ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Hidden Markov Model for Duplicate Detection in Web Markup Language Data Mining

[I]Dr R.S. Vetrivel, [II]S . Priya, [III]P Jeyanthi Rani
[I]Professor, [II]Research Scholar, [III]Assistant Professor
[I,II,III]Computer Science, Subramanya College of Arts & Science, Palani.

Abstract
Data mining is the process of extracting the possible information from huge number of database. Data mining technology is used in the field of duplicate detection to discover the multiple representations in the same or different databases. Duplicate detection process is mainly related with the data cleaning and data integration scenarios.
Existing method was presented a technique for efficient and effective duplicate detection of hierarchical data. The Probabilistic duplicate detection algorithm (PDD) is introduced for hierarchical XML data (XMLDup). The PDD algorithm employed with Bayesian Network to determine the probability of two XML objects being duplicates. The pruning strategy is also used in this method to enhance the efficiency of XMLDup runtime. The accuracy rate of duplicate detection can be reduced for multiple similarity pair.
To overcome these problems, the proposed method is developed with Hidden Markov Model for efficient detection of duplicates in web markup language Data Mining. In this method, the Hidden markov model is used for obtaining the all information about data in database. According to this information, the similar data are occurred. The duplicate data elements classifier is developed for organizing the similar data which is used to detect the duplicate data elements. This proposed method can able to work with multiple similarity scores. The HMM method improves the efficiency and runtime for duplicate detection.

Key Terms
Data mining, Duplicate detection, Data cleaning, Hidden Markov Model, Web Markup Language.

I. Introduction

A. Data Mining

The main goal of data mining is the discovery of knowledge (information) from databases (KDD). Data mining process mines the data from the web effectively with less time consumption.
Data mining is used to examine the data from various perspectives and for summarizing it into valuable data. It permits to users for evaluating data from many different dimensions and categorize it and to identify the summarized data.
The data mining is introduced for discovering correlations or patterns between numbers of fields in huge databases. In data mining, uncertain information is discovered efficiently by recognizing the patterns and trends in data collected using classification, association, and clustering rules.
Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- To business analysts and information technology professionals, it gives the permission to access the data in web.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

B. Data cleaning

Data cleaning is the process of enhancing the quality of data by detecting and removing errors from data**.** Data quality can be affected by misspellings through data entry, illegal values, and duplicates, information loss and invalid data.
Data cleaning process is particularly required while combining the heterogeneous data sources. In data management, data cleaning is a most significant task which called as ETL (extraction, transformation, loading) process.
Data cleaning process which is called as data scrubbing is used to identify the corrupt and unrelated parts of the data from a table and database for correcting and deleting the dirty.
Data cleaning involves the following process

Data analysis:
Data analysis is required to identify which kinds of errors are needed to be removed and corrected.
ETL process:
Based on the number of data sources, errors can be removed by ETL process.
Verification:
For identifying the copy (duplicate) of the source data, the correctness and efficiency of a transformation workflow should be tested through verification process.
Transformation:
Data transformation can be done by running the ETL workflow for loading and refreshing a data.
Backflow of cleaned data:
After errors are removed, the cleaned data can be replaced the error data in the original sources which neglect the redoing cleaning process for future data extractions.

C. Hidden Markov Model

The Hidden markov model performed as markov chain for discovering duplicates from database. Hidden markov model is the process of creating an unobservable sequence. The future state in hidden markov model depends on the current states. But future states are assumed as hidden states. HMM gives information about the sequence of states which helps to find out the duplicates from original data.

D. Duplicate Detection

The detection and elimination of duplicated data is main task in the wide area of data cleaning and data quality in database. Duplicate detection is the process of comparing pairs of elements by evaluating similarity score based on their feature values.

When the similarity is greater than a predefined threshold, two elements are classified as duplicate. This outlook rejects the other available linked data information i.e., data stored in relational table relevant to data in other tables through foreign keys. The quality of data can be improved by elimination of duplicate data information.

