

## 基于 Web 挖掘的主页多主题更新模型

张顺香,朱广丽,陆 奎

(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

(sxzhang@aust.edu.cn)

**摘 要:**针对改善网站主页中多个主题更新的合理性问题,提出了一种新颖的基于 Web 数据挖掘技术的主页主题更新模型。对当前主页主题更新方法进行分析,指出目前一些大学校园网站主页主题更新的不合理性,然后通过数据挖掘,从安徽理工大学网站 Web 日志中提取有效数据,分析各个主题的点击率随时间的变化趋势,进而提出基于点击率的网站主页主题更新模型。实验结果证明,模型能够实现对主页主题的合理更新,有效提高网站主页的受欢迎程度。

**关键词:**主题更新;数据挖掘;点击率;更新模型

**中图分类号:** TP301.6 **文献标志码:** A

## Model for updating multi-topics on homepage based on Web data mining

ZHANG Shun-xiang, ZHU Guang-li, LU Kui

(College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan Anhui 232001, China)

**Abstract:** To improve the reasonableness of updating multi-topics on homepage, this paper proposed a novel model based on Data Mining (DM). Firstly, after deep research of the present methods of updating topics, the irrationality of updating multi-topics on homepages, especially in college campus network, was pointed out. Then some valid data were extracted from Web log of Anhui University of Science and Technology (AUST) was used to analyze the continuously changing trend of click rate of every topic. Based on the obtained trend, a model for updating multi-topics on homepage was proposed. Experimental results show that the proposed model can reasonably update multi-topics, and enhance the popularity of homepages efficiently.

**Key words:** theme updating; data mining; click rate; updating model

### 0 引言

网站主页是一个网站的标志和门户,网站的更新内容需要在主页上突出显示。也就是说,网站主页上的主题的具有很高的时新性。因此,考虑到网站服务群体和网站主页的时新性两个因素,提出主题更新的合理性的问题,这也就是我们要研究的问题。

目前,大多数网站对于主页主题的更新只是考虑主题的时新性,这种主题更新方式从算法角度来看,多数采用类似队列形式进行,即将新出现的网页主题插入队列的一端(放置在主页上所排列主题的顶端),而将相对时间存在最长的网页从队列中删除,即从网站主页上删除相应的网页主题。但这种网页主题更新方法存在一个明显的缺点,可能会导致一些网页主题虽然是新近出现但并不受本网站的大多数用户欢迎的主题留在网站主页上,而出现时间相差不多但受本网站的大多数用户欢迎的主题被过早地替换,这显然是不合理的。

解决这个问题最直接而有效的方法是通过跟踪浏览器客户端的鼠标行为,按区块对页面进行点击行为的分析。也就是说,我们需要通过数据挖掘技术对用户行为进行正确的统计和分析。对于主页上的主题更新问题需要研究三个方面的问题:1)在一段时间内主页上哪些主题经常被用户点击;2)某用户在点击该主题页后,停留在该主题页上的时间是多少;3)在一段时间内,每个主题被本网站用户群点击的变化趋势是什么。所幸的是,WWW 上每一个提供信息资源的服务器上都有一个结构比较好的记录集,即 Web 访问日志。从 Web

