基于Web结构分区的协同过滤推荐算法研究

邵延振[1],蒙 韧[2],袁鼎荣[1],李新友[1]
(1.广西师范大学 计信学院,广西 桂林 541004;
2.广西师范大学 财务处,广西 桂林 541004)

摘 要:Intenet的快速增长导致了个性化服务的需求急剧增加。基于页面结构的信息提取与推荐是Web数据挖掘中三大研究领域之一。该研究的关键技术是识别Web页面的组织形式,从中挖掘所需要的个性化页面信息。基于Web数据挖掘的个性化信息推荐系统可以满足互联网未来发展趋势的需要。与传统的以页面为单位的Web信息提取相比,基于页面结构分区的信息推荐更符合实际情况,粒度优势明显。以一组数据为实例阐述了基于Web挖掘的协同过滤推荐算法是如何进行数据表示、近邻查询以及产生推荐页面分区信息的。

关键词:Web数据挖掘;推荐系统;协同过滤;页面分区;个性化信息
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2010)06-0067-03

Collaborative Filtering Recommendation Algorithm Research Based on Web Blocks

SHAO Yan-zhen[1], MENG Ren[2], YUAN Ding-rong[1], LI Xin-you[1]
(1. School of Computer Information and Eng., Guangxi Normal University, Guilin 541004, China;
2. Financial Department, Guangxi Normal University, Guilin 541004, China)

Abstract: With development of Internet, personalized service demand rapidly increased. Information extraction and recommendation based web page structure is one of three web data mining's research fields. Key technology of the research is how to recognize web page's organization form and mine the needed information. Personalized recommender system based web data mining can meet the need of the Internet's future development trends. Compared with based on web page, the based web block is more accordant to the fact and the advantage of granularity is evident. In this paper, with one set of data as an example collaborative filtering recommendation algorithm was elaborated based on web data mining how to work in the progress of data expression, neighbors queries and recommend generation.

Key words: web data mining; recommendation system; collaborative filterring; web block; personalized information

0 引 言

当前,Web站点通常以Web的形式展现产品和服务信息以供用户浏览,是一种典型的"one-size-fits-all"的方法[1],这样使得Web站点提供给用户的信息根本没有考虑用户个性化的需求,而是以一种方式对待所有用户。近年来在国外兴起的个性化推荐成为解决这一问题的重要方法。个性化推荐就是在通过收集和分析用户信息来预测用户行为,进而实现主动推荐的服务模式,其实是以用户需求为中心的Web服务。它可以根据用户的偏好、历史访问信息以及其他用户的相关信息,通过推荐个性化信息,为用户提供定制的Web体验。它不仅能减轻用户的"信息过载",而且可以帮助企业建立良好的客户关系。

1 Web数据挖掘

Web挖掘是将数据挖掘技术与互联网相结合的一项综合技术,简单地说"Web挖掘就是从Web文档、Web活动中抽取感兴趣的、潜在的有用模式和隐藏信息[2~5]。Web上有海量的数据信息,怎样对这些数据进行复杂的应用成了现今数据库技术的研究热点,数据挖掘就是从大量的数据中发现隐含的规律性的内容,解决数据的应用质量问题。Web上数据基本上都是来源于异构数据库环境,而且大多是半结构化的数据结构,这就给数据处理带来了许多挑战。它面对的信息常常为文本、图形、图像数据等半结构化的数据,

收稿日期:2009-10-28;**修回日期**:2010-01-08
基金项目:广西自然科学基金(桂科自0640069)
作者简介:邵延振(1983-),男,山东泰安人,硕士生,研究方向为数据挖掘,Web信息挖掘与检索;袁鼎荣,副教授,研究方向为数据挖掘。

甚至是异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。网络信息挖掘大致分为四个步骤：

①资源发现，即检索所需的网络文挡；

②信息选择和预处理，即从检索到的网络资源中自动挑选和预先处理得到专门的信息；

③概括化，即从单个的 Web 站点以及多个站点之间发现普遍的模式；

④分析，对挖掘出的模式进行确认或解释。

一般地，Web 数据挖掘可以分为三类：Web 内容挖掘(Web content mining)，Web 结构挖掘(Web structure mining)，Web 使用记录挖掘(Web usage mining)。如图 1 所示。

