



Crawling the Website Through Keyword Optimization and Evaluating Its Quality Analytics

Rohini*, Dr. Indu Chhabra

Department of Computer Science and Applications, Panjab University,
Chandigarh, India

Abstract— Success and quality of website is directly related with hits on the website. More the number of hits more will be the number of visitors successfully retrieving the relevant information from the website. Website and visitors are directly dependent on each other. Web crawlers act as inter mediator between visitor and the website. Crawlers find out the relevant match of websites as per the information desired by the visitor is entered. The visitors, crawlers and website works in the dynamic web learning environment. The dynamic environment information can be tracked using analytical tool. The statistics generated from the recorded information leads to better management and help the crawler to work in the right and relevant direction. The footfall of the visitor on the website and ultimately enhance the performance and the popularity index of the covered web site. This paper is focused on the implementation of the proposed solution in which keywords are monitored using analytical tool and solution is implemented to decrease the bounce rate on relevant keyword match. Effective assessment of these keywords is helpful in tuning the overall performance of dynamic web based learning environment and the analytics confirm the visitor behavior towards the website.

Keywords— Web crawler; Analytics; Dynamic Web Learning; Bounce Rate; Website

I. INTRODUCTION

Today, websites are important communication channels that reach a massive audience. Measuring the effectiveness of these web-sites has become a key issue. However, there is no consensus on how to define web site effectiveness and which dimensions need to be used for the evaluation of these websites [1]. Effectiveness of information driven web sites are defined by the success of their information architecture in the literature. Effectiveness of the website is the true reflection of its quality. Better the quality effective will be the website. Web learning environment is a framework by which behavior of a website can be traced by the third party application. In this scenario the third party program is helpful in tracing the information relating the static and dynamic behavior of the web site. Web learning environment contains Web Server, Website, Users, Site Administrator and Tracking Application as component.

Web learning environment when tracked gives information about dynamic metrics of the website. This information is related with visitors of the website, their behavior and the website response to those users/visitors. Website and visitors relationship is depicted in Figure 1.

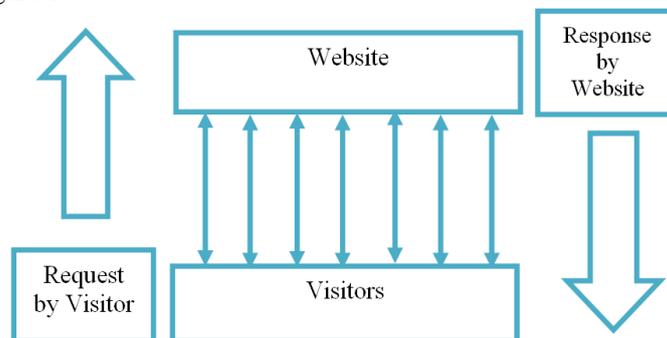


Fig 1: Website-Visitors Relationship

Dynamic environment of website deals with the changing properties of the attributes which are related to the website when it is being used by the visitor and processing is going on in the background.

II. QUALITY ANALYTICS

Millions of websites are visited by millions of visitors daily. In order to provide the browse information of World Wide Web to the visitors search engines are used. Search engines are the efficient tools for end users to search relevant information provided by millions of websites. The responsibility of the search engine is to provide the user with the list of options available to find the relevant information. To generate the list according to the choices entered by the user, search

engines uses a technique referred to as web crawling. Web crawling technique gives the insight into the working of web crawler and enlightens about how visitor gets the hyperlinks to go to the relevant information. The relevance of the visitor interest and the information available on the website can only be tracked using analytical tool and be checking its quality analytics [2].

Quality analytics is the data analytics in which the data patterns formed are the values of those parameters which are contributing in the quality assessment and assurance of the given application [3]. Analytics is the four step process as shown in Fig. 2. These four steps include prediction, monitoring, analysis and reporting. Prediction deals with short listing of the parameters to be analyzed. This prediction is the base for the correct baseline for developing good quality analytics. These predicted parameters and monitored for the fixed time domain to have a data pattern of values. The values are captured and then intensively analyzed [4]. Analytics application runs parallel with the learning web application to be tracked. These tracking results are analytics which gives the clear picture regarding the dynamic behavior of the learning application as well as learner's behavior. The complete analytics scenario is depicted in Fig. 3. Analytics gathers the data patterns which are useful form recording information of dynamic learning environment to optimization of the application for improving its performance wherever possible as predicted by the analytics [5]. Analytics of gathered data can be carried out at four levels as Descriptive, Diagnostic, Predictive and Prescriptive analytics.