访问日志中可以提取到解决这几个问题所需要的基础数据。

数据挖掘是从海量的数据中自动、高效地提取有用知识的一种新兴的数据处理技术,包括分类、聚类、关联规则挖掘、特征与偏差、时序模式发现、趋势分析等<sup>[1-3]</sup>。Web 挖掘可分为 3 类:Web 内容挖掘、Web 结构挖掘和 Web 使用记录的挖掘<sup>[3]</sup>。对于 Web 使用记录的挖掘主要目标则是从 Web 的访问记录中抽取感兴趣的模式<sup>[3-4]</sup>,分析这些数据可以帮助理解用户的行为,从而改进站点的结构,或为用户提供个性化的服务。许多文献利用分析得到的用户访问模式和倾向,进行个性化服务和网页推荐。文献[5]综合考虑了具体领域和用户兴趣,以及网页和用户兴趣的相似程度,来为用户提供高效的个性化网页推荐。文献[6]提出了一种增量式信息更新方法,很大程度上提高了大型网站搜索引擎网页的更新效率。文献[7]从搜索引擎更关心搜索结果的“重要性”以及搜索结果的“时新性”角度出发,提出了一种改进的结合重要性与时新性的页面分类更新策略。文献[8]提出在估算个性化搜索的效率时,最好的方法就是通过分析短时间内用户的点击行为。

本文通过对服务器访问日志文件的分析,提取主页上不同主题的点击次数及用户在相应主题页面停留的时间数据,计算有效点击率,利用我们提出的主题更新模型,得到主页的主题实时更新策略,从而提高网站主页的受欢迎程度。

### 1 网页主题更新的基本概念

**定义 1** 主题点击次数(Click Number of Topic, CNT)。

主题点击次数就是自某个主题被放置于主页的  $t_0$  时刻

收稿日期:2009-04-27;修回日期:2009-06-22。 基金项目:安徽省教育厅自然科学研究项目(2005kj085)。

作者简介:张顺香(1970-),男,安徽无为,人,讲师,硕士,主要研究方向:软件理论与工程、计算理论;朱广丽(1971-),女,安徽淮南人,讲师,硕士,主要研究方向:网络信息监控;陆奎(1962-),男,安徽淮南人,教授,博士,主要研究方向:网络控制。

开始到当前时刻 $t_j(j \geq 0)$ , 在 $N$ 个终端(以 IP 来区分)点击该主题次数的总和。第 $i$ 个主题在 $t_0$ 到 $t_j$ 时间内被第 $k$ 个终端点击次数为 $CN_k$ , 则当前 $t_j$ 时刻第 $i$ 个主题点击次数为所有用户(主题浏览者)点击次数之和:

$$CNT(i, t_j) = \sum_{k=0}^N CN_k \quad (1)$$

式中 $N$ 表示至 $t_j$ 时刻点击某个主题总的用户数。这里充分考虑到每个终端可能有多个用户使用, 单个用户对某个主题特别有兴趣而重复点击几次, 因此 $CN_k$ 可能大于 1。忽略单个用户对主题无兴趣但故意多次点击的小概率情况, 且这种特殊情况可以通过定义 2 中的页面停留时间的过滤也能够得到一定程度的减弱。

**定义 2** 主题有效点击次数 (Valid Click Number of Topic, VCNT)。

考察第 $k$ 个终端对第 $i$ 个主题的一次点击 $u_k^i$ , 用户在进入相应主题页面停留的时间(阅读时间, 单位秒)为 $t_k^i$ , 规定某次点击为有效点击见式(2):

$$u_k^i = \begin{cases} 1, & t_k^i \geq 10 \\ 0, & t_k^i < 10 \end{cases} \quad (2)$$

定义式(2)有两个目的, 一个是它可有效防止主题的各种意外点击, 另一个是保证用户对该网页的内容有阅读的兴趣。结合定义 1 可知, 第 $i$ 个主题在 $t_0$ 到 $t_j$ 时间内被第 $k$ 个终端有效点击次数为 $VCN(i, k)$ :

$$VCN(i, k) = \sum_{t_0}^{t_j} u_k^i \quad (3)$$

则至 $t_j$ 时刻第 $i$ 个主题的有效点击次数:

$$VCNT(i, t_j) = \sum_{k=0}^N VCN_k = \sum_{k=0}^N \sum_{t_0}^{t_j} u_k^i \quad (4)$$

由式(1)和(4), 显然存在 $VCNT(i, t_j) \leq VCN(i, k)$ , 且有效主题点击次数比主题点击次数更能真实反映该主题受用户欢迎程度。

