IJESMR
International Journal OF Engineering Sciences & Management Research

PREDICTION OF NEXT ACCESSED WEB PAGE USING CLOSPAN ON DHMM

Gourav Kumar Sharma*[1] & Dr. R. K. Gupta[2]
*[1&2] Madhav Institute of Technology & Science, Gwalior , M.P.

Keywords: Web Mining; HMM; Web Prediction; CloSpan; Sequential Pattern Mining.

ABSTRACT

Web page prediction is based on user navigation patterns which are retrieved from the web logs of servers from various places that acts as standard for the prediction. The disadvantage of markov model was low coverage which was corrected by developing dynamic markov model. Also the orders of the markov model are defined based on the levels specified. The drawback of DHMM was high execution times and redundant links between the nodes which consumes times in processing. The proposed paper consist of new web page prediction based on CloSpan on DHMM. The graphs are plotted to show the comparison and the time execution difference between the base and the new proposed algorithm. The results obtained are far better than the base.

INTRODUCTION

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large **data** sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Web mining is the use of data **mining** techniques to automatically discover and extract information from **Web** documents and services. There are three general classes of information that can be discovered by **web mining**: **Web** activity, from server logs and **Web** browser activity tracking.

- Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.
- Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information. The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining.
- Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query[1].

CloSpan

Clospan: Design and Implementation

In this section, we formulate our clospan based on the early termination techiniques. CloSpan can be outlined as two manor steps: (1) it generates the LS set, a superset of closed frequent sequences. And stores it in a prefix sequence lattice; and (2) it does post pruning to eliminate non-closed sequences [2].

Algorithm 1 closed mining (D, min_sup, L)

Input: A database Ds, and min_sup .

Output: The complete closed sequence set L.

```
Remove infrequent items and empty sequences, and sort each itemset of a sequence in Ds;
$S^1$← all frequent 1-item sequence;
S← $S^1$;
For each sequence sequence s € $S^1$ do
CloSpan (s, Ds, min-sup, L);
climinate non-closed sequences from L;
```

Algorithm 1, closed Mining, illustrates the framework which includes the necessary preprocessing step. It first sorts every itemset and removes infrequent items and empty sequences. Then it calls CloSpan recursively by doing depth first search on the prefix search tree and building the corresponding prefix sequence lattice. Finally, it eliminates non closed sequences. Algorithm 3, CloSpan, is similar to PrefixSpan, however, it performs a major improvement using the search space pruning techniques developed above. That is, before exploring a discovered sequence and its corresponding projected database to mine its successive super sequences, CloSpan first checks whether a discovered sequence S^1 exists, s.t. either s S^1 or S^1 and I(Ds) = I(S^1). If the condition is satisfied, based on Lemma 3, it is unnecessary to continue expansion since all its possible descendants have been discovered before. Algorithm 3 outlines the pseudo code of CloSpan.

Algorithm 2 CloSpan (s, Ds, min_sup, L)

Input: A sequence s, a projected DB Ds, and min_sup.

Output: The prefix search lattice L.

```
1.  check whether a discovered sequence $S^1$ exists s.t. either s  $S^1$ or  s, and (Ds) =I(D s);
2.  if such super pattern or sub exists then
3.  modify the link in L, return;
4.  else insert s into L;
5.  scan Ds once, find every frequent item α such that
    (a)  s can be extended to (s I α), or
    (b)  s can be extended to (s I α);
6.  if no valid α available then
7.  return;
8.  for each valid α do
9.  call clospan (s  I α, $Ds_{0i}$ α, min_sup, L);
10. for each valid α do
11. call clospan (s  I α, $Ds_{0i}$ α, min_sup, L);
12. return;
```

Now one problem remains: how to do line 1-4 of Algorithm 2 efficiently. There are two approaches to check the condition of Theorem 1 since the condition has two components: (1) the containment, If we testing is involved with a large testing space. If we first check the containment i.e., finding all the sequences which are sub-sequences or super sequences of the current sequence, it is expensive. Although when a new sequence is extended from the current sequence, its sub sequence and super sequence set can be directly computed from the current set, it is still costly based on our testing. Thus, we devised an alternative approach which uses a hash index on the size of projected database. Then only the sequences whose projected database size is the same as that of the current sequence are tested, we found this approach significantly improves the performance and makes the cost of such checking nearly negligible compared to the total running time.

