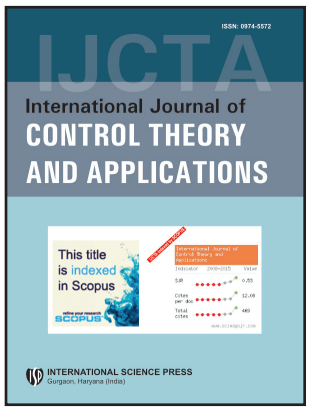

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 12 • 2017

Semantic Web Mining using Web Crawler and DOM Tree with Esvm-Modified SOM for Advanced Medical Information Retrieval System

P. Sumathi[a] and R. Manickachezian[b]

[a]*Ph.D Research Scholar, Research Department of Computer Science, NGM College, Pollachi, Coimbatore, Tamil Nadu, India*
[b]*Associate Professor, Research Department of Computer Science, NGM College, Pollachi, Coimbatore, Tamil Nadu, India*

Abstract: Information retrieval has emerged to be a significant research area. Content-Based Information Retrieval (CBIR) systems are the commonly used retrieval systems in biomedical fields though the challenge with those systems is the problem of semantic gap. Therefore in this research work, a technique is introduced for both the image and text retrieval for the biomedical query words. The search engine proposed comprises of pre-processing phase and semantic search phase for the respective processes of data collection and retrieval. The Extended WordNet is utilized rather than forming a new ontology. The Semantic web crawler is used for collecting data that are in the HTML format. These meta-data consists of both images and texts that are segmented employing DOM tree. Thereafter the images and texts are simultaneously processed. When the texts are processed and partitioned employing the k-Nearest Neighbor (kNN) algorithm, the image features are acquired employing efficient extraction methodologies and clustered making use of Particle Swarm Optimization (PSO). Thereafter, the Modified Self-Organizing Maps (MSOM) training process is carried out which is then followed by the labeling process. At last, the Enhanced Support Vector Machine (ESVM) is exploited for matching the information retrieved to the input biomedical query. The experimental results indicate that the retrieval system proposed yields a better performance for the medical information retrieval.

Keywords: Content-Based Image Retrieval, Extended WordNet, Semantic web crawler, DOM tree, *k*-Nearest Neighbor, Particle Swarm Optimization, Self-Organizing Maps, Support Vector Machine.

1. INTRODUCTION

With the evolution of the Internet, and the accessibilityto image capture devices like digital cameras, image scanners, the digital image collection size is rapidly increasing. Effective image searching, browsing and retrieval tools are needed by users from different domains, inclusive of remote sensing, fashion, crime prevention, publishing, medicine, architecture, etc. Especially in the biomedical fields, the information retrieval is of greater importance. For this, several general purpose image retrieval systems have been designed. There are two frameworks available: text-based and content-based. The origin of the text-based scheme can be tracked back to 1970s. In systems such as this, the images get manually annotated by text descriptors that a reutilized by a

database management system (DBMS) for performing image retrieval. There are two demerits with this scheme. Firstly, a significant degree of human labor is necessary for manual annotation. Secondly is the inaccuracy in the annotation owing to the subjectivity relative to human perception [1], [2]. In order to surpass the above limitations in text-based retrieval system, content-based image retrieval (CBIR) was proposed in the early 1980s. In the case of CBIR, images get indexed by means of their visual content, like color, texture, shapes [3]. A remarkable work was developed by Chang in 1984, where the author introduced a picture indexing and abstraction mechanism for the retrieval of pictorial database. The pictorial database contains picture objects and picture relations. For constructing the picture indexes, abstraction operations are designed for performing picture object clustering and classification. In the last decade, some commercial products and experimental model systems have been designed, like QBIC, Photobook, Virage, VisualSEEK, Netra, SIMPLI city [4]. In the case of CBIR technology, some potential applications can be recognized to be architecture design [5], art & craft museums [6], archaeology [7], medical imaging and geographic info system [8], [9], trademark databases, weather forecast [10], criminal investigations [11], image classification [12], [13], image search over the Internet [14] and remote sensing field for the indexing of biomedical images through contents [15].

The CBIR system depends on color, texture and shape that are low level image features. The low level features are obtained from the database images and then saved in a feature database. In a similar way, the extraction of the low level features are carried out from the query image and then the query image features are compared with that of the database image employing the distance measure. Images with the least distance having the query image are generated as the result. The chief disadvantage of the CBIR system is that the images having similar low level features might differ from the query image with respect to the semantics that are perceived by the user. This problem is referred to as the Semantic gap problem [16], [17]. Therefore, reduction of the semantic bridge between the low level image features and the high level image concepts have become a very intriguing and area of research with challenges. It is known as the Semantic Content Based Image Retrieval (SCBIR). The CBIR system has to be unaffected by the geometric transformations.

