Open access Journal International Journal of Emerging Trends in Science and Technology

Website Structure Improvement using User Navigation Monitoring and Web Mining

Authors

Pratik Kawale[1], Urvi Pimparkar[2], Pooja Singh[3], Amol Potgantwar[4]

[1,2,3]BE IT, Sandip Institute of technology And Research Centre, Nashik, Maharashtra, 422213 INDIA

[4]HOD, IT Dept, Sandip Institute of Technology And Research Centre, Nashik, Maharashtra, 422213

Email: *writwick007@gmail.com[1], Urvi.pimparkar@rediffmail.com[2],550singhpooja@gmail.com[3] amol.potgantwar@sitrc.org [4]*

Abstract

To style effective information processing system user navigation may well be an enormous challenge. A main reason is that website|internet site|site|computer|computing machine|computing device|data processor|electronic computer|information processing system} developers have to be compelled to style a web site otherwise than the other users. completely different strategies have to be compelled to be generated to relink the WebPages, so the user navigation ar typically merely done. The new organized ways ar unpredictable, disorienting worth can't be analyzed. This paper introduces information processing system whereas not substantial changes. User navigation may be reduced victimization mathematical programming and together alteration to Structure. The Result shows that it improved its user navigation and together it's effectively solved. to boot we've got an inclination to outlined 2 all completely different methods: assess performance of information processing system and real information set. The result together shows that user navigation has been improved greatly. The observation is disoriented user ar quite that of improved structure users.

Index Terms— *net Personalization, net Transformation,DOM Tree, greatest Forward Reference, mini Sessions, Out- Degree Threshold, web site style, User Navigation, Web Mining, Mathematical Programming.*

INTRODUCTION

The month before Christmas of information superhighway has on condition that associate unexampled flat structure for folk to induce data and have a look for knowledge. There unit of measurement net Users all over on earth as of September an increase of a neighborhood of 100 since The tightly growing type of net Users additionally presents very nice business chances to companies harmonically with to Grau the U.S.A. retail e business sales keeping out (away from) journey destroyed and might get stretched. therefore on free from doubt the increasing demands from connected customers companies unit of measurement heavily stroke money into inside the event and support of their internet-sites. net merchant reports that the internet-site operations making payments enlarged in with one third of place operators going for long walk making payments by a minimum of a neighborhood of a hundred created a comparison to it in Despite the weighty and increasing money place into business in internet-site vogue it's still let be seen however that having experience desired knowledge in associate internet-site isn't easy, not exhausting Associate in Nursingd developing with operative well websites isn't associate unimportant or everyday work Galletta et giving a {thought|a concept|a plan|an inspiration} of that connected

sales lag manner behind those of brick and armed forces fighting device stores and a minimum of a neighborhood of the nothing can be explained by a significant trouble Users meeting once taking grass for food connected stores linksman marks that poor internet-site vogue has been a key [*fr1] in an extremely type of high seen from the side place returning in would like of one's hopes McKinney et Al together discover that Users having trouble in giving position of the persons marked unit of measurement very most likely to dropping of associate internet-site though its knowledge is of high quality. a primary reason behind poor internet-site vogue is that worldwide net of Associate in Nursing insect ones that makes getting even of but associate electronic computer ought to be structured is also considerably utterly completely different from those of the Users Such amounts, degrees, points utterly completely different outcome if where Users cannot merely provide position of the specified knowledge in associate web-site this robust question is difficult to remain from as a results of once making inherit existence Associate in Nursing web-site internet ones that makes may not have a clear obtaining even of Users desires and would possibly alone place into order pages supported their own choices. but the live of web-site good impact have to be compelled to be the pleasure of the Users instead of that of these that makes therefore WebPages have to be compelled to be place into order terribly} very approach that generally matches the Users style to be derived of but pages have to be compelled to be place into order earlier studies on internet-site has place at purpose at that rays close on an expansion of issues like getting justly internet of Associate in Nursing insect structures having experience on the aim pages of a given page mining giving data structure of a replacement internet-site and getting from example from WebPages Our work on the other hand is closely related to the literature that appears at the thanks to recover internet-site suitability through the utilization of User keeping direction embarrassed data for computers completely different works have created a tricky work to subsume this question and that they are going to be generally place therefore as into two groups to assist one User by with motion reconstituting pages based on his outline and traversal strategies typically has connexion as personalization and to vary the building land structure to rest the keeping direction embarrassed for all Users typically has regard to as nice modification throughout this paper we've got an inclination to face live had an area in primarily with nice modification moves near The literature giving tho't to as nice changes moves near primarily provides one's mind to a thought on undergoing growth ways to completely reorder the association structure of Associate in Nursing internet-site though there unit of measurement supporters for internet-site reorganization moves near their unhealthy points unit of measurement clearly and promptly seen initial since a full reorganization could greatly modification the place of everyday things the new internet-site would possibly discombobulate Users Second the reordered internet-site structure is extraordinarily ineffectual to say for sure and thus the worth of disorienting Users once the changes remains raw. this will|this may} be as a results of Associate in Nursing internet-sites structure is representatively designed by specialists and comes as business or to undertake and do with organization reasoning but this reasoning would possibly now not have existence inside the new structure once the net web site is totally reordered to boot to know before studies have value placed on the instability of a very reordered internet-site leading to doubts on the employment of the reorganization moves near finally since internet-site reorganization moves near could with abrupt, gorgeous modification the current structure they can't be typically did to urge higher the suitability.

