ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Literature Survey on Noise and Redundant Data on Web Pages

[1]Sekhar Babu.Boddu, [2]Prof.Rajasekhara Rao.Kurra

[1]Research Scholar, Dept. of CSE, SCSVMV University, Kanchipuram, Chennai, India

[1]Asst. Professor, Dept.CSE, KL University, AP, India

[2] Professor, Dept.CSE & Director, Usharama Engg College, AP, India

ABSTRACT: Filtering noisy data and getting needful data have become the key issues for web mining, finding and accessibility. These days web technology is reaching appearance significance in day to day life, each one is browsing on the web uploading main data on the web. A business websites normally contains noisy data blocks with key content, it usually such blocks as navigation panels, copyright, privacy notices and advertisements. There are a variety of researches which are focusing on the mine appropriate in sequence from the web pages.

The quick development of the internet has made the www a well-liked position for collecting in turn

.

KEYWORDS: Web Mining, Web Content mining

I. INTRODUCTION

A web mining has important assignment to ascertain advantageous ability or advice from the web. Web mining can be divided in to three parts: web content mining, web structure mining and web usage mining. Web structure mining is the action of advertent hyperlink and certificate structure advice from the web. Web usage mining is the appliance of data mining techniques for award absorbing and advantageous usage patterns from web abstracts which make it added ambitious for web based applications. Web content mining is the action of extracting advantageous advice from the capacity of web documents. Current techniques are mainly based on apparatus acquirements and accustomed accent processing approaches to learner generalization rules from manually labelled examples. Nowadays abounding advisers are proposing structure based babble abolishment or amount agreeable extraction. Basically their adjustment splits the Web abstracts into small sections by application various tags, eventually chief which area is amount and non-core by using various information metrics. However, the adjustment that we proposed does not use any structure analysis. Instead we focus on the back-up of Web abstracts from abovementioned URL pavement and on the abolishment of long-winded advice for affection extraction. Though our access can be activated as pre-processing for either Web generous or Web monitoring, this cardboard focuses on the Literature Survey on assorted Research affidavit, and they are focused on Dom Tree, Web ecology and certificate Classification context, because in this environment the URL paths that transparent abstracts abolishment arrangement action is a bound set and the admeasurements of set is abate set than a Web sufficient arrangement could be process. The afterward table will accord the data about what are the techniques and models are acclimated by the authors and accommodate the collection Like Blocks, Data, Precision & Recall.

II. LITERATURE SURVEY ON NOISE

Hassan F. Eldirdiery [1]The developed algorithm called BDBNE (Block Density-Based Noise Extractor), processes the target web document and segments it into many blocks. Then it analyzes the extracted blocks in order to distinguish the noisy blocks from not noisy blocks using pre-computed threshold value. The algorithm works on the HTML file of the web page. It used the sequence of characters outside HTML tags to construct the blocks and ignores the sequence of characters inside HTML tags. The process of detecting noisy blocks based on the text density of the block.

DOI: 10.15680/IJIRCCE.2017. 0505170

Ms. Shalaka B. Patil,[2]
Webpage Content Searching (WCS) algorithm will reflect the detail work of the system design after providing a text query as a primary input. Algorithm shows the classification of HTML file into three separate segments viz. namely presentation, structure and content based on their data type. The content within DIV tags are compared with the text query of the user. If the content is similar to the text query within the web-page, it will be stored in Similarity Dataset. Using this dataset, system will search for frequent occurrence of text query in each DIV for extracted webpage and calculates similarity (frequency) within those DIV tags of each page for user required content.

Rajni Sharma, Max Bhatia [3]
DOM tree has some disadvantages like it has high complexity and time consuming process. To overcome this Problem a new algorithm is used in page replacement that is least recently used (LRU). A good approximation to the best possible algorithm is based on the surveillance that pages that has been greatly used in the last. A small number of instructions will most likely be a lot used all over again in the next the minority.

