IJCSI International Journal of Computer Science Issues, Volume 13, Issue 6, November 2016
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201606.6875

Graph Theoretic and Genetic Algorithm-Based Model for Web Content Mining

Moses Akinjide Adelola[1], Olumide Sunday Adewale[2], Gabriel Babatunde Iwasokun[2]

[1]ICTA Centre, Adekunle Ajasin University, Akungba,
Akungba, Ondo State, Nigeria

[2]Department of Computer Science, Federal University of Technology,
Akure, Ondo State, Nigeria

Abstract

The World Wide Web (www) is arguably the largest and the most heterogeneous repository of data and has continued to expand in size and complexity. With consistency in expansion, retrieval of required web pages and information has become a herculean task for web users due to information overload and worst still, existing web content retrieval techniques have not exhibited enough efficiency in areas of speed and accuracy. This paper presents a Graph Theoretic (GT) and Genetic Algorithm (GA)-based technique for mining of web documents. The technique utilizes graph representations of document content to address the problems of initialization, convergence to local minimal and failure to handle large datasets. The technique works in three phases; namely contents extraction, preprocessing and database formulation while Maximum Common Sub-graph (MCS) was used to calculate the distance between clusters. Results of the web-based experimental study on Pentium 4 with 2GHz processor and 1GB RAM running on Window 7 operating system platform with web scraper (import.io) as front-end and PHP 6 and MySQL5 as back-ends show the applicability and the superiority of the new techniques over some existing ones.

Keywords: *Web mining, graph theory, genetic algorithm, knowledge discovery*

1. Introduction

Internet is a global computer network that has played major roles in present day information dissemination and knowledge discovery. It provides varieties of information and communication facilities with interconnected networks that are based on standardized communication protocol. With its large and dynamic information source that is structurally complex and ever growing, the Internet offers a fertile ground for data (web) mining principles. Web mining generally involves the discovery of user access patterns and useful knowledge from the structure of hyperlinks on web usage logs and mining, extraction and integration of useful data, information and knowledge from web page contents [1]. Due to its heterogeneity and lack of structure, the challenges confronting web data mining include inaccuracy of the "mined document" and the time taken to mine useful information. These challenges prompted the need for improved way to accessing web documents and shift in paradigm from demand of information to demand for knowledge. This also explains why the World Wide Web is gradually transforming into semantic web, which is knowledge oriented and when subjected to mining, leads to discovery of end users-oriented knowledge.

The web mining field encompasses a wide array of issues that centered on deriving actionable knowledge from the web as well as information retrieval (IR), database technologies and artificial intelligence-based researches [2]. Information retrieval (IR) is the process of obtaining relevant information from a collection of resources. It may rely on searches based on metadata or full-text (or other content-based) indexing and can be seen as a standard optimization problem [3]. IR has search space S represented by the set of documents, a set of possible solutions S+ (that is, the possible documents related to the user query), and evaluation function, *f* for the investigation of the relevance of each of the possible documents related to the user query. Finally, a search engine tries to output documents that maximize *f*. The optimal solution is a document or set of documents that have the maximum score returned by function *f*. Such an optimization problem can be solved efficiently using genetic algorithm, which offers a probabilistic model that borrows from the process of natural selection in living organisms. Moreover, genetic search algorithms enable intelligent and efficient Internet searches and are especially useful when the search space is relatively large, as is the case with the web [4].

Web content

Today, there are information on business transactions, scientific data, medical and personal data, satellite sensing and pictures, text reports, surveillance video and pictures, games, digital media, Computer Aided Design (CAD), software engineering data, Virtual Worlds (VW), text reports and memos (e-mail messages), www repositories and military intelligence on the web [5-7]. Based on these, web contents (data) appear in form of text, images, sounds (audio), videos, animations, applications, e-services, personal web pages, archived e-mail messages and more [8]. Generally, the more the contents of a website are

dynamic, useful and interesting, the more is their patronage and success.

Data Mining

Data mining is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases and its basic steps in an iterative knowledge discovery process is presented in Figure 1 [9]. Data mining is applicable to any kind of information repository with different data, algorithms and approaches and is being put into use and studied for relational, object-relational, object-oriented and transactional databases as well as unstructured and semi-structured repositories such as the www and spatial, multimedia, time-series and textual databases [10-11].

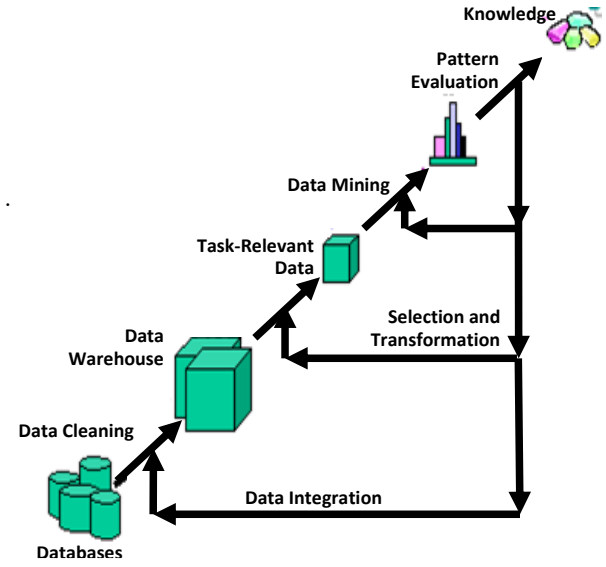

Figure 1: Data mining as the core of knowledge discovery process

The iterative process of data mining consists of data cleaning (or cleansing), data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation [12]. Data mining functionalities include data characterization, association analysis, classification, prediction, clustering, outlier analysis and evolution and deviation analysis [13-14]. Existing data mining tools include some traditional data mining programs that are available in both Windows and UNIX operating systems, dashboards and text-mining tools [15]. Data mining techniques include artificial neural networks, decision trees, the nearest-neighbor method, clustering, classification, association, sequential patterns and predictions among others [16].