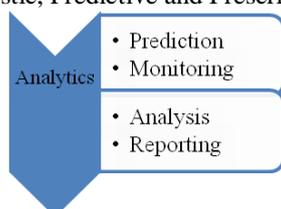


Fig 2. Analytics Process

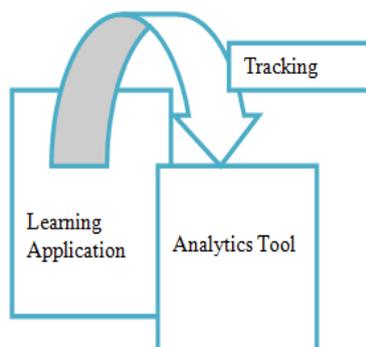


Fig 3. Analytics Scenario

Descriptive analytics gives the current information about the dynamic environment that is what happened in the current scenario. In Diagnostic analytics the values are diagnosed to identify the reasons for the results generated. Depending on the available pattern future actions can be predicted in predictive analytics to guess what will happen and finally in prescriptive analytics an action plan will be framed in order to make some conditions [6]. Analytics are helpful in the collection of very detailed information about visitors, hits, search details, locations, page views etc which includes details about the time spent on the site by the visitors as well. Analytics approach collects data for the given time span second by second [7]. The five website components that are to be critically analyzed are market, website, visitor, visit and page [8]. Recording of data about these five components can analyzed to retrieve five types of details. These details are described in Table 1.

Table I. Data Analysis Components for Website

Component	Domain	Data Analysis
Visibility	Market	Using audience share data
Popularity	Website	Using unique visitor data
Loyalty	Visitors	Using visit per person data
Depth	Visits	Using pages per visit
Stickiness	Page	Using time spent on per page

III. IMPLEMENTATION

Web analytics has many potential benefits as data collection, behavioral analysis, keyword success rate, page views, session limits. These benefits not only help in making decision but are also helpful in optimizing the success rate of web sites. Data pattern captured through tracking are the backbone for utilizing the benefits of analytics. Analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. The implementation of complete scenario is possible by following four steps [9]. These steps are performed in order to get the descriptive analytics. The steps are as follows.

1. Select the learning application and analytical tool.
2. Generate track code and track the application using analytical tool.
3. Set the timeline for which applications is to be tracked.
4. Predict the parameters for which tracking is to be done, monitor them for the set timeline, analyze the results and finally create a report.

In our implementation, the application is monitored and data patterns are generated for different keywords. The data patterns are generated for different keywords for which the bounce rate is high and analyzed for the results. The key observation in this implementation is very high bounce rate. The high value of bounce rate indicates that application is failed to open and bouncing many times for many learners. In order to identify the reason for this high bounce rate more factors are cross examined such as network traffic, referrals, page views and keywords [10]. The complete data patterns of one month timeline are examined for the month of April 2016. The problem is identified miss-matched keywords from the metadata file. It was found that metafile contains very less combinations of keywords. The flowchart for the proposed solution is in Figure 4.

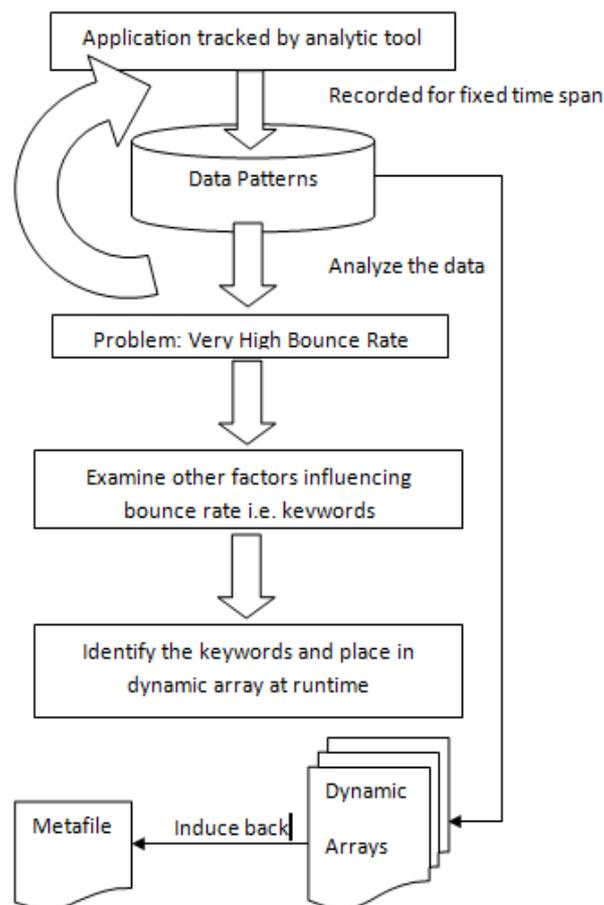


Fig 4. Flowchart for proposed solution

In order to troubleshoot the problem diagnostic analytics are used. The data patterns are analyzed for that defined timeline and keywords are identified from the analytics data whose bounce rate was high for the application. As a solution, a script is designed to induce the newly generated keywords back to the metafile. Through this solution it is possible to pick the keywords from the running application in the dynamic environment and induce them back to the metafile. The solution is implemented using concept of dynamic arrays in javascript. This work is implemented on academic website and the results are tracked using analytical tool Google Analytics [11].

IV. RESULTS

The implementation procedure mentioned in the above section is applied on one academic website. The result shows tremendous change in the success rate of the website runtime behaviour. It is found that many keywords entered by the visitors on the crawling tool gives the high bounce rate for the tested website.

