

An Improved Recommendation System via Web Usage Mining and Content Mining

Author : Shweta Mishra^{#1}, Dr C.S. Satsangi²

[#]Information Technology, Medicaps Institute of Science and
Technology Indore, India

smmishrashweta09@gmail.com , *cssatsangi@yahoo.com*

ABSTRACT

In most of the recommendation system design the web usages mining techniques are used. In web usages mining the web access logs are evaluated for computing the nature of web browsing by the given user. But in order to find the actual context of user or the requirements of the user it is required to have knowledge about the contents which is utilized by the user. Therefore using this concept for improving the traditional techniques of the recommendation system design a new model is introduced. This model usages the concept of web usage mining and web content mining for designing the user context aware approach. In order to perform this task the web access log is analyzed in first step by which the refined attributes based on target IP address is computed. Based on these personalized attributes the frequent access patterns of the web access and content access are extracted. These factors are providing the information about the user behavior of web navigations in further two more features namely web page contents and their probability on the URLs are also computed. Finally using the computed four different features the URLs weights are prepared. These weights are rescheduled according to the user's current query for finding the best fit URLs in the entire data traversing domain. The proposed technique's implementation is performed using the JAVA technology. Additionally for the performance analysis and comparative results analysis four parameters are used. These parameters are precision, satisfaction, time and space complexity. According to the results the proposed technique is more effective than the traditional technique of recommendation system.

Keywords: Data mining, web mining, web usages mining, web content mining, recommendation system design, results.

1. INTRODUCTION

The recommendation system [1] design is an application of web mining. The recommendation systems are helpful for suggesting the relevant objects according to the user requirements or their past habits. Therefore the uses of recommendation system are frequently performed on the e-commerce web applications, data pre-fetching and other similar applications. In this domain the three main components are used first the previous user behavior, personalization of web data and the current user intend. Using three essential components the recommendation models are designed. In this presented work the user's next web page recommendation system is design using the concept of web usages mining and the web content analysis [2, 3, 4].

Basically the web mining is a technique for analyzing the web data. In order to perform this data mining algorithms are implemented with the web data. The selection and modification of the data mining algorithms are performed according to the nature of data which is used for any application [5]. In this context the web mining can be categorize in three main categories web usage mining, web content mining and the web structure mining. In the web structure mining the web link structure and their organization is considered for the mining purpose [6]. Similarly in the web content mining the web contents and the available material on the web pages are considered for mining. Finally in the web usages mining the web usages patterns with available on server logs are analyzed. In this presented work the web usage mining and the web content mining is combined for providing the web recommendation system. That features of our proposed technique make it different from the traditional web recommendation systems. Here web usage mining is used for evaluating the navigational behavior observations and the content mining is implemented to find out the interest of the end user [7, 8].

A. Web-Page Recommendation System

Recommender systems have become an important study area recently. The increasing amount of web content on web sites makes the recommender system an essential part of web sites. Recommender systems try to direct users to where they would like to go without getting the user lost in the huge amounts of information on the web site. The recommending of books, CDs and other products at Amazon.com [9] is an example of such a system. Web-page recommendation has proved in recent years to be a valuable means of helping Web users by providing useful and effective recommendations or suggestions. The core techniques in web-page recommendation are the learning and prediction models which learn users' behaviour and evaluate what users would like to view in the future. In particular, it can suggest interesting items from a large set of items based on the knowledge gained about an active user. Web-page recommendation can automatically recommend Web-pages that are most interesting to a particular user based on the user's current Web navigation behaviour. Good Web-page recommendations can improve website usage and Web user satisfaction [10].

Web Recommendation system is a specific type of information filtering system technique that attempts to predict the user next browsing activity then recommend to the user web pages items that are likely to be of interest to the user. A recommender system is a typical software solution used in e-commerce for personalized services. Based on the customer preferences, it helps to find the products they would like to purchase by providing recommendations and is particularly useful in e-commerce sites that offer millions of products for sale [11].