Hui Xiong et.al [4]
In order to enhance data analysis in the presence of high noise levels, While our noise removal techniques are based on outlier detection, these techniques are different from outlier detection techniques in two significant ways. First, the notion of an anomaly or outlier implies rareness with respect to the majority of normal objects. However, as this paper demonstrates, eliminating a substantial fraction of all data objects can enhance the data analysis. Second, outlier detection techniques seek to avoid classifying normal objects as outliers, the elimination of irrelevant or weakly relevant (normal) objects is often essential for enhancing the data analysis.

Rekha Garhwal [5]
The Web mining process is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. Data mining on the Web thus becomes an important task for discovering useful knowledge or information from the Web.

Erdinc Uzun et.al [6]
Eliminating noisy information and extracting informative content have become important issues for web mining, search and accessibility. This extraction process can employ automatic techniques and hand-crafted rules. Automatic extraction techniques focus on various machines learning methods, but implementing these techniques increases time complexity of the extraction process. Conversely, extraction through hand-crafted rules is an efficient technique that uses string manipulation functions, but preparing these rules is difficult and cumbersome for users.

Shine N. Das [7]
Here we propose a novel and efficient frame work for the true detection and well suited removal of noisy blocks. Noise elimination can be implemented as a pre-processing step for web content mining and especially for web page classification. Our objective is to find how to identify noisy blocks or irrelevant blocks from an input record, the web page to be processed, with a reduced complexity and increased efficiency. A three stage algorithm is proposed with phases featuring, modelling and pruning.

Xin Qi and JianPeng Sun [8]
Web page parsing is the process of establishing the corresponding DOM tree based on a given HTML page, but the HTML language itself has a high degree of flexibility, allowing the realization of the process becomes very complicated. There are already many excellent open source project can help us accomplish this work. This paper used a very popular HTML parser Neko HTML can often revised HTML coding errors, and allows developers to use standard XML interface to operate on the HTML document.

Thanda Htwe et.al [9]
The Document Object Model (DOM) specification is an object-based interface developed by the World Wide Web Consortium (W3C) that builds an XML and HTML document as a tree structure in memory. An application accesses the XML data through the tree in memory, which is a replication of how the data is actually structured. The DOM also allows the user to dynamically traverse and update the XML document . It provides a model for the whole document, not just for a single HTML tag.

Byeong Ho Kang [10]
We classified the Web document data as following three types: core information, redundant information and hidden information. Core information is the content that a user wants to view from a Web page. For example, the main article in the news article Web page is core information. This information is mainly used for text classification tasks. Redundant information is added to enhance Web content accessibility or business attractiveness. Inserting this information is promoted officially by W3C or profit seeking companies. Web documents also contain the ‘hidden information’ like HTML tags, script language and programming comments, which is called ‘hidden’ because it is not seen by end users. Users only can see it by performing the “view source” action.

Thanda Htwe[11]
Artificial Neural network (ANN) model can be known as a good problem-solving method for problems that can’t be solved using conventional algorithms. Neural networks are very good at pattern-recognition and pattern-matching tasks. If the input is one it has never seen before, it produces an output similar to the one associated with the closest matching training input pattern. next layer.

T. Sun et.al[12]
Created a DOM tree on the visual blocks of a web page and for each block, an information block matching ratio is calculated and checked against a threshold to identify the level of relevancy. But the algorithm was mainly based on an assumption that the same site are often made from a different page with an HTML template generation, their structure is very similar to the same or only part of the theme of data with different contents.

Kang[13]
Built a tree alignment model representing HTML structure and a vector model representing the features of the blocks. They stated that the blocks of a web page might be related to different categories even though they are structurally similar. Since it is difficult to classify the blocks into accurate categories through building one classifier, multiple classifiers are built, one for each training domain, and the block classification proceeded through combining them. Through block classification, relevant and irrelevant blocks are identified.