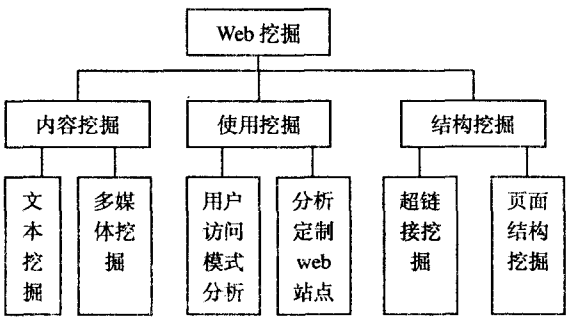

图 1 Web 挖掘分类

2 页面分区

以整个的 Web 页面作为最小的信息单元的方式已逐渐不能适应 Web 页面信息挖掘的快速发展，正如一个大的门户网站要将自己的页面分为很多区块，如新闻，体育，财经，娱乐，文化等等，因此，把页面按照一定的算法划分为若干个区域(Block)，把这些区域作为基本的信息处理和提取单元，提出该方法的理由如下：

(1)当今的 Web 页面大部分是由多个不相关的主题页面 Block 构成，以页面 Block 为最小单位来个性化推荐 Web 信息更符合实际情况，粒度上优势明显。

(2)基于页面 Block 进行信息推荐，可以适当忽略稳定的且内容无关的区域，节省处理和存储代价。

Block 是页面中在内容和显示上独立的、闭合的矩形区域。Web 页面可以分割为若干互不相交的 Block，把这个过程称为页面分区。典型的分区如图 2 所示。

对于页面分区的 DOM 树，VIPS 算法见参考文献[6~10]，这里不再重复。

3 协同顾虑推荐算法

用户相似性度量、最近邻居查询和预测评分是整个协同过滤推荐算法的主要工作，相应地，协同过滤推荐算法可以划分三个阶段：

(1)数据表示：对用户已经浏览的页面分区进行进模，从而可以有效度量用户之间的相似性。

(2)最近邻查询：搜索当前用户的最近邻居。

(3)推荐产生：根据当前用户最近邻居的访问情况产生推荐页面分区集。

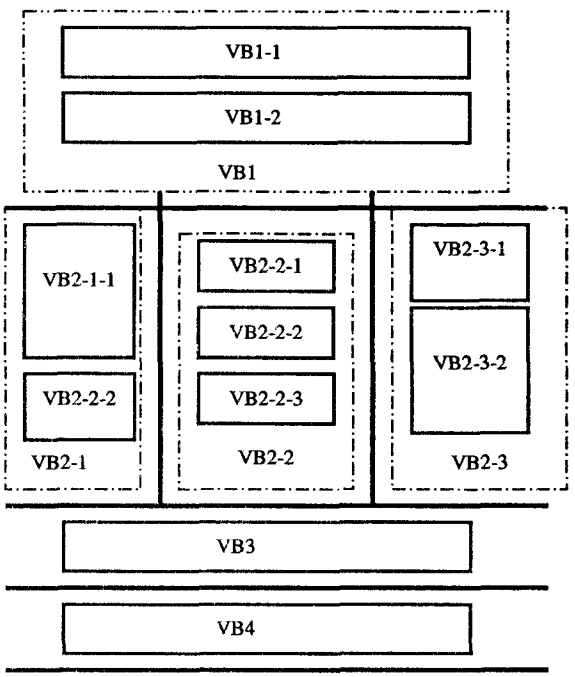

图 2 VIPS 页面分区实例

3.1 数据表示

协同过滤根据用户的行为(如用户注册信息、用户评分数据、用户浏览行为等)建立用户的行为模型，然后利用建立的行为模型向用户推荐有价值的页面分区，用户数据的收集在协同过滤推荐算法中占有重要地位，如何有效收集高质量的用户数据直接关系到推荐算法的推荐效果。