RELATED WORK

Min Chen and Young U. Ryu [1] planned AN approach of mathematical programming model to enhance the navigation result of the web site minimizing changes to its current structure. Their model was notably appropriate for informational websites whose contents are comparatively stable over time. It improves the performance of web site

instead of reorganizes and so appropriate for web site maintenance on a progressive basis The Mathematical Programming model was ascertained to proportion all right, optimally resolution large-sized issues in an exceedingly few seconds in most cases on a desktop laptop Perkowitz and Etzioni [02] describe AN approach that mechanically synthesizes index pages that contain links to pages touching on specific topics supported the co-occurrence frequency of pages in user traversals, to facilitate user navigation. but this methodology is internet personalization The ways planned by Mobasher et al. [13], and Yan et al. [16] produce clusters of users profiles from weblogs then dynamically generate links for users UN agency are classified into totally different classes supported their access patterns. These ways are internet personalization primarily based. Nakagawa and Mobasher[13] develop a hybrid personalization system that may dynamically switch between recommendation models supported degree of property and therefore the users position within the web site. For reviews on internet personalization approaches, see[18] and [19]. internet transformation, on the opposite hand, involves dynamic the structure of a web site to facilitate the navigation for an outsized set of users [28] rather than personalizing pages for individual users. Fu et al. [29] describe AN approach to reorganize WebPages therefore on offer users with their desired data in fewer clicks. However, this approach considers solely native structures in an exceedingly web site instead of the positioning as a full, therefore the new structure might not be essentially best. Gupta et al. [19] propose a heuristic methodology supported simulated annealing to relink WebPages to enhance suitableness. This methodology makes use of the combination user preference knowledge and might be accustomed improve the link structure in websites for each wired and wireless devices. However, this approach doesn't yield 5 optimal solutions and takes comparatively a protracted time (10 to fifteen hours) to run even for a small web site. architect [20] develops whole number programming models to reorganize a web site based on the cohesion between pages to cut back data overload and search depth for users. Additionally, a 2-stage heuristic involving two integer-programming models is developed to cut back the computation time. However, this heuristic still needs very long computation times to unravel for the best resolution, particularly once the website contains several links. Besides, the models were tested on haphazardly generated websites solely, therefore its pertinence on real websites remains questionable. Lin and Tseng [20] propose AN hymenopter colony system to reorganize web site structures. Although their approach is shown to produce solutions in an exceedingly comparatively short computation time, the sizes of the artificial web sites and real website tested in are still comparatively little, posing queries on its quantifiability to large-sized websites. There are many outstanding differences between internet transformation and personalization approaches. First, transformation approaches produce or modify the structure of a web site used for all users, while personalization approaches dynamically structure pages for individual users Hence, there's no predefined/built-in internet structure for personalization approaches In order to know the preference of individual users, personalization approaches need to collect data related to these users (known as user profiles). This computationally intensive and long method isn't needed for transformation approaches. Transformation approaches create use of mixture usage knowledge from weblog files and don't need chase the past usage for every user whereas dynamic pages are usually generated supported the users traversal path. Thus, personalization approaches are additional appropriate for dynamic websites whose contents are additional volatile and transformation approaches are additional acceptable for websites that have a inherent structure and store comparatively static and stable contents. This paper is concerning survey of internet structure mining and clustering techniques over sites and hyperlinks, as structure mining is beneficial for organization if done according to user want, thus to facilitate user we tend to thought of structure mining by playacting data processing techniques on weblogs also referred

