WEB 用户的视图*

阳小华[1,2] 周龙骧[1]

[1](中国科学院数学研究所计算机科学研究室 北京 100080)

[2](中南工学院计算机系 衡阳 421001)

摘要 视图不仅是数据库中的一个重要概念,也能够在 Web 系统中发挥重要的作用.但是,Web 视图不能完全照搬数据库视图的概念,而应该体现出 Web 特色.文章提出了浏览区域的概念,能较好地刻画 Web 用户活动的特征.在此概念的基础上,给出一个能体现 Web 特色的用户视图的定义,初步探讨了 Web 用户视图的实现方法和一些可能的应用.

关键词 视图,万维网,浏览,查询.

中图法分类号 TP393

视图是数据库中的一个重要概念.在关系数据库中,它是从一个或几个基本表(或视图)导出的表[1].数据库管理员可以为特定的用户或用户群定义一个或多个视图,以限制或引导他们对数据库的访问.用户也可以定义视图,以把自己的视野集中在有意义的范围内.

World Wide Web(万维网)是一个信息的海洋,用户要从中获取所需的信息,有时如同大海捞针一样困难.而事实上,虽然 Web 资源丰富,但是每个用户都有自己特定的爱好和需求,他只会对某些特定的资源感兴趣.因此,如果能够像数据库系统那样,引入视图的概念,使用户能够以自己特定的方式来看待 Web 系统,把注意力集中在所关心的范围内,将十分有利于用户获取所需的信息.

虽然在 Web 上有许多索引查询系统[2],使用户可以通过关键字检索等手段来帮助寻找所需的信息.但是,从本质上看,访问 Web 文档的基本手段是直接通过 URL,或间接通过超链.也就是说,浏览是访问 Web 的基本方式,而查询只不过是在浏览方式的基础上提供的一种辅助手段.因此,Web 视图不宜像数据库视图那样依赖于特定的查询操作,而应该体现出 Web 用户获取信息的浏览特征.

本文主要从 Web 用户的角度来考察视图.我们着眼于客户端,通过对 Web 用户获取信息的过程进行分析,提出了浏览区域的概念,较好地刻画了 Web 用户活动的本质特征.在此概念的基础上,给出了一个能体现 Web 特色的用户视图(在客户端)的定义,初步探讨了 Web 用户视图的实现方法和一些可能的应用.

1 Web 用户视图的定义

1.1 Web 用户的浏览区域

Web 是一个由众多文档组成的超文本信息系统,它有两个基本成分:文档(只考虑 HTML 文档)和超链(只考虑文档之间的超链).如果我们把文档看作结点,把文档间的超链看作从源文档指向目标文档的有向边,则可以把 Web 简单地看作一个有向图.

如果忽略有向边的方向,即把有向边可作无向边,则有向图变成无向图,称之为(原有向图的)底图.如果在一个无向图中,任意两个结点之间都有路径可达,则称之为连通图.如果一个有向图的底图是连通图,则称该有向图为弱连通图.

* 本文研究得到国家自然科学基金资助.作者**阳小华**,1963 年生,博士生,副教授,主要研究领域为 WWW 智能查询,数据库技术.**周龙骧**,1938 年生,研究员,博士生导师,主要研究领域为数据库实现技术,WWW 智能查询,数据采掘.

本文通讯联系人:阳小华,北京 100080,中国科学院数学研究所计算机科学研究室

本文 1998-05-29 收到原稿,1998-08-14 收到修改稿

定义 1. Web 中的弱连通子图称为 Web 的区域.

简单说来,一个 Web 用户的工作就是从 Web 丰富的资源中不断地发现和获取自己感兴趣的信息. 在 Web 中,各种信息散落在众多的文档中,而大量的相关文档又常常通过超链链接,构成一个个区域. 因此,Web 用户活动的基本特征是浏览,通过浏览一个个 Web 区域来满足自己特定的信息需求. 一个用户在不同的时候会有不同的需求,从而在其活动中就会表现出不同的特点. 根据这些特点,我们可以把用户的活动划分为一个个活动单元,称之为用户的浏览过程.