The two retrieval systems which are, content and text based retrieval systems are different in the way that the human interaction constitutes essential portion of the latter system. High level features such as keywords, text description are used by humans for measuring the similarity and image interpretation. On the other side, the low level features with semantics [18], [19] generally color, shape, texture extraction is carried out automatically employing computer vision methods. System introduced in [20] was developed for coping with audiovisual queries [21] merging the general approach to any original valued similarity measure to be embedded in the recent CBIR systems [22]. For the purpose of eliminating the semantic gap, CLUE technique is introduced for retrieving the image clusters that are coherent semantically. In other CBIR system, target images that are top matched are presented to users. After providing the image in the form of the query, target image collections are selected close to or identical to query image. These target images can then be clustered by employing NCut clustering into various semantic classes by gathering the image of same semantics into one cluster. Thereafter the image clusters is presented by the system and the similarity measure model is adjusted as per the feedback from the user [23].

Even though the CBIR and SCBIR systems carry out image retrieval with efficiency, the semantic gap problem is an important limitation. This issue is dealt with in the earlier work [24].But, the usage of PCA for dimensionality reduction can be improved while considering various images features. Therefore in this research work, a technique is demonstrated for both the image and text retrieval. The search engine proposed comprises of pre-processing phase and semantic search phase for the respective processes of data collection and retrieval process. The Extended WordNet is employed in place of making a new ontology. While the data gathered has images in addition to the texts, their segmentation is done employing the Document Object

Model (DOM) tree algorithm. After this, the images and texts are separately processed exploiting effective techniques and the images are clustered before they are trained employing MSOM training module. Next, the matching process is performed employing ESVM for a resourceful image and text retrieval for the biomedical query words. The rest of this research work is organized as below: section 2 presents the earlier research techniques that are relevant to the work proposed. Section 3 describes the proposed technique in detail. Section 4 demonstrates the performance evaluation results when the section 5 provides the conclusion about this research work.

2. LITERATURE REVIEW

SPARK [25] is basically a retrieval system which can do the query of the semantic data with SPARQL queries that are created from keywords. It comprises of three important steps: term mapping, query graph construction and query ranking. The step of term mapping attempts matching the query terms to ontological resources. In the step of query construction, the user query is evaluated and then the terms are connected along with the missing relations for getting SPARQL queries. Generally more than one query is built because of the ambiguities. Hence in the final step, the queries are analyzed with the help of a probabilistic ranking model such that the more possible SPARQL queries obtain greater rates.

Q2Semantic [26] attempts at bridging the gap between keyword queries and formal queries. The authors tackle with three issues for achieving their target, which are term matching, ranking and scalability. The first problem is dealt through the enrichment of the user queries with the terms which are extracted from Wikipedia, such that the query terms are matched easily with the entities present in the ontology. In order to resolve the second problem, a ranking mechanism is implemented which considers several factors like length of the query, the relevance and significance of ontology elements, etc. At last, for the scalability problem, a clustered graph structure is proposed in order to represent the summaries of RDF graphs. But still they require the expensive graph construction and traversal procedures at execution-time.

Sem Search [27] is similar kind of a system which produces formal queries from the keywords. It is different from the systems mentioned above in the manner it builds the queries. It employs a number of predetermined templates which are filled with different combinations of query terms. As every keyword in the query could be a class, an instance or a property, the number of combinations is actually large that impacts the query response times adversely. After the queries get generated, they are directly utilized for having a search over the semantic data. Then the ranking is carried out once the results of all the queries get retrieved.

OWLIR [28] is basically a semantic retrieval system that adapts a similar approach. It comprises of an information extraction module, an inference module and also a retrieval module. They employ the AeroText1 system for extracting the key phrases from free-text and then specify them in the form of RDF triples. This data is full with reasoning and the semantic rules for obtaining the inferred RDF triples. At last, these triples are indexed in addition to the respective free-text. The details regarding the indexing mechanism and ranking are neglected. In the part of evaluation, the improvements are shown by making a comparison between the three indices: free-text, free-text with RDF triples and free-text with inferred RDF triples,

Quiz RDF [29] integrates the keyword-based querying with RDF browsing, such that users can start a query with simpler keywords and go on with their search by browsing through the semantic knowledge base till the required information is reached. The knowledge base is constructed again through the indexing the RDF triples along with the free-text. As an indexed document has only one property which is associated with the entry, the retrieval system does not provide support to complicated queries having more than one property of the entity searched.

