以一个邻接图开始构建, 因为通常在考虑页面之间的链接特性的时候, 很自然的都只考虑到本地的一个局部的页面链接, 这在公共

收稿日期: 2010-12-10; 修回日期: 2011-03-07

基金项目: 广东省自然科学基金项目 (8151064007000004); 珠海市科技计划项目 (PC20082010)

作者简介: 何拥军 (1976-), 男, 湖南邵阳人, 硕士研究生, 讲师, 研究方向为智能计算、数据挖掘、数据库。

文献引用方法中尤为突出。但从整个 web 链接图来说,web 页面之间的链接图不仅相当巨大,而且应该是动态发展和变化的,要把整个 web 链接图一次性装入进行分析是不可能的而且也是无效的,因为必然会存在一些内容很好的页面在当前的某一时刻并没有被链接,也就无法进入到链接图里进行分析,而且由于搜索用户的主观性,对于搜索需求的不断变化,如果链接图在某一段时间内是固定不变的话,搜索的结果往往不尽人意<sup>[3-6]</sup>。

基于以上存在的问题分析,文中提出以邻接范围内的相似链接图为基础,不断地在整个 web 图中进行迭代传播,动态地更新这个链接图,可以较好地解决以上问题。同时还考虑到用户对于搜索需求的主观性,利用用户的辅助估计来设定相应的相似性阈值,使得相似性值不再是一个固定的值,在搜索到的一系列具有相关性的页面中,可以根据用户的要求来改变页面的相似性,从而更好地提高用户对搜索页面的质量要求。

## 1 相关研究工作

搜索具有相似性结果页面的方法从分类上来说主要有基于关键字和基于链接分析的两大类。基于链接分析方法是近年来众多研究领域都在研究的一个思想,以链接分析为基础的相似性搜索方面主要有公共文献引用(co-citation),文献二分匹配法(bibliographic coupling)和 Amsle 方法<sup>[5-8]</sup>等。

### 1.1 公共文献引用(co-citation)

一个 web 页面的创建人总是会在页面中插入一些与自己页面内容或者主题相似的其他相关页面的链接,类似在科技文献的参考引用中,类似主题的文章总会同时引用相同的文献。

假定  $P$  是一个 web 页面, $I(p)$  为链接到  $P$  的所有页面集合,也即页面  $P$  的入链接集合,那么公共引用的两个页面  $p_1$  和  $p_2$  之间的相似度可以用如下公式计算:

$$\text{cit}(p_1, p_2) = \frac{|I(p_1) \cap I(p_2)|}{|I(p_1) \cup I(p_2)|} \quad (1)$$

上面公式说明,如果  $p_1$  和  $p_2$  之间的公共入链接数越多,那么它们的相似度值就越大,当然也可以看出该值符合规范化,介入 0 和 1 之间。如果  $I(p_1)$  和  $I(p_2)$  都为空,那么这两个页面的公共引用相似度为 0<sup>[5,6]</sup>。

### 1.2 文献二分匹配法(bibliographic coupling)

基于完全二分图核的算法是由 Kumar 等人<sup>[1]</sup>提出的。它建立在 web 页面上集中页面与权威页面的二分图关系上。它从二分有向图的角度对互联网上的主题社区给出了一种明的定义描述。根据随机二分图的理论,一个足够大而稠密的随机二分图将以很高的概

率包含一个完全二分有向图,那么如果将某个主题的链接结构看作一个大而稠密的二分有向图,主题社区的核就可以用一个完全二分有向图  $K_{ij}$  来表示。具体到互联网环境中,可以对上述概念有如下直观的理解:如果在互联网上存在一个某种主题社区,那么这种二分的核必将包含在其中。图 1 给出了一个  $K_{43}$  二分核的例子。图中左边四个节点均存在指向三家著名飞机制造商主页的链接。

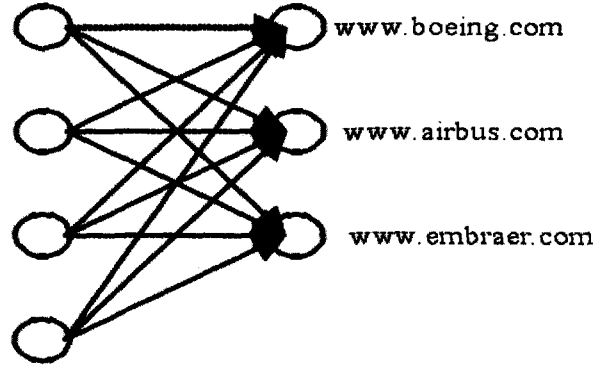


图 1  $K_{43}$  二分图核

假定  $P$  是一个 web 页面, $O(p)$  为链接到的所有页面集合,也即页面  $P$  的出链接集合,那么基于二分匹配法的两个页面  $p_1$  和  $p_2$  之间的相似度可以用如下公式计算<sup>[9]</sup>:

$$\text{bib}(p_1, p_2) = \frac{|O(p_1) \cap O(p_2)|}{|O(p_1) \cup O(p_2)|} \quad (2)$$