定义 2. 设 $D=\langle d_1,d_2,\ldots,d_n\rangle$是一个 Web 用户已经访问过的全部文档(按时间先后顺序)所组成的序列(若同一个文档在不同的时候被重复访问,则它在 D 中不同的位置重复出现),$S=\langle d_i,\ldots,d_j\rangle$是 D 的某个子序列,称 S 为该用户的第 $k(1\leqslant k)$次浏览过程,如果

(1) 对 S 中的任一元素 $d_q(i<q\leqslant j)$,都存在着元素 $d_p(i\leqslant p<q)$,使得在 d_p 中有指向 d_q 的超链,而且用户是通过这条超链访问到 d_q 的;

(2) 用户没有通过 S 中某个文档里的超链访问 $d_{j+1}(j<n)$;

(3) 用户没有通过第 $k-1$ 次浏览过程中的某一元素里的超链访问 $d_i(k>1)$.

定义 3. 设 $S=\langle d_i,\ldots,d_j\rangle$是一个浏览过程,则称 S 中包含的元素的个数为它的长度.

定义 4. 设 S 是一个 Web 用户的某一次浏览过程,称以 S 中的元素为结点,以这些元素之间用户使用过的超链为边构成的有向图为该用户的一个浏览区域.

浏览过程 S 中的第 1 个元素 d_i 称为其浏览起点. 由定义 2 可以看出,从浏览起点到浏览过程中的其他元素之间存在着由用户使用过的超链构成的有向路径. 因此,浏览区域是 Web 的弱连通子图,即为 Web 的区域. 由于用户的每一次浏览都有其特定的目的和需求,在一个浏览过程中,用户之所以会从起点顺着特定的超链序列访问某些文档,是因为他认为这些文档中有他此时想要的东西. 因此,每一个浏览区域通常都有自己的特性,即浏览区域中的文档都有某些共同的性质. 浏览区域可大可小,最小的可以只有 1 个结点(起点). 由于超链链接表示的是文档之间的逻辑联系,所以一个浏览区域可以包含位于不同服务器上的文档.

1.2 Web 用户的视图

虽然 Web 用户在不同的时候通常会有不同的信息需求,但是用户一般都会有一些相对稳定的兴趣,比如说对某些特定的主题、Web 域或者 Web 站点. 因此,在一些不同的浏览过程中,用户会表现出相同或相似的需求,从而相应的浏览区域会表现出相同或相似的特性.

如果有一个适当的判别规则,那么就可以把浏览区域分为一个个相似类,使得在同一个相似类中的浏览区域彼此都是相似的,它们对应于用户相同的兴趣. 若用户对某个相似类的兴趣强度超过了特定的阈值,我们就可以把它定义为一个视图.

定义 5. 设 W 是一个 Web 用户访问过的浏览区域组成的集合,R 是一个给定的相似性判别规则,S 是 W 关于 R 的相似类的集合,F 是一个从 S 到非负实数的函数(用来度量用户对相似类的兴趣强度),H 是一个给定的阈值,C 是 S 在关系 R 下的一个相似类,若 $F(C)>H$,则称 C 为该用户的一个视图.

2 Web 用户视图的获取

2.1 相似性判别规则

相似性判别规则是浏览区域集合上的一个关系(称为相似关系). 显然,每一个浏览区域应该是与自己相似的. 其次,若浏览区域 A 与浏览区域 B 相似,则浏览区域 B 也应与浏览区域 A 相似. 也就是说,相似关系应该是自反的、对称的,因而是一个相容关系.

由于每一个浏览区域中的文档都有某些共同的性质,因此,如果两个浏览区域之间有重迭部分,则由于重迭部分兼有两个区域的特性,我们有理由认为这两个区域很可能有相似的特性或者说包含的信息有某种程度的相似性. 另外,若浏览区域 A 和 B 都与 C 重迭,即使它们彼此之间并没有重迭,A 和 B 也会有某种程度的相似性,这种相似性可以看做是由 A 与 C 和 C 与 B 的相似性传递得到的. 所以,我们有如下的相似性判别规则:

设 A,B 是两个浏览区域,我们说它们是相似的,如果下面的条件(1)或者(2)成立:

(1) A 和 B 包含有重迭的部分,即有一些共同的文档;

(2) 存在着浏览区域 $A_0=A,A_1,\ldots,A_k=B$,使得 A_{i-1} 与 $A_i(i=1,\ldots,k)$ 是重迭的,其中 $k\leqslant m$(m 是一个设定的常量),此时,我们称 A 和 B 在不超过 m 步传递的意义下是重迭的.