Data Clustering (DC)

With DC, logically similar information is physically stored together for minimal disk access and increased database efficiency. The available DC methods are Probabilistic Clustering (PC), Hierarchical Clustering (HC), Text-based Clustering (TC), Graph-based Clustering (GC), Neural Network-based Clustering (NNC), Fuzzy Clustering (FC) and Partitioning [5, 16-17]. The choice of a particular method depends on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

Web Mining

Web mining is the application of data mining techniques on the web data to solve the problem associated with extracting useful information [13, 17-18]. Most of the existing search engines lack the efficiency of providing relevant and required information as the information on the Internet increases [19-21]. The huge, diverse and dynamic nature of the Web has resulted in information overload and heightened the need for an intelligent software agent for finding, sorting and filtering the available information [6, 22-23]. Web mining can be decomposed into resource finding, information selection and preprocessing, generalization and analysis [18]. Resource finding means the process of retrieving the data that is either online or offline from the web sources like text, relational data and semi structural data like XML. The information selection and preprocessing step is any kind of transformation processes of the original data retrieved in the IR process. Machine learning or data mining techniques are used for generalization [17]. Based on which part of the Web to mine, web mining is divided into three areas of interest; namely Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) as shown in Figure 2.

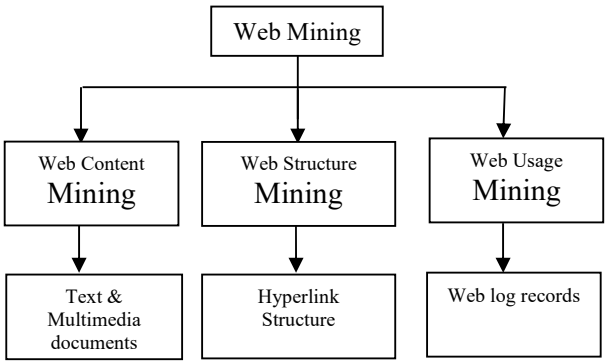

Fig. 2: Web Mining Divisions

Web Data Extraction Technique (WDET)

A WDET provides a broad class of software applications targeting at extracting information from the web. If the source is an HTML Web page, the extracted information could consist of elements in the page as well as the full-text of the page itself [24]. The extracted data might be post-processed, converted in the most convenient structured format and stored for further usage [25-29]. WDETs find extensive use in a wide range of applications including the analysis of text based documents like e-mails, support forums, technical and legal documentation and so on and its importance depends on the fact that a large (and steadily growing) amount of information is continuously produced, shared and consumed online. A company can probe the Web to acquire and analyze information about the activity of its competitors, identify the opportunities provided by the market and learn from faults and successes [30].

Graph Theory

Graphs are mathematical constructs for representing objects or systems which contain structural (or relationship) information and have been used in many domains, from software engineering to artificial intelligence. Graphs are more robust than typical vector representations as they can model structural information that is usually lost when converting the original web document content to a vector representation. With graph, information such as the location, order of proximity or term of occurrence which is discarded under the standard document vector representation models is captured. A graph G is defined as $G=(V,E)$ where V is a set of nodes (also called vertices), and E the set of edges connecting the nodes. It is also defined as a 4-tuple: $G= (V, E, \alpha, \beta)$, where V is a set of nodes (vertices), $E \subseteq V \times V$ is a set of edges connecting the nodes, α: $V \rightarrow \sum v$ is a function labeling the nodes, and β: $V \times V \rightarrow \sum e$ is a function labeling the edges. $\sum v$ *and* $\sum e$ are the sets of labels that can appear on the nodes and edges, respectively. For brevity, G may be referred to as $G= (V, E)$ by omitting the labeling functions [31-33].

2. Literature Review

Every Internet user desires satisfactory results from web search engine in the sense that all the retrieved results are relevant and all relevant documents are retrieved. This implies that the web user is mostly satisfied when the information retrieval system retrieves all and only the relevant documents within a reasonable response time. Despite significant improvement on the existing search techniques, web users still encounter some problems relating to the retrieval of irrelevant documents from the web. Most of the early researches on web content mining focused on the retrieval of textual contents with the traditional text representation method and vector space model presenting several noticeable weak points of inability to capture text structure and the semantic information of text content.