HMM

In 1913, A. A. Markov asked a less controversial question about Pushkin's text: could we use frequency counts from the text to help compute the probability that the next letter in sequence would be a vowel? In this chapter we introduce a descendant of Markov's model that is a key model for language processing, the hidden Markov model or HMM [3].

The HMM is a sequence model. A sequence model or sequence classifier is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels. An HMM is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), they compute a probability distribution over possible sequences of labels and choose the best label sequence.

The hidden Markov model is one of the most important machine learning models in speech and language processing. To define it properly, we need to first introduce the Markov chain, sometimes called the observed Markov model. Markov chains and hidden Markov models are both extensions of the finite automata.

A Markov chain is useful when we need to compute a probability for a sequence of events that we can observe in the world. In many cases, however, the events we are interested in may not be directly observable in the world. For example, in part-ofspeech tagging, we didn't observe part-of-speech tags in the world; we saw words and had to infer the correct tags from the word sequence. We call the partof-speech tags hidden because they are not observed. The same architecture comes up in speech recognition; in that case we see acoustic events in the world and have to infer the presence of "hidden" words that are the underlying causal source of the acoustics. A hidden Markov model (HMM) allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model [4].

A first-order hidden Markov model instantiates two simplifying assumptions. First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:
Markov Assumption: $P(qi |q1...qi-1) = P(qi |qi-1)$

A HMM Model is specified by: - The set of states $S = \{s1, s2, . . . , sNs \}$, (corresponding to the three possible weather conditions above), and a set of parameters $\Theta = \{\pi, A, B\}$:

- The prior probabilities $\pi i = P(q1 = si)$ are the probabilities of si being the first state of a state sequence. Collected in a vector π. (The prior probabilities were assumed equi-probable in the last example, $\pi i = 1/Ns$.)
- The transition probabilities are the probabilities to go from state i to state j: $ai,j = P(qn+1 = sj |qn = si)$. They are collected in the matrix A.
- The emission probabilities characterize the likelihood of a certain observation x, if the model is in state si . Depending on the kind of observation x we have:
- for discrete observations, $xn \in \{v1, . . . , vK\}$: $bi,k = P(xn = vk|qn = si)$, the probabilities to observe vk if the current state is qn = si . The numbers bi,k can be collected in a matrix B. (This would be the case for the weather model, with K = 2 possible observations v1 = and v2 = .)
- for continuous valued observations, e.g., $xn \in R D$: A set of functions $bi(xn) = p(xn|qn = si)$ describing the probability densities (probability density functions, pdfs) over the observation space for the system being in state si . Collected in the vector B(x) of functions. Emission pdfs are often parametrized, e.g, by mixtures of Gaussians.

The operation of a HMM is characterized by

- The (hidden) state sequence $Q = \{q1, q2, . . . , qN \}$, $qn \in S$, (the sequence of the weather conditions from day 1 to N).
- The observation sequence $X = \{x1, x2, . . . , xN \}$. A HMM allowing for transitions from any emitting state to any other emitting state is called an ergodic HMM. The other extreme, a HMM where the transitions only go from one state to itself or to a unique follower is called a left-right HMM.

LITERATURE SURVEY

Edgar F. Black (2014) et al present that the Hidden Markov Model, an unsupervised learning data mining technique, is used to automatically determine the postoperative day (POD) corresponding to a decrease of graft function, a possible sign of transplant rejection, on nonhuman primates after isolated islet cell transplant. Currently, a decrease of graft function is being determined solely on the experts' judgment. Further, information gathered from the evaluation of construted Hidden Markov Models is used as part of a clustering method to aggregate the nonhuman subjects into groups or clusters with the objective of finding similarities that could potentially help predict the health outcome of subjects undergoing postoperative care. Results on expert labeled data show the HMM to be accurate 60% of the time. Clusters based on the HMMs further suggest a possible correspondence between donor haplotypes matching and loss of function outcomes [5].