Lan Yi et.al[14]
The proposed cleaning technique is based on the analysis of both the layouts and the actual contents (i.e., texts, images, etc.) of the Web pages in a given Web site. Thus, our first task is to find a suitable data structure to represent both the presentation styles (and layouts) and the actual contents of the Web pages in the site. We propose a Style Tree (ST) for this purpose. Below, we start by giving an overview of the DOM (Document Object Model) tree, which is commonly used for representing the structure of a single Web page, and showing that it is insufficient for our purpose.

III. LITERATURE SURVEY ON REDUNDANT DATA

Duplicate content on a site is not grounds for action on that site unless it appears that the intent of the duplicate content is to be deceptive and manipulate search engine results. If your site suffers from duplicate content issues, and you don't follow the advice listed above, we do a good job of choosing a version of the content to show in our search results.

Broder et al (1999) has proposed a technique for the estimation of the degree of similarity among pairs of documents , which was known as shingles, does not rely on any linguistic knowledge, other than the ability to tokenize documents into a list of words, i.e., It is merely syntactic. In shingling, all word sequences of adjacent words are extracted. If two documents contain the same set of shingles they are considered equivalent and if their sets of shingles appreciable overlap, they are exceedingly similar.

Deniel (1999) have proposed in the duplicate document detection, various works have been performed. Their techniques have been utilized by many applications. However, the investigation of the performance and scalability of Duplicate Document Detection (DDD) is modes performed a systematic study of parameter correlations in DDD and evaluated numerous most important parameters of DDD.

Steinbach et al (2000) has proposed that web pages are grouped into clusters of incredibly similar documents. They identified that in their dataset almost one third of the pages are near duplicates of other pages. The Grainy Hash Vector has (GHV) representation, which can be deployed in cooperative DIR systems for efficient and accurate merge-time duplicates detection.

Arasu et al (2001) have proposed a study on the diverse technique to eliminate the duplicates and near duplicate objects in the MyLifeBits personal storage system results of near-duplicate detection for personal contents like emails, documents and web pages visited, are efficient. The number of documents and the number of web pages that a user must consider was reduced by 21% and 43% respectively, by the duplicate and near duplicate detection.

Ilyinsky et al (2002) have proposed the process of managing these during incremental growth, acquisitions, mergers, and integration efforts inevitably results in some duplication are defined, this natural entropy by using a process that mines the repository for partially duplicated material, helping to maintain the highest quality control of the content. Although the overall process is satisfactorily efficient with computer resources, practically, human attention to consider the many results is the bottleneck.

Jack G Conrad et al (2003) have proposed to investigate the phenomenon and determine one or more approaches that minimize its effect on search results. The determination of the extent and the types of duplication existing in large textual collections was their chief objective. In addition, one or more approaches that minimize its deleterious impact on search results in an operational environment were devised. The issues of computational efficiency and duplicate document detection (and, by extension, “deducing”) are effectiveness in relying on “collection statistics” to recognize consistently document replicas in full-text collections.

Tripathy and Singh (2004) have proposed a technique which was employed to eliminate noise. A tree structures, called the Pattern Tree was proposed to capture the general presentation styles and the definite essential of the pages in a specified Web site. A Pattern Tree called the Site Pattern Tree (SPT) was put up for the site, by sampling the pages of the site.

Cesario et al (2005) have proposed a hybrid queries-dependant duplicate detection scheme that combines the advantages of both online and offline methods. The solution provided in duplicate detection of the hybrid method was effective and in addition scalable. Precisely, the method initially conducts offline processing for popular queries. Then to improve additionally the performance for unpopular queries, it does additional work at run time. The scalability problem of traditional offline methods could be effectively dealt with such a strategy, if the performance problem of traditional online methods is avoided.