Squiggle [30] is a basic retrieval framework which makes use of the semantic indexing for efficiently accessing semantic knowledge. In place of making use of the entire knowledge base for the purpose of searching, it carries the indexing and then searches just those entities which are identified at the time of the semantic annotation phase. Again, the unit of indexing is triples. Dissimilar to the systems which are explained above, Squiggle maintains free-text in an index separately and then both indices are searched for the query submitted by the user. The retrieval is carried out primarily employing the free-text index and the semantic index is exploited for making suggestions to the user based on the ontological knowledge.

Johnson et. al., [31] seeks to resolve the problems of irrelevancy through the annotation of images with a graph-based semantic representation referred to as a scene graph that explicitly extracts the objects present in an image, their attributes and the association between the objects. It is plausibly argued that paragraph-long image descriptions expressed in natural language are presently too complicated to be automatically mapped to images and rather they indicate that elaborate image descriptions as scene graphs can be got through crowd sourcing. They also reveal that they can conduct semantic image retrieval over un-annotated images utilizing partial scene graphs. But, one huge limitation of their model is that it needs the user to input a query which is a scene graph in place of an image description in the natural language that has much less possibility to receive popular adoption between the prospective users.

Pourghassem & Ghassemian [32] introduced a two-level hierarchical medical image classification technique. The first level was employed for classifying the images into the merged and non-merged classes. This algorithm was validated on medical X-ray images consisting of 40 classes. Even though this is basically a two-level hierarchical classification, it varies from this approach since the merged classesonly were assessed in the second level in order to have the classification with multilayer perceptron (MLP) classifiers into 1 of 40 classes.

Mehta et. al., [33] suggested a region-specific retrieval system which is dependent on sub-image query search on whole-slide images through the extraction of scale invariant features on the points of interests detected and 80% of match was accomplished with the manual search for the case of prostate H&E images in the searches in top five. In another work, image-level retrieval of four special kinds of skin cancer [34] was carried out by building a visual word dictionary making use of a bag-of-features technique for representing a relationship between the visual patterns and the semantic concepts.

Zheng et. al., [35] demonstrated a CBIR system which is dependent on the weighted similarities of four feature kinds which are color histogram, image texture, Fourier coefficients, and wavelet coefficients. The retrieval performance of this system was verified employing the agglomerative cluster evaluation for various pathology image groups and then the best retrieval performance was examined for prostate query images.

Yang et. al., [36] presented a Web-based system referred to as Path Miner, that comprises automatic segmentation, CBIR, and classification modules for assisting diagnostics in pathology. The classification performance of their system was evaluated over five various blood cells like chronic lymphocytic leukemia, mantle cell lymphoma, follicular center cell lymphoma, and acute lymphocytic leukemia and acute myelogenous leukemia by employing SVM classifiers having text on histogram features and 87.27% of classification accuracy was accomplished over an open set with huge difference in staining characters.

3. PROPOSED METHODOLOGY

The proposed technique is focused over designing an information retrieval system for the separate retrieval of the images and the texts from the web data which is extracted. The retrieval system proposed comprises of two important phases: the pre-processing phase and the semantic search phase.