类似上一节,上面公式说明,如果  $p_1$  和  $p_2$  之间的公共出链接数越多,那么它们的相似度值也就越大,当然也可以看出该值同样是规范化的,总是介入 0 和 1 之间。如果  $O(p_1)$  和  $O(p_2)$  都为空,那么这两个页面的二分匹配相似度为 0<sup>[9-12]</sup>。

### 1.3 Amsle 方法

Calado 等人<sup>[9,10]</sup>提出一个叫 Amsle 的相似度计算方法,该方法结合了文献公共引用法和二分图匹配法,试图利用二者的优点,去除二者的缺点,依据 Amsle 算法,两个页面  $p_1$  和  $p_2$  如果具有相关性,只要满足如下三个条件之一即可:

- 1)  $p_1$  和  $p_2$  同时被第三个页面链接;
- 2)  $p_1$  和  $p_2$  同时链接了第三个页面;
- 3)  $p_1$  链接页面  $p_3$ ,  $p_3$  同时也链接了  $p_2$  页面。

Amsle 方法计算两个页面相似度的公式如下:

$$\text{ams}(p_1, p_2) =$$

$$\frac{|(I(p_1) \cup O(p_1)) \cap (I(p_2) \cup O(p_2))|}{|(I(p_1) \cup O(p_1)) \cup (I(p_2) \cup O(p_2))|} \quad (3)$$

由上面的公式可以看出, Amsle 方法把两个页面之间的出链接和入链接结合起来评价它们的相似性,当两个页面之间的出入链接的数越大,那么页面之间的相似度就越大。

## 2 用户辅助估计相似性搜索方法

### 2.1 SAHN 聚类方法

为了获得一系列相关页面的相关性排序列表,采取 SAHN 方法对相关页面排序、聚类 and 分级,该聚类方法相对于常用非结构化的一些聚类方法(如 K 均值, EM 聚类方法等)可以产生更好一些分级的结果,因为该算法一个非常重要的属性就是聚类簇之间距离的单调性<sup>[12,13]</sup>。假定  $d_1, d_2, \dots, d_k$  是聚类簇之间的距离,那么  $d_1 \leq d_2 \leq \dots \leq d_k$  一定是收敛的,因为这种单调收敛的属性,簇的合并总能找到最短的距离。在这种单调性的促使下,那些距离最短的页面会最先进行合并,在这种层次合并的每一步中,保留的都是距离最短的页面对,如此反复的重复合并,便可以获得一系列相关页面之间的相似性分数的排序。Lance 等人<sup>[5]</sup>对该聚类算法通过设定一个阈值( $0 < \alpha \leq 1$ )给出了一个灵活的方法,定义如下:

$$d_{hk} = \alpha d_{hk} + \alpha d_{hj} + (1 - 2\alpha) * d_{ij} \quad (4)$$

这里的  $h, i, j$  分别是包含  $n_h, n_i, n_j$  个元素的三个不同团体,团体之间的距离分别为  $d_{hi}, d_{hj}$ , 和  $d_{ij}$ 。

通过对参数  $\alpha$  的设置,范围从 0 到 1 可以对相关页面进行分级排序。

### 2.2 改进的方法

文中提出的改进方法基于以上三种常用方法的优点,克服了以上三种方法在链接分析中并没有考虑两个页面之间的直接联系的缺点,改进以后的相似度公式如下:

$$\text{newsim}(p1, p2) = \frac{|L(p1) \cap L(p2)| + \text{direct}(p1, p2)}{|L(p1) \cup L(p2) \cup p1 \cup p2|} \quad (5)$$

这里的  $L(p1)$  表示页面  $p1$  公共文献引用方法里入链接数,二分匹配方法里的出链接数,或者 Amsler 方法里的入链接数和出链接数的结合。 $\text{direct}(p1, p2)$  分别取值如下:

$$\text{direct}(p1, p2) = \begin{cases} 0, & \text{如果 } p1 \text{ 和 } p2 \text{ 页面之间没有直接链接} \\ 1, & \text{如果 } p1 \text{ 链接了 } p2 \text{ 或者 } p2 \text{ 链接了 } p1 \\ 2, & p1 \text{ 和 } p2 \text{ 之间互相链接} \end{cases}$$