**定义 3** 主题点击变化率 (Click Rate of Topic, CRT)。

若将从时刻 $t_0$ 到 $t_j$ 的连续时间作等间隔 $t$ 划分, 使得 $t = t_j - t_{j-1} = t_{j-1} - t_{j-2} = \dots = t_2 - t_1 \geq t_1 - t_0$ 。据此, 主题点击变化率定义为在某个时间段内所有终端用户对主题的有效点击次数, 它表现为主题的有效点击次数的增加量, 可用来衡量某个主题被用户点击的动态变化趋势。第 $i$ 个主题在 $t_j$ 时刻的主题点击变化率 $CRT(i, t_j)$ 为:

$$CRT(i, t_j) = VCNT(i, t_j) - VCNT(i, t_{j-1}) \quad (5)$$

根据 Web 日志的记录结果, 提取我校网站新闻类主题中 5 个新闻主题的用户点击数据按照上面的定义进行分析处理, 得到每个主题的主题点击次数和有效主题点击次数在不同时间段的点击变化情况, 如图 1 所示。图中 $t_0 = 8$ 点,  $t_j = 24$ 点,  $t = 2$ h, (时间间隔的选取应该由实际应用情况而定, 由于校园网中的新闻更新较慢, 因此选取的时间间隔较大)。

进一步, 将图 1 所示的 5 个重要新闻主题在不同时间段有效主题点击次数与主题点击次数比例列表, 见表 1。

从表 1 看出, 每个主题有效点击次数与主题点击次数之间没有确定的变化规律(不存在统一的函数关系), 而主题有效点击次数更能直接反映网站用户群对各个主题的兴趣, 因此本文在研究网页主题更新模型时, 是基于主题有效点击次数, 而不是主题点击次数。

## 2 网页主题更新模型

### 2.1 几个假设

为了方便明确表达该模型的思路, 简化模型的建立过程,

在描述网页主题更新模型之前, 首先给出以下基于该模型的 4 个假设条件。

1) 主题的位置无关假设: 该假设条件是指位于主页上的某个分类主题, 其点击次数和有效点击次数只与浏览该主题的用户本身的兴趣有关, 不管其被放置于网站主页的上面还是下面, 左边还是右边。

2) 主题的重要性等同假设: 该假设条件是指所有置于主页上的主题, 其重要性是相同的, 不论其对某个领域或者该网站的用户群更重要或者不重要。

3) 主题之间无关性假设: 该假设条件是指所有置于主页上的主题之间不存在相关性, 即不挖掘某个用户在浏览某个主题网页后必定还浏览另一个主题网页, 同一用户浏览该主页上的几个主题是独立进行的过程。

4) 用户按时间段有效分布假设: 该假设条件是指在同一时间段内存在一定数量对主题 1 有浏览兴趣的用户, 也存在一定量对主题 $n$ 有浏览兴趣的用户, 且与其本身趋势是一致的。例如, 图 1(b) 中在 8:00~12:00, 各个主题的有效点击次数相对都有一个高峰。