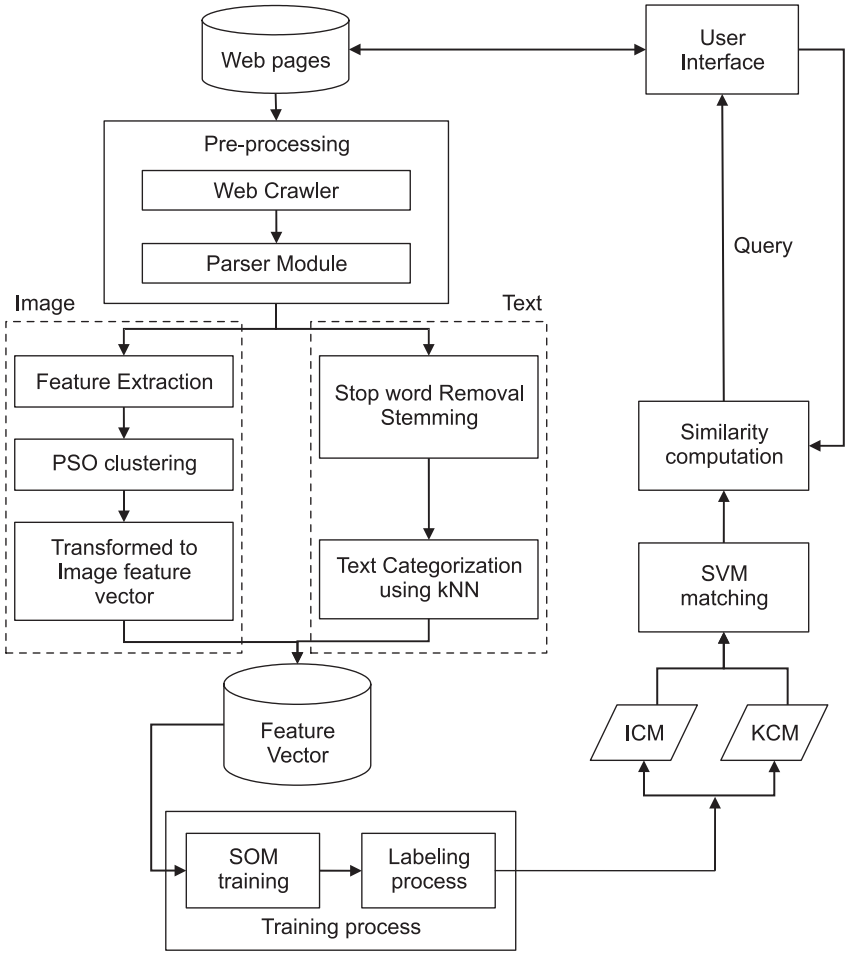

Figure 1: Proposed Information retrieval system

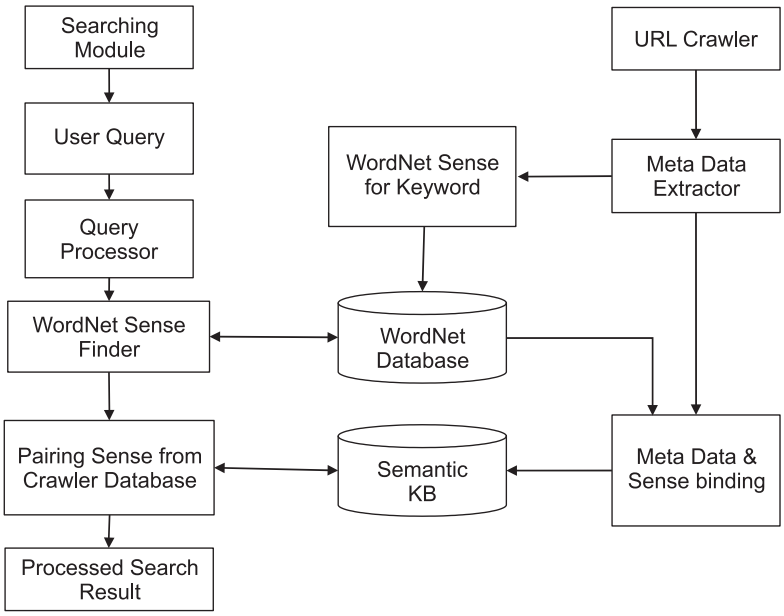

Figure 2: Proposed Web Crawler Architecture

The pre-processing phase holds the responsibility for web data collection and the extraction of the images and texts whereas the next phase does the processing of the images and texts and then does their accurate retrieval.

The URL Crawling module will be utilized for crawling the web pages for the extraction of the meta data; User inputs the URL that must be indexed, thereafter the meta data are stripped the from web page that contains keyword, description and author name (if exists) employing the Meta Data Extractor process. Description and author name are sent directly to meta data and sense binding module while the keywords get redirected to a particular process in which all the keyword sense are matched from Extended WordNet Lexical Database that information is again sent to meta data and sense binding is done there by keyword sense. Both description and author name are brought together and stored in Semantic Knowledge Base. In the searching module, the query processor considers the input query from the user where the Stop Words are elimination and Stemming is carried out for the query given and thereafter the refined query gets forwarded to WordNet Sense Finder module from where word sense is got for the refined query from the WordNet Database and then the received Sense gets directly referred to Semantic Knowledge Base (KB) by means of a relational database query that does the retrieval of most related results which are processed for displaying to the User. The procedure of stemming and stop word elimination are also carried out in this phase for the minimization of the poor indexing of the search results and improving the efficiency of the query processing.

Parser Module

Still, the web data extracted are in HTML format indicating that just the web pages get extracted employing the web crawler. Once the crawler and cleaning of the extracted web data is over, their processing is carried out for segmenting them into images and texts in separate. This is accomplished by employing a DOM tree algorithm. DOM Tree is actually a webpage segmentation algorithm which does the automatic segmentation of the web pages into sections, with every section having a web image and its contextual information (i.e. image segment), and thereafter extracts the text and images through the traversal across the DOM tree.