The authors in [34] developed a web content mining system using a graph-based representation technique and the k-means-based clustering algorithm to overcome the problem of vector space model. Although, the work presented some good results, its solution often converges to local minima. In [33], graph-theoretic method formed the platform for the analysis of protein structural information on the basis that sub-graph and maximum common sub-graph isomorphism algorithms from graph theory provide effective and efficient way of identifying structural relationships between objects. Although the platform justified extension of graph theory to other areas of machine learning base on its sufficient identification of the sequence relationships between biological macromolecules, it is however only restricted to graph theory-based analysis for which three-dimensional crystallographic or nuclear magnetic resonance (NMR) structures are available. The need for a system with improved and optimal representation of web contents motivated the work of the authors in [31]. A hybrid web document representation methodology based on vector space and graph models was presented for classifying web document contents. The calculation of the classification rate for each candidate sub-graph is however too complicated for the methodology and the ensued solution is often not optimal.

The need to track and analyze the usage patterns of web users in the deep web motivated the research in [35]. The research identified enormous effort, time wasting and avoidable extra monetary cost as part of the challenges facing the Internet users in their bid to extract useful knowledge from the web. Extensible Markup Language (XML) format of web pages and XML document object model were used for information extraction. Correlation mining approach was used for finding the correlation attributes in query interfaces while Jaccard measurement technique was used to measure the degree of similarity between items. Though the proposed algorithm ably found correlated attributes in query interfaces with greater accuracy and speed compared to some existing algorithms, it however miss out with its trial by error method, due to absence of standard or general way of finding the most suitable measurement. Also, as a convectional vector representation technique, vital information such as proximity of word occurrence and or the location of a word within a document were not captured thereby leading to loss of information.

The research work presented in [14] revealed how increased redundancy and duplication in web pages result in indexing space and retrieval and removal time complexity. This prompted the authors to develop a web outlier mining system that offers speedy and accurate retrieval of information from structured and unstructured web documents. Obtained results presented some vital structural information such as the order and proximity of terms occurrence and the location of word within a document as not captured by the system. The research in [36] was motivated by the need to facilitate knowledge-based response to the user and also to discover hidden patterns from the web. An agent based system was developed as a solution for mining semantic web contents and to provide context based knowledge oriented results to the user. The shortfall of the work is that in the process of discovering hidden patterns from the web, some unrequested information are retrieved, which may lead to information overload and time wastage. A composite graph model and maximum common sub-graph-based technique for web document extraction is proposed in [37]. The research focused on the development of a standard method for representing web documents as graphs and graph-based classification of web documents. Tag and context sensitive graph models on extracted web pages samples were used alongside graph distance computations for similarity comparison and measurements.

In [38], an optimal graph theoretic approach to data clustering was presented. The research used network flow theory as data clustering technique and graph theoretic approach to image segmentation with a view to handle maximum flows computation in an undirected graph. The proposed system suffers in its inability to handle moving images. A generalized graph-theoretic mesh optimization model is proposed in [39]. Direct derivation of mesh optimization model, primitives and multi-pole components method and algebraic multi-grid principles based on coarsening technique were used. With this model, heuristics assumptions of analogy of FEM stiffness and cutest matrices made in the graph theoretic method which required rigorous theoretical validation were used. The stability of the coarse/fine interpolation between meshes in

Garlekin operator and the limits of the AMG-type coarsening can as well not be predicted.

3. Proposed GT AND GA-Based Technique

The proposed system combines the strength of existing works proposed in [34, 38-39]. Its conceptualization is presented in Figure 3 with three phases; namely contents extraction, preprocessing and database of mined data. Based on the user query, the content extraction phase uses the search engine to extract and store raw web pages from the Internet. The second phase uses graphic theoretic and genetic algorithms to process the extracted web document base on tag separation, stop word removal and stemming.

Graph Theoretic and Genetic Algorithms

In order to design the graph theoretic model for web content mining, the graph similarity and the frequent sub graph extraction algorithms were used. The input of the algorithm is a set of training examples; a query instance q, and a parameter k, defining the number of neighbors to use. The output is labeled to indicate the class of the query instance q. First, determine the closest training examples to q according to the distance. The second step is to select the nodes that are closest to the center. In order to get the distance between the clusters, the maximum common sub-graph distance measure was used. The data for the nearest neighbours was determined using the method of graph techniques to model web documents.

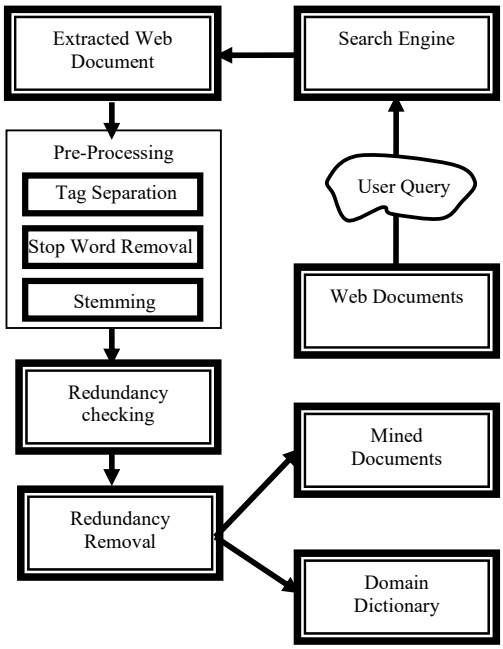

Figure 3: Proposed System Architecture

The distance is measured using Equation 1.