II. Literature Survey

1. Web Service Diagnoser Model for managing faults in web services

In this paper [1], the author proposes a technique as WISDOM (Web Service Diagnoser Model). During execution of web services, the faults can be detected by this proposed method. The proposed strategy illustrates the intended behavior of web services. Imperfect behavior can be considered as deviations or inconsistencies with respect to the specified behavior.

During publishing, discovery, binding and execution of web services, run-time errors can be detected by examining the components in service registries and service providers with help of WISDOM Model. For the specified web service policies, the individual monitoring components can be organized by developing independent fault diagnoser.

2. Fast and robust duplicate image detection on the web

In this paper [2], the author proposes a technique with two different datasets by using different sets of distractor images. This proposed method leads to fully search a large-scale image collection (up to 100 million images) for duplicates in half a second on a 16-core processor.

The compact size (< 100 bytes) and the use of efficient Hamming distance computation allow us to mine a descriptor for one image. The fast and robust image description is achieved for indexing and searching with image data streams.

This method uses the efficient inverse index structure, for enhancing better accuracy in duplicate detection.

3. Crawling the Hidden Web: An Approach to Dynamic Web Indexing

In this paper [3], the author introduces a technique for dynamic web indexing. Through the integration of Hadoop- Mapreduce, possible future scope is represented to update and maintain the index.

The proposed work includes dynamism in content, not dynamism in appearance or user interaction. For dynamic web contents which are the part of hidden web, this method is developed with automatic indexing mechanism.

4. Analysis of accounting models for the detection of duplicate requests in web services

The author introduces a technique as cookie based accounting model in this paper [4]. Cookie based accounting model is developed to record each and every client request in the cookie and the hash value of the cookie in the server database.

Duplicate request attacks detection, accounting the client history (i.e., client request detail) is very essential in the web services. Client's misbehavior like modifying the cookie information or resending (replay) the prior request cookie with the current request are detected by accounting model which is used in this paper.

5. Near-Duplicate Segments based news web video event mining

The Near-Duplicate Segments technique was proposed in this paper [5] by the author for video event mining. The spatial and temporal information is effectively integrated by this proposed method.

Every video can be divided into segments which segments are arrived from different videos. But they are sharing similar visual content which are clustered into groups. Every group is named as an NDS, which concludes the latent content relation among videos.

The spatial-temporal local features are extracted which is used to represent each video segment. This proposed method is developed to captures the main content of news web videos and omit the noise efficiently.

6. Retrieve Main Content using Vision-base Web Page Segmentation with Gomory-Hu Tree

In this paper [6], the author proposes a technique as Gomory-Hu tree based Vision-based Page Segmentation (VIPS) algorithm for wed page segmentation. Web page segmentation included as Ranking, Duplicate detection, Content extraction process.

This proposed method is used to extract main content from web pages by using the vision and structure information which helps to make a weighted undirected graph.

7. Feature evaluation for web crawler detection with data mining techniques

In this paper [7], the author presents a technique as Feature evaluation for web crawler detection with the help of data mining techniques. According to following points, the mining techniques are established on web server.

(1) Based on an automated web crawlers or human visitors, the user sessions are classified.
(2) Identify which of the automated web crawlers sessions are demonstrate the 'malicious' behavior and feasibly participants in DDoS attack.

This proposed work established with web-session features as consecutive sequential request ratio and standard deviation of page request depth. The performance of classification is measured in terms of classification accuracy, recall, precision and F1 score. The value of new feature is measured as information gain and gain ratio metrics.

8. Improvised Architecture for Distributed Web Crawling

In the paper [8], the author introduces a technique for distributed web crawling. The scalable web crawling system and addressing the challenges is established by this proposed method. They are helps to resolve the problem which is related to the structure of the web, distributed computing, job scheduling, spider traps, canonicalizing URLs and inconsistent data formats on the web.

9. Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context

The author presents a technique as Real-Time Near-Duplicate Elimination in this paper [9]. Based on the content analysis derived from color and local points, the proposed work combine the contextual information from time duration, number of views which helps to attain a real-time near-duplicate elimination.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

The proposed work is developed for effectively integrating the content and context to reach real-time novelty re-ranking of web videos. The majority of duplicate scan run rapidly to detect and eliminate from the top rankings.

