Journal of Wuhan Polytechnic

基于校园 web 浏览行为的师生关注度计算方法

李　冬

（商丘职业技术学院 计算机系，河南 商丘 476000）

摘　要：了解师生对校园网不同主题信息的关注度，是校园网建设的重要一步，也是决定校园网服务质量的关键因素。通过分析网页浏览行为，结合用户关注度衰减的规律，提出一种基于校园 web 浏览行为的师生关注度计算方法，利用 K-means 算法对师生浏览主题进行聚类，计算出师生主题关注度，最后通过实验证明所提出的师生关注度计算方法是有效的。

关键词：校园 web；浏览行为；师生关注度

中图分类号：TP393.18　　**文献标识码：**A　　**文章编号：**1671-931X（2016）01-0070-04

一、引言

校园网是不少院校传达信息的重要平台，甚至一些院校的信息发布只针对校内用户，校外用户无法进行浏览，通过对不同院校网页浏览的数据信息可以得知，无论查看校园 web 信息的用户权限是校内还是校外，浏览校园网页的主体都是本院校的教师和学生，因此，及时了解师生对校园网不同主题信息的关注度，是校园网建设的重要一步，也是决定校园网服务质量的关键因素。师生自己标注关注的校园网页是一种可信度很高的获得师生关注度的方法，但大多数师生在浏览校园网页时并不太在意参与标注，因此就需要通过校园 web 浏览行为来计算师生关注度的方法来获得此项数据信息。

二、用户浏览行为分析

根据师生通过校园网对相关内容进行浏览的行为进行综合分析，大致可以分为以下几种类型。

标记行为：用户增加、删除书签；对页面进行保存；直接打印页面等[1]。

操作行为：剪切、粘贴、复制、拖拽滚动条以及点击网页连接等[2]。

重复行为：重复对同一个页面进行访问等。

由于师生对校园网页的处理动作属于即时的心理反应动作，能更好的反映出师生对网页是否关注，师生如果对此页面非常关注，希望以后再次进行阅读和查询，较多的会采用保存页面或收藏页面的操作方式，即使在条件允许的情况下，采用打印页面的操作方式也比较少，而拉动滚动条、点击链接等操作行为在浏览网页的时候一般都使用的比较频繁，并不能很准确的反映师生是否对网页关注。此外，由于浏览网页的时间跟网页内容的大小关系十分密切，不能单纯通过网页浏览时间的长短去判断师生对网页关注程度的大小。结合上述分析，本文通过保存页面、收藏页面和网页浏览速度这三种最能体现用户关注度的浏览行为，来计算师生对校园网信息的关

收稿日期：2015-11-10

基金项目：河南省基础与前沿技术研究项目“云计算资源调度优化研究”（项目编号：132300410445）。

作者简介：李冬（1982-），男，河南商丘人，硕士，商丘职业技术学院计算机系讲师，研究方向：计算机应用技术。

注程度。

用户关注度是指用户对一个网页上的内容关注程度的大小，用 0-1 之间的实数表示，0 表示不关注，1 表示最大关注程度[3]。

目前，通常用来计算用户对网页关注（兴趣）度的方法主要可以分为两类，一类是基于浏览内容的方式，一类是基于用户行为的方式[4]。基于浏览内容的方式计算的对象完全来自于网页面本身的内容，一般采取通过关键词频率和权重计算来衡量浏览者关注度，这种方法可以帮助得到用户浏览内容的概括，但它没有考虑用户浏览页面时能反馈心理的动作信息，基于此本文采用分析浏览网页行为的方式进行师生关注度的计算。设定初始师生关注度为 Interest×[0，1]。对每个校园网页 l 的师生关注度采用三种浏览行为来计算，即保存页面表示为 save(l_i)；收藏页面表示为 bookmark(l_i)；在页面上的驻留时间表示为 time(l_i)，校园网页内容的大小会对驻留网页的时间产生实际影响，因此重新考虑三种浏览行为的方式，采用保存网页 save(l_i)、收藏页面 bookmark(l_i)和网页浏览速度 speed(l_i)作为本文所考虑的网页浏览行为。其中页面用 l 表示，网页数用 i，师生所访问的第 i 个网页页面表示为 l_i。本文在计算师生对校园网页面的关注度时做了如下规定。

1.基于保存和收藏页面的特殊性，只要发生保存和收藏页面其中一个行为，就表明师生对这个页面的关注度很高，在此种情况下浏览网页速度就不再作为计算因子。此时收藏和保存页面与用户关注度的关系为一个二值函数。

$$h(l_i)=\begin{cases}0 & save(l_i)+bookmark(l_i)=0\\ 1 & save(l_i)+bookmark(l_i)>0\end{cases} \quad (1)$$