$$dist_{mcs}(G_1, G_2) = \frac{|\sum d\pm(mcs(G_1,G_2))|}{Max(\sum d(|G_1|)),Max(\sum d(|G_2|))} \quad (1)$$

mcs is the minimum common sub-graph, G_1 is graph 1 and G_2 is graph 2. The measured distance is based on mcs and to make use of the information held by the composite model, all measurements may be modified to incorporate the vertex degree instead of simply taking the number of nodes. The parameter for Equation 1 is derived from Equations (2) through (6)

$$d_{mcs}(G_1, G_2) = 1 - \frac{|mcs(G_1,G_2)|}{Max(|G_1|,|G_2|)} \quad (2)$$

The graph union-based distance measurement algorithm proposed in [40] is adopted as follows:

$$d_{WGU}(G_1, G_2) = 1 - \frac{|mcs(G_1,G_2)|}{|G_1|+|G_2|-|mcs(G_1,G_2)|} \quad (3)$$

With graph union, the denominator represents the size of the union of the two graphs in the set theoretic sense; specifically, adding the size of each graph ($|G_1| + |G_2|$) then subtracting the size of their intersection ($|mcs(G_1,G_2)|$) leads to the size of the union. This distance measure behaves similarly to *mcs*. The use of graph union in the denominator is premised on the need for allowing for changes in the smaller graph to exert some influence over the distance measure, which does not happen with *mcs*. This measurement is a metric created in the interval [0,1]. A similar distance measure not normalized to the interval [0, 1] is presented as follows [41]:

$$d_{UGU}(G_1, G_2) = |G_1| + |G_2| - 2|mcs(G_1, G_2) \quad (4)$$

The distance based on the maximum common sub-graph and the minimum common super-graph is measured based on the formula [42]:

$$d_{mmcs}(G_1, G_2) = |G_1| + |G_2| - |MCS(G_1, G_2) \quad (5)$$

This distance measure is informed by the concept that the maximum common sub-graph provides a "lower bound" on the similarity of two graphs, while the minimum common super-graph is an "upper bound". If two graphs are identical, then both their maximum common sub-graph and minimum common super-graph are the same as the original graphs and $|G_1| = |G_2| = |mcs(G_1,G_2)| = |mcs(G_1,G_2)|$, which leads to $d_{mmcs}(G_1,G_2) = 0$. As the graphs become more dissimilar, the size of the maximum common sub-graph decreases, while the size of the minimum common super-graph increases. This in turn leads to increasing values of $d_{mmcs}(G_1,G_2)$. For two graphs with no maximum common sub-graph, the distance becomes $|mcs(G_1,G_2)| = (|G_1| + |G_2|)$. *mmcs* is a metric that does not produce values normalized to the interval [0,1], unlike the *mcs* or *wgu*. If it holds that $|mcs(G_1,G_2)| = |G_1| + |G_2| - |mcs(G_1,G_2)| \; G_1, G_2$, then $d_{MMCS}(G_1,G_2) = |G_1| + |G_2| - 2|mcs(G_1,G_2)|$ is computed. This is much less computationally intensive than computing the minimum common super-graph. A version of this distance measure normalized to [0, 1] is created as follows:

$$d_{mmcsn}(G_1, G_2) = 1 - \frac{|MCS(G_1,G_2)|}{MCS(|G_1|,|G_2|)} \quad (6)$$

With this algorithm, several parameters are defined to control the properties of resulting cluster hierarchy. The Maximum Terms Threshold (MTT) is the first parameter and it restricts the maximum number of vertices in the resulting graph representations of documents. There are two options with the first option as the default choice and involves the use of thirty of the most frequent terms in the pages while the second option is to use the maximum

threshold terms that are most occurring. The second parameter is the Minimum Pages Threshold (MPT), which is used to eliminate the clusters having fewer pages than the maximum point assigned to them using a default value of 3. The other parameter is the Maximum Distance Threshold (MDT), which is used to restrict the expansion of the hierarchy. Clusters whose difference in size from their parents is greater than MDT are not added to the hierarchy [43] and the default distance is two, which has the capability to accept one new word to an existing phrase that is one node and one side.

The Maximum Cluster Threshold (MCT) is another parameter and it limits the total size of the hierarchy. The hierarchy construction phase of the algorithm is stopped once it has created MCT clusters or there are no candidate graphs remaining. The default value is 50. Finally, there is the Base Cluster Size Threshold (BCST) which limits the size of the base clusters. No new base cluster is created if its size is more than the BCST. The default BCST is three and is usually large enough to allow a phrase with two terms, which are nodes connected by an edge as a base cluster.