10. Pattern-based Near Duplicate Video Retrieval and Localization on Web-Scale Videos

The author introduces a technique for efficient and effective near-duplicate video retrieval and localization. Based on the hierarchical filter-and-refine framework, the spatiotemporal pattern-based approach was developed in this paper [10].

By the efficient data structure as Pattern-based Index Tree (PI-tree), non-near-duplicate videos are fast filtered out. The m-Pattern-based Dynamic Programming (mPDP) algorithm was performed with proposed method to discover the near-duplicate segments and to re-rank the videos retrieved. The time-shift m-pattern similarity (TPS) measurement enhances the influence of time shift misalignment.

11. Concept-based near-duplicate video clip detection for novelty re-ranking of web video search results

In this paper [11], the author proposes a technique as concept-based near-duplicate video clip (CBNDVC) detection for novelty re-ranking. By use of semantic features (events/concepts) and re-rank the top results, semantic NDVC are discovered which helps to enhance the content as well as semantic novelty. Videos are signified as a multivariate time series of confidence values of relevant concepts. CBNDVC clusters are identified with the help of conceptual clustering.

12. Anomaly detection techniques for a web defacement monitoring service

In this paper [12], the author introduces a technique for web defacement monitoring service. This proposed method was developed to perform with augmenting availability and monitoring services with defacement detection capabilities.

By the performance of several anomaly detection approaches, web defacements are detected automatically. The proposed approach develops a profile for monitored page automatically, with the help of machine learning techniques, and raises an alert while the page content not fit the profile

We evaluated the proposed method performance in terms of false positives and false negatives on a dataset composed of 300 highly dynamic web pages which examine for 3 months and a set of 320 real defacements.

13. An Efficient Duplication Record Detection Algorithm for Data cleansing

The author proposes a technique for data cleansing in this paper [13]. This proposed method is used to review, analyze and compare algorithms which enhance the efficiency and accuracy.

Initially, the relevant research papers are collected with the query of duplication record detection from IEEE database. Based on the different techniques proposed in the literature, these papers are categorized. After performing comparative analysis, the best selected algorithm is implemented to detect duplicate record effective

14. Security Testing Methodology for Vulnerabilities Detection of XSS in Web Services and WS-Security

In this paper [14], the author develops an approach with the help of two Security Testing techniques such as Penetration Testing and Fault Injection. These techniques are used to emulate XSS attack against Web Services. The proposed work combined with WS Security (WSS) and Security Tokens which discover the sender and guarantee the legitimate access control to the SOAP messages exchanged.

In this paper, the method uses the vulnerability scanner soap UI which is one of the most recognized tools of Penetration Testing. WS Inject is the fault injection tool, which establishes faults or errors on Web Services to analyze the behavior in an environment not robust.

15. Finding Similar Identities among Objects from Multiple Web Sources

In this paper [15], a new approach was developed to find out the Similar Identities among Objects from Multiple Web Sources. The object identification performs like as relational join operation. The similarity function was designed based on the information retrieval techniques which takes the place of the equality condition. This proposed work used to identify the objects which more complexly structured (e.g., XML documents) and not only objects with a flat structure such as relations.

III. Methodology

The Hidden Markov Model is developed for duplicate detection in web markup language data mining. The proposed method is used to improve the efficiency of detection in duplicates elements which provides accurate search results.

The efficient and effective duplicate detection method was developed for hierarchical data. The existing method was introduces a probabilistic duplicate detection algorithm (PDD) for detecting duplicate in hierarchical data (XMLDup).

The hierarchical relationships in XML provide us additional information that helps to improve the runtime and quality of the duplicate detection. XMLDup is used with Bayesian network to determine probability of two XML elements which consisting information about elements structure of XML data.

The PDD algorithm enhances the efficiency of the network evaluation with pruning strategy. It considers the similarity of the attribute contents and relevant meaning of descendant elements which support for similarity scores. In recall, improvement of efficiency in slight drop is very important. It helps to discover the number of identified duplicates is manually tuned or performed automatically using the known duplicate objects.

