International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 4 Issue 3 March 2017

Improving the Web Application Organization Using Web Structure Mining and Content Mining Techniques

Abhijeet Kour Bagga[1], Er. Anand S. Sen[2], Er. Harish Patidar[3]
Master of Technology (C.S.E), Lakshmi Narain College of Technology, Indore Madhya Pradesh India[1]
Department of Computer Science Engineering, Lakshmi Narain College of Technology, Indore Madhya Pradesh India[2, 3]
abhijeetbagga13@gmail.com[1], senanand05@gmail.com[2], harish.cs@lnctindore.com[3]

Abstract: *The problem of organizing web application's web contents in a semantically and systematically is considered under the web structure mining. In addition of that organization of contents according to their importance is subject of content mining in web mining. In this presented work both the techniques are fussed for improving the website/ web application reachability, scalability and search engine friendly structure. Therefore a new model for organizing the web documents on server space using the web mining technique is presented. Basically the web mining is a domain where the web data is analyzed for improving the web structure, web contents and others. In this context proposed data model is an effective effort for improving the performance of web applications.In order to perform the required improvement in automatic manner the web site data is analyzed in three different parameters. First the web content analysis technique is implemented for recognizing the contents and their importance therefore two parameters are computed first the TF-IDF features, in addition of that syntactic features are also computed. On the other hand for obtaining the third parameter the web structure mining concept is used, by which the link probability is estimated. Using these parameters the weights for the page and their links are computed which helps to organize the website structure.The implementation of the proposed work is performed with the help of java technology. Additionally for computing the performance precision, recall, f-measures, time complexity and space complexity is estimated. The measured results show the performance of the proposed technique is efficient and provide much precise information for web system organization.*

Keywords: *Web Mining, WWW, Semantic, Synaptic, Web Link, Information Extraction, Probability, TF-IDF.*

1. INTRODUCTION

The growth of internet [1] users increases rapidly as the number of active users for Facebook and Twitter increases worldwide. Some of the tools were used by social media intelligence to collect information about the company [2] and its product. But complaints regarding the new features were unpredictable as they were based on user defined keyword search. A product may be identified with a specific name, whereas the username which is same as that of product name does not have both context and specific names. The advancement in the technology covered faster communications [3] [4] [5]. The previous decade experienced a dramatic development in computer technology, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet and this technique was named as web mining. Web mining is further classified into three types which are Web content mining, Web Structure Mining and Web Usage mining. Using the objects like text, pictures, multimedia etc. content mining is done in the web. In Web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users and the study of this navigational pattern will trace out the interest of the users [6]. In this project we are going to develop web content of user link prediction using relevant information extraction [7].

This section provides the basic overview of the proposed web mining based link prediction for information retrieval. In further sections the additional information regarding the system development is provided.

2. PROPOSED SYSTEM

This section provides the detailed description of the proposed working model. Therefore, the system overview is provided first then the methodology of the system is described. Finally, the proposed algorithm steps are provided.

A. *Domain Overview*

Mining of information is a task of extracting information from the rich data sources. In this work the web data mining technique using an application is described. The web mining is a domain where the data mining algorithms are used for analyzing the web data. Among a number of different kinds of data sources the web a more complex and huge data sources. That is a source of heterogeneous information where different kinds of information are spread over the world wide networks. On the other hand each and every day a number of new applications and information is updated in this infrastructure. But reflecting the web page importance and their reachability are depends on the organization of web application. Therefore the different techniques are applied to improve the web page ranking or improving the organization of web applications.

In the available literature it is observed that the organizations of the web pages are required in such manner by which the contents and their relativity can be reflexive. Therefore in this presented work for improving the web application organization a new technique is proposed. The proposed technique is a weighted technique for evaluation of impotence of the web page and their links. In order to do this, the web structure mining and web content mining concepts are merged. Additionally using the content analysis and link structure analysis the weighted ranking system developed for the web site page organization. This technique incorporates the content importance and the incoming and outgoing link directions to validate the web site structure. In this section the overview of the proposed concept is provided and the detailed methodology of the proposed concept is described in the further sections.

B. *Problem Domain*

The proposed work is motivated from the research article listed on [3]. In this given work the web pages are considered as the nodes and the links that organize the web pages are considered as links. These links are developed on the basis of the content associated on the web pages and provides the user's guidelines for searching the relevant contents on the web application. The proposed work is aimed to extend the work for the e-commerce relevant data search and using for improving the web pages association.