to as a part of internet usage mining. About internet usage mining, author in [1] explains concerning weblogs like WHO accessed order of page request, total time for page read. This paper includes many pre-processing like; 1: knowledge cleaning-It is technique of removing digressive items or logs like removing of file with .gif and .jpg extensions. 2: User identification-It involves USER ID for each user to supply individualism even completely different users area unit on same IP. 3: Session identification- this is often defines consistent with time i.e. time between page request and page shut or time out. 4: Path completion- it's outlined as if some information or page is very important and principally accessed however not recorded in logs and not coupled cause drawback .5: Formatting- it's technique of changing transactions or logs it to a format of information mining like removal of numeric worth for determinant association rules. In[2]Author specialize in needs and problems with internet structure mining and what area unit the parameters and knowledge mining techniques may be applied, Author projected here k-means Clustering formula and Apriori association rule mining formula, they additionally uses probabilistic cluster algorithm referred to as abstract cluster formula like COBWEB .They additionally introduced the agglomerate and hierarchical cluster formula this all accustomed solve index page drawback in reconciling websites, they thought of 2 parameters sequence of page views and links clicked throughout each sessions author used here 2 quality measures primarily based on variety of times user get correct page and efforts done by user. Author here outline drawback supported contents, hyperlinks and title, formula uses by author supported frequent incidence of pages in user logs. We targeted on structure mining and lots of work done on structure mining as in [3]Author projected manner of finding browsing potency ,here they 1st performed design of information processing system as graph of pages as nodes and links between pages as edges then on basis on logs, proxy server information and user cookies potency is calculated. Author in [4] projected 0-1 programming model for reorganizing netsites supported cohesion between web pages. They uses 2 approaches grouping of comparable session and grouping of pages with co-occurrence frequency on that they performed cluster and association rule mining for pages concerned within the session. They conjointly used constraint as length of shortest path from home page to every page. In[5]Author projected structure mining supported range of links traversed in a very session, here instead of directly changing structure they extra a lot of links between net pages that square measure a lot of oftentimes browsed. In[6] Author projected reorganization by classification techniques supported variety of file extension, range of links page, quantitative relation of session on last page to the whole session on web site and average time that user on websites or

user is login. In [7] Author projected some a lot of parameter for web site transformation and here author uses sessions that divided in to mini sessions and user traversing path, author conjointly uses 2 threshold path threshold-length of pages from begin page to focus on page in mini session and out-degree threshold-number of links allowed from pages at the time of reorganization In this paper our second concentrate on data processing technique, survey includes several approaches on weblogs as in [8] author provides k-means cluster formula on co-citation of pages, here they thought of common links shared between pages and similarity measures they contemplate is trigonometric function similarity measures.

In [9] Author projected weighted page rank formula based on rank assign on page that is most well liked according to user behaviour, anon this k-means is performed. They conjointly concentrate on parameters like in-links, out-links on every page. In[10]author steered cluster on frequent item-sets and their frequent item-sets square measure user session and access patterns In [11] Author projected cluster and association rule mining individually. For association rule mining Apriority algorithm is employed supported pre-processed logs and for clustering co-occurrence of pages is employed. Author in [12] projected cluster supported page views by user and that they projected complete linkage cluster algorithm supported user dealings. A. cluster