此时，师生关注度与式 h(l_i)存在如下关系：

$$\omega 1=Interest(l_i)\cdot h(l_i) \quad (2)$$

2.在保存和收藏页面这两个行为都没有发生的情况下，只计算浏览网页速度所反映的师生关注度。

师生对校园网页的浏览网页速度 speed(l_i)取决于网页内容的大小 size(l_i)和页面驻留时间长短 time(l_i)，然而当师生浏览页面时间太短，少于 5s 时，则认为师生对此校园网页不感兴趣、不关注；而驻留页面时间太长，大于 3min 时，则考虑为页面驻留时间异常。所以为避免以上两种情况，在 5s<time (l_i)< 3min 时，才考虑为正常页面驻留时间。而页面浏览速度 speed(l_i)依据页面驻留时间 time(l_i)和网页大小 size(l_i)产生的兴趣度函数定义为：

$$speed(l_i)=\frac{size(l_i)}{time(l_i)} \quad (3)$$

师生浏览网页速度越快，表明对此校园网页越不关注，为了避免值过度偏离 1，则师生关注度与浏览网页速度存在如下关系：

$$\omega 2=Interest(l_i)\cdot\frac{1/speed(l_i)}{\max\{1/speed(l_i)\}}(\omega\in i) \quad (4)$$

基于上述理论，可以得到每个校园网页的师生关注度计算公式如下：

$$Interest(l_i)=\omega 1+[1-\omega 1]\times\omega 2 \quad (5)$$

三、师生关注度的衰减

时间对师生关注度有着深入而广泛的影响。师生关注度是会随着时间而变化的。基于此，一些国内学者就假设人对某一事物的关注（兴趣）如果在没有外界刺激去加强它的情况下会随时间而衰减，艾宾浩斯就此提出了这种记忆保存量与时间关系的函数，即关注（兴趣）衰减因子：

$$K(x)\cdot\frac{k}{(\lg T)^c+k} \quad (6)$$

其中，表示记忆保存量，T=cur-per（cur 表示当前时间，per 表示第一次浏览该网页的时间），单位为分钟，c=1.25、k=1.84 时公式（6）所代表的遗忘函数与人的遗忘规律比较匹配。因此，如果要准确预测师生现在的关注情况，就应该注意师生最近的关注行为，但是，考虑师生最近的关注只能针对渐变的师生关注情况，而对突变的师生关注情况是很难起作用。根据以上的理论，得到了随着时间的流逝师生关注度变化的计算公式：

$$Interest^*(l_i)=Interest(l_i)\times K(x) \quad (7)$$

其中，Interest(l_i)为用户原来对每个网页的关注度，Interest*(l_i)为衰减后的关注度。

根据以上公式，得出用户对每个网页的关注度公式如下：

$$Interest''(l_i)=Interest^*(l_i)+newInterest(l_i) \quad (8)$$

其中，newInterest(l_i)为师生浏览一个新的校园页面后对该页面对应的关注度的变化值，也可当做用户对该文档的关注度。为了避免单一的方法所造成的局限性，决定通过师生的浏览动作即收藏页面和保存页面、浏览速度两者来对它进行计算，得到公式如下：

$$newInterest(l_i)=\omega 1*f1(h(l_i))+\omega 2*f2(speed(l_i)) \quad (9)$$

其中，ω1 为收藏和保存页面所对应的权重，ω2 为浏览网页速度所对应的权重。

最后利用 K-means 算法依据网页内容聚类成各个主题，在每个主题中随机选取 10 篇师生关注的网页进行实验，所以根据式（7）得到主题的关注度函数：

$$Interest(p_m)=\sum_{i=1}^{j}Interest''(l_i) \quad (10)$$

其中，p 表示主题，m 表示主题数目，j 表示第 m 个主题的总共网页数目。由于关注度的值的维持在 0 和 1 之间，为了避免页面关注度值偏大，所以要进行归一化处理，进而得到归一化的主题关注度函数：

$$Interest(p_m)=\frac{Interest(p_m)}{\max\{Interest(p_\omega)\}}(\omega\in m) \quad (11)$$

四、实验分析

表 1 整理后的部分师生浏览行为

院系、网页	字节 byte	字节[kb]	时间[s]	速度 kb/s	保存	收藏
1	106844	104.3398	90.56	1.152163		
2	111657	109.04	147.59	0.738804	是	
3	123257	120.3682	252.75	0.476234		是
4	106553	104.0557	50.33	2.067468		
5	106701	104.2002	96.24	1.082712		是
6	119359	116.5615	39.03	2.98645		
7	118745	115.9629	149.23	0.777075		
8	120653	117.8252	221.03	0.533073	是	
9	117006	114.2637	63.18	1.808542		
10	131964	128.8711	75.63	1.703968		
11	119889	117.0791	52.45	2.232204		是
12	125407	122.4678	390.31	0.313771		
图书馆	字节 byte	字节[kb]	时间[s]	速度 kb/s	保存	收藏
1	106553	104.0557	50.33	2.067468	是	
2	106844	104.3398	90.56	1.152163		
3	123257	120.3682	252.75	0.476234		
4	111657	109.04	147.59	0.738804		
5	106701	104.2002	96.24	1.082712		
6	106650	104.1501	66.25	1.572081		
7	107480	104.9609	49.81	2.107226	是	
8	106463	103.9678	81.31	1.278659		
9	107827	105.2998	83.19	1.265775		是
10	106641	104.1416	51.2	2.034016		
11	107714	105.1895	47.8	2.200616	是	
12	106136	103.6484	50.1	2.068831		