C. *Methodology*

This section provides the understanding about the proposed system which is required to develop for the demonstration of the performance study of User Web Link Information Retrieval system.

TF-IDF

TF-IDF stands for term frequency-inverse document frequency, and the TF-IDF weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a link is to a website in a collection. The importance increases proportionally to the number of times a link appears in the website but is offset by the frequency of the word in the corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a website relevance given a user query.

Typically, the TF-IDF weight is composed by two terms: the first computes the normalized Term Frequency (TF). The number of times a words appears in a website, divided by the total number of words in that website; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the website in the corpus divided by the number of website where the specific term appears.

TF: Term Frequency, which measures how frequently a term occurs in a Website. Since every website Link is different in length, it is possible that a term would appear much more times in long website than shorter ones. Thus, the term frequency is often divided by the website length (the total number of terms in the website) as a way of normalization:

$$\mathrm{TF\ (t)} = \frac{\text{Number of Times t Appears in a Documents}}{\text{Total Number of the Terms in the Documents}}$$

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$\mathrm{IDF\ (t)} \\ = \log\_e \frac{\text{Total Number of Website Links}}{\text{Number of the Websites Links with Term t in it}}$$

Proposed Flow Architecture

In order to improve the traditionally available model the following solution is proposed for implementation. The basic over of the proposed solution is demonstrated using the figure 2.1. In this diagram the web application is provided as the input to the system and the system extract the two primary things from all the web pages namely web contents and the links on the web pages. Using the extracted contents similarity among the web pages are computed. The similarity computation of the web pages is performed on the basis of the token frequency and the statement development frequencies.

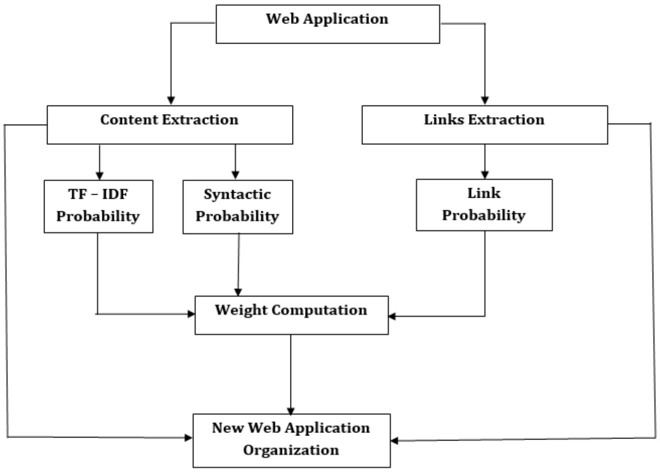

Fig 2.1: Proposed System Architecture

On the other hand for evaluation of links incoming and outgoing links based probability is computed. The computed four different factors of links and contents are combined using the weighted technique to produce the link's weights for the web pages. This may help the user to find the relevant contents on the web applications. This new organization of the web application is represented using the graphical user interface that provides the web pages and their content similarity in terms of the web page organization.

D. Proposed Algorithm

The given section provides the detailed description of the proposed implemented Semantic Synaptic Web User Link information Extraction systems. The entire system processes for improving the basic user accessing relevant information extraction is summarized using the algorithm steps as given in table 2.1.

Table 2.1 Proposed Algorithms for Web Link Information Extraction

```
Input: User Link $U_{link}$ or URL
Output: Web Content and Relevant Links $R_L$
Process:
1: $\mathrm{Link} = \mathrm{Input}\ (\mathrm{U_{link}})$
2: $\mathrm{Data, Link} = \mathrm{extractData}\ (\mathrm{Link})$
3: $\mathrm{TF - IDF\ Probability} = \mathrm{TF - IDF\ Probability.link}\ (\mathrm{Data, Link})$
4: $\mathrm{synapticLinkProbability} = \mathrm{synapticProbability}\ (\mathrm{Data, Link})$
5: $\mathrm{InlinkProbabilith} = \mathrm{InlinkProbability}\ (\mathrm{Data, Link})$
6: $\mathrm{OutLinkProbability} = \mathrm{OutLinkProbability}\ (\mathrm{Data, Link})$
7: $\mathrm{linkWeight} = \mathrm{findWeight}\ (\mathrm{TF - IDF\ Porb., Synptic\ Prob., Inlink\ Prb., OutLink\ Prpb.})$
8: $\mathrm{Return\ R_L}$
```

In table 2.1 shows the proposed algorithm of web link information extraction for user relevant data. To find the link relevancy of the website according to user input query we prepare the algorithm of which summarization of the stepwise process. Firstly user gives the input query link of the website. After that we have to extract the link in order to provide data and extracted link. Now we calculate probability of different constraints i.e. link, TF-IDF and synaptic.