The PDD technique has the following disadvantages.

- The existing method was performed for single similarity scores not for multiple similarity scores.
- It reduces the accuracy rate for multiple similarity pair duplication.
- The increase in XML structure shows poor duplicate detection.
- The rate of speed of duplicate detection was not appreciable.

To overcome these problems we propose a technique as Hidden Markov Model in web markup language data mining to detect the duplicates.

In this proposed work, the hidden markov model is developed with hidden states which are considered as future states. The present states are fully depends on the future states. Then the HMM model having entire information about states of elements like structure and quantity of data. This information's are very essential to detect

the duplicate effectively from database.
Advantages of the proposed method

- The proposed method can work with multiple similarity scores.
- By this proposed method, the accuracy rate of duplicate detection is increased.
- The speed in duplicate detection also improved by HMM method.

Figure 3.1 represents the process of Hidden Markov Model for duplicate detection. Initially the user (client) sent a query request for finding the duplicates. The Hidden Markov model is assumed as Markov chain for which the state is only partially observable. But HMM gives all information about the sequence of states from the database.
Based on the information getting from HMM, the two possible similar elements (data) are discovered. If the similarity is greater than the predefined threshold, then the two elements are indicated as duplicates. Comparing data stored in relational table relates to data in other tables by the duplicate data elements classifier which identifies the duplicate data efficiently.

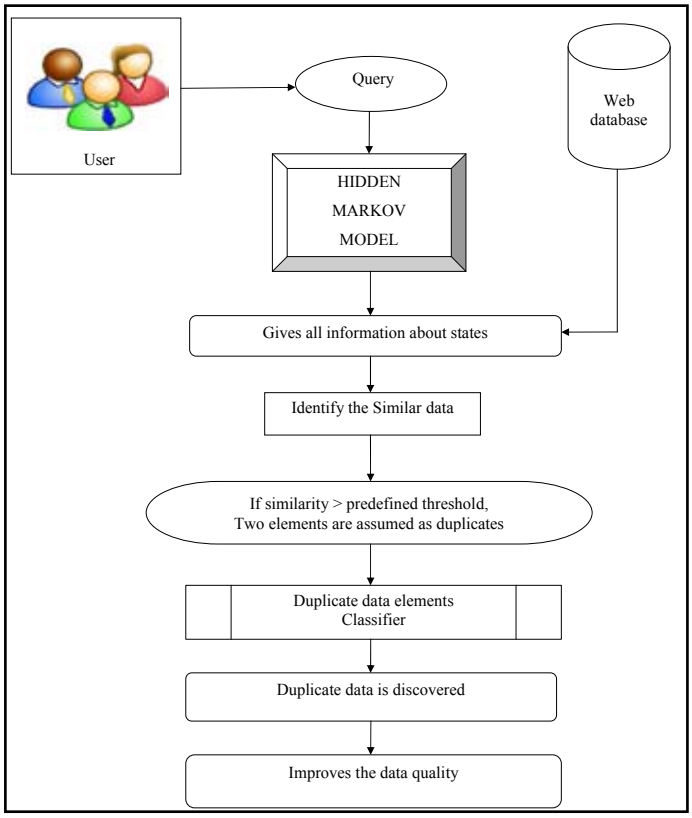

Fig. 3.1: Architecture of the HMM model for duplicate detection

A. Modules

a) Hierarchical Web Markup Language Data model
b) Hidden Markov Model
c) Duplicate data elements classifier

B. Modules Description

1. Hierarchical Web Markup Language Data model

Hierarchical and Web Markup Language Data are most important in the duplicate detection. For duplicate detection in single relation do not directly apply to markup data.
Instances of a same object type have a different structure at the instance level. Tuples within relations always have same structure. A hierarchical relationship in markup data provides valuable information which enhances the runtime and quality of duplicate detection.