Hui Yang et al (2006) have proposed the problem of duplicate and near-duplicate text has become increasingly important owing to the growth of the text collection in size and various sources from which it is gathered. An instance level constrained clustering was proposed as a solution to near duplicate detection for notice and comment rulemaking. The ability of Instance-level constrained clustering to express the varied information based on document attributes, information extracted from the document text, and structural relationships among pairs of documents as constraints on cluster contents is its advantage. Thus accuracy and efficiency are improved as the search space is narrowed. They conducted experiments with EPA and DOT datasets. They demonstrated that at less computational cost than competing methods, their approach in detection of near-duplicate was almost efficient as high quality manual assessment.

Manku et al (2007) have proposed made two research contributions in developing a near-duplicate detection system intended for a multi-billion page repository. Initially, they demonstrated the appropriateness of Charikar's fingerprinting technique for the objective. Subsequently, they presented an algorithmic technique to identify the existing f-bit fingerprints that varies from a given fingerprint in at most k bit positions, provided that the value of k is small. Both online queries (single fingerprints) and batch queries (multiple fingerprints) are aided by this technique. The expediency of the experimental evaluation confirmed their design over real data.

PrasannaKumar and Govindarajulu (2009) have proposed a problem of finding all documents-pairs swiftly whose similarities are equal to or greater than a given threshold is known as duplicate document detection. A multi-level prefix-filter, which is reduce the number of similarity calculation more efficiently and maintains the advantage of the current prefix filter by applying multiple different prefix-filters.

Poonkuzhali et al (2009) have proposed a method of distinguishing the redundant links from the web documents that utilized set theory (classical mathematics) such as subset, union, intersection etc., and proposed an algorithm for mining the web content. Then for obtaining the required information, the redundant links were taken out from the original web content.

Alpuente and Romero (2010) have proposed an approach to identify similar documents based on a conceptual tree-similarity measure. They used the concept associations obtained from a classifier to represent each document as a concept tree. Subsequently, they computed the similarities between concept trees by using a tree-similarity measure based on a tree edit distance. They conducted experiments on documents from the Site-Seer collection and illustrated that when compared to the document similarity based on the traditional vector space model, the performance of their algorithm was significantly better.

Ranjna Gupta et al (2010) have proposed a new approach that performs copy detection on web documents are copying detection approach determines the similar web documents, similar sentences and graphically captures the similar sentences in any two web documents. Besides handling a wide range of documents, their copy detection approach is applicable to web documents in different subject areas as it does not require static word lists.

Syed Mudhasir et al (2011) have proposed a problem of improving the stability of I-Match signatures with respect to small modifications to document content instead of using just one IMatch signature, they employed numerous that I-Match signatures all which were derived from randomized versions of the original lexicon, in their proposed solution. The proposed
scheme does not involve direct computation of signature overlap regardless of employing multiple fingerprints.

Huda Yasin and Mohsin Mohaammad Yasin (2011) have proposed a Jaccard Coefficients for calculating similarity of text attributes. Analytical Hierarchy Processed (AHP) is used to obtain the weights of entity. Using the sum of weights, the entity similarity is calculated and it needs to integrate duplicate entity for achieving the entity identification.

IV. DETECTION ALGORITHMS ON WEB DATA

Daniel P Lopresti (1999) has proposed a significant amount of network bandwidth for clarifying and formalizing the duplicate document detection problem. They used uncorrected OCR output to study several issues related to the detection of duplicates in document image databases. They presented four distinct models for formalizing the problem and in each case they present algorithms that determine the best solution. The algorithm most suited to a particular a solid highlighted problem dot ( ). Whereas the algorithm that will find not only such duplicates but other types as well, is showed by a hollow dot ( ). They conducted experiments by using data reflecting real-world degradation effects to illustrate the robustness of their techniques.

Broder et al (2000) have proposed an identification of near exact duplicate web page shingling algorithm random projection based aloft he were co-utilized state-of-the-art" algorithms. These two algorithms are compared, on a very large scale, specifically for a set of 1.6 Byte distinct web pages. In accord number of the results, if of identifying the near duplicate pairs on the same site, neither of the algorithms works well, whereas if of dissimilar sites, they both obtain high precision.