The recommendation system is a response which is generated by observation of Web requests according to their interest. Such kind of single or multiple outcome generation is termed as recommendation system. However, to describe the issues in designing a recommendation system two examples are [12]:

- ❖ Now in these days for different online service such as NEWS papers offering news articles to their readers based on reader interests.
- ❖ For another example on-line retailer which is based e-commerce application recommend the products to buy based on user navigational history and product search.

that computes the user behavior and the content utilization behavior, additionally by using the correlation between both the components and the relevant similar behaviors the upcoming seeds are generated for recommendation to the user. Therefore in order to obtain an accurate data model the solution is designed. This technique analyses the

B. Categories of Web Page Recommendation System

A recommendation system can be developed in a number technique. According to behaviour of recommendation system can be classify into following categories [13].

Content based Recommendation: Similar items to the ones the user preferred in the past are generated as a recommendation.

Collaborative Recommendation: Items preferred by the people who have the similar taste to the user are generated as a recommendation.

Hybrid Approach: The above recommendation methods are combined in this approach.

PROPOSED WORK

This section intended to improve the traditional technique of recommendation system design. Therefore first the overview of the proposed system is provided and then the complete solution formulation methodology is described.

A. System overview

The recommendation systems are mostly utilized with some kinds of e-commerce web applications. These systems are help full to the e-commerce users for selection of the produces in the website. But now in these days such kinds of recommendation models are becomes popular with various other application domain i.e. movies, songs, images and others. In this presented work the user's next web page prediction is the main area of application. In this system the user wants the required contents from a text based website additionally the recommendation system helps to find out precise and required contents form the entire website pages (i.e. articles). In this context need to analyze the web contents which is requested by user and also need to compute the user browsing behavior by which the desired contents can be extracted from the website.

In order to design such kind of recommendation system two different streams of the web mining approaches are combined together, namely web content analysis and the web usages mining. In web usages mining the user's past web navigational patterns are evaluated for defining the user's behavior. In addition of that web content mining is useful for approximating the contents of the web pages. In this context a technique is required

web access log file and URL contents for designing the proposed recommendation system. This section provides the overview of the proposed system and the next section the detailed solution design is described.

B. Methodology

The proposed system for user next web prediction technique is organized using the Figure 1. In this diagram the different intermediate processes are defined using the blocks and their working is explained in this section.

Web access log: as discussed previously the proposed technique is combination of the web usage mining and the web content mining. Therefore initially a dataset is required by which the entire model perform functioning. In this context the web access log file is used as the initial input to the system. This file contains the different user attributes and the web access attributes in text file format. In most of web access log files the navigated URL, time stamp of navigation, method, protocol, browser and other various attributes are available. Among some of them are required by the proposed system thus the accepted data is first extracted from the log file and produces to the next phase of processing.

Pre-processing: the main motive of the data pre-processing is to filter the meaningful contents form the raw input file. Therefore using the different data filtering or processing techniques the data is transformed and the required features or attributes are recovered from the web access log file. Therefore in this phase the time stamp, navigated URL, method, User IP address is recovered additionally the remaining data is cleared.

Temp database: the previous phase recovered data is stored in a database table for further use with the system components.

Extract URLs: in this phase the two key points are extracted from the temp database table namely navigated URLs and the user IP address. Because the proposed model is a user oriented data model therefore the IP based navigated URLs are grouped in first step

In further the unique URLs that navigated by an IP address is concluded. Both of the features are used by further processes therefore both the parameters are essential.

Extract URL contents: after extraction of the user IP address and the related URLs from the temp database. Now need to analyze the contents of the web pages, therefore all the visited web pages by the selected user the web page contents are extracted from the web. These contents are help to identify on which kind contents the user having the interest.

Term frequency: in order to evaluate the web page contents some text features are need to be compute. In various feature extraction techniques in text or content mining the term frequency is a popular technique. The term frequency of the available text in the web page can be computed using the following formula.

$$T_f = \frac{\text{number of times a term available in document}}{\text{total terms available in document}}$$

After computing the term frequency from the data it is required to obtain similar size of feature list from all the web pages. Because the length of the extracted terms are their frequency is depends on the length of the content written in the web page. Therefore to regulate the amount of terms in a web document only top 20 most frequent terms are considered in this work.