以上四个假设条件中, 前三个是为了简化主题更新模型的建立过程, 而第四个假设是建立本模型的前提条件。

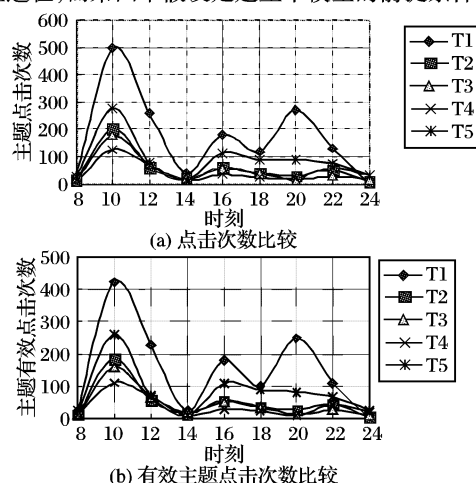


图 1 主页中 5 个主题点击次数在不同时间段的变化趋势

表 1 不同时间段主题有效点击比例

时间段	主题有效点击比例				
	主题 1	主题 2	主题 3	主题 4	主题 5
6:00~8:00	0.82	0.80	0.76	0.92	0.88
8:00~10:00	0.84	0.90	0.88	0.88	0.93
10:00~12:00	0.85	0.91	0.89	0.91	0.93
12:00~14:00	0.84	0.89	0.89	0.90	0.92
14:00~16:00	0.87	0.90	0.90	0.89	0.93
16:00~18:00	0.86	0.90	0.90	0.90	0.93
18:00~20:00	0.87	0.90	0.90	0.86	0.93
20:00~22:00	0.87	0.89	0.89	0.83	0.93
22:00~24:00	0.87	0.89	0.89	0.82	0.92

### 2.2 网页主题更新模型

#### 2.2.1 模型基本思想

网站主页上的各个主题对于用户的受欢迎程度是随着时间推移而动态变化的。可能某个主题总的有效点击次数很大, 但由于网页的时新性较低变得不再受欢迎; 可能某个主题由于本身内容的原因, 即使是最新出现的, 但其有效点击次数有限, 导致其受本网站用户群的欢迎程度不高。因此, 建立模型要充分考虑到各个主题的受欢迎程度随时间的动态变化过程。一般情况下, 当一个主题刚刚出现后, 其受欢迎程度会逐

渐变高,在达到一定高度后会渐渐降低。

建立网页主题更新模型以提高整个主页的受欢迎程度为目标,它具体涉及到主页中每个主题的受欢迎程度,因此,当有新的主题出现时,如何选择适当的主题,用新的主题取而代之是关键问题。这需要研究当前已有主题受欢迎程度的变化趋势。从而建立模型的基本思想是以每个主题在某个时间段内的点击变化率(有效点击的增加次数)作为该主题在该时间段内的受欢迎程度的标准,以当前时刻  $t_j$  向前推  $x$  个时间段点击变化率的平均值作预测,它可以相对合理地、实时地预测在下一个时间段内主题的受欢迎程度(Degree of Reception, DR)。第  $i$  个主题在  $t_j$  时刻的受欢迎程度  $DR(i, t_j)$  为:

$$DR(i, t_j) = \frac{1}{x} \sum_{h=j-x+1}^j CRT(i, t_h) \quad (6)$$

这里  $x$  的选择要视实际情况而定,如果太大导致结果倾向于历史数据,不能反映受欢迎程度的动态变化的特点;反之如果太小(如等于 1),可能会导致数据不能反映实际情况(如在校园网中,在学生上课期间,大多数学生感兴趣的主题有效点击次数就不能真实反映该主题的受欢迎程度)。

进一步,根据式(6),可以得到第  $l$  个主题类的主题在  $t_j$  时刻的受欢迎程度  $CDR_l$ :

$$CDR_l = \sum_{i=1}^{nl} DR(i, t_j) = \sum_{i=1}^{nl} \left( \frac{1}{x} \sum_{h=j-x+1}^j CRT(i, t_h) \right) \quad (7)$$

其中  $nl$  表示第  $l$  个主题类所包含的主题个数。相应地,若一个主页包含  $m$  个主题类,根据公式(7),则主页在  $t_j$  时刻的受欢迎程度  $HDR$  可定义为:

$$HDR = \sum_{l=1}^m w_l CDR_l \quad (8)$$

式(8)中: $HDR$  表现为主页中每个主题类受欢迎程度的函数,它是衡量主题更新策略好坏的综合性指标; $w_l$  为第  $l$  个主题类在主页中的重要程度。若设  $w_l = 1$ ,则主页的受欢迎程度可简化为:

$$HDR = \sum_{l=1}^m \sum_{i=1}^{nl} \left( \frac{1}{x} \sum_{h=j-x+1}^j CRT(i, t_h) \right) \quad (9)$$