For the purpose of analyzing a web page for extraction of content, initially the page is passed into an HTML parser which does the HTML correction and generates a Document Object Model tree representation of the web page. After the processing is over, the resultant DOM document can be effortlessly revealed to be a webpage to the end-user like if it were HTML. This procedure achieves the steps of structural analysis and structural decomposition carried out by Rahman's [37], Buyukkokten's [38] and Kaasinen's [39]. The DOM tree is arranged hierarchically and can be evaluation in parts or completely, rendering a broad range of flexibility for the extraction algorithm. Then the content extractor does the navigation of the DOM tree in a recursive manner, employing a sequence of various filtering methodologies for the removal and modification of certain nodes and discards only the content behind. Every one of the filters could be turned on and off easily and personalized to a particular degree.

There are two categories of filters, with various degrees of granularity. The first set of filters just neglects the tags or the particular attributes within the tags. With the help of these filters, images, links, scripts, styles, and several other elements can be removed quickly from the web page. But, the second set of filters is more sophisticated and algorithmic, offering a greater level of extraction rather than provided by the conversion of HTML to WML. This set that can be extended, comprises of the advertisement remover, the link list remover, the empty table remover, and the removed link retainer. The advertisement remover exploits a resourceful methodology for the removal of advertisements. While the DOM tree is being parsed, the values of the "src" and "href" attributes across the page undergo survey with the purpose of determining the servers to which the links associate. When an address is a match against a list consisting of common advertisement servers, the node

of the DOM tree which are present in the link is eliminated. This procedure is identical to using an operating systems-level "hosts" file for preventing a computer from linking to advertiser hosts.

The link list remover uses a filtering method which eliminates all the "link lists" that are actually table cells for which the proportion of the number of links to the number of non-linked words is higher than a particular threshold (referred to as the link/text removal ratio). If the DOM parser comes across a table cell, the Link List Remover does the tallying of the number of links along with the number of non-linked words. The number of non-linked words is obtained by considering the number of letters which are not present in a link and then dividing it by the average number of characters per word. In case the ratio is bigger than the ratio of user-determined link/text removal, the table cell content (and, optionally, the cell itself) is eliminated. This algorithm attains success in the removal of most long link lists which are observed to reside across the sides of web pages when keeping the text-intensive parts of the page untouched.

The empty table remover does the removal of tables which are empty without any "substantive" information. The user decides, by means of settings, which HTML tags must be taken into consideration to be a substance and how many of the characters in a table are required to be seen as substantive. Then the table remover does a check over a table for substance once it has been parsed after passing through the filter. In case a table contains no substance, it is eliminated from the tree. This algorithm efficiency removes any tables that are leftover from earlier filters which have meagre quantities of unnecessary information and undesired images.

When the filters above eliminate non-content from the site, the removed link retainer will add the link information back at the document end to maintain the page browsing ability. The removed link retainer keeps trace of every text links which are eliminated all through the filtering process. Once the DOM tree is wholly parsed, the list of the links removed is added at the bottom of the page. In this manner, any significant navigational links which were earlier eliminated stays available. Once the whole DOM tree gets parsed and changed suitably, it can be displayed in either HTML or just as images and plain text. The plain text output eliminates all the tags and keeps solely the text of the site, when removing most of the white space. The output is an image document and text document which contains the chief content of the site in a format appropriate for summarization, or storage.

Once the processing through the parser model is completed, the extracted data is now in the form of separate images and associated texts. These data are required to be processed at the same time for retrieving the necessary information with respect to images and texts. In the image processing, the features like color, shape and texture are then extracted and they get reduced to create feature vectors. Thereafter the clustering is carried out for grouping the similar featured segments. In a similar manner, the texts are processed at the same time.

Color Features

The extraction of the color features of images are done by using the color histogram feature extraction. A color model is represented in terms of 3-D coordinate system and a sub space within that system in which every color is denoted by a single point. The generally applied color models are *RGB* (red, green, blue), *HSV* (hue, saturation, value) and Y,Cb,Cr (luminance and chrominance). This way, the characterization of the color content is done by 3-channels from a color model. One of the representations of color content of the image is by means of using the color histogram. Statistically stated, it represents the joint probability of the intensities of the three color channels. In the case of a three-channel image, three such histograms will be created. The histograms are usually categorized into bins in an attempt to roughly denote the content and reduce the dimensionality of the next matching phase. Then, a feature vector is created by the concatenation of the three channel histograms into one vector. For the retrieval of image, the histogram of query image is now matched against the histogram of every image in the database making use of any similarity metric.