2. Hidden Markov Model

Markov model is the arbitrarily changing model being modeled as future states depends on the present state. In Markov models, the state is directly visible to the observer. The state transition probabilities are done by parameters.
In Hidden Markov model, the future states assumed as unobserved (hidden) states. The Hidden Markov model is considered as Markov chain which the state is partially observable. In a Hidden Markov model, the future state is not directly visible. Every state has the probability distribution among the sufficient output tokens. The sequence of tokens created by an HMM which gives information about the elements of states. The Viterbi and forward algorithms gives probability of the most likely sequence of states. The Viterbi or forward algorithms are used to identify the duplicates based on the similarity of data.

3. Duplicate data elements classifier

Duplicate data elements classifier classifies the duplicate from the original data .The duplicate classifier is used to identify the duplicate data with the help of HMM. The detection of duplicate is directly deal with the quality of data.
Based on the information fetching from HMM, similarity measure is developed to classify pairs of objects as duplicate or non-duplicate. Duplicate detection is the process of comparing pairs of elements by estimating similarity score based on their aspects values.
If the similarity is greater than a predefined threshold, two elements are denoted as duplicates. This analysis helps to ignore the other available correlated information i.e., data stored in relational table relates to data in other tables through foreign keys. Finally, duplicate data is separated from original data in database. Through the detection and discard the duplicate information's from database which enhances the quality of data automatically.

Input: Database 'D', Number of queries 'Q'
Output: Identify the duplicate data.
Step 1: Begin Step 2: Enter the required text Step 3: Hidden Markov model gives the information about the elements Step 4: Based on information, similar data are discovered Step 5: If similarity > predefined threshold, two elements are assumed as duplicates Step 6: Duplicate data elements classifier classifies the data Step 7: Duplicate data is identified Step 8: End

Fig. 3.2: Algorithm for duplicate detection by using HMM method

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

IV. Experimental Evaluation

In this section, we evaluate performance of Hidden Markov Model for Duplicate Detection in Web Markup Language Data Mining through java environment. In this paper worked on performance evaluation in terms of Searching speed, Multiple similarity scores, Accuracy in duplicate detection. The performance measures of the proposed work analyzes with following metrics.

1. Searching Speed
2. Multiple Similarity Scores
3. Accuracy in duplicate detection

A. Searching Speed

The Searching speed is defined as the rate of detecting the duplicate data based on the information which gives by HMM model.
The Searching Speed is measured in terms of milliseconds (ms).

Table 4.1: Tabulation for Searching speed

Document size	Searching Speed (ms)	
	Probabilistic duplicate detection algorithm (PDD)	HMM for Duplicate Detection
5	10	12
10	20	22
15	30	34
20	40	44
25	50	56

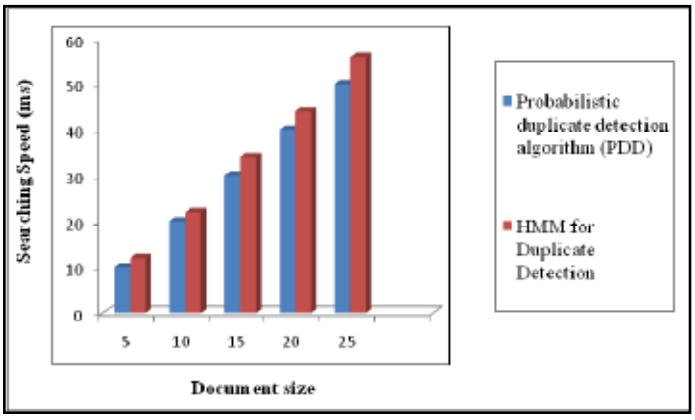

Fig 4.1: Measure of Searching speed

Figure 4.1 demonstrates the searching speed. From the figure X axis represents the document size whereas Y axis denotes searching speed using HMM method. The documents taken for the experimental consideration is varied from 5 to 25. From the figure it is clearly evident that the proposed HMM for duplicate detection technique improves searching speed results than the Existing PDD method. Hence, the searching speed is improved up to 13% by using proposed HMM technique than the existing PDD method.

B. Multiple Similarity Scores

HMM technique uses a similarity measure to provide learning to discriminate between positive and negative members of a given class of n-dimensional vectors operates by mapping given training set of data into a possibly high dimensional feature space and attempting to locate in space plane separates the duplicate from the original data.