Ghazaleh Khodabandelou et al present that for several decades, discovering process models is a subject of interest in the Information System (IS) community. Approaches have been proposed to recover process models, based on the recorded sequential tasks (traces) done by IS's actors. However, these approaches only focused on activities and the process models identified are, in consequence, activity-oriented. Intentional process models focus on the intentions underlying activities rather than activities, in order to offer a better guidance through the processes. Unfortunately, the existing process-mining approaches do not take into account the hidden aspect of the intentions behind the recorded user activities. Discover the intentional process models underlying user activities by using Intention mining techniques. The aim of this paper is to propose the use of probabilistic models to evaluate the most likely intentions behind traces of activities, namely Hidden Markov Models (HMMs). The focus on this paper on a supervised approach that allows discovering the intentions behind the user activities traces and to compare them to the prescribed intentional process model [6].

Priyanka S. Panchal (2013) et al present that Web Mining consists of three different categories, namely Web Content Mining, Web Structure Mining, and Web Usage Mining (is the process of discovering knowledge from the interaction generated by the users in the form of access logs, browser logs, proxy-server logs, user session data, cookies). This paper present mining process of web server log files in order to extract usage patterns to web link prediction with the help of the proposed Markov Model. The approaches result in prediction of popular web page or stage and user navigation behavior. Proposed technique cluster user navigation based on their pairwise similarity measure combined with Markov model with the concept of apriori algorithm which is used for Web link prediction is the process to predict the Web pages to be visited by a user based on the Web pages previously visited by other user. So that Web pre-fetching techniques reduces the web latency & they predict the web object to be pre-fetched with high accuracy and good scalability also help to achieve better predictive accuracy among different log file The evolutionary approach helps to train the model to make predictions commensurate to current web browsing patterns[7].

Abdelghani Guerbas (2013) et al present that Accurate web log mining results and efficient online navigational pattern prediction are undeniably crucial for tuning up websites and consequently helping in visitors' retention. Like any other data mining task, web log mining starts with data cleaning and preparation and it ends up discovering some hidden knowledge which cannot be extracted using conventional methods. In order for this process to yield good results it has to rely on some good quality input data. Therefore, more focus on this process should be on data cleaning and pre-processing. On the other hand, one of the challenges facing online prediction is scalability. As a result any improvement in the efficiency of online prediction solutions is more than necessary. As a response to the aforementioned concerns proposing an enhancement to the web log mining process and to the online navigational pattern prediction. Contribution contains three different components. First, proposing a refined time-out based heuristic for session identification. Second, suggesting the usage of a specific density based algorithm for navigational pattern discovery. Finally, a new approach for efficient online prediction is also suggested. The conducted experiments demonstrate the applicability and effectiveness of the proposed approach [8].

Lin Jianhui (2009) et al present that Recommender system can predict and recommend end users their probably most interesting commodities after abstracting their interest and preference through their browsing information. In this paper, accordingly abstract user preference to predict the users' coming access webpage to prefetch them. Through an optimized clustering and sequence mining algorithm, a recommender system is realized to predict user preference webpage. Experiments prove the precision and applicability of the system [9].

Sha Jin (2011) et al present that The CloSpan algorithm first suggested that the closed set of sequential patterns is more compact and has the same expressive power with respect to the full set. Based on the PrefixSpan algorithm, CloSpan added two pruning techniques, backward sub-pattern and backward super-pattern, to efficiently mine the closed set. This paper proposed a new closed sequential pattern mining algorithm. However, instead of depth-first searching used in many previous methods, adopt a breadth-first approach. Besides, previous methods seldom utilize the property of item ordering to enhance efficiency. A list of positional data to reserve the information of item ordering. By using these positional data, developed two main pruning techniques, backward super pattern condition and same positional data condition. To ensure correct and compact resulted lattice, manipulated some special conditions. From the experimental results, algorithm outperforms CloSpan in the cases of moderately large datasets and low support threshold [10].

V. Purushothama Raju (2014) et al present that Closed sequential pattern mining is an important data mining task because it produces a more compact result set and it is more efficient than sequential pattern mining. In general closed sequential patterns are generated from large data sets by applying algorithms like CloSpan and BIDE which require more execution time to compute all the closed sequential patterns. By using genetic algorithms reduce the execution time. The advantage of using a genetic algorithm in finding the closed sequential patterns is that it performs global search and it has less time complexity compared to other algorithms. This paper proposes a novel genetic algorithm G-CSPM to find closed sequential patterns. To improve the performance, develop an effective fitness function and a pruning method. The algorithm is the first method that utilizes genetic approach for closed sequential pattern mining. The results indicate that the proposed algorithm G-CSPM outperforms CloSpan [11].