The Color histogram defines the distribution of colors inside a complete or within an interested region of image. The histogram is unaffected by rotation, translation and scaling of an object though the histogram does not have semantic information, and the two images with identical color histograms can have diverse contents. A color histogram H for an image given is expressed as a vector H = {h[1], h[2], ..., h[N]} in which i denotes a color in the color histogram, $h[i]$ stands for the number of pixels present in color i in that image, and N refers to the number of bins present in the color histogram, i.e., the number of colors existing in the adopted color model. For the purpose of comparing images of various sizes since they are downloaded from several sites, color histograms has to be normalized. Then the normalized color histogram H′ is expressedas $h'[i] = \frac{h[i]}{XY}$ in which XY refers to the total number of pixels present in an image. The standard measure of similarity utilized for the color histograms is that the normalized color histogram which is stored in the database is actually matched with the testing image so that it matches against all the probable targets present in the database.

Texture Features

The extraction of the texture features are carried out by making use of the co-occurrence matrix technique of Gabor filter. The co-occurrence matrix C(i, j) numbers the co-occurrence of pixels with the gray values i and j at a specific distance d. The distance d is then expressed in polar coordinates (d, θ), along discrete length and orientation. The co-occurrence matrix C (Imp) can be expressed as:

$$C(i,j) = \text{card}\left\{\begin{array}{c}((x_1,y_1),(x_2,y_2)) \in (XY)\times(XY) \text{ for } f(x_1,y_1)=i,\ f(x_2,y_2)=j \\ (x_2,y_2)=(x_1,y_1)+(d\cos\theta, d\sin\theta); \quad \text{for } 0<i,j<N\end{array}\right\} \tag{1}$$

Where card {.} represents the number of elements in the set.

Let G refer to the number of gray-values in the image, then the dimension of the co-occurrence matrix C (i,j) would be N × N. Hence, the computational complexity of the co-occurrence matrix is quadratically dependent on the number of gray-scales which is utilized for quantization. Features can get extracted from the co-occurrence matrix for reducing the feature space dimensionality and the formal definitions of the five features from the co-occurrence matrix are below:

$$\text{Energy} = \sum_i \sum_j C(i,j)^2 \tag{2}$$

$$\text{Inertia} = \sum_i \sum_j (i-j)^2 C(i,j) \tag{3}$$

$$\text{Correlation} = \frac{\sum_i \sum_j (i-j)C(i,j) - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{4}$$

$$\text{Difference Moment} = \sum_i \sum_j \frac{1}{1+(i-j)^2} C(i,j) \tag{5}$$

$$\text{Entropy} = \sum_i \sum_j C(i,j) \log C(i,j) \tag{6}$$

where,

$$\mu_i = \sum_i i \sum_j C(i,j) \tag{7}$$

$$\mu_j = \sum_j j \sum_i C(i, j) \tag{8}$$

σ_i is defined as

$$\sigma_i = \sum_i (i - \mu_i)^2 \sum_j C(i, j) \tag{9}$$

σ_j is defined as

$$\sigma_j = \sum_i (j - \mu_j)^2 \sum_j C(i, j) \tag{10}$$

Shape Features

Shape is an essential visual feature and it is one among the primitive features for the description of image content. Shape content description is hard to be defined as the measurement of the similarity between shapes is tedious. The extraction of the shape descriptors are done by using employing contour based technique that computes the features such as circularity, aspect ratio, discontinuity angle irregularity, length irregularity, complexity, right-angleness, sharpness, directedness from the image contour. There are always possibilities of extracting the image contours from the extracted edges. The shape information is obtained from the object contour.

Circularity: $$cir = \frac{4pA}{P^2} \tag{11}$$

Aspect Ratio: $$ar = \frac{p_1 + p_2}{C} \tag{12}$$

Discontinuity Angle Irregularity: $$dar = \sqrt{\frac{\Sigma\left|\theta_i - \theta_{i+1}\right|}{2\pi(n-2)}} \tag{13}$$

Length Irregularity: $$lir = \frac{\Sigma\left|L_i - L_{i+1}\right|}{K} \tag{14}$$

Complexity: $$com = 10^{-\frac{3}{n}} \tag{15}$$

Right-Angleness: $$ra = \frac{r}{n} \tag{16}$$

Sharpness: $$sh = \sum \frac{\max\left(0, 1 - \left(\frac{2\left|\theta - \pi\right|}{n}\right)^2\right)}{n} \tag{17}$$