Table 4.2: Tabulation for Multiple Similarity Scores

Document size	Multiple Similarity Scores	
	Probabilistic duplicate detection algorithm (PDD)	HMM for Duplicate Detection
5	45	50
10	50	57
15	55	62
20	60	66
25	65	70

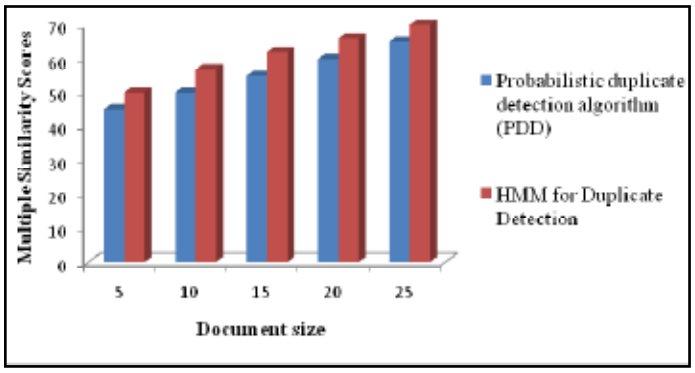

Fig. 4.2: Measurement of Multiple Similarity Scores

Figure 4.2 demonstrates the Multiple Similarity Scores. From the figure X axis represents the document size whereas Y axis denotes Multiple Similarity Scores using HMM method. The documents taken for the experimental consideration is varied from 5 to 25. From the figure it is clearly evident that the proposed HMM for duplicate detection technique improves Multiple Similarity Scores results than the Existing PDD method. Hence, the Multiple Similarity Scores is improved up to 11% by using proposed HMM technique than the existing PDD method.

C. Accuracy in duplicate detection

The Accuracy is defined as the rate of the number of queries which is identified as duplicates through the similarity measure by using HMM model.
Accuracy of duplicate detection is measured in terms of percentage (%).

Table 4.3: Tabulation for Accuracy in duplicate detection

Document size	Accuracy in duplicate detection (%)	
	Probabilistic duplicate detection algorithm (PDD)	HMM for Duplicate Detection
5	30	36
10	35	40
15	42	46
20	50	52
25	55	60

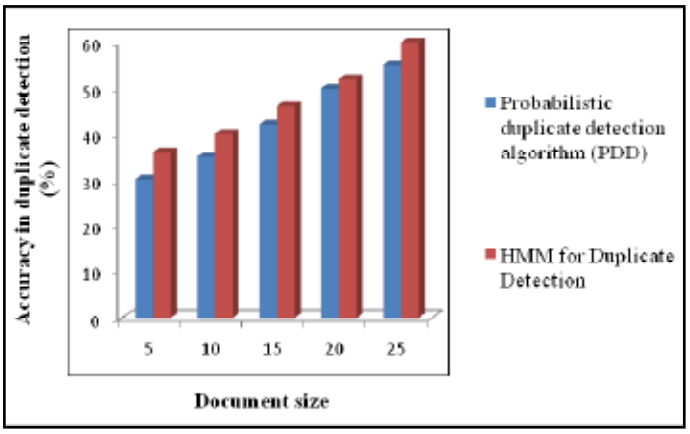

Fig. 4.3: Measurement of Accuracy in duplicate detection

Figure 4.3 demonstrates the Accuracy in duplicate detection. From the figure X axis represents the document size whereas Y axis denotes Accuracy in duplicate detection using HMM method. The documents taken for the experimental consideration is varied from 5 to 25. From the figure it is clearly evident that the proposed HMM for duplicate detection technique improves Accuracy in duplicate detection than the Existing PDD method. Hence, the Accuracy in duplicate detection is improved up to 11% by using proposed HMM technique than the existing PDD method.

V. Conclusion & Future Scope

In this paper, the author proposes a technique as Hidden Markov Model to detect duplicates in Web Markup Language Data Mining. HMM method is developed in this paper, for obtaining the whole information about the data elements. HMM is used to reduce the runtime for detecting duplicates with enhanced accuracy in duplicate detection. The main objective of this proposed work is identifying the duplicates efficiently through the similarity measure. The identification of duplicates improves the quality of data with high accuracy.