on net logs Clustering or cluster analysis is outlined as technique of grouping a group of objects in such the way that objects within the same cluster (called a cluster) square measure a lot of similar (in some sense or another) to every aside from to those in alternative teams (clusters). it's a main task of alpha data processing, and a common technique for applied math knowledge analysis, used in many fields, together with machine learning, pattern recognition, image analysis, data retrieval, and bioinformatics etc. There square measure several cluster techniques, this paper contain comparison of 3 cluster technique. We will begin with most simple k-means cluster formula, The term "k-means" was 1st utilized by James MacQueen in 1967 this formula kind k-cluster for n-objects according to nearest mean randomly generating k points within the knowledge house. This is typically done by generating a price uniformly indiscriminately within the vary for every dimension. every iteration of Kmeans consists of 2 steps: i) cluster assignment, and ii) centroid update. another imp facet of k-mean is that the total of square errors evaluation operate.

METHODOLOGY

In the Methodology we have discussed the whole flow of the project through which we have implemented our project it gives you the clear idea what process it took to get the project implemented we have discussed every phase in detail so it is easy to understand the flow of the project.

Communication

The developer will observe the user navigation path and gather required information from them.
– Navigation path.
– Recently visited pages.
– Dynamic web pages.
– User id.

Planning

Here we all developer gathered together and discuss the issues occurred in communication phase.
– what will be the flow of project
– What will be the platform of the project
– What kind of Server is going to be use
– which language will be used in Application development
– which communication technology will be used.
– How the testing will be performed
– What documentation will be provided

Designing

We will discuss the following points:
– GUI (graphical user interface) should be user friendly and easy to understand
– Backend of the System
– Frontend of the System

Construction

– Language: Java
– Web designing language: PHP
– Server: oracle 11g
– Internet The communication between the Web application and the server, Vice versa is going to hold through the Internet
– Testing: Android virtual tool.

Deployment

Here we deploy our system to the customer and provide the customer with manual and proper documentation as well as will get the feedback

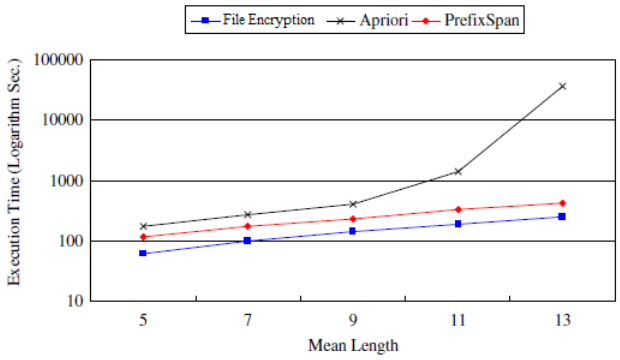
RESULTS AND ANALYSIS

Depicts the path traversal graph corresponding to the four Web browsing sessions in where the notations and represent edges and via-links respectively. For simplicity, the edges of the vertices except vertex a are omitted. Suppose the minimum support is 50%. After all the edges and via-links with supports below the minimum support are removed and those vertices unconnected by any edge or via-link are deleted, the remainder is the frequent path traversal graph.

All throughout-surfing patterns identified from the data set in Table 2.

ID	Throughout-surfing pattern
P_1	$\langle a, c, d, j, n, s \rangle$
P_2	$\langle a, c, g, j, n, s \rangle$
P_3	$\langle a, c, g, l, p, n, s \rangle$
P_4	$\langle a, c, g, l, p, t, g, l \rangle$
P_5	$\langle a, d, j, n, s \rangle$
P_6	$\langle w, x, y, z, w, x \rangle$

The synthetic Web browsing sessions are created as follows. First, the length |P| of a Web browsing session is determined. Second, the Web browsing session begins at the root node and one node is uniformly selected from the lower level in the Web structure graph as its successive node. Then, the level of the following node is determined by the probability Pu or Pd. The parameter Pu indicates the probability of branching to the upper levels in the graph-expanding stage, and Pd indicates the probability of branching to lower levels in the graph-shrinking stage. While the level of the following node is determined, a node is picked uniformly from the level. The next nodes in this Web browsing session are generated in the same way until the length of the Web browsing session is equal to |P|.