Directedness: $$dir = M / \Sigma P_i \tag{19}$$

where, n refers to the number of sides of polygon bounded by segment boundary, A is area of polygon bounded by segment boundary, P indicates the perimeter of polygon bounded by segment boundary, C refers to the length of the longest boundary chord, p_1, p_2 indicates the greatest perpendicular distances from longest chord to boundary, in every half-space on either side of line through the longest chord, θ_i refers to the discontinuity angle

between $(i-1)$-th and i-th boundary segment, r indicates the number of discontinuity angles that is equal to a right-angle inside a given tolerance, and M refers to the total length of straight-line segments which are parallel to mode direction of the straight-line segments inside a given tolerance.

Clustering

The color, texture and shape features of the images are acquired to be clustered employing an optimization algorithm for the purpose of storing them in the form of clustered feature vectors. The particle swarm optimization (PSO) along with two new fitness functions are proposed for this. The fitness functions use the quantization error, intra-cluster distance and inter-cluster separation.

The quantization error J_e is defined as:

$$J_e = \frac{\sum_{j=1}^{N_c}\left[\sum_{\forall Z_p \in C_j} d(Z_p, m_j)\right] \Big/ |C_j|}{N_c} \tag{19}$$

where, $d(Z_p, m_j)$ indicates the Euclidean distance between the pth pixel Z_p and the centroid of j-th cluster m_j.

The intra-cluster distance is computed by

$$\bar{d}_{\max}(Z, x_i) = \max_{j=1,\ldots,N_c}\left\{\sum_{\forall Z_p \in C_{ij}} d(Z_p, m_{ij}) \Big/ |C_{ij}|\right\} \tag{20}$$

The inter-cluster separation is defined by

$$d_{\min}(x_i) = \min_{\forall j1, j2, j1 \neq j2}\{d(m_{ij1}, m_{ij2})\} \tag{21}$$

A modified quantization error is introduced referred to as the weighted quantization error J_{e2}

$$J_{e2} = \left\{\sum_{j=1}^{N_c}\left[\left(\sum_{\forall Z_p \in C_{ij}} d(Z_p, m_{ij})/|C_{ij}|\right).(|C_{ij}|/N_0)\right]\right\} \tag{22}$$

The first fitness function is dependent on the quantization error J_e

$$f_1(x_i, Z) = w_1 \bar{d}_{\max}(Z, x_i) + w_2(z_{\max} - d_{\min}(x_i)) + w_3 J_e \tag{23}$$

For the purpose of providing an effective clustering of the images, the second fitness function is presented that is based on the weighted quantization error J_{e2}

$$f_2(x_i, Z) = w_1 \bar{d}_{\max}(Z, x_i) + w_2(z_{\max} - d_{\min}(x_i)) + w_3 J_{e2} \tag{23}$$

The new fitness function will make an improvement on the fitness function that utilizes only the weighted quantization error. It will also reveal that J_{e2} must be utilized together along with $\bar{d}_{\max}$ and $d_{\min}$ forgetting compact clusters and big inter-cluster separation. This compactness could be measured by means of the mean squared distance of the pixels from its corresponding cluster centroid.

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{K}\sum_{Z_p \in C_j}(Z_p - m_j)^2 \tag{25}$$

Where n refers to the total number of pixels in the image, z_p indicates the p-th pixel, K indicates the number of clusters, m_j refers to the centroid of the j-th cluster C_j. This way, the clustered image feature vectors are got and are saved in the feature vector space.

Text Processing

Just like the processing of images, the processing of the texts are also done separately. Once the web data get extracted in the form of plain texts, they are simultaneously processed. At first, the process stemming and stop word removal is conducted and thereafter the plain text data are grouped by using the *k*-Nearest Neighbor (kNN). Thereafter the text and image feature vector are trained by employing the Modified SOM module.

Existing SOM

1. Randomize the map's nodes' weight vectors
2. Grab an input vector $D(t_i)$ from $D(t)$
3. Traverse each node in the map
4. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector
5. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
6. Update the nodes in the neighborhood of the BMU by pulling them closer to the input vector

$$W_{ij}(t+1) = W_{ij}(t) + \eta(t)h(v, t)(V_i - W_{ij}(t))$$

7. Increase t & repeat from step 2 until $t < \lambda$

Enhanced SOM

1. Randomize the map's nodes' weight vectors
2. Split input vector set $D(t)$ into p parts;

$$k \gg N \times p$$

where, k is the number of input vector, N is the number of neurons and p is the number of parts.