This algorithm has three basic steps; namely original hierarchy construction, document assignment and bottom-up clustering pruning. The first step is the initial hierarchy construction. The candidate chart having the minimum size for a graph G, $|G|$ is the sum of the quantity of edges, E and the vertices, V and are shown by $|V|+|E|$ with this cluster taken as the cluster candidate. If there is a tie, one of the graphs is selected at random. The likely parents of a candidate in the hierarchy are then identified, so that any parent that is defined here is the lowest. If the candidate of the cluster has no parents cluster and if the size of the graph is not greater than or equal to the BCST, then the cluster should be added. If it is otherwise, then the cluster candidate is added to the hierarchy. The cluster candidate is then removed from the set of candidate graph. After this, if the clusters are less than the MCT and a candidate graph is still remaining; proceed to the initial original assignment phase. As the nodes are added, the cluster in the hierarchy that has the smallest distance is determined according to the *mcs* distance measure as follows:

$$d_{MCS}(G_1, G_2) = \frac{|MCS(G_1,G_2)|}{Max(|G_1|,|G_2|)} \quad (7)$$

If there is 1 minimum distance, then this page has to be skipped. The pages are assigned to the clusters that have a minimum distance as a native page. The inherited pages continue to give up the cluster from the child to the parent until the base cluster is reached. In the next phase, starting with the lowest level in the hierarchy, all clusters at that level are eliminated from the level that has fewer pages than the MPT pages assigned to them. Given the new hierarchy, all the pages are re-aligned from the deleted clusters. The orphaned clusters are fixed by updating the parent information. This is repeated until the top level is reached. For each cluster, the longest simple paths are displayed first. Isolated nodes are displayed as single terms. If a cluster happens not to be a base one, then those on the hierarchy are shown.

The basic step is an additional determination of the cluster centers. This is a good way to determine the initial centers for the clusters in the algorithm. Starting at the case of one cluster, the center of a cluster is set as the centroid of the whole data set. For the k-cluster, the centers are derived by obtaining the centers of that cluster and then determine the location of the new centers of the cluster which are the most optimum. To achieve this, the data items are considered each at a time. The algorithm is then applied on these data items, and the results examined. The new centers are then determined and used. The determination of the new centers is based on the formula:

$$E(m_1 \dots m_M) = \sum\sum_{i=1}^{N} I(x_i \in ck)||x_i - m_k||^2 \quad (8)$$

N is the number of data items, M is the number of clusters, x_i is the data item i and m_k is the cluster center k. I(x) = 1 when x is true and 0 otherwise. For this reason, a faster way to calculate the centers is stated as follows:

$$b_n = \sum_{j=1}^{N} \max(dj_{k-1}||x_n - x_j||^2, 0) \quad (9)$$

dj_{k-1} is the distance between data item x_j and the center of the cluster that is closest when using this clustering algorithm. A new cluster is then selected to be the input for the algorithm. Before the genetic algorithm is applied on a web page, there has to be some pre-processing to be done. This involves searching all the pages that contain the terms in the query of the user. A weighted web tool is applied and the genetic algorithm is designed and iterated to come up with the optimal solution.

When a web user enters a list of key words, then a new document is created containing the documents with the key words and its frequency in the search. Through the chromosome of a gene, each document is allocated a reference number. Chromosome is made up of sets of genes, g_i such that $g_i \in [d_{1\dots\dots}d_n]$. The number of genes in the chromosome does not, at any one point, exceed the maximum number of documents that are selected at random. The chromosomes that are produced have varying lengths. The solution generated is mostly between five and a specified maximum number. The first chromosome, $g_{1,}$ is then generated and $S=S\cup g_1$. The loop is then iterated until the chromosomes generated are equal to the length requested [44]. The first generation of chromosomes is randomly generated and has a fixed number. In the subsequent generations, two individuals are selected randomly, and the one with a higher fitness value is selected. In this algorithm, the crossover is used to bring forth offspring from the existing population. It is mostly operated with a probability of 0.8. Two randomly selected parents of different lengths were selected, and a point of the crossover is also selected randomly. The first point to be selected is based on the length of the parent and the other point is also selected in accordance to the length of the parent. The first offspring is obtained by switching the tail of the second parent from the first selected point [45]. The second offspring is obtained from the first parent from the second point. Any duplicate gene is removed.

4. System Implementation

The implementation environment is a **w**indow 7 operating system on a 2.4GHz Core i3 processor with 4GB RAM. Import.io (web scraper) served as the frontend while PHP 6 and MySQL5 served as the backend. Web Scraping (also termed Screen Scraping, Web Data Extraction or Web

Harvesting) is very useful in the extraction of large amounts of data from websites tabular or spreadsheet format. Data displayed by most websites can only be viewed using a web browser. Examples are data listings at yellow pages directories, real estate sites, social networks, industrial inventory, online shopping sites, contact databases and so on. Since most websites do not offer functionality for saving data and displaying on computer, Web Scraping is therefore used to automate the tedious and slow process of manually copying and pasting website (browser) data to a local file.