The problem encountered is the absence and variation of keywords in the metafile of the website which misleads the crawler and gives irrelevant results. After implementing proposed solution the keywords are analysed in the dynamic time and induced back to the metafile at run time. The results are discussed in the following tables as Table 2, 3, 4 and 5 based on parameters such as sessions, repeat visitors, page views and bounce rate respectively.

Session: A session is a group of hits recorded for a user in a given time period. A hit is a user's interaction (pageview, screen view, event, transaction etc) with your website that results in data being sent to the Google Analytics server.

Table 2. Comparative Study for Sessions

Technique Time Period	Existing	Proposed
1 st Week	3	7
2 nd Week	5	13
3 rd Week	8	14
4 th Week	11	20

Page Views: A request for a file, or sometimes an event such as a mouse click, that is defined as a page in the setup of the web analytics tool.

Table 3. Comparative Study for Page Views

Technique Time Period	Existing	Proposed
1 st Week	3	9
2 nd Week	8	16
3 rd Week	11	17
4 th Week	14	25

Repeat Visitor: A visitor that has made at least one previous visit. The period between the last and current visit is called visitor recently and is measured in days.

Table 4. Comparative Study for Repeat Visitors

Technique Time Period	Existing(%)	Proposed (%)
1 st Week	66.7	57.1
2 nd Week	60	69.2
3 rd Week	62.5	71.4
4 th Week	54.5	70

Bounce Rate: The percentage of visits that are single page visits.

Table 5. Comparative Study for Bounce Rate

Technique Time Period	Existing(%)	Proposed(%)
1 st Week	100.0	71.43
2 nd Week	80.0	76.92
3 rd Week	87.50	78.57
4 th Week	90.91	80.0

V. CONCLUSION

Quality analytics produce data statistics for website which are helpful in quality assessment and assurance. Further, these statistics are used to optimize the application for its improved performance. The present research and implementation has designed and evaluated various quality analytics capable of tracking the data patterns of dynamic web based learning application. The data patterns are further scanned to identify the key factors where values have shown the influential gap in performance. The present research has identified the problem of high bounce rate which is lowered down by implementing dynamic arrays of keywords at run time and induced back to metafile.

The results of implementation of the proposed work on the selected website reveals that the proposed solution is helpful to identify those keywords for the website at run time whose bounce rate is very high and the crawler for these keywords mislead the crawler. It is observed that the method opted gives better performance and outcome. This work is useful for those who intend to improve the quality of web based learning environment at run time.

REFERENCES

- [1] Nanduri, Babu, Jain, Sharma V., Garg V., Rajshekhar, Rangi V., "Quality Analytics Framework for E-Learning Application Environment", IEEE Fourth International Conference on Technology for Education, 2012, pp 204-207.
- [2] Wu H. Y., Lin H. Y., "A hybrid approach to develop an analytical model for enhancing the service quality of learning", Computers and Education, vol. 58, 2012, pp. 1318-1338.
- [3] Wu H. Y., Lin H. Y., "A hybrid approach to develop an analytical model for enhancing the service quality of learning", Computers and Education, vol. 58, 2012, pp. 1318-1338.
- [4] Ehlers, "Understanding quality culture", Quality Assurance in Education, 2009, v17, pp 343-363.

- [5] D. A. Menase, "QoS Issues in Web Services", IEEE Internet Computing, vol. 6, issue 6, December 2002, pp. 72-75.
- [6] Pawlowski, J.M., "The Quality Adaptation Model: Adaptation and Adoption of the Quality Standard ISO/IEC 19796-1, for Learning, Education and Training", Educational Technology & Society, vol. 10, issue 2, 2007.
- [7] Rohini Arora, Indu Chhanra , "Investigating Attributes for Performing Quality Analytics in Web Learning Environment", International Journal of Computer Applications (0975 – 8887), Volume no. 140, April 2016.
- [8] Kaisler S.H., Armor, "Advanced Analytics -- Issues and Challenges in a Global Environment", proceedings of 47th International Conference of HIISS, 2014, pp 729-738
- [9] Rohini, Indu Chhabra, 2014, "Quality Analytics for Evaluation of Dynamic Web Based Learning Environment", published in IEEE Digital Library IEEE Xplore (Catalog Number: 978-1-4799-6876-3) through conference proceedings of IEEE International Conference on MOOCs, Innovation and Technology in Education held on 19-20 Dec, 2014.
- [10] Rohini Arora, Indu Chhabra, "Extracting Components and Factors for Quality Evaluation of e-Learning Applications" in IEEE Digital Library through conference proceedings of International Conference on Recent Advances in Engineering and Computational Sciences held at UIET, Panjab University, Chandigarh from 06-08 March, 2014. (ISBN no. 978-1-4799- 2291-8) , 2013.
- [11] Kumar, L. (2009a, November 21). *Does Google Analytics affect SEO*. Retrieved April 2, 2016, from Google Analytics Help Forum: <http://www.google.com/support/forum/p/Google+Analytics/thread?tid=2b810d1cf1680e>