After calculation of the different probability of the website link, finally, we calculate link weight of above all the probabilities to pass them in to weight. Finally we get the extracted data of the particular website which is relevant to the old link. And generate a tree graph which is show that all link information with website words. So, the above process of the algorithm produces result on the basis of different performance parameter that are precision, recall, f-measure, time and measure. This parameter is show the performance of the proposed system that more adoptable and efficient.

3. RESULT ANALYSIS

The given section includes the performance analysis of the implemented algorithms for the Proposed Semantic Synaptic Web user Link Extraction Systems. Therefore, the performance of algorithms is evaluated in this section.

A. Precision

In any data retrieval or search applications the precision is a fraction of search results which is most relevant to the input data query. The provided precision of the proposed Web User Link Information Extraction system are given using figure 3.1. This can be evaluated using the user feedback basis and can be evaluated by the following formula.

$$\mathrm{Precision} = \frac{\mathrm{Relevant\ Link} \cap \mathrm{Retrieved\ Link}}{\mathrm{Retrieved\ Link}}$$

The precision rate of the implemented system is described in the figure 3.1 and table 3.1, the computed precision values are demonstrated using the Y axis of the given figure and the X axis shows the link data. According to the obtained results the performance of the proposed system is showing efficient result when number of input links is retrieved. In addition of the precision rate is growing continuously as the similar kinds of high links are associated to their probability

Table 3.1 Precision Rate

Number of Input Links	Proposed Approach
1	0.6666
2	0.8095
3	0.6666
4	0.7116
5	0.7521

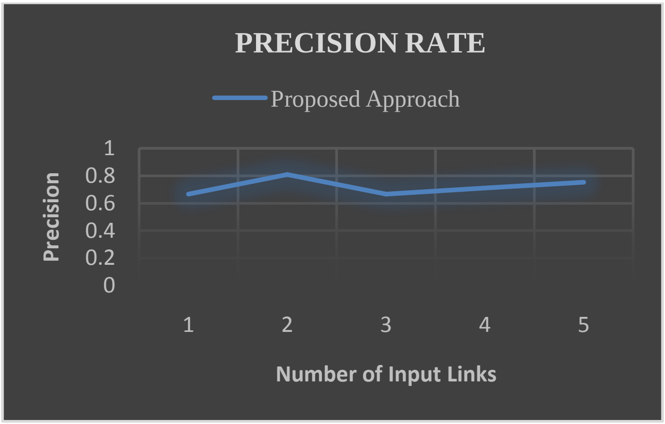

Fig. 3.1: Precision Rate

B. Recall Rate

In data retrieval application or the search application recall values are measured for accuracy measurement in terms of relevant document retrieved or relevant data obtained according to the input user query. This can be evaluated using the following formula.

$$\text{Recall} = \frac{\text{Relevant Link} \cap \text{Retrieved Link}}{\text{Relevant Link}}$$

The figure 3.2 and the table 3.2 show the recall values of the proposed Information Retrieval System. In order to represent the performance of the proposed work, the X axis contains the number of experiments to run the input link and the Y axis reports the obtained recall rate of the implemented system. According to the obtained results the performance of the proposed system is enhances as if we increase the experiment length. Proposed concept is adoptable for the e-commerce applications. Therefore the performance of the proposed system is much efficient for different types of web user link data.

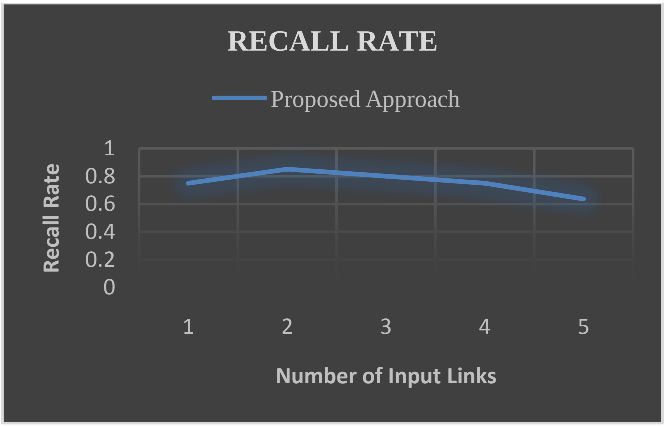

Fig. 3.2: Recall Rate

Table 3.2 Recall Rate

Number of Input Links	Proposed Approach
1	0.75
2	0.85
3	0.8
4	0.75
5	0.6363

C. F-Measures

The f-measures of the system demonstrate the fluctuation in the computed performance in terms of precision and recall rates. The f-measures of the system can be approximated using the following formula.