协同过滤的实现一般分为两步：首先，获得用户信息，即获得用户对某些分区项的评分；其次，分析用户之间的相似性并预测特定用户对某一信息的评价。用户评分数据可以用一个 $m \times n$ 阶矩阵 $\mathbf{A}(m,n)$ 表示，m 行代表 m 个用户，n 列代表 n 个分区项目，第 i 行的第 j 列元素 R_{ij} 代表用户 i 对分区项目 j 的评估数值，评估值与项的内容有关，如果项是 Web 文挡，则表示浏览与否，用户对它兴趣有多高，这样的评估值可能分为几个等级比如 1 ~ 5 等。如果项是电子商务中的商品，则表示用户订购与否，比如 1 表示订购，0 表示没有订购。

表 1 是一个用户访问站点页面分区的表，表中的数字对应于用户浏览该页面分区的兴趣度(以 0~9 的兴趣度作为划分的标准)，与之相对应的是用户页面分

区兴趣度矩阵。考虑到用户的浏览模式不是一成不变的,用户浏览购买过程是动态的,按照访问时间和最近访问的原则对用户进行聚类,从而能够保证发现的用户聚类模式符合用户当前的浏览行为。

表 1　用户访问站点分区兴趣度表

访问者访问	开始访问时间	新闻	娱乐	体育	财经	科技	军事
用户 1	09:07:20	9	7	5	0	3	9
用户 2	09:08:36	6	3	9	7	0	2
用户 3	09:20:22	5	9	9	0	8	0
用户 4	09:25:23	9	6	4	1	4	8
用户 5	09:30:24	8	6	4	0	3	9
用户 6	09:37:25	9	7	5	0	2	7
用户 7	09:45:10	8	6	4	1	4	8
当前用户	10:00:00	9	7	5	1	3	?

根据设定的时间范围,由用户访问站点的页面分区(block)的兴趣度形成了矩阵:

$$M_{7\times6}=\begin{bmatrix}9&7&5&0&3&9\\6&3&9&7&0&2\\5&9&9&0&8&0\\9&6&4&1&4&8\\8&6&4&0&3&9\\9&7&5&0&2&7\\8&6&4&1&4&8\end{bmatrix}$$

在表 1 所示的用户兴趣度矩阵中,协同过滤算法需要对当前用户对页面军事分区的兴趣度进行预测。通过数据发现当前用户与用户 1 的兴趣度评分非常相似。当前用户前五个页面分区的兴趣度分别是 9、7、5、1、3,而用户 1 对五个页面分区的兴趣度是 9、7、5、0、3,他们二者的相似度最高,所以用户 1 是当前用户的最近邻居,而其他用户不是当前用户的最近邻居。在实际的推荐过程中,可以根据用户 1 的访问情况对当前用户进行推荐。

3.2　最近邻查询

搜集完人们的偏好数据后,需要确定一种方法来确定人们在品味方面的相似程度。协同过滤系统的核心是为当前用户寻找相似的"最近邻居"集。即:对一个用户 u,要产生一个依相似度大小排列的"邻居"集合 $N=\{N_1,N_2,\cdots,N_t\}$,u 不属于 N 从 N_1 到 N_t,距离 $d(u,N_k)$ 为用户 u 和其邻居 N_k 的距离,用来度量二者之间的相似性。度量用户 u 和其邻居 N_k 之间相似性的的方法如下:这里 N_k 是其他不同的用户。首先得到用户 u 和 N_k 评分的所有项,然后通过欧氏距离度量形式如下:

$$d(x,y)=\left[\sum_{i=1}^{n}|x_i-y_i|^2\right]^{1/2}\qquad(1)$$

从直观上讲,属于同一类的用户事物对象在空间中越接近越好,而不同类的用户事物对象之间的距离越大越好,所以,用户间的距离越小,他们的相似性越大。以表 1 的数据作为数据集合,根据欧氏距离进行距离矩阵的计算如表 2 所示,以每一个用户之间的距离作为其相似度的依据进行聚类的划分。