Chowdhury et al (2001) have proposed a novel similar document detection algorithm called I-Match. They used multiple data collections to evaluate their performance. The employed document collections were different in terms of size, degree of expected document duplication, and document lengths. NIST and Excite@Home were the source of the data employed.

Ahmad (2004) have proposed a novel data reduction algorithm employing the concept analysis which can be used as a filter in retrieval systems like search engines to eliminate redundant references to the similar documents. A study was performed on the application of the algorithm in automatic reasoning which effected in minimizing the number of stored facts
without loosing of knowledge, by the authors. Their results illustrate that besides reducing the user time and increase his satisfaction; there was a good increase in the precision of the retrieval system.

Fetterly et al (2004) have proposed a study on the evolution of web pages over time during which many machines-generated “spam” web pages emanating from a handful of web servers in Germany, was discovered. The grammatically well-formed German sentences drawn from a large collection of sentences were stitched together to assemble dynamically these spam web pages. The development of techniques to find other instances of such “slice and dice” generation of web pages, where pages are automatically generated by stitching together phrases drawn from a limited corpus, is aggravated by the discovery.

Muhammad Sheikh Sadi et al (2004) have proposed an efficient algorithm to measure relevance among web pages using hyperlink analysis (RWPHA). RWPHA searched the web using the URL of a page rather than the set of query terms, given as input to the search process. A set of related web pages is the output. A web pages that address the same topic as the original page is known as the related page. RWPHA does not employ the content of pages or usage information rather only the connectivity information on the web (i.e., the links between pages) is utilized. The extended Co citation analysis is the basis of the algorithm. The superior performance of the algorithm over some dominant algorithms in finding relevant web pages from the experimental results illustrated linkage information.

Yang and Callan (2006) have proposed a work for exact-near duplicate detection, for which the process of identifying near duplicates of form letters is the focus. They defined the most near and exact-duplicates that are appropriate to eRulemaking and explored the employment of simple text clustering and retrieval algorithms for the task. The effectiveness of the experiments illustrated the method in the public comment domain.

Yang and Callan (2006) have proposed a refinement of a prior near duplicate detection algorithm. DURIAN (DUplicate Removal In large collection N), identifies form letters and their edited copies in public comment collections by employing a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure. In accordance with the experimental results, DURIAN was almost as effective as human assessors. They discussed the challenges in moving the near-duplicate detection into operational rulemaking environments, in conclusion.

Deng and Rafiei (2006) have proposed establishing a simple algorithm known as Stable Bloom Filter (SBF), which is based on the following idea: Given that there was no way to store the whole history of the stream, the stale information is removed by SBF to provide space for those more recent elements. They systematically identified some properties of SBF and consequently illustrated a guaranteed tight upper bound of false positive rates. The authors conducted experiments to compare the SBF with the alternative methods. If a fixed small space and an acceptable false positive rate were given, the outcome illustrated that their method was superior in terms of both accuracy and time efficiency.

Deng and Rafiei (2006) have proposed an efficient and elegant probabilistic algorithm to approximate the number of near-duplicate pairs are created. The algorithm scans the input data set once and uses only small constant space, independent of the number of objects in the data set, to provide a provably accurate estimate with high probability. They performed a
theoretical analysis and the experimental evaluation on real and synthetic data. They illustrated that in reasonably small dimensionality, the algorithm significantly outperforms the alternative random-sampling method.

Huffman et al (2007) have proposed a problem of duplicate detection as a part of such evaluation was added. Their results illustrate that the combination of multiple text-based signals and its computation over both fetched and rendered bodies significantly improve the accuracy of duplicate detection algorithms. They deemed that by
1. Detecting and removing boilerplate from document bodies
2. More fine-grained feature selection
3. Using more sophisticated URL-matching signals, and
4. Training over large data sets, the quality of the model can be improved additionally.