把以上计算相似度公式与 SAHN 聚类方法结合起来应用,具体分三步执行:

步骤 1: 页面相似度分数的计算。

在如此多的 web 页面中并不是所有的页面都是适合聚类的,因为 web 上往往存在很多镜像页面,为了更好地计算页面的相似度分值,通过过滤掉那些无关的页面将可以大大提高最终结果的精度。通过设定一个相似度阈值,比如 0.95,这些保留下来的页面构成一个链接图 G,当然这里的链接图 G 和原始的 web 链

接图是不一样的,或者说是从原始 web 链接图里抽取出来的一个具有一定相似度的链接图,该图里的节点对应相应的 web 页面,但边都按照我们的策略分配了权重,更有可能的,这样的图很可能是一个非连通的图,所以可以继续细分成一些相应的连通子图。

步骤 2: 抽取相似度邻接子图。

由于以上图 G 并非是连通的,可以抽取相应的一些邻接连通的相似度子图,这些连通子图里的节点都来自于图 G,为了把以上图 G 转换为一些连通的子图,在不相交的两个图的集合中计算边的相似度值,当处在不同集合中边的分数值达到或超过某一规定的阈值的时候,就把它加入到同一个图集合,如此反复进行,最后通过图的一些计算,得到所有连通的邻接子图。

步骤 3: 借助用户的辅助估计分级相关页面。

最终的目的是要借助于用户的辅助估计得到相似性页面的分级排序,在获得了以上相应的邻接子图以后,应用 SANH 聚类算法,给定一个由用户决定的阈值  $\alpha$ ,该值的获得基于上面公式(5),以邻接子图里的节点页面作为输入数据,这些页面作为最终相似性结果页面的原始页面,输出的时候对这些页面进行排序,也即输出一组有序的结果页面,具体算法描述如下:

SAHN 聚类和分级机制

输入: 相似邻接子图里的所有节点页面,如公式(5)定义的由用户决定的阈值  $\alpha$

输出: 一组进行了分级排序的有序结果页面  
距离矩阵

1. for  $k = 1$  to  $N$
2. for  $l = 1$  to  $N$
3.  $D[k][l] = \text{dis}(pk, pl)$
- 初始化
4.  $H[N]$  (簇点之间结合的距离)
5.  $M[N][2]$  (簇点合并序列)
6.  $O[N]$  (输出的有序页面链表)
7. for  $k = 1$  to  $N$
8.  $I[k] = 1$  (跟踪的活动群集)

计算聚类

9. for  $k = 1$  to  $N$
10. begin loop
11.  $(i, j) = \underset{(i,j) \in M, I[i]=1, I[j]=1}{\text{argmin}} D[i][j]$
12.  $M.append(<i, j>)$
13.  $H.append((i, j))$
14. for  $h = 1$  to  $N$
15. begin Loop
16.  $d_{h(i,j)} = \alpha * D[h][i] + \alpha * D[h][j] + (1 - 2 * \alpha) * D[i][j]$
17.  $D[i][h] = D[h][i] = d_{h(i,j)}$
18. End Loop
19.  $I[j] = 0$  (停用集群)
20. End Loop

页面分级

```
21. if M[i][j] == -ip
22. d[ip] = H[i] (第一次合并 ip)
23. for cp=1 to N (cp = ip)
24.   Being Loop
25.   if M[i][j] == -cp
26.   d[cp] = H[i] (第一次合并 cp)
27. if row i of M[N][2] are clusters that include cp
   and ip respectively
28. d[(ip, cp)] = H[i]
29. O[cp] = |d[ip]-d[(ip, cp)]| + |d[cp]-d[(ip,
   cp)]|
30. End Loop
31. Sort O[N] (结果集合以升序输出)
```

3 试验及其评价

3.1 试验数据的收集与清理

数据集:由于搜索中相似性概念的主观性,往往很难对其作出精确的判断,为了提高实验结果的客观性,实验采取了 Google 网页目录集数据作为原始数据集,并进行一定的清理。

种子节点:为了对几种聚类方法进行验证比较,首先抽取 135 个页面,以 Google 网页目录分成三类:数据挖掘(C1),知识发现(C2)和机器学习(C3),这些页面作为初始化核心页面集。并对其中的每一个页面应用 Google API 工具获得链接它的页面的入链接数大于 60 的所有页面,并同时获得它自身链接的所有其它页面,以这些页面作为该原始页面可以迭代增长的一个初始数据集,以该初始数据集构建邻接子图。