### 2.2.2 基于主题的主题更新模型

根据 2.2.1 提出的模型基本思想,我们提出的主题更新模型可描述为 4 个主要部分:1)从 Web 访问日志文件中提取各个主题的点击次数、点击时间和停留时间等基本参数;2)对提取的基本数据按照用户浏览停留时间完成过滤,得到每个主题的有效点击次数;3)计算每个主题在不同时间段有效点击次数,得到每个主题的点击变化率,由主题的点击变化率计算  $t_j$  时刻每个主题的  $DR(i, t_j)$  和  $CDR_l$ ;4)根据 3)计算得到的结果,按照以下算法中第 4~6 步完成更新决策。

主题更新算法:当某个时间段内需要进行主题更新,按照以下步骤可实现单个主题或多个主题的更新。

第 1 步 根据 Web 日志数据的分析统计,得到当前时刻所处的时间段内主页上所有主题的有效点击次数;

第 2 步 若前  $x-1$  个时间段所有主题的有效点击次数已计算,直接读取,否则计算前  $x-1$  个时间段所有主题的有效点击次数;

第 3 步 按照式(6)计算每个主题的当前时刻的受欢迎程度  $DR$ ;

第 4 步 将每个主题用两个数据域进行存储:前一个数据域存储主题的受欢迎程度,后一个数据域存储主题在主题类中的相对位置。分别对  $m$  个主题类中所有主题的受欢迎程度完成排序,记第  $l$  个主题类中主题受欢迎程度最大值为  $DR_{upper}^l$ ,最小值为  $DR_{lower}^l$ 。

第 5 步 若需要更新主题的个数  $> 0$ ,选择  $m$  个  $DR_{lower}^l$  最小者  $\text{Min}\{DR_{lower}^l \mid l = 1, \dots, m\}$ ,转第 6 步;否则转第 7 步;

第 6 步 判断更新的主题所属的主题类,并把该主题类各个主题受欢迎程度的均值作为更新主题的受欢迎程度初始值,从主页上移走  $\text{Min}\{DR_{lower}^l \mid l = 1, \dots, m\}$  对应的主题,并修改移走主题对应主题类的  $DR_{lower}^l$  值,转第 5 步;

第 7 步 结束,完成主题更新。

该模型在执行过程中的时间代价很低,只有第 4 步对每个主题类排序以及第 5~6 两步循环时,时间复杂度较高,为  $O(n^2)$ ,且每个主题类的主题个数和需更新的主题个数往往都是非常有限的,因此该模型适于实时的动态计算。

## 3 模型有效性分析与验证

### 3.1 有效性分析

从这个主页受欢迎程度的角度,对比网页主题更新模型和单纯考虑主题时新性的更新方法,分析网页主题更新模型的有效性。显然只存在两种情况。

1)两种方法替换了相同的主题。

不管是单个主题更新还是多个主题更新,当两种替换了的主题完全相同时,最后计算得到的主页  $HDR$  肯定是一致的。

2)两种方法替换了不同的主题。

当两种方法替换了不同的主题时,由于我们的模型对于每个主题类都计算了  $DR_{lower}^l$ ,根据式(8),采用我们提出的模型计算所得  $HDR$  肯定相对较高。

事实上,情况 2)发生的概率很大,这是由于同一时间段经常会更新几个主题,且不同时间段更新的主题个数不相同,这样就导致单纯考虑主题时新性的更新方法在选择被更新的主题时的随机性,而我们的模型可以有效地克服这个缺点。