The future work includes the some other cloud based application to be implement and also enhance the real time application in further investigation.

References

[1] *K. Jayashree , Sheila Anand " Web Service Diagnoser Model for managing faults in web services" ELSEVIER, Computer Standards &Interfaces, Volume 36, Issue 1, November 2013, Pages 154-164.*

[2] *Etienne Gadeski , Herve Le Borgne, Adrian Popescu " Fast and robust duplicate image detection on the web" Multimed Tools Applications, Multimedia Tools and Applications , May 2016, pp 1–20.*

[3] *Moumie Soulemane, Mohammad Rafiuzzaman, Hasan Mahmud "Crawling the Hidden Web: An Approach to Dynamic Web Indexing" International Journal of Computer Applications (0975 – 8887) Volume 55– No.1, October 2012.*

[4] *S. Venkatesan , M.S. Saleem Basha , C. Chellappan , Anurika Vaish , P. Dhavachelvan " Analysis of accounting models for the detection of duplicate requests in web services" Journal of King Saud University –Computer and Information Sciences, Volume 25, 2013, Pages 7–24.*

[5] *Chengde Zhang , DiantingLiu , XiaoWu , GuiruZhao , Mei-LingShyu , Qiang Peng "Near-Duplicate Segments based news web video event mining" ,ELSEVIER, Signal Processing, Volume 120, March 2016, Pages 26–35.*

[6] *Khaing Wah Wah Linn " Retrieve Main Content using Vision-base Web Page Segmentation with Gomory-Hu Tree" International Journal of Computer Applications (0975 – 8887) Volume 108 – No 17, December 2014, Pages 34-37.*

[7] *Dusan Stevanovic, Aijun An, Natalija Vlajic "Feature evaluation for web crawler detection with data mining techniques" ELSEVIER, Expert Systems with Applications, Volume 39, Issue 10, August 2012, Pages 8707-8717.*

[8] *Tilak Patidar, Aditya Ambasth "Improvised Architecture for Distributed Web Crawling" International Journal of Computer Applications (0975 – 8887), Volume 151 – No.9, October 2016.Pages 14-20.*

[9] *Xiao Wu, Chong-Wah Ngo, Alexander G. Hauptmann, Hung-Khoon Tan "Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context" IEEE Transactions on Multimedia, Vol. 11, No. 2, February 2009, Pages 196-207.*

[10] *Chien-Li Chou, Hua-Tsung Chen, Suh-Yin Lee "Pattern-based Near Duplicate Video Retrieval and Localization on Web-Scale Videos" IEEE Transactions on Multimedia, Volume 17,Issue 3,2015,Pages 382-395.*

[11] *Chidansh A. Bhatt ,Pradeep K. Atrey , Mohan S. Kankanhalli " Concept-based near-duplicate video clip detection for novelty re-ranking of web video search results" Multimedia Systems , Volume 18, Issue 4, July 2012, pp 337–358.*

[12] *G. Davanzo , E. Medvet, A. Bartoli, "Anomaly detection techniques for a web defacement monitoring service" ELSEVIER, Expert Systems with applications, Volume 38, Issue 10, 15 September 2011, Pages 12521–12530.*

[13] *Arfa Skandar, Mariam Rehman, Maria Anjum "An Efficient Duplication Record Detection Algorithm for Data Cleansing" International Journal of Computer Applications (0975 – 8887), Volume 127 – No.6, October 2015, Pages 28-37.*

[14] *M.I.P. Salas, E. Martins "Security Testing Methodology for Vulnerabilities Detection of XSS in Web Services and WS-Security" ELSEVIER, Electronic Notes in Theoretical Computer Science , Volume 302 ,25 February 2014, Pages 133–154.*

[15] *Joyce C. P. Carvalho , Altigran S. da Silva " Finding Similar Identities among Objects from Multiple Web Sources" November 2003, Pages 90-93.*