Lei Chang (2008) et al present that Previous studies have shown mining closed patterns provides more benefits than mining the complete set of frequent patterns, since closed pattern mining leads to more compact results and more efficient algorithms. It is quite useful in a data stream environment where memory and computation power are major concerns. This paper studies the problem of mining closed sequential patterns over data stream sliding windows. A synopsis structure IST (Inverse Closed Sequence Tree) is designed to keep inverse closed sequential patterns in the current window. An efficient algorithm SeqStream is developed to mine closed sequential patterns in stream windows incrementally, and various novel strategies are adopted in SeqStream to prune search space aggressively. Extensive experiments on both real and synthetic data sets show that SeqStream outperforms PrefixSpan, CloSpan and BIDE by a factor of about one to two orders of magnitude [12].

PROPOSED WORK

Proposed pseudo code

The markov model is constructed based on the web user navigation log that is based on sessions. The data obtained if from the server of a company is based on the user sessions that have been browsed in past sessions. In HMM, nodes are drawn which represents web pages. The transactions of the users are considered and the inlinks are shown based on previous and next nodes.

The nodes are represented and the connections between them is shown in the graphical form using HMM.

The process consist of:

1. Preprocessing- it consists of the removal of the robotic elements and images with extension GIF, JPEG, CSS etc.
2. Applying pattern mining algorithm: Clospan in HMM()
   The clospan algorithm here, reduces the number of links that connect to various nodes. The total necessary count is selected only which removes the redundancy of the count of the links of the nodes.

```
i.    Constructing patterns of length upto 6.
          pattern_2=final{p,1};
          pattern_3=final{p+1,1};

ii.   Select the unique patterns
          uarr=unique(npattern);
```

3. Applying HMM model and save the previous and current nodes in of the browsed sessions by the user.

```
data(loc).prev=[data(loc).prev;data(1).current];
data(loc).current=npattern(i,j);
data(loc).count(length(data(loc).count)+1)=1;
data(1).next=[data(i).next,npattern(i,j)];
prev=loc;
```

4. Creation of the matrix

```
j=1:length(enddata.prev)
mnodes(k,1)=enddata.prev(j);
```

```
        mnodes(k,2)=enddata.current;
        mnodes(k,3)=enddata.count;
```

5. Assign names to the nodes as P(i), S for start and E for end.
6. Plot the graph.
7. Show the graph and the total time taken by the code for execution.

RESULT ANALYSIS

The results are shown as in graphical form of the base paper and the proposed algorithm applied on the HMM concept using cloSpan sequential pattern mining.

The graphs obtained from the base and the proposed algorithm are:

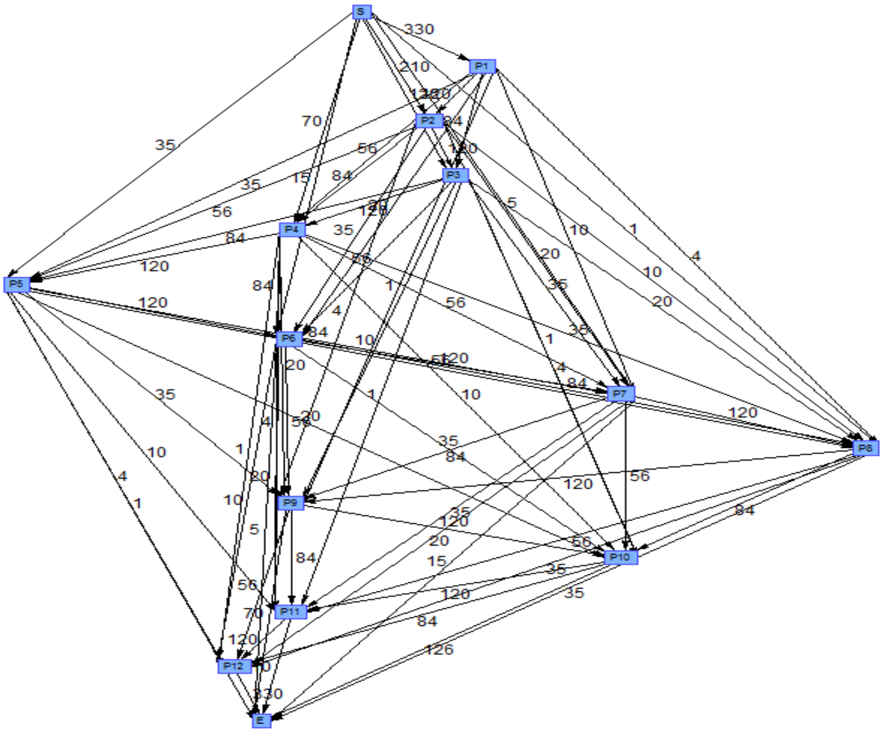


Fig . Graph of Base code

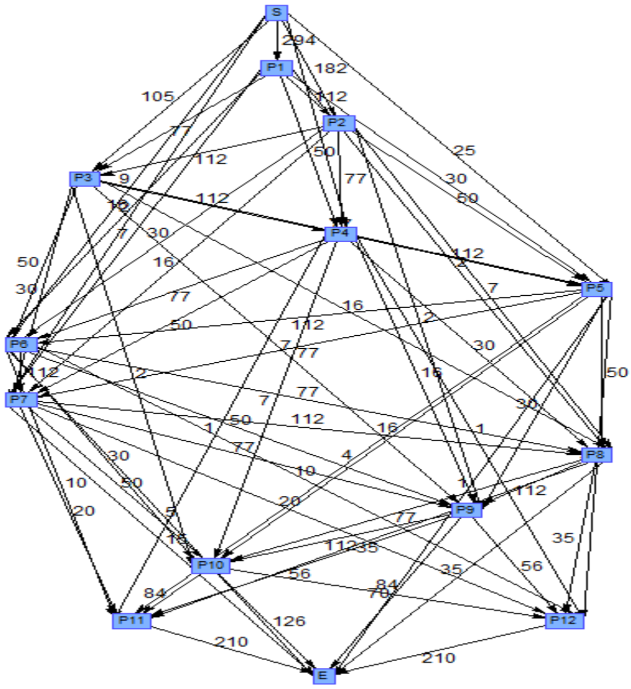


Fig . Graph of Proposed code

From the graphs above, we can see that the number of counts of the connecting nodes shows a significant difference. This also leads to the less time consumption for the processing.

A comparative analysis of the execution time is shown:

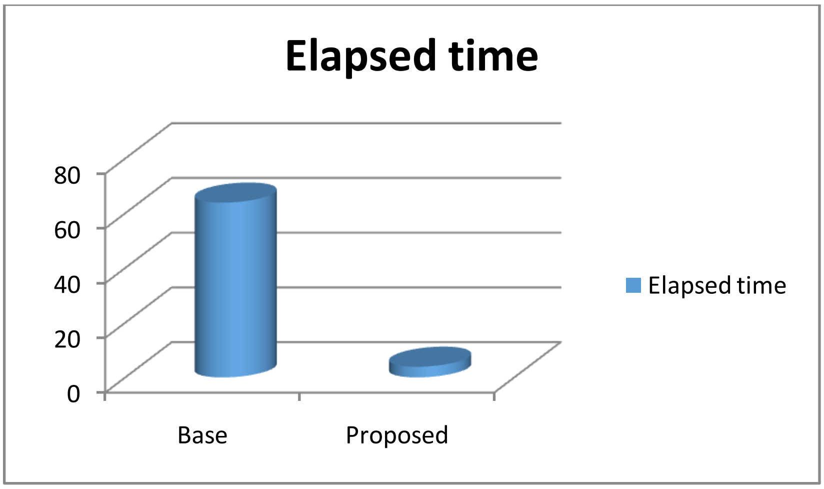
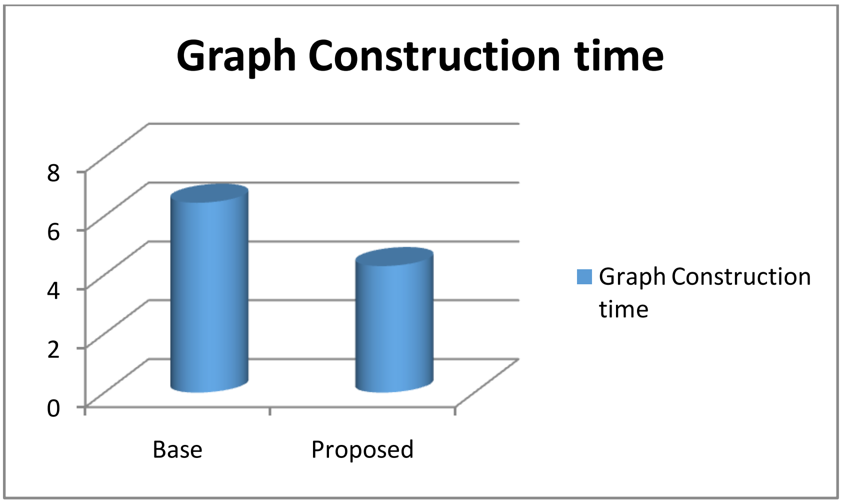