CONCLUSION

In this paper, we've got created a suggestion a mathematical programming style to be traced to induce higher the keeping direction bewildered sensible result of Associate in Nursing internet-site whereas creating appear unimportant changes to its current structure, a full of danger question underneath discussion that has not been place inquiries to within the literature. Our style to be copied is especially right for knowledge-sorting internet-sites whose what's in square measure comparatively arduous to maneuver over time. It gets higher Associate in Nursing internet-site instead of reorders it and for this reason is true for internet-site support on a progressive base. The tests on a real internet-site showed that our style to be traced may build prepared important enhancements to User navigation by adding solely few new links. best answers were quickly got, suggesting that the planning to be traced is extremely effective to real world internet-sites. Additionally, we've got tested the MP style to be traced with variety of created by uniting knowledge puts that square measure abundant larger than the most important data place thought out as in connected studies in addition because the true knowledge place. The MP style to be traced was determined to proportion okay, optimally obtaining answer to, reply of large-sized issues during a few seconds in most cases on a work surface laptop. to form bound the doing a play of our style to be traced, we've got fashioned 2 metrics and used them to price the got higher internet-site victimization simulations. Our results created doubtless that they got better structures indeed greatly helped User keeping direction bewildered. additionally, we tend to found a taking from lower to higher authority outcome that heavily disoriented Users , i.e., those with the next however probable to forgoing the internet-site, square measure a lot of doubtless to assist from the got higher structure than the less disoriented Users.

ACKNOWLEDGMENT

We would like to thank our guide Prof. Amol Potgantwar and Prof. Vivek Patil for the guidance and support. We will forever remain grateful for constant support and guidance extended by guide, for the completion of paper.We also thank to prof.Vijay Sonawane for their valuable

REFERENCES

1. Pingdom, Internet 2009 in Numbers, [Online] Available: http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/, 2010.

2. J. Grau,US Retail e-Commerce: Slower But Still Steady Growth, [Online] Available: http://www.emarketer.com/ Report. aspx?code=emarketer 2000492, 2008.
3. Internetretailer, Web Tech Spending Static-But High-for the Busiest E-Commerce Sites, [Online] Available: http://www. internetretailer. com/ dailyNews.asp?id = 23440, 2007.
4. D. Dhyani, W.K. Ng, S.S. Bhowmick,A Survey of Web Metrics, ACM Computing Surveys, Vol. 34, No. 4, pp. 469- 503, 2002.
5. X. Fang, C. Holsapple,An Empirical Study of Web Site Navigation Structures Impacts on Web Site Usability, Decision Support Systems, Vol. 43, No. 2, pp. 476-491, 2007.
6. J. Lazar,"Web Usability: A User-Centered Design Approach", Addison Wesley, 2006.
7. D.F. Galletta, R. Henry, S. McCoy, P. Polak,When the Wait Isnt So Bad: The Interacting Effects of Website Delay, Familiarity, and Breadth, Information Systems Research, Vol. 17, No. 1, pp. 20-37, 2006.
8. J. Palmer,Web Site Usability, Design, and Performance Metrics, Information Systems Research, Vol. 13, No. 2, pp. 151-167, 2002.
9. V. McKinney, K. Yoon, F. Zahedi,The Measurement ofWeb- Customer Satisfaction: An Expectation and Discon-firmation Approach, Information Systems Research, Vol. 13, No. 3, pp. 296-315, 2002.
10. T. Nakayama, H. Kato, Y. Yamane, Discovering the Gap between Web Site Designers Expectations and Users Behavior, Computer Networks, Vol. 33, pp. 811-822, 2000
11. M. Perkowitz, O. Etzioni,Towards Adaptive Web Sites: Conceptual Framework and Case Study, Artificial Intelligence, Vol. 118, pp. 245-275, 2000.
12. J. Lazar,"User-Centered Web Develop-ment", Jones and Bartlett Publishers, 2001.
13. Y. Yang, Y. Cao, Z. Nie, J. Zhou, J. Wen,Closing the Loop in Webpage Understanding, IEEE Trans. Knowledge and Data Eng., Vol. 22, No. 5, pp. 639-650, May 2010.
14. J. Hou, Y. Zhang,Effectively Finding Relevant Web Pages from Linkage Information, IEEE Trans. Knowledge and Data Eng., Vol. 15, No. 4, pp. 940-951, July/Aug. 2003.