URL term sequence: in this phase the previous phase information and data is used for approximating one more feature. Thus the web page features (i.e. high frequent terms) are compared with the URL terms and computed what amount of terms are correlated with the URL. In order to compute this feature the following formula is used:

$$T_s = \frac{\text{total words matched in URL}}{\text{total words in URL}}$$

That is a kind of probability distribution for finding content in a specific URL. Thus this feature is computed and preserved separately.

Frequent viewed text pattern: in this phase the text patterns that are frequently visited by the target user is computed. Therefore the total visited web pages are needed to compare for finding the similar text patterns that are in behavior of the user. In order to perform this task, the extracted features are again used for computing the similarity score. But here the possibility is build up for the semantically similar contents are similar but the terms may be different from each other. Therefore the possible synonyms are used

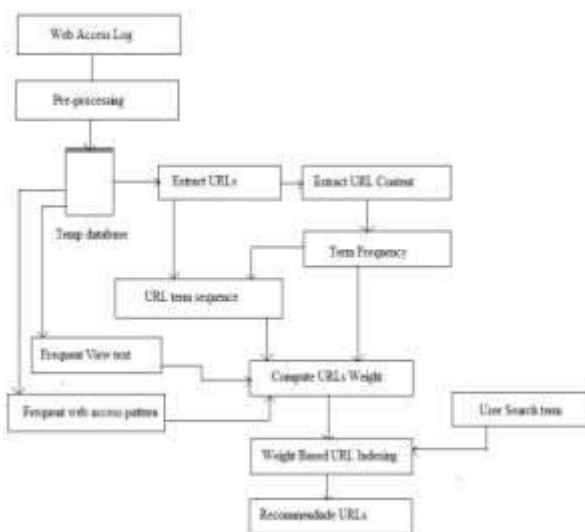


Figure 1: Proposed System

to find the equivalent terms for all the web pages and the similarity score between two web page's features are computed in following manner.

$$D(f_1, f_2) = \sqrt{(f_1 - f_2)^2}$$

Here f_1 and f_2 are the features and the distance between two features are denoted by function $D(f_1, f_2)$. Thus the similarity score of these features are given by:

$$Sim = 1 - D(f_1, f_2)$$

This similarity score is used for computing the amount of contents which is similar to each other and which kind of content is required to serve for the target user. In order to improve more the next step defines the frequently accessed URLs.

Frequent web access pattern: as discussed previously the URLs are grouped according to the user IP address. In this context a similar or same user can access the same URLs multiple times. Thus in this phase the amount of frequently accessed web URLs are computed as the fourth feature of the recommendation engine. The computations of the frequently accessed web pages are provided by the following formula:

$$F_{wap} = \frac{\text{total amount of times a URL visited}}{\text{total amount of URL visited}}$$

Compute URLs weight: after computing the different features from the input web access log file and the relevant contents need to approximate their importance of a URL for different content search scenarios. Therefore for each unique URLs the weights are computed. To compute the weight of the URLs the following formula is used.

$$W = F_{wap} * sim * T_s * T_f$$

User search term: after conducting the whole process of computation need to provide the precise contents to the end user. Therefore a query interface is needed to design that helps to provide the input to the system for suggesting the more precise outcomes.

Weight based URL indexing: using the input text query the key words from the query is extracted and according to similar matched information and weights of the URLs are rescheduled.

Recommended URLs: finally the top ranked URLs are recommended as the outcome of the previous phase is returned as output of the system.