由于相似性的传递随着传递步数的增加,相似性一般会越来越弱.因此,需要一个限制传递步数的上界 m,只有当两个浏览区域在不超过 m 步传递的意义下是重迭的,我们才能认为它们是相似的.很容易发现,当 $m=1$ 时,不超过 m 步传递的意义下重迭就是区域本身重迭.m 的选取应视具体情况而定,不要太大,一般取 2 或 3 可能比较适当.

上面的相似性判别规则简单易用,有一定的准确性(不完全准确).但是,由于它没有深入考虑浏览区域的信息内容,因而有很大的局限性(有些相应于相同信息主题的浏览区域不能够满足我们的规则).寻找更加准确实用的相似性判别规则是我们需要进一步努力的工作.

2.2 兴趣强度度量

如果用户对某些类型的信息有较大的兴趣,那么他就会花费较大的精力来探寻、获取这些信息.由于 Web 用户探寻、获取信息的基本特征是浏览,用户通过一次次浏览过程来满足自己的一个个信息需求.因此,如果用户对某些信息有特别的和持续的兴趣,那么他就会对这些信息所在的地方进行比较深入的探索.有两个迹象可以反映出这一点.一方面,如果某个浏览过程的长度远远大于其他浏览过程的长度,那么我们有理由认为用户在这个浏览过程中花费了大量的精力,因此可以认为用户对相应的浏览区域有较大的兴趣.另一方面,如果有许多个浏览过程都对应于某种信息需求,那么可以认为这是用户的一个持续的兴趣所在.

相似类由彼此相似的浏览区域组成,与用户的某种特定的兴趣和信息需求相对应,每个浏览区域对应于一个浏览过程.因此,我们可以通过综合考虑相似类中所包含的浏览过程的个数和这些浏览过程的长度来度量用户对它的兴趣强度.下面,我们给出一个简单的计算公式.

$$F(C)=(l/L+n/N)/2. \tag{1}$$

其中 F 是用户兴趣强度的度量函数,C 是一个相似类,l 是相似类 C 的长度(C 中包含的浏览过程的长度之和),L 是用户轨迹的长度(所有浏览过程的长度之和),n 是 C 中含有的浏览过程的个数,N 是用户浏览过程的总数.从公式(1)可以看出,一个相似类包含的浏览过程越长、越多,则意味着用户对它的兴趣越大、越持久.显然,我们有

$$\Sigma_{C\text{是相似类}}F(C)=1. \tag{2}$$

2.3 阈 值

由于视图是 Web 中用户特别感兴趣的地方,因此,阈值一般应明显地高于平均兴趣强度值.

2.4 视图的获取

给定相似性判别规则、用户兴趣强度函数和阈值,就可以从用户的轨迹中自动地抽出视图了.考虑到用户活动的复杂性,在抽取的过程中允许适当的人工干预.一般地说,抽出的视图要经过用户的确认(编辑),才能真正作为视图保存起来.另外,还允许用户人工调整某些相似类的兴趣强度值,以满足某些特殊的需要,比如说,通过明确表示出自己对某些资源的强烈兴趣来定义视图.

由于用户的轨迹在不断地变化,因此,视图的获取不会是一劳永逸的.随着时间的推移,可能会有新的视图产生,而旧的视图也需要调整或删除,以反映新的变化.相关的一个问题是何时进行视图的获取及调整.由于可能需要用户的参与,因此还是由用户在感到需要和方便的时候进行获取或调整为好.

例 1:设 $W=\langle A,B,A,U,D,E,G,Y,T,U,I,U,J,G,X,G,Y,M,V,M,O,P,R,A,I,K,R,A,B\rangle$ 是某个用户访问过的文档序列,在 W 中有 10 个浏览过程:$\langle A,B,A,U\rangle$,$\langle D,E\rangle$,$\langle G,Y\rangle$,$\langle T\rangle$,$\langle U,I,U,J\rangle$,$\langle G,X,G,Y\rangle$,$\langle M,V,M,O,P,R\rangle$,$\langle A,I\rangle$,$\langle K,R\rangle$,$\langle A,B\rangle$.按照相似性判别规则($m=2$),相应的浏览区域被划分为 5 个相似类 (C_1,C_2,C_3,C_4,C_5),每个类对应于一个 Web 区域,如图 1 所示.