Experimental Setup

A total of 2100 web pages were downloaded with various web data types. The downloaded pages were divided into training and testing pages using the structured query language (SQL) statements shown in Figure 4. The web data extraction is based on the use of a web scraping software *import.io* on selected websites and the extracted web page is converted to comma separated values (csv) format which serves as input to the next phase. At the data preprocessing phase, the data obtained from different sources such as HTML documents, browser logs are cleaned before processing and grouping in accordance to physical location. The input data to the preprocessing algorithm consists of the web pages accessed in a session by a web user. The data is in a tree-like format and listed in order of access.

```
SQL statement for Training dataset
SELECT v1.* FROM ( -- randomly divide members of the population into subgroups based on target classes
SELECT a.*, row_number() OVER (partition by {target column} ORDER BY ORA_HASH({case id column})) "_partition_caseid" FROM {input data} a
) v1, ( -- get the count of subgroups based on target classes
SELECT {target column},
COUNT(*) "_partition_target_cnt"
FROM {input data} GROUP BY {target column} ) v2
WHERE v1. {target column} = v2. {target column}
-- random sample subgroups based on target classes in respect to the sample size
AND ORA_HASH(v1."_partition_caseid", v2."_partition_target_cnt"-1, 0) <= (v2."_partition_target_cnt" * {percent of training dataset} / 100)

SQL statement for Test dataset
SELECT v1.* FROM ( -- randomly divide members of the population into subgroups based on target classes
SELECT a.*, row_number() OVER (partition by {target column} ORDER BY ORA_HASH({case id column})) "_partition_caseid" FROM {input data} a
) v1, ( -- get the count of subgroups based on target classes
SELECT {target column},
COUNT(*) "_partition_target_cnt"
FROM {input data} GROUP BY {target column}
) v2
WHERE v1. {target column} = v2. {target column}
-- random sample subgroups based on target classes in respect to the sample size
AND ORA_HASH(v1."_partition_caseid", v2."_partition_target_cnt"-1, 0) > (v2."_partition_target_cnt" * {percent of training dataset} / 100)
```

Figure 4: structured query language (SQL)

The fitness level of the data is then calculated from each user based on the sum of three factors. The first factor, f_1, is the ratio of the frequency of the data of a given web user when compared to all the other users in the database and is obtained from the formula:

$$f_1 = \frac{f(l+r)}{d} \quad (10)$$

l is the left side pages, r is the right side pages while d is the total dataset in the database.

The next factor, f_2 is described by the ratio of the frequency of the left and the right side pages and the pages on the left of the same user and is obtained from:

$$f_2 = \frac{f_1}{l} \quad (11)$$

Finally, the third factor, f_3 is the ratio between the pages on the left side of a particular user and a page from the left side together with the first page, p on the right and is obtained from the equation:

$$f_3 = \frac{f(l)}{d(l+\mathrm{p})} \quad (12)$$

The time duration, T for the computation of f_1, f_2 *and* f_3 is obtained from:

$$\mathrm{T} = \sum_{i=1}^{l_u} \mathrm{N_u} \quad (13)$$

l_u *and* $\mathrm{N_u}$ are the left side pages and the time for user u respectively. If obtained duration exceeds the threshold, then the cross over process takes place and the total quality, $\mathrm{Q_t}$ is calculated based on the formula:

$$\mathrm{Q_t} = \sum_{i=1}^{l} N_i^t \quad (14)$$

N_i^t represents the node in a tree i and l is the total number of nodes in a tree. The crossover stage is a process of interchanging the sequence of nodes in a particular tree. This process deals with those trees that have the right fitness values above the threshold and the selected nodes move on to the mutation process. Mutation is an iterative process for node transformation and selection of chromosomes with the best fitness levels.

Results

The extracted data is tested using three standard measures; namely precision (p), applicability (a) and hit ratio (h), which are measured based on Equations 15, 16 and 17 respectively:

$$P = \mathrm{xy}^{-1} \quad (15)$$
$$a = \mathrm{yz}^{-1} \quad (16)$$
$$h = ap \quad (17)$$

The initial population for the real and generated data is 800 and the obtained results for GA and training samples are presented in Table 1 with superior performance for the genetic algorithm.

Table 1: Comparison of GA-generated data and training samples

Measure	T1 (GA)	T2 (Training data)
Precision	94.4225	89.025
Applicability	100.0000	87.500
Hit ratio	94.4225	89.000

For performance evaluation, three experiments were conducted with notations F, J and K series for all the available web documents that can be represented with graphs as well as truth value. The F-series originally contained 93 web pages, each of which were subdivided

into four major groups; namely manufacturing, finance, business and education. The J-Series contained 185 pages and had 10 classifications while the K-Series contained 400 web pages in 20 categories. 800 pages were randomly selected from the original 2100 pages and the number of vectors was set to 40, which doubled the number of categories in the K-Series experiment. The choice of this number of clusters is premised on the fact that it is the most natural number based on the initial tests and observations. The number of maximum nodes per graph was set to be higher to provide improved baseline for the results as shown in Table 2.

Table 2: Performance of Graphs with increasing nodes

Max. Nodes/Graph	A^M (average)
150	0.2218
120	0.2142
90	0.2074
75	0.2045
60	0.1865
45	0.1758
30	0.1617
15	0.1540
5	0.1326

Each row in Table 2 provides results for 10 experiments using the same data sample of 800. The variation in the results is due to randomization in the first stage of the algorithm. Previously obtained data were represented in the graph for better visualization and with a 2.2 GHz processor, it took 7 minutes to represent five nodes per graph. Euclidian distance, δ for point (x,y) was also determined with a view to measuring the vector distance metrics as follows:

$$\delta(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - x) + (y_i - y)^2} \quad (18)$$

x_i and y_i are the i^{th} components of the x and y vectors respectively. The cosine equivalent, β of the distance is obtained as follows:

$$\beta(x,y) = 1 - \frac{x * y}{||x||.||y||} \quad (19)$$

$*$ is the dot operator, and $||$ shows the magnitude of the vector being considered. Comparison of the results obtained from the graphs with those from other techniques is presented in Table 3.