RESULT DISCUSSION

This section is provides the experimental details

and the outcomes obtained after the experimentation. Therefore the computed results in different parameters are listed in this chapter. The detailed discussion on the results is given as:

A. Precision

The precision of the web page recommendation system is the measure of the accurate recommendation made by the system. Let the R_c is the sub-set of R where the R_c denoted by the correct recommendation rule for the web page recommendation is computed using the following formula:

$$\text{precision} = |R_c|/|R|$$

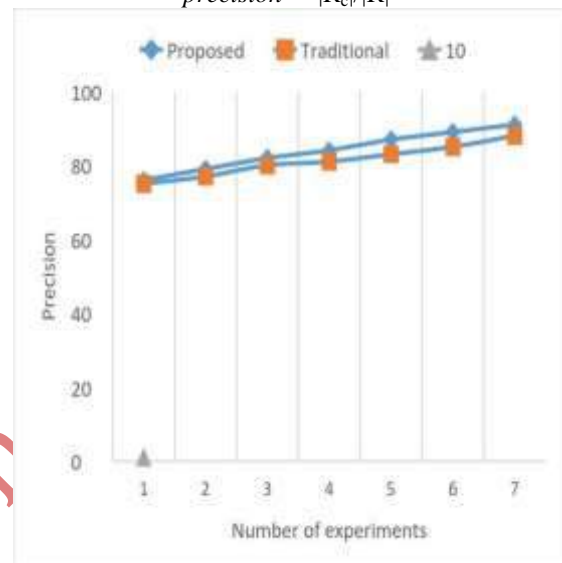


Figure 2: Precision level comparison of proposed & traditional

The comparative precision of the proposed and the traditional technique is given in figure 2 and table 1. In this diagram the X axis contain the different experiments performed with the system and the Y axis shows the obtained precision values of the system.

Table 1: Precision level comparison of Proposed & Traditional Methods

Experiment Number	Proposed Method	Traditional Method
1	76	75
2	79	77
3	82	80
4	84	81
5	87	83
6	89	85
7	91	88

In this diagram the performance of proposed technique is demonstrated using the blue line and the traditional techniques performance is given by red line. According to the demonstrated results the proposed technique's precision ratio is higher as compared to the traditional technique in all the experiments. Therefore the proposed technique is more adoptable as compared to the traditional technique of web page recommendation system.

A. Satisfaction

The satisfaction is the measurement of the recommendation correctness therefore that is defined using the following formula

$$Satisfaction = \frac{|R_s|}{|R|}$$

Where the R_s is a subset of R which correctly satisfy the recommendation rules. The figure 3 and table 2 includes the comparative satisfaction of the recommendation rules.

Table 2 : Satisfaction

Experiments Number	Proposed Method	Traditional Method
1	0.79	0.782
2	0.801	0.795
3	0.82	0.814
4	0.836	0.822
5	0.848	0.836
6	0.853	0.841
7	0.867	0.849

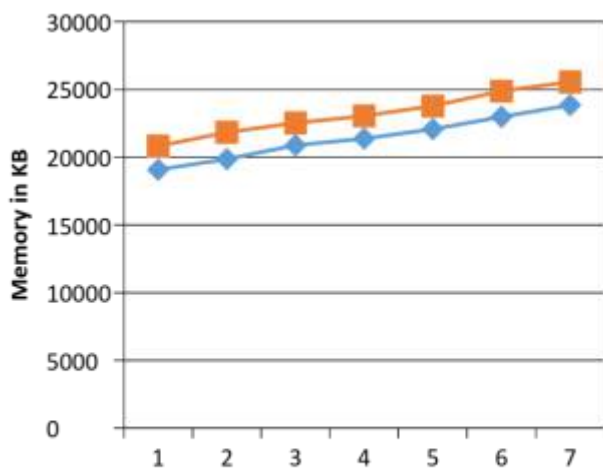


Figure 3: satisfaction level comparison of proposed & tradition

The comparative satisfaction ratio of both the models are given in figure 3, the performance of the proposed recommendation system is given using blue line and the traditional technique is given using red line. In order to represent the performance of the techniques the X axis contains the different experiments performed and the Y axis shows the satisfaction ratio of the techniques. According to the obtained performance the proposed technique results higher degree of satisfaction as compared to the traditional technique. Thus the proposed technique is more accurate than the previously available technique.