Gong et al (2008) have proposed the SimFinder, an effective and efficient algorithm to identify all near duplicates in large-scale short text databases. The three techniques, namely, the ad hoc term weighting technique, the discriminative-term selection technique, the optimization technique are included in this SimFinder algorithm. It was illustrated that the SimFinder was an effective solution for short text duplicate detection with almost linear time and storage complexity by the experiments conducted.

Xiao et al (2008) have proposed an efficient similarity joins algorithms that exploit the ordering of tokens in the records. Various applications such as duplicate web page detection on the web have been provided with efficient solutions. They illustrate that the existing prefix Filtering technique and the positional filtering, suffix filtering are complementary to each other. The problem of quadratic growth of candidate pairs when the data grows in size was effectively solved. They evaluated their algorithms on several real datasets under a wide range of parameter settings and proved to be superior when compared to the existing prefix filtering based algorithms. In order to improve the result quality or accelerate the execution speed, their method can additionally be modified or incorporated with existing near duplicate web page detection methods.

Kanhaiya Lal and Mahanti (2010) have proposed a novel algorithm, Dust Buster, for uncovering DUcorrespond trials with similar Text, Dust Buster, for uncovering DUcorrespond trials with similar Text. They intended to discover rules that transform a given URL to others that are likely to have similar content. Dust Buster employs previous crawl logs or web server logs instead of probing the page contents to mine the dust efficiently. It is
necessary to fetch few actual web pages to verify the rules via sampling. Search engines can increase the effectiveness of crawling, reduce indexing overhead, and improve the call-in the form of popularity statistics such as Page Rank, which are the benefits provided by the information about the DUST.

Bassma (2011) have proposed a method that can eliminate near duplicate documents from a collection of hundreds of millions of documents by computing independently for each document a vector of features less than 50 bytes long and comparing only the vectors rather than entire documents are provided that m is the size of the collection, the entire processing takes time O (mlogm). The algorithm illustrated, has been successfully implemented and isemployed in the context of the AltaVista search engine, currently.

V. CONCLUSION

Authors are used many machine learning and web base techniques are used for extracting data from web pages and identifying the blocks. Which are contain the content, blocks are analyzed identify the useful data and noisy data. But there is still questioning on efficiency and quality of web data. All web data are incorporate with the search engines, so we need to analyze and examine our data with respect to search engine optimization techniques. a special handling of duplicates and a way to reduce a frequent source of false alarms-template similarity is provided. Internet Search

DOI: 10.15680/IJIRCCE.2017. 0505170

Engines are posed with challenges owing to the growth of the Internet that flood more copies of web documents over search results making them less relevant to users suggested a method of "descriptive words" for definition of near-duplicates of documents, which is because of the choice of N words from the index to determine a “signature" of a document. Any search engine based on the inverted index can apply this method. The method based on "shingles" and the authors compared the suggested method. At almost equal accuracy of algorithms, their method in the presence of inverted index was more efficient. The need for various forms of duplicate document detection has increased due to the accelerated growth of massive electronic data environments, both web-based and proprietary. This detection can take any of several forms based on the nature of the domain and its customary search paradigms; however either identical or non-identical deducing can be used to characterize basically them.