Fig . Comparison of Elapsed Time

A comparative analysis of the graph construction time is also shown in order to show the difference between them.



The above results show that the proposed algorithm seems better in all aspects.

CONCLUSION

The paper showed the implementation of the HMM model using closed sequential pattern mining algorithm. The proposed algorithm proved as better algorithm in terms of efficiency and execution time. The coverage of the model is 100%. The dynamic concept is new in the field of web prediction and HMM using Clospan technique has proved to be a better algorithm.

REFERENCES

1. *LI Yue and WANG Xiao-Gang "Web Mining Based on User Access Patterns for Web Personalization" ISECS International Colloquium on CCCM 2009, @IEEE PP.194-197*
2. *X.Yan, and R.Afshar. CloSpan: Mining Closed Sequential Patterns in Large Datasets. Proceedings of the 2003 SIAM International Conference on Data Mining (SDM'03),. 2003.*
3. *F. Khalil, J. Li, and H. Wang "A framework of combining Markovmodel with association rules for predicting web page accesses", Proc.Fifth Australasian Data Mining Conference (AusDM2006), vol. 61,2006, pp 177–184.*
4. *J. Borges, and M. Levene, "A clustering-based approach formodelling user navigation with increased accuracy", Proc. Second Int'l Workshop Knowledge Discovery from Data Streams, Oct. 2005,pp. 77–86,*
5. *Edgar F. Black, Luigi Marini, Ashwini Vaidya, Dora Berman, Melissa Willman, Dan Salomon, Amelia Bartholomew, Norma Kenyon and Kenton McHenry," Using Hidden Markov Models to Determine hanges in Subject Data over Time, Studying the Immunoregulatory effect of Mesenchymal Stem Cells", 2014 IEEE 10th International Conference on eScience, pp: 83-91.*
6. *Ghazaleh Khodabandelou, Charlotte Hug, Rébecca Deneckère and Camille Salinesi," Supervised Intentional Process Models Discovery using Hidden Markov Models", IEEE.*
7. *Priyanka S. Panchal and Prof. Urmi D. Agravat," Hybrid Technique for User's Web Page Access Prediction based on Markov Model", ICCCNT 2013.*
8. *Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley and Reda Alhajj," Effective web log mining and online navigational pattern prediction", Knowledge-Based Systems 49 (2013) 50–62.*
9. *Lin Jianhui and Zhao Bingjie," A Web Prediction Pattern Recommendation Algorithm", 2009 International Conference on Networking and Digital Society, pp: 263-266.*
10. *Sha Jin, Hu Yingxin and Jia Lianjuan," Efficient Sequential Pattern Mining Algorithm by Positional Data", 2011 International Conference on Internet Computing and Information Services, pp:419-422*
11. *V. Purushothama Raju and G.P. Saradhi Varma," Mining Closed Sequential Patterns Using Genetic Algorithm", 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (lCACCCT), pp: 634-637.*

12. *Lei Chang, Tengjiao Wang, Dongqing Yang and Hua Luan," SeqStream: Mining Closed Sequential Patterns over Stream Sliding Windows", 2008 Eighth IEEE International Conference on Data Mining, pp: 83-92*