其中 C_1 包含 4 个浏览过程:$\langle A,B,A,U\rangle$,$\langle U,I,U,J\rangle$,$\langle A,I\rangle$,$\langle A,B\rangle$,其长度为 12;C_2 只有 1 个浏览过程:

$\langle D,E\rangle$,其长度为 2;C_3 也只有 1 个浏览过程:$\langle T\rangle$,其长度为 1;C_4 含有两个浏览过程:$\langle G,Y\rangle$,$\langle G,X,G,Y\rangle$,其长度为 6;C_5 包含两个浏览过程:$\langle M,V,M,O,P,R\rangle$,$\langle K,R\rangle$,其长度为 8.$W$ 的长度为 29.

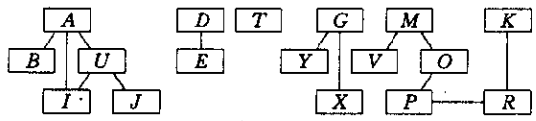

图1

根据公式(1),用户对这 5 个相似类的兴趣强度分别为:

$$F(C_1)=(12/29+4/10)/2=0.510,$$
$$F(C_2)=(2/29+1/10)/2=0.084,$$
$$F(C_3)=(1/29+1/10)/2=0.067,$$
$$F(C_4)=(6/29+2/10)/2=0.203,$$
$$F(C_5)=(8/29+2/10)/2=0.238.$$

因此,如果我们设定阈值为 0.3(平均兴趣强度为 0.2),则 C_1 将被抽出来作为(候选)视图提交给用户.

3 小　结

在本文中,我们把 Web 用户的视图定义为相似的浏览区域的集合.与其他关于 Web 视图的工作相比[3,4],本文的工作有两个突出的特点:一个是 Web 用户视图不依赖于特定的查询系统,不是建立在查询操作的基础上,而是以浏览操作为基础.由于浏览是 Web 用户获取信息的基本手段,因此我们的定义具有一般性,体现了 Web 特色.另一个是 Web 用户视图不需要预先定义,而是从用户的活动轨迹中自动提取.因此,它不会对用户提出额外的要求,适合所有的 Web 用户.

Web 用户视图有许多的应用,例如,它可以用于缓冲管理、浏览引导,还可以用来提高查询的质量等.

参考文献

1 周龙骧.数据库管理系统实现技术.北京:中国地质大学出版社,1989.17～18
(Zhou Long-xiang. The Techniques for Implementing Database Manager System. Beijing: University of Geology of China Press, 1989. 17～18)

2 阳小华,周龙骧.World Wide Web 索引与查询技术.计算机科学,1997,24(6):26～34
(Yang Xiao-hua, Zhou Long-xiang. The index and query techniques of world wide web. Computer Science, 1997,24(6): 26～34)

3 Dan Suciu. Query decomposition and view maintenance for query language for unstructured Data. In: Vijayarman T M, Buchmann A, Mohan C eds. Proceedings of the 22nd VLDB Conference. San Francisco: Morgan Kaufmann Publishers, Inc., 1996. 227～238

4 David Konopnicki, Oded Shmueli. W3QS: a query system for the world-wide web. In: Dayal Umeshwar, Gray Peter M D, Nishio Shojiro eds. Proceedings of the 21st VLDB Conference. San Francisco: Morgan Kaufmann Publishers, Inc., 1995. 54～655

The View of Web Users

YANG Xiao-hua[1,2]　ZHOU Long-xiang[1]

[1](*Department of Computer Science　Institute of Mathematics　The Chinese Academy of Sciences　Beijing*　100080)
[2](*Department of Computer Science　Central-South Institute of Technology　Hengyang*　421001)

Abstract　View is an important idea of database and it can also be used effectively in the Web. But the view of Web is not the same as that of database, since the Web is different from the database. In this paper, the authors present an important idea, the area of navigation, which describes the characteristics of Web users' activities. Based on this idea, the definition of Web user's view is given and its implementation and some possible applications are also discussed.

Key words　View, world wide web, navigation, query.