REFERENCES

[1]. Hassan F. Eldirdiery, A. H. Ahmed,” Detecting and Removing Noisy Data on Web document using Text Density Approach “,International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 5, February 2015.
[2]. Ms. Shalaka B. Patil, Prof. Rushali A. Deshmukh,” , Enhancing Content Extraction from Multiple Web Pages by Noise Reduction”, International Journal of Scientific & Engineering Research, Volume 6, Issue 7, July-2015 ISSN 2229-5518
[3]. Rajni Sharma, Max Bhatia,” Eliminating the Noise from Web Pages using Page Replacement Algorithm, International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3066-3068.
[4]. Hui Xiong, Member, IEEE, Gaurav Pandey, Michael Steinbach, Member, IEEE,and Vipin Kumar, Fellow, IEEE,” Enhancing Data Analysis with Noise Removal”, IEEE Transactions On Knowledge And Data Engineering.
[5]. Rekha Garhwal,” Improving Privacy in Web Mining by eliminating Noisy data & Sessionization”, International Journal of Latest Trends in Engineering and Technology (IJLTET).
[6].Erdinc,uzun,et.al,” A hybrid approach for extracting informative content from web pages”, Information Processing and Management: an International Journal archive Volume 49 Issue 4, July, 2013,Pages 928-944.
[7].Shine N. Das,et.al.” Eliminating Noisy Information in Web Pages using featured DOM tree, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 2– No.2, May 2012. [8]. Xin Qi and JianPeng Sun, Eliminating Noisy Information in Webpage through Heuristic Rules, 2011 International Conference on Computer Science and Information Technology (ICCSIT 2011*)*
[9]. Thanda Htwe,” Cleaning Various Noise Patterns in Web Pages for Web Data Extraction”, Iinternational Journal of Network and Mobile Technologies Issn 1832-6758 Electronic Version Vol 1 / Issue 2 / November 2010.
[10]. Byeong Ho Kang and Yang Sok Kim,” Noise Elimination from the Web Documents by Using URL paths and Information Redundancy**”.**
[11]. Thanda Htwe, Nan Saing Moon Kham,” Extracting Data Region in Web Page by Removing Noise using DOM and Neural Network”, 2011 3rd International Conference on Information and Financial Engineering.
[12] .Tieli Sun, Zhiying Li, Yanji Liu, Zhenghong Liu, “Algorithm Research for the Noise of Information Extraction Based Vision and DOM Tree”, International Symposium on Intelligent Ubiquitous Computing and Education, pp 81-84, May 2009.
[13].Jinbeom Kang, Joongmin Choi, “Block classification of a web page by using a combination of multiple classifiers”, Fourth International Conference on Networked Computing and Advanced Information Management, pp 290 -295, September 2008.
[14]. Lan Yi,” Eliminating Noisy Information in Web Pages for Data Mining”.2003.
[15]. Bassma, S., Alsulami, Maysoon, F., Abulkhair and Fathy E. Eassa, “Near Duplicate Document Detection Survey”, International Journal of Computer Science & Communication Networks, Vol. 2, No. 2, pp. 147-151, 2011.
[16].Huda Yasin and Mohsin Mohammad Yasin, “Automated Multiple Related Documents Summarization via Jaccard’s Coefficient”, International Journal of Computer Applications, Vol. 13, No. 3, pp. 12-15, 2011.
[17]. Syed Mudhasir, Y., Deepika, J., Sendhilkumar, S., Mahalakshmi, G. S, “Near- Duplicates Detection and Elimination Based on Web Provenance for Effective Web Search”, (IJIDCS) International Journal on Internet and Distributed Computing Systems, Vol. 1, No. 1-5, 2011.
[18]. Kanhaiya Lal & N.C.Mahanti ,”A Novel Data Mining Algorithm for Semantic Web Based Data Cloud” International Journal of Computer Science and Security (IJCSS), Volume (4): Issue (2),Pg.160-175,2010.
[19].Ranjna Gupta, Neelam Duhan, Sharma, A. K. and Neha Aggarwal, “Query Based Duplicate Data Detection on WWW”, International Journal on Computer Science and Engineering Vol. 02, No. 04, pp. 1395-1400, 2010.
[20] .Alpuente, M. and Romero, D. “A Tool for Computing the Visual imilarity of Web pages”, Applications and the Internet (SAINT), pp. 45-51, 2010.
[21].PrasannaKumar, J. and Govindarajulu, P. “Duplicate and Near Duplicate Documents Detection: A Review”, European Journal of Scientific Research, Vol. 32 No. 4, pp. 514-527, 2009.
[22]. Poonkuzhali, G., Thiagarajan, G. and Sarukesi, K. “Elimination of Redundant Links in Web Pages - Mathematical Approach”, World Academy of Science, Engineering and Technology, No. 52, p. 562,2009.