B. Memory Requirements

In order to execute the processes the data is placed in the main memory for computation. This required space is termed as the memory usages or space complexity of the algorithms. In java technology for computing the processes consumed memory or utilized main memory the following formula can be used:

$$memory\ usage = total\ memory - free\ memory$$

Table 3 memory usages

Experiment's Number	Proposed	Traditional
1	19083	20827
2	19862	21837
3	20885	22535
4	21374	23048
5	22073	23785
6	22981	24883
7	23847	25562

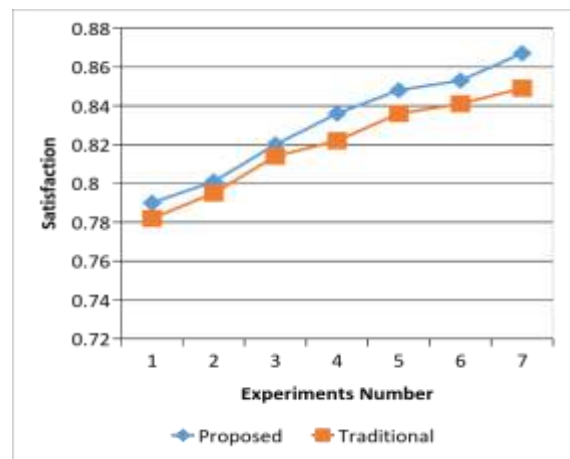


Figure 4: memory requirements

Table 4: Run Time Comparison

Experiment's Number	Proposed Method	Traditional Method
1	90	102
2	108	183
3	188	225
4	274	348
5	322	485
6	481	573
7	547	624

The comparative memory consumption of the proposed and traditional technique is described using table 3 and figure 4. In order to present the memory usages of the proposed approach the blue line is used and the red line is used for demonstrating the performance of traditional approach. The X axis of the diagram shows the number of experiments performed with different sets of data and the Y axis shows the obtained memory usages of the algorithms. According to the performance the proposed technique consumes less memory as compared to the traditional technique, thus proposed model is superior to the traditional approach.

C. Run time comparison

In computational domain every process of the system need to some execution time to generate the outcomes. This time requirement of process or the algorithm is termed as the time consumption for the system. The total time required for completing the task can be computed using the following formula:

$$\Delta t = T_{end} - T_{start}$$

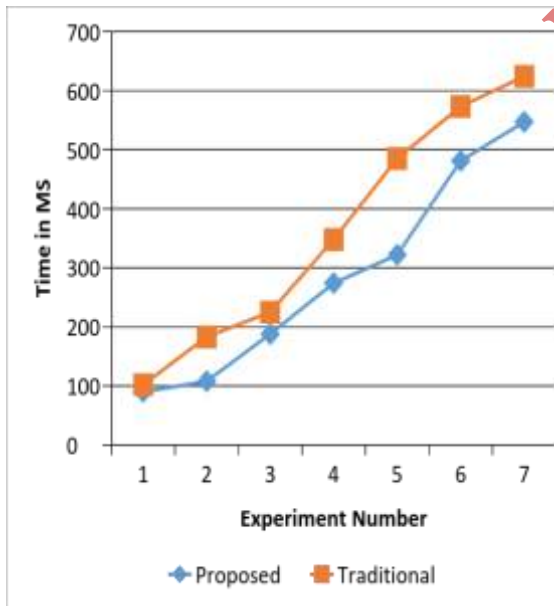


Figure 5: Run Time Comparison

III CONCLUSION

The main aim of the proposed study to design an improved recommendation system is achieved successfully. Using the implemented technique the different experiments are performed and these facts are reported as the conclusion of work. In addition of that the future research extensions are also suggested in this chapter.

A. Conclusion

In web the information and data is available but the quantity of data is significant. Therefore users interacted with the search engines for finding the required data from web. Not only data in this domain is directly available some of the data is hidden from the public to preserve the security and privacy. But such kind of data is very valuable for different web masters and the business promoters. Different techniques and methods are applied on such kind of data to refine the contents to understand the user's upcoming needs. In this context the recommendation systems are a very useful technique by which the user behavior is estimated form the user past navigational behavior and by using this outcome the precise data is suggested to use.