Table 3: Comparison of Graph theoretic approach with other techniques at an instance

Method	A^M (average)
Graphs (current study)	0.222
Extended Jaccard Similarity [34]	0.184
Pearson Correlation [34]	0.178
Cosine Measure [31]	0.178
Random [31]	0.066
Euclidean [31]	0.046

Since larger graphs hold more data, the mutual information is seen to increase as the graph increases in size. The random baseline was used to provide a basis for comparison in the experiment. The Jaccard means was based on the Jaccard similarity and the cosine and Pearson measures were omitted for improved clarity. The graphical representation of experimental values for Graphs, Random, Euclidian and Jaccard methods with the same experimental conditions is shown in Figure 5. It is revealed vividly that the graph theoretic and genetic algorithm-based technique outperforms other techniques especially with increased graph nodes. In other words, as the complexity of the web contents increases, other reviewed techniques could not match up with the proposed technique in the area of mutual information index.

5. Conclusion

With standard tools for web content mining, there is opportunity for extracting only the relevant text from web while unrelated textual noise like advertisements, navigational elements, contact and copyright notes are reliably suppressed. The reported research hybridized graph theoretic and genetic algorithm to formulate a web content mining technique for achieving this purpose. The new technique provides timely search and discovery from large web datasets and experimental results had shown its superiority over other techniques. These suggest the new technique will be very useful in areas where knowledge discovery, web structure and web analytics are required. It is of note that the applicability of the new technique on complex and large number of parameters has not been investigated.

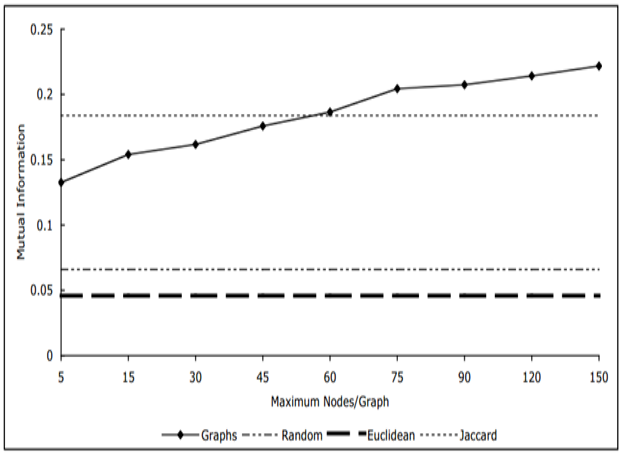

Figure 4.3: Comparison of graph with other techniques

REFERENCES

[1] Liu B., Structured data extraction: Wrapper generation and Web Data Mining, Editorial Issues on Web content Mining. SIGKDD Explorations –Vol. 6, 2005, pp. 363 - 423

[2] Abdelhakim H., Khentout C. and Djoudi M., Overview of Web Content Mining Tools, *The International Journal of Engineering and Science (IJES), Vol. 2, 2013.*

[3] Marghny M. H. and Ali A. F., Web mining based on genetic algorithm, Proceedings of AIML '05 Conference, Cairo, Egypt, 2005, pp 19-21

[4] Ammar S. A., Enhancing recall and precision of web-Search using genetic algorithm, A thesis submitted for the degree of Doctor of Philosophy, School of Information Systems Computing and Mathematics, Brunel University, UK, 2012

[5] Zaiane R. O., Introduction to Data Mining: Principles of knowledge discovery in databases, 1999.

[6] Han J. and Kamber M., Data mining: concepts and techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000