3.2 实验及性能评价

基于以上的邻接子图,应用 SAHN 聚类 and 分级机制,并灵活设定相应的  $\alpha$  值,比如 0.5,实验分别测试了三种聚类方法,结果比较如表 1。

表 1 三种算法的搜索精度与增长率的比较

E1(30,1000)	搜索精度(%)			迭代增长率(%)		
	C1	C2	C3	C1	C2	C3
Cit	83.82	65.5	84.36	53	36.33	69.85
Bib	80	0	65.67	25.45	0	35.66
Ams	67.66	54.56	78.57	57.55	41.67	82.35
E1(135,1000)	搜索精度(%)			迭代增长率(%)		
	C1	C2	C3	C1	C2	C3
Cit	87.8	67.9	94.5	71.3	43.6	78
Bib	82	36.4	69.9	35	9.43	35.6
Ams	85.6	65.8	86.8	83.4	53	89.6

其中搜索精度是指在某一类页面当中所有聚类符合用户满意度的页面所占比例,迭代增长率是指在所有页面当中符合用户搜索满意度的页面所占比例。从

以上实验结果可以看出,Cit(公共文献引用)方法虽然在迭代增长率方面略差于 Ams 方法,但在搜索精度上要远远高于其它两种方法。结果也表明,混合使用入链接和出链接,虽然可以在帮助提升迭代增长率的但却降低了搜索精度;同样,Bib 方法虽然在搜索精度上偶尔会超过 Ams 方法,但是随着入链接页面数的增加,Ams 的搜索精度就开始超过了 Bib 方法。更进一步可以看出,虽然在初始化数据集合中加入的出链接页面比入链接页面数多,但在以出链接为主的 Bib 方法中,迭代增长率却显得更加低。所有这些都表明,出链接页面的增加会带来更多的松散页面和噪音页面,而更多的人链接页面的增加在搜索精度上会收到更好的效果。

4 结束语

文中提出了搜索相关网页上的基于用户相似性估计策略,为了更好地解决搜索精度的问题,提出了一种新方法来鉴别交叉主题的网页如何计算相似性分数的值,并通过让用户设定有效的阈值。基于 SAHN 聚类方法进行顺序,聚集,分级,结合自适应距离排名,以 Cit 的入链接为主要参考,来聚类相关网页。实验结果表明,Cit 的措施主要考虑入链接的方法,在精确度方面普遍好于其他两个链接方法。此外,应用 SAHN 聚类方法使用户在选择他们的搜索结果需求上可以通过设置相应的参数来灵活的选择相关网页。实验结果收到了较好的效果。

下一步希望能通过实验抓取更加多的网页信息,进一步检验以上方法。

参考文献:

[1] 王晓宇,周傲英. 万维网的链接结构分析及应用综述[J]. 软件学报,2003,14(10):1768-1780.

[2] 何拥军,骆嘉伟,孙星明. 应用链接分析的 web 搜索结果聚类[J]. 计算机工程与应用,2005,41(1):179-183.

[3] 何拥军,龚发根. 最大流算法发现 web 社团的改进[J]. 计算机工程与应用,2007,13(1):170-173.

[4] 于洪涛,段军义,杜照丰. 一种基于聚类技术的个性化信息检索方法[J]. 计算机工程与应用,2008,44(8):187-189.

[5] 刘馨月,赵明砚,张宪超,等. 基于最大流 HITS 的改进算法[J]. 计算机工程与应用,2008,44(17):141-143.

[6] 何国斌,赵晶璐. Web 页面主题相关性排序算法的研究[J]. 计算机工程与应用,2009,45(23):149-151.

[7] 邵兰洁,李光忠. Web 使用挖掘的数据采集技术探究[J]. 计算机技术与发展,2010,20(3):225-229.