### 3.2 有效性验证

为了对我们提出的更新模型进行有效性验证,选择更新速度较慢的校园网和更新速度较快的天涯论坛进行实验。

1)选择我校校园网上重要新闻、领导讲话和学子风采三个主题类(每个主题类均包含 5 个主题)的主题为实验对象。每个主题的不同时间段的有效点击次数从 Web 日志中获得。在计算过程中,可调参数设置情况为: $t = 2 \text{ h}$ ,  $x = 3$ ,  $w_l = 1$ 。如图 2 所示,分别采用我们的主题更新模型(方法 1)和单纯考虑主题时新性(方法 0)两种更新方法,计算在不同时间段主页  $HDR$ 。

2)选择天涯论坛中天涯社区热贴榜(包含 8 个主题)的主题为实验对象。以对某个主题的一次回帖作为一次有效点击与有效点击的定义是一致的,同时,由于论坛的 Web 日志不方便获得,因此,我们编写程序获取不同时间段对某个主题的回帖数,作为主题的有效点击次数。在计算过程中,可调参数设置情况为: $t = 5 \text{ min}$ ,  $x = 3$ ,  $w_l = 1$ 。如图 2(b)所示,同样分别采用我们的主题更新模型(方法 1)和单纯考虑主题时新性(方法 0)两种更新方法,计算在不同时间段天涯论坛首页的  $HDR$ 。

在图 2 中,对每个主题类中的主题受欢迎程度作归一化处理(除以该主题类中主题最大值),再由所有主题的受欢迎程度计算主页的受欢迎程度  $HDR$ 。

从图 2(a)可以看出,对更新速度较慢的校园网新闻主题,在 10 次更新计算结果中,只有在第 3 次计算中,两种方法的计算结果相同,而另外 9 次计算中,主题更新模型方法都使主页的受欢迎程度得到一定程度的提高。从图 2(b)可知,对

(下转第 2815 页)

择的维数分别为 500 维、1 000 维、1 500 维、2 000 维和 2 500 维。实验结果的精确率 precision、查全率 recall 如表 4 所示,  $F1$  值如表 5 所示。

表 4 分类实验 Precision 值及 Recall 值

特征维数	MI		本文改进的 MI	
	准确率	查全率	准确率	查全率
500	0.38	0.25	0.55	0.4
1 000	0.48	0.53	0.67	0.73
1 500	0.61	0.49	0.71	0.83
2 000	0.62	0.77	0.75	1.00
2 500	0.76	0.65	0.84	0.89

表 5  $F1$  值

特征维数	MI/%	本文改进的 MI/%
500	30.20	46.30
1 000	50.40	69.90
1 500	54.30	76.50
2 000	68.70	85.70
2 500	70.10	86.40

由表 5 中的  $F1$  值可得到如图 1 的直观表示。

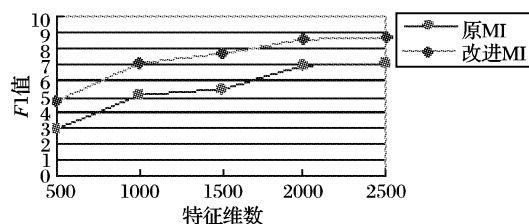


图 1 原 MI 与改进的 MI 算法的  $F1$  值比较

(上接第 2798 页)

更新速度较快的论坛主题,在 20 次的更新计算中,有 3 次两种方法的计算结果相同,其余 17 次的计算中,主题更新模型均能够使首页受欢迎程度得到提高。

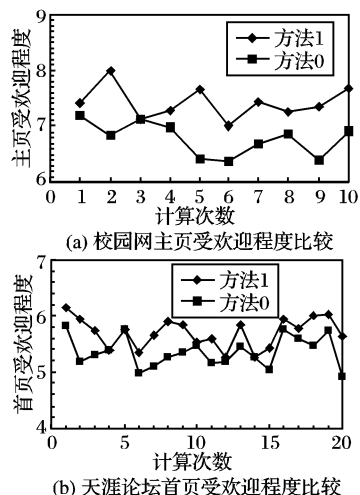


图 2 两种更新方法的受欢迎程度比较

## 4 结语

不同于个性化推荐,本文给出面向整个网站用户群基于 Web 挖掘技术的多个主题更新策略,通过这种更为合理的更新,力求达到主页中主题有更大浏览点击次数,提高主页的受欢迎程度。