[7] Imielinski T. and Mannila H., A database perspective on knowledge discovery, Communications of ACM, Vol. 39, pp. 58-64.
[8] Rosenfeld L. and Morville P., Information architecture for the World Wide Web, 1st edition, CA, 1998
[9] Callan J., System and method for filtering a document strem", US Press, Patient 6.105.023, 2000
[10] Chen M. S., Han J. and Yu P. S., Data mining: An overview from a database perspective, IEEE Trans. Knowledge and Data Engineering, Vol. 8, 1996, pp 866-883
[11] New York Stock Exchange, 2000. Available at http://www.ecgi.org/codes/documents/nyse_cgreport_23sep 2010 en.pdf. Accessed October 12, 2013.
[12] Piatetsky-Shapiro G., Fayyad U. M. and Smyth P., From data mining to knowledge discovery, An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, pp. 1-35
[13] Sivaramakrishnan J. and Balakrishnan V., Web Mining Functions in an Academic Search Application, *Informatica Economica, Vol.* 13, No. 3, 2009.
[14] Poonkuzhali G., Sarukesi K. and Uma G. V., Web Content Outlier Mining Through Mathematical Approach and Trust Rating, Recent Researches in Applied Computer and Applied Computational Science, 2012.
[15] Silltow J., Data Mining 101: Tools and Techniques", paper presented at The Institute of Internal Auditors (IIA), 247 Maitland Avenue, Altamonte Springs, Florida U.S.A., 2006
[16] Colet E., Clustering and Classification: Data Mining Approaches". Virtual Gold Incorporated, 2002
[17] Galeas P., Web Mining, 2005. Available at: http://www.galeas.de/webmining.html, Accessed January, 2016.
[18] Cooley R., Mobasher B. and Srivastava J., Web mining: information and pattern discovery on the World Wide Web. *Proceedings of 9th IEEE International Conference,* pp. 558 – 567, 1997
[19] Arvind K. S. and Gupta P. C., Exploration of efficient methodologies for the improvement in web mining techniques - A survey, *International Journal of Research in IT & Management,* Vol. 1, No. 3, 2011
[20] Arvind K. S. and Gupta P. C., Study and Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining, *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, Vol. 1 ,No. 8, 2012
[21] Baumgartner R., Gatterbauer W. and Gottlob G., Web data extraction system: Encyclopedia of Database Systems, 2009, pp 3465-3471.
[22] Kosala R. and Blockeel H., Web mining research: A survey," SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data (SIGKDD) Mining, ACM, Vol. 2, 2000
[23] Gore M. M. and Mishra A. K., Algorithm for Data Mining. *Proceedings of Winter School on Data Mining*, Allahabad, India, 2001
[24] Ferrara E., De-Meob P., Fiumarac, G. and Baumgartnerd, R., Web Data Extraction, Applications and Techniques: A Survey, 2014. Available at http://www.sciencedirect.com/science/article/pii/S09507051 14002640, Accessed on September, 2015.
[25] Irmak U. and Suel T., Interactive wrapper generation with minimal user effort". *Proceeding of 15th International Conference on World Wide Web*, Edinburgh, Scotland, 2006, pp 553-563
[26] Wang P., Hawk W. and Tenopir C., Users' interaction with World Wide Web resources: an exploratory study using a holistic approach, Information Processing Management, 2000, pp 229 - 251,
[27] Furche T., Gottlob G., Grasso G., Gunes O., Guo X., Kravchenko A., Orsi G., Schallhart C., Sellers A. J. and Wang C., Domain-centric, intelligent, automated data extraction methodology, Companion, Vol. 10, 2012, pp 267-270.
[30] Chen H., Chau M. and Zeng D., Spider: a tool for competitive intelligence on the web, Decision Support System, Vol. 17, No. 34, 2002
[31] Marcov A. Last M. and Kandel A., Model-Based Classification of Web Documents Represented by Graphs, Proceedings of WEBKDD'06, Philadelphia, Pennsylvania, USA, 2006
[32] Cormen T. H., Leiserson C. E., Rivest R. L. and Stein C., The algorithms of Kruskal and Prim: *Introduction to Algorithms"*, 3rd edition. MIT Press, Vol. 23, No. 2, 2009, pp. 631-638.
[33] Artymiuk P. J., Spriggs R. V. and Willett P., Graph theoretic methods for the analysis of structural relationships in biological macromolecules, *Journal of the American Society for Information Science and Technology,* Volume 56, 2005, pp 518 – 528
[34] Schenker A., Last M., Bunke H., and Kandel A., Graph Theoretic Techniques for Web Content Mining, PhD Thesis, College of Engineering, University of South Florida, 2003
[35] Shoreh A. and Mohammad D. J., Deep Web Content Mining, *The Journal of World Academy of Science, Engineering and Technology, 2009,* pp. 49.
[36] Sighn A., Agent Based Framework for Semantic Web Content Mining, *International Journal of Advancements in Technology* (*IJoAT)*, Vol. 3, No. 2, 2012. Available at http://ijict.org/. Accessed September, 2014.
[37] Kaushik M. and Phukon S., A composite Graph Model for Web Document and the MCS Technique, *International Journal of Information Technology and Knowledge Management,* Vol. 4, No.1, 2012, pp. 211-215
[38] Zhenyu W. and Richard L., An Optimal Graph Theoretic Approach to data Clustering, theory and its application to Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 15, No. 11, 1993.
[39] Andrey A. M., A Generalized Graph-Theoretic Mesh Optimization Model, *Proceedings of the 26th International Meshing Roundtable, South Lake Tahoe*, 2005
[40] Wallis W. D., Shoubridge P., Kraetzl M. and Ray D., Graph distances using graph union. *Pattern Recognition Letters*, Vol. 22, No. 6, 2001, pp 701-704.
[41] Bunke H., Recent Development in graph matching. *Proceedings of 15th International Conference on Pattern Recognition,* Vol. 2, 2000, pp 117-124.
[42] Mirtha-Lina F. and Gabriel V., A graph distance metric combining maximum common subg--raph and minimum common super-graph, Pattern Recognition Letters, Vol. 22, 2001, pp 753-758.
[43] Sandhya Chaturvedi M. and Shrotriya A., Graph Theoretic Techniques for Web Content Mining. *The International Journal of Engineering and Science*, Vol. 2, No. 7, 2013, pp. 35-41.
[44] Vikrant S. and Thakur R. S., GA Based Model for Web Content Mining", *International Journal of Computer Science Issues (IJCSI)*, Vol. 10, No. 2, No 3, 2013
[45] Ammar S. A. and Shaker R., Genetic Algorithm Mining for HTML Documents, School of Information Systems Computing and Mathematics (SISCM), Brunel University, UK, 2009