In this presented work the web recommendation system is the key area of investigation and system design. The proposed system is usages the two different concepts of web mining, namely web usages mining and the web content mining. By the combining both the techniques a new model is proposed that technique evaluates the frequently accessed web patterns, content patterns, textual features and URL based text probability. These features are helps to identify the user behavior and their contents needs. Additionally according to these features the weighted recommendation is designed. Finally for providing the

recommendations for the user the current search keywords and the previously computed keyword oriented weights are used with the indexing process. That helps to improve the recommendation time requirements and the memory resource consumption. In addition of that the only likely patterns are suggested for the user, therefore the outcomes are also much accurate and refined for utilizing as the recommendations.

The implementation of the presented technique is performed using JAVA technology. Additionally web server log files are used for recommending the text articles. After the implementation the experiments are performed and based their experiments the performance of this technique is compared with the similar model. The comparative performances of both the techniques are given using the table 5.

Table 5: Performance Summary

S No.	Parameter	Proposed technique	Traditional technique
1	Precision	High	Low
2	Satisfaction	High	Low
3	Memory consumption	Low	High
4	Time comparison	Low	High

According to the obtained performance during the different experiments and their mean performance outcomes the proposed technique is much efficient and accurate than the traditional recommendation systems. Thus this model is recommended for different aspects of e-commerce and the text content based recommendations.

A. Future Works

The main aim of the proposed work to improve the traditional approach of recommendation system design is completed and achieved high degree of accuracy with low resource consumption. In future works the following research directions are feasible for work.

1. This technique is suitable for accurate outcomes thus need to optimize this technique for big data analytics
2. The technique is an effective predictive technique for recommending the text articles thus for future that is also improves for the cross domain modeling

3. The technique is works on simple text and web log data thus in near future that can be extended for the video and other data formats recommendations

REFERENCE

- [1] Nguyen, ThiThanh Sang, Hai Yan Lu, and Jie Lu, "Web-page recommendation based on web usage and domain knowledge." *IEEE Transactions on Knowledge and Data Engineering* 26.10 (2014): 2574-2587.
- [2] Ida Mele, "Web usage mining for enhancing search-result delivery and helping users to find interesting web content", *Proceedings of the sixth ACM international conference on Web search and data mining*, ACM, 2013.
- [3] Sawadsky, Nicholas, Gail C. Murphy, and Rahul Jiresal. "Reverb: Recommending code-related web pages", *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013.
- [4] Chen, Xi, et al. "Web service recommendation via exploiting location and QoS information." *IEEE transactions on parallel and distributed systems* 25.7 (2014): 1913-1924.
- [5] Suguna R., and D. Sharmila, "An efficient web recommendation system using collaborative filtering and pattern discovery algorithms", *International Journal of Computer Applications* 70.3 (2013).
- [6] Parra Arnau, and Jordi Forné. "Measuring the privacy of user profiles in personalized information systems", *Future Generation Computer Systems* 33 (2014): 53-63.
- [7] Mishra Rajhans, Pradeep Kumar, and Bharat Bhasker, "A web recommendation system considering sequential information", *Decision Support Systems* 75 (2015): 1-10.
- [8] Adeniyi D. A., Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method." *Applied Computing and Informatics* 12.1 (2016): 90-108.
- [9] Gediminas Adomavicius and Er. Tuzhilin. *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.* *IEEE Transactions on Knowledge and Data Engineering*, pages 734–749, 2005 Volume 17.
- [10] G. Linden, B. Smith, and J. York, *Amazon.com Recommendations: Item-to-Item Collaborative Filtering*, *IEEE Internet Computing*, Jan./Feb. 2003
- [11] AlMurtadha, Yahya, et al. "Ipaact: Improved web page recommendation system using profile aggregation based on clustering of transactions." *American Journal of Applied Sciences* 8.3 (2011): 277.
- [12] Chapter 9- Recommendation Systems, available online at: <http://i.stanford.edu/~ullman/mmds/ch9.pdf>
- [13] P. N. Vijaya Kumar, Dr. V. Raghunatha Reddy, "A Survey on Recommender Systems (RSS) and Its Applications", *International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)*, Vol. 2, Issue 8, August 2014