1) 通过对提取实际新闻主题点击次数、用户停留时间或回帖等数据的分析,提出基于主题有效点击次数来建立主题

通过图 1 中的对比可以看出,本文提出的特征选择方法有效提高了分类能力,并且较高的维数比低维数的分类效果更佳。由于本文是在中文邮件数据集上进行实验,鉴于预处理过程中的词典选择等因素,本文实验的绝对结果受到一定影响,但这并不影响验证相对效果。

## 4 结语

本文对邮件过滤中的特征选择算法进行了研究,对其中的互信息方法结合邮件过滤的特点进行了具体分析,得出了其在频度、集中度以及分散度上对分类效果的影响,并进一步针对这三方面进行了改进,提出了使用类别贡献比  $r(t_i)$  来衡量特征对分类的贡献,给出了改进后算法的具体计算公式,最后在实际环境中的真实邮件集上进行了实验验证。由于中文分词以及预处理效果比英文分词效果相对较差,本文实验的绝对结果不够理想,中文分词及预处理技术的改进将是以后的研究方向。

### 参考文献:

- [1] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26 - 32.
- [2] 陈平,刘晓霞,李亚军. 文本分类中改进型互信息特征选择的研究[J]. 微电子学与计算机, 2008, 25(6): 194 - 196.
- [3] 卢新国,林亚平,陈治平. 一种改进的互信息特征选取预处理算法[J]. 湖南大学学报: 自然科学版, 2005, 32(1): 104 - 107.
- [4] CCERT Data Sets of Chinese Emails( CDSCE) [EB/OL]. [2009 - 05 - 01]. <http://www.ccert.edu.cn/spam/sa/datasets.htm>.
- [5] ICTCLAS [EB/OL]. [2009 - 05 - 01]. <http://ictclas.org/>.
- [6] WEKA [EB/OL]. [2009 - 05 - 01]. <http://www.cs.waikato.ac.nz/ml/weka/>.

更新模型。

2) 提出基于主题点击率的主题更新模型,并结合模型指出如何进行单个主题和多主题更新,从而提高主页的受欢迎程度。

本方法也可通过参数的调节推广应用到对时新性要求更高的门户网站的主页更新中。本主题更新模型未来的改进方向是如何合理地将主题重要性以及主题关联两个因素融入模型中。

### 参考文献:

- [1] 涂承胜,鲁明羽,陆玉昌. Web 挖掘研究综述[J]. 计算机工程与应用, 2003, 39(10): 90 - 93.
- [2] HAN J, KAMBER M. Data mining: Concepts and techniques[M]. San Mateo, CA: Morgan Kaufmann, 2000.
- [3] 韩家炜,孟小峰,王静,等. Web 挖掘研究[J]. 计算机研究与发展, 2001, 38(4): 405 - 414.
- [4] SRIVASTAVA J, COOLEY R, DESHPANDE M, et al. Web usage mining: Discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2000, 1(2): 12 - 23.
- [5] 邵华,高风荣,邢春晓,等. 基于 VSM 的分层网页推荐算法[J]. 计算机科学, 2006, 33(11): 86 - 88.
- [6] 谭艳霞,徐珂. 基于大型网站的搜索引擎网页更新方法研究[J]. 微计算机信息, 2005, 21(11-3): 125 - 127.
- [7] 吕韩飞,王申康. 一种重要性与时新性结合的网页更新策略[J]. 计算机应用研究, 2005, 22(11): 212 - 214.
- [8] DOU ZHI-CHENG, SONG RUI-HUA, WEN JI-RONG, et al. Evaluating the effectiveness of personalized Web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(8): 1178 - 1190.