表 2　用户之间的欧氏距离表

	用户 1	用户 2	用户 3	用户 4	用户 5	用户 6	用户 7
用户 1	0						
用户 2	12.16	0					
用户 3	11.92	12.41	0				
用户 4	2.24	11.44	11.44	0			
用户 5	1.73	12.04	12.20	2.00	0		
用户 6	2.23	10.90	11.00	2.83	2.83	0	
用户 7	2.45	10.72	11.13	1.00	1.73	3.00	0

从表中可以看出第 1 个用户和第 2 个用户之间的距离为 12.16,而第 1 个用户和第 3 个用户之间的距离为 11.92,第 2 个用户和第 3 个用户之间的距离为 12.41,依次类推。输入聚类数目为 3,随机选取前 3 个用户为初始点。由于距离越近相似度越高,分析距离矩阵可知:用户 4,用户 5,用户 6,用户 7 都与用户 1 距离最近,所以形成的聚类为{1,4,5,6,7},{2},{3},然后计算第一个用户到这 3 个聚类中心的距离和,找出距离最近的对象作为新的聚类中心,重新计算距离矩阵。再反复进行迭代。

3.3　推荐产生

通过上面提出的相似性度量方法得到目标用户的最近邻居,下一步需要产生相应的推荐。首先根据用户当前的访问集合得到与他具有类似访问集合的最近邻居,那么这些最近邻居的其它页面分区访问就可以作为该用户的推荐集合,因此可以向当前用户推荐有关军事方面的信息。

4　结束语

根据信息提取和信息推荐的发展趋势[11],文中对基于 Web 数据挖掘的网页分区推荐系统所使用的算法进行了研究。文中的研究内容主要包括信息推荐及 Web 数据挖掘的技术研究,以及页面分区的信息推荐系统的算法研究。

参考文献:

[1] 谢　中.基于 Web 数据挖掘商务网站的推荐系统的研究[D].重庆:西南师范大学,2002.

[2] Apte C, Weiss S M. Data mining with decision trees and decision rules[J]. Future Generation Computer Systems, 1997, 13

(下转第 73 页)

时间是1.37秒,使用CPU亲和力后,程序执行时间明显减少。

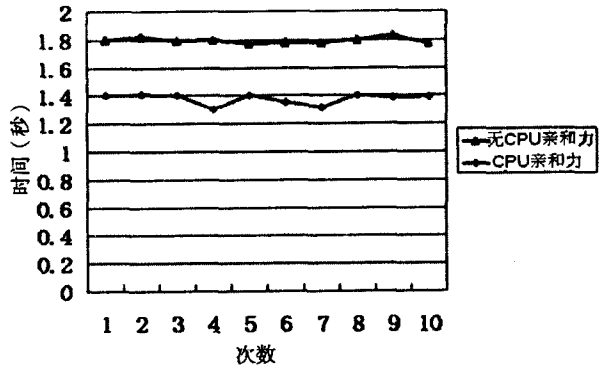

图3　是否使用CPU亲和力的时间比较

在上例中,计算fool.a和fool.b的两个线程没有绑定到同一个CPU核心,而fool.a和fool.b由于在同一结构体中,所以在内存中是相邻的数据,现在将计算fool.a和fool.b的线程绑定到同一个核心上。

如图4所示,将fool.a和fool.b绑定到同一核心后,程序平均运行时间由原来的1.37秒减少到了0.89秒。这是因为将访问相同数据或相邻数据的一组线程绑定到一个CPU核心上,可以提高一组线程的Cache命中率,从而提高程序执行效率。