$$\text{F} - \text{Measures} = 2.\frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}}$$

Table 3.3 F-Measures

Number of Input Links	Proposed Approach
1	0.7085
2	0.8292
3	0.7272
4	0.6666
5	0.6086

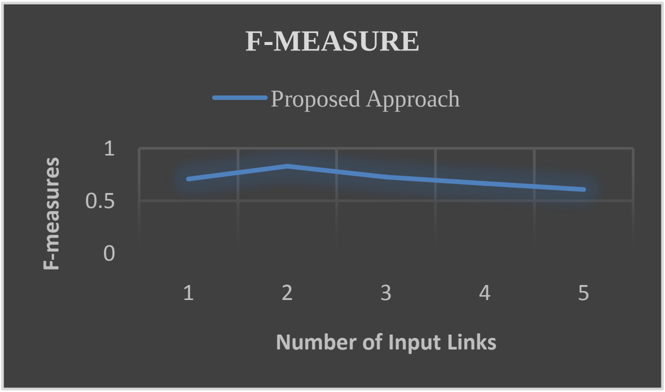

Fig. 3.3: F-Measures

The figure 3.3 and the table 3.3 show the performance of proposed systems in terms of f-measures. To demonstrate the performance of the system the X axis shows the number of experiments and the Y axis shows the obtained performance in terms of f-measures. According to the obtained results the performance of the proposed system is much stable and provides ease of use of website. Link organization and extraction process show the user web link to achieve relevant link In addition of that the results are in more progressive manner as if we increase number of experiments. Thus the obtained results are adoptable and efficient for semantic and synaptic web link information extraction.

D. Time Consumption

The amount of time required to process the link extraction time during running of algorithm is known as the time consumption. That can be computed using the following formula:

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

The time consumption of the proposed algorithm is given using figure 3.4 and table 3.4.In this diagram the X axis contains the number of experiments of input links and the Y axis contains time consumed in terms of milliseconds. According to the produced results analysis the performance of the proposed technique shows the low time consumption. But in most of the time it consumed constant amount of time which is deliver constant result.

Table 3.4 Time Consumption

Number of Input Links	Proposed Approach (Time in Millisecond)
1	52
2	164
3	33
4	21
5	63

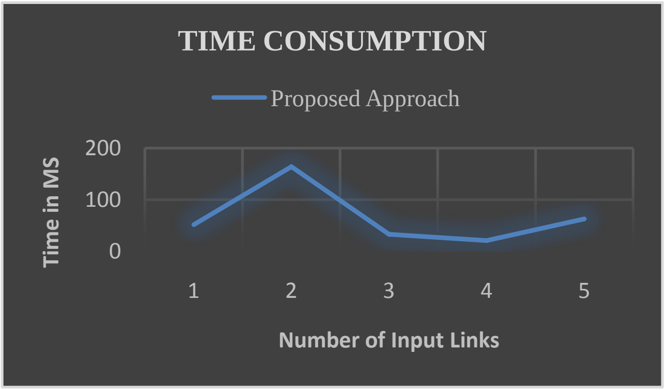

Fig. 3.4: Time Consumption

E. Memory Consumption

The memory consumption shows the amount of main memory required to process the algorithm with input amount of data to be processed. That is also known as the space complexity of algorithm. To compute the memory consumption, the following formula is used.

$$\text{Memory Usage} = \text{Total Memory} - \text{Free Memory}$$

The figure 3.5 and table 3.5 shows the memory consumption or space complexity of the system with increasing the number of runs of the input links. The amount of data is given using the X axis and the Y axis shows the amount of consumed memory during experimentation with respective amount of data in terms of kilobytes. According to the experimented results the amount of memory is similar and not more fluctuating. So, this graph shows the proposed system is not highly consumed memory other than traditional system.