[8] Calado P,Cristo M,Goncalves M A,et al. Link-based simi-

到可审计、可跟踪、可追溯,同时可以在一定程度上检测出系统的非法访问、入侵和攻击。

#### 5) 系统数据备份和恢复。

可以制定完整的备份规划。实现备份数据的多机、多地存放,不仅在服务器的磁盘上保存,同时利用现有的磁带库进行数据备份工作,并实现备份数据的异地存放。

#### (2) 高扩展性。

##### 1) 跨平台特性和伸缩性。

由于采用 J2EE、B/S 等编程技术,可移植性、适应性强,系统可以在很多流行的平台上运行(Windows 系列、Linux、Unix 等),特别适合事业机关、企业内部网络。同时,根据客户对安全性、稳定性等需求,系统可以灵活的部署,从一般的小型系统到数据库或应用服务器群集等多种形式。

##### 2) 客户化定制及二次开发。

基于基础构件的平台和应用,具有高度的柔性,提供多种形式、多种方式的定制功能和开发接口。可以根据客户的需求灵活的定制业务应用,比如:在公文处理系统中,公文所有的表单都可以按实际的纸质表格的样式定制,所见即所得。通过 workflow 建模可以建立最符合实际业务的流程。

##### 3) 良好的开放性。

系统的对外接口都采用符合规范的协议、标准,具有良好的开放性,很多系统模块比如邮件服务、全文检索服务、用户认证服务、加密和数字签名、公文手写签名和批阅技术等都可以很好兼容各厂商的产品或做外挂集成。

## 4 结束语

南京水科院 3G 移动办公平台采用最新信息技术的发展成果和手段,实现科研综合管理的实时化、异地化、信息化、流程规范化和自动化,全面提升南京水科院的科研综合管理水平。为南京水科院的发展提供了很好的技术支撑。

由于国内三大运营商的 3G 网络和服务刚刚部署

和面市,移动办公的产品还有待成熟。市场和用户对于 3G 服务的接受还有一段时间,移动办公近期之内可能还看不到大规模的爆发。但是可以预见,人们对于不受空间限制的办公需求是巨大的,随着 3G 网络覆盖得更加全面,3G 终端的样式更加丰富,3G 服务的资费更加便宜,移动办公业务必将在不久的将来迎来高峰。

#### 参考文献:

- [1] 高 铭. 电子政务中移动办公的研究[J]. 电脑知识与技术:学术交流,2009,11(2):182-184.
- [2] 孙 宇,唐绮薇. 移动邮件系统的关键技术研究[J]. 数字通信世界,2007,4(12):52-54.
- [3] 陆剑江. 通用模式的移动办公平台设计方案研究[J]. 计算机工程与设计,2006(4):162-165.
- [4] 唐 宁,蒋红源,杨 恒. 基于 3G 运营商的移动办公系统应用和分析[C]//2009 年信息通信网络技术委员会年会. [出版地不详]:[出版者不详],2009:107-109.
- [5] Wilson C, Doak P. Creating and Implementing Virtual Private Networks: The All-encompassing Resource for Implementing VPNs[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(1): 22-30.
- [6] White J. An introduction to Java2 micro edition (J2ME) Source [C]//Waveform Diversity and 17 ~ signConference. [s.l.]:[s.n.], 2007:204-208.
- [7] Skonnard A, Gudgin M. Essential XML Quick Reference: a Programmer's reference to XML, XSLT, XML Schema, SOA, and more[M]. US: Pearson Education, Inc, 2002:304-307.
- [8] 杨 悦,许 琪. 浅谈企业办公自动化网络安全[J]. 信息技术,2009(22):251-253.
- [9] 李成严,冯惠灵. 基于开源技术的 web 应用架构研究[J]. 计算机技术与发展,2009,19(8):27-33.
- [10] 张在东,盛步云. 基于 Web Services 和本体的信息集成框架[J]. 计算机技术与发展,2009,19(3):134-140.
- [11] 况 旭,刘 波. XML 的面向对象语言特性[J]. 计算机技术与发展,2010,20(1):54-57.
- [12] 李 苏,刘建勋. Web 服务的绑定与调用方法研究[J]. 计算机技术与发展,2010,20(6):31-35.
- [13] search[C]//In Proc. of the 16th International Conference on World Wide Web(WWW'07). [s.l.]:[s.n.], 2007:131-140.
- [12] Lance G N, Williams W' T. A generalized sorting strategy for computer classifications[J]. Nature, 1966, 7:212-218.
- [13] Lance G N, Williams T W T. A general theory of sorting strategies[J]. hierarchical the computer Journal, 1967, 9:373-380.
- [9] 蒙 韧,邵延振,袁鼎荣. 一种基于页面 Block 的 Web 信息提取方法[J]. 计算机技术与发展,2010,20(1):197-200.
- [10] 胡国晴,李建华. 一种基于可信度分析的 Web 页面新属性发现方法[J]. 计算机技术与发展,2009,19(1):56-59.
- [11] Bayardo R J, Ma Y, Srikant R. Scaling up all pairs similarity

(上接第 115 页)

ity measures for the classification of web documents[J]. JASIST, 2006, 57(2):208-221.