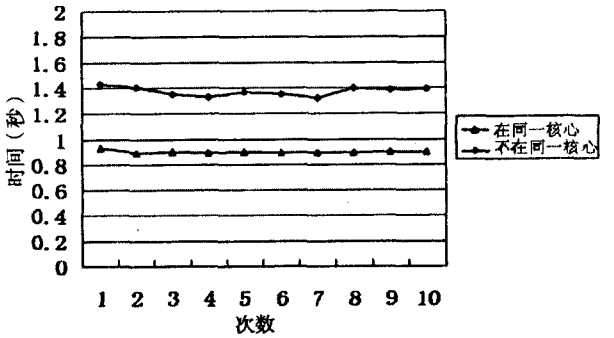

图4　fool.a和fool.b绑定到不同核心的比较

4　结束语

文中针对多核编程环境的特点,在使用传统的多线程并行编程技术的基础上,介绍了基于Cache优化和CPU亲和力的多线程程序优化思路,在实际实验中,单线程原始程序平均耗时0.76秒,最快的三线程优化方案为0.44秒,效率提高了约42%。基于Cache和CPU亲和力的优化方法具有一定的通用性,在软件开发过程中,可以根据实际情况采用这两种优化方法,从而有效地提高软件的运行效率。

参考文献:

[1] Reinders J. Programming For Parallelism[EB/OL]. 2007. http://www.cajcd.edu.cn/pub/wml.txt/980810-2.html.

[2] Stevens W R, Rago S A. Advanced Programming in the UNIX Environment[M]. 北京:人民邮电出版社,2006.

[3] 杨　静,李　炜,万峰松,等. Linux2.6内核进程调度分析与改进[J]. 计算机技术与发展,2009,19(7):105-107.

[4] 王　晶,樊晓桠,张盛兵,等. 多核多线程结构线程调度策略研究[J]. 计算机科学,2007,34(9):256-258.

[5] Doweck J. Inside Intel Core Microarchitecture and Smart Memory Access[EB/OL]. 2006. http://download.intel.com/technology/architecture/s-ma.pdf.

[6] 李晓明,臧斌宇,郑纬民,等. 多核程序设计[M]. 北京:北京大学出版社,2007.

[7] 金国华,陈福接. 简单访问模式下假共享Cache行抖动的消除[J]. 计算机学报,1994(6):435-445.

[8] Love R. CPUaffinity[EB/OL]. 2003. http://www.linuxjournal.c-om/article/6799.

[9] Bovet D P, Cesati M. Understanding the Linux Kernel[M]. 北京:中国电力出版社,2007.

[10] 杨　磊,石　磊,张铁军,等. 多核系统中共享cache的动态划分[J]. 微电子学与计算机,2009,26(5):56-59.

(上接第69页)

(2-3):197-210.

[3] 谢　榕. 数据挖掘与商业智能系统[J]. 计算机系统应用,1999(8):9-10.

[4] 王晓宇,熊　方,凌　波,等. 一种基于相似度的主题提取和发现算法[J]. 软件学报,2003,14(9):1578-1585.

[5] 李晓明,朱家稷,阎宏飞. 互联网上主题信息的一种收集与处理模型及其应用[J]. 计算机研究与发展,2003,40(12):1667-1671.

[6] Cai D, Yu S, Wen J R, et al. Block-based Link Analysis [C]//in 27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, South Yorkshire, UK: [s. n.], 2004.

[7] 宋　杰,王大玲,鲍玉斌,等. 基于页面Block的Web档案采集和存储[J]. 软件学报,2008,19(2):275:275-290.

[8] Cai D, Yu S, Wen J R, et al. VIPS: a version-based page segmentation algorithm[R]. US: Microsoft, 2003.

[9] Cai D, Yu S, Wen J R, et al. Block-based Web Search[C]// in 27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, South Yorkshire, UK: [s. n.], 2004.

[10] 张　敏,高剑锋,马少平. 基于链接描述文本及其上下文的Web信息检索[J]. 计算机研究与发展,2004,41(1):221-226.

[11] 宋聚平,王永成,尹中航,等. 面向主题的网页搜索系统[J]. 上海交通大学学报,2003,37(3):401-403.