Table 3.5 Memory Consumption

Number of Input Links	Proposed Approach (Memory in KB)
1	43155
2	26092
3	40071
4	34876
5	28061

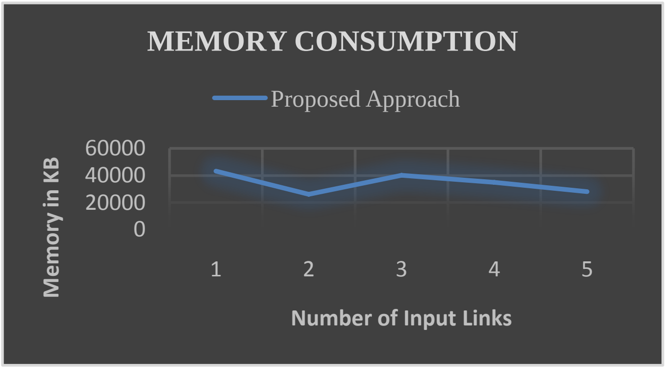

Fig. 3.5: Memory Consumption

4. CONCLUSION AND FUTURE WORK

This section draws the conclusion of the entire investigation work performed on the web mining domain. In addition of that the future extension of the work is also listed in this chapter.

A. Conclusion

The web mining is a technique where the web data i.e. links, links structure, web page contents and the web access logs are analyzed. According to the nature of data analysis the web mining is also distinguished in three different manners namely web usage mining, web content mining and the web structure mining. In this presented work the web structure mining and web content mining is the key area of study and investigation. In most of the structure mining concepts only the links and their structure is analyzed for recovering the essential pattern and their structural analysis. On the other hand the impacts of contents are also has importance in organization of web applications. So, to address the problem to develop an effective solution need to make additional efforts. Therefore a significant amount of literature is collect for the problem and solution formulation.

The main aim of the proposed approach is to analyze the existing web application organizational structure and produce most optimal structure for improving the user's and search engine reachability. Therefore the two different web mining concepts namely structure mining and content mining technique is employed to design a new technique. This technique works on the weighted link organization concept. In order to compute the weights for web pages and their correspondence the content mining features TF-IDF and the syntactic features are used. In addition of that the computed features are obtained in different scales thus need to normalize the computed features for computing the weights. The normalization technique involves the weighted factor assumption for computing the effective weights by which the direction of link structure and their advantages are become representable. Additionally for their existing structural organization the incoming and outgoing links probability is used. By using both the features the weights are estimated for organizing the web applications pages and their links.

The implementation of the proposed work is performed with the help of JAVA technology. Additionally for providing their performance the different parameters i.e. precision, recall, f-measures, time complexity and space complexity is computed. The obtained mean performance of the proposed system is given using table 4.1.

Table 4.1 mean performance

S. No.	Parameters	Mean performance
1	Precision	34.04
2	Recall	42.83
3	f-measures	41.68
4	Time consumption	11.2 MS
5	Memory usages	37809.2 KB

According to the obtained performance summary the proposed method is suitable for analyzing the web structure according to their contents and their relativity. Therefore the proposed model is adoptable for automated web structure analysis and generation of new organization of web applications.

B. Future Work(s)

The proposed work is intended to improve the performance of web applications by using the concept of web mining, more specifically the web structure mining and the web content mining. The primary goal of the proposed work is achieved successfully. In further for the more improvements and their different application feasibility the following suggestions are made for extensions.

- Incorporate the technique of web page ranking
- Implement the technique with the help of social media based feature too
- Improvement on the existing model by selecting more effective parameters for web structure and contents.

REFERENCES

[1] S. Monisha and S. Vigneshwari, “A Framework for Ontology Based Link Analysis for Web Mining”.

[2] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, PP. 1-15.

[3] Kumar, Sarowar, Kumar Abhishek, and M. P. Singh. "Accessing Relevant and Accurate Information using Entropy", Procedia Computer Science 54 (2015): PP. 449-455.

[4] Hiteshwar Kumar Azad and Kumar Abhishek, “Semantic-Synaptic Web Mining: A Novel Model for Improving the Web Mining”, Fourth International Conference on Communication Systems and Network Technologies, 2014, P. 454 – 457.

[5] Gurpreet Kaur and ShrutiAggarwal, “A Survey- Link Algorithm for Web Mining”, International Journal of Computer Science & Communication Networks, Volume 3(2), PP. 105-110.

[6] V. Lakshmi Praba and T. Vasantha, “Evaluation of Web Searching Method Using a Novel WPRR Algorithm for Two Different Case Studies”, ICTACT Journal on Soft Computing, APRIL 2012, Volume: 02, Issue: 03.

[7] About Web, available:http://www.technicalsymposium.com/web_mining_notes.html.