本文选取浏览一周的商丘职业技术学院师生关注的校园网页作为处理数据，利用 K-means 算法最终聚类成 5 个主题，并且包含了其相关文本内容，随机选取每个主题的 10 篇文章作为实验数据。使用 dynaTrace AJAX Edition 软件来得到师生校园网浏览行为和操作，然后通过计算师生浏览行为的方法获取其对每个网页的关注程度，进而得到每个主题的关注度。一周后再对浏览这些网页的师生进行统计调查，让师生自己评价对每个主题的关注度，将师生主观的评价结果与计算得到的结果进行比较。

实验步骤：①采用 dynaTrace AJAX Edition 软件获取用户浏览行为，在本文中，获取的是用户访问的网页链接、浏览时间、网页大小、保存和收藏页面等操作。经整理后的师生浏览校园网页行为如表 1 所示。②利用 Java 程序计算用户关注度，随机选取其中部分网页关注度经整理如表 2 所示。

表 1 包括本文前面提到所需的的浏览行为，即师生浏览网页的大小(字节数)、停留时间、浏览网页的速度(经计算得到)、保存和收藏页面的操作动作。

表 2 师生关注度估计

主题 \ 页面	院、系网页	图书馆	学生处(社团)	科研处	教务处
1	0.1518	0.6589	0.2956	0.1849	0.2773
2	0.2898	0.2723	0.2358	0.0778	0.3737
3	0.5886	0.2453	0.3115	0.1511	0.1321
4	0.4038	0.4247	0.3233	0.0960	0.1298
5	0.6589	0.2898	0.2786	0.1298	0.2581
…	…	…	…	…	…
主题关注度	2.0929	1.8910	1.4448	0.6396	1.1710
归一化	1	0.9035	0.6903	0.3056	0.5595
预估关注度	1	0.90	0.7	0.22	0.5
绝对误差	0	0.0035	0.0092	0.0856	0.0595

从表 2 可以看出，根据前面提到的关注度计算方法计算得到的用师生关注度与师生自评得到的关注度值用绝对误差值来验证，绝对误差值控制在 9% 以内，由此可以验证本文基于校园 web 浏览行为的师生关注度计算方法是合理和有效的。

五、结束语

本文主要通过师生用户对校园网浏览行为分析的基础上，提出了一种基于校园 web 浏览行为的师生关注度计算方法，考虑到关注度会随时间衰减的规律，对计算方法进行了相应的改进，然后通过 K-means 算法将师生所浏览的校园网页内容聚类成不同主题，然后获取主题关注度，为下一步师生关注模型的构建提供了基础。最后通过实验对计算方法进行了验证。计算师生对校园网页关注度并不是最终的目的，如何在此基础上，利用得来的结果，进一步完善校园网建设，提高校园网服务师生的质量，是下一步研究的内容。

参考文献：

[1] 伊雯雯，孙涌，尹春晖.集群环境下个性化检索系统的研究与实现[J].苏州大学学报：自然科学版，2008，24(3)：45-48.

[2] 尹春晖，邓伟. 基于用户浏览行为分析的用户兴趣获取[J].计算机技术与发展，2008，18(5)：37-39.

[3] 王洪伟，张艺伟. 基于百度指数的网页用户关注度研究[J].情报学报，2012，31(8)：837 -845.

[4] 管金才. 基于个人网页数据挖掘模型的研究与构建[D].上海：华东师范大学，2007.

[责任编辑：胡大威]

Computing Method for Teacher and Student Attention Degree based on Campus Web Browsing Behavior

LI Dong

(Computer Department, Shangqiu Polytechnic, Shangqiu476000, China)

Abstract: To optimize campus web, the first step is to learn about teacher and student attention degree on the themes of campus web, which are also critical factors determining the campus web service quality. After analyzing the web browsing behavior and law of users´ attention degree attenuation, the paper proposes a new calculation of teacher and student attention degree based on research on campus web browsing behavior; it then uses K-means algorithm to cluster teacher and students´ browsing content into themes and teacher and student attention degree is obtained. Finally, experiments prove that calculation method of teacher and student attention degree is effective.

Key words: campus web; browsing behavior; teacher and student; attention degree

(上接第 65 页)

Research on Information Extraction of Webpage Based on Text Tagging Attributes

SHEN Na

(Suqian Open University, Suqian223800, China)

Abstract: With the rapid development of Internet, online information resources present a blowout situation. At the same time, it has become increasingly difficult to extract information from huge amounts of the information we need. After studying the existing Web information extraction technology and the challenges faced, we design a Web news information extraction model based on text tagging attribute. This paper mainly introduces the Web information extraction technology based on the attribute of text tag, presents the specific implementation process of information extraction, describes the traversal algorithm of the filtering text labels based on DOM tree node, and chooses the mainstream news sites to carry out the extraction experiment and to verify the feasibility of the algorithm.

Key words: HTML DOM Tree; text tagging attributes; Web news; information extraction