3. Grab an input vector $D(t_i)$ from each p part of $D(t)$
4. Perform steps 3 to 7 as in SOM
5. Obtain weight vectors in every p part of $D(t)$
6. Merge all p parts with the unused neurons and their weight vectors specified together
7. Update SOM results

Thus Modified SOM training module and labeling process does the mapping of the images and text into image cluster map (ICM) and keyword cluster map (KCM) with high precision. After this, the matching process is performed employing the Enhanced SVM.

Matching Process

Enhanced SVM does the retrieval of the images and texts on the basis of the query data. It marks the top n images into two classes: relevance set I^+ and irrelevance set I° whereas the texts are marked into relevance set T^+ and irrelevance set T°. Then the general matching process is carried out.

$$(x_i, y_i), x_i \in I^+ \cup I^\circ, y_i = \begin{cases} +1, \text{if } x_i \in I^+ \\ -1, \text{if } x_i \in I^\circ \end{cases} \quad (27)$$

$$(x_i, y_i),\ x_i \in \mathrm{I}^+ \cup \mathrm{I}^\circ,\ y_i = \begin{cases} +1, \text{if } x_i \in \mathrm{T}^+ \\ -1, \text{if } x_i \in \mathrm{T}^\circ \end{cases} \tag{28}$$

The classification function applying SVM algorithm is $b.\ f(x) = \sum_i \alpha_i y_i k(x_i, x) + b,$ In order for the similarity distance to be output to the query means, the function sign (.) present in the classifier $f(x)$ is neglected. Thus the score is computed for every image I_i in the database score $(\mathrm{I}_i) = f(x_i)$ and all the images are sorted on the basis of the score and returns a new classification result. Thereafter the semantic similarity is calculated as

$$\mathrm{S}'_{q,x} = \mathcal{B}\left(\|\mathrm{G}(q) - \mathrm{G}(x)\| + \frac{|q \cdot x|}{\|q\| \|x\|} \right) \tag{29}$$

where function G: $\mathrm{R}^N \rightarrow \mathrm{R}^2$ returns the two-dimensional grid coordinates of its argument in the ICM. Therefore the images and the text are retrieved efficiently employing the retrieval model proposed.

4. EXPERIMENTAL RESULTS

In this section, the new technique is validated with a collection of web pages that were gathered manually based on the Yahoo! Or Google web site directory. The Yahoo! directory hierarchy was considered to be best since it has remained a standardized test bed for categorization and semantics evolution of web pages, and several other research works have exploited the Yahoo! hierarchy in their experiments. One benefit of the Yahoo! hierarchy is that it is built by human linguists and domain specialists, thus rendering it "semantically correct''. Which means, web pages which have been allocated to the same category must usually have semantic relevance. In addition, the relationships among the directories are also cautiously shown and allocated such that the directories in the same hierarchy follow their intrinsic semantic structures. The extended WordNet is actually a project at the University of Texas at Dallas (and also funded by the National Science Foundation) which is aimed at improving WordNet through semantic parsing of the glosses, this way offering the information present in these definitions to be available for automated knowledge processing systems. Every gloss is first tagged employing Brill's tagger. The glosses are thereafter parsed utilizing both Charniak's parser and an in-house Collins' style parser. Every parsed gloss is now allocated a level of quality.

The performance of the retrieval system proposed employing Web crawler & DOM tree is then compared with the image retrieval system utilizing ESVM-MSOM is compared with ESVM-SOM, SVM, ontological based image retrieval (OIM) and picSOM based image retrieval system. The techniques are assessed with respect to accuracy, precision, recall, *f*-measure, mean absolute error and execution time.

The analysis of the proposed technique, is done for drawing the conventional precision versus recall curves over all the queries. Initially, the precision versus recall curve approach is defined. The recall here refers to the percentage of correct documents which are being retrieved over all correct documents associated with a query, i.e. recall = $\frac{|\mathrm{Ra}|}{\mathrm{R}}$ where Ra indicates the set of correct documents which are being retrieved and R refers to the set of correct documents for a query q. The precision indicates the percentage of correct documents in the retrieved documents in response to q, i.e. precision = $\frac{\mathrm{Ra}}{\mathrm{A}}$ where A represents the set of retrieved documents for q. In order to simplify this task, the images which are relevant to a query image are described as those images which are in the same category like the query image, and all the remaining images are treated to be irrelevant. All the images are exploited as queries and the average precision versus recall curve is plotted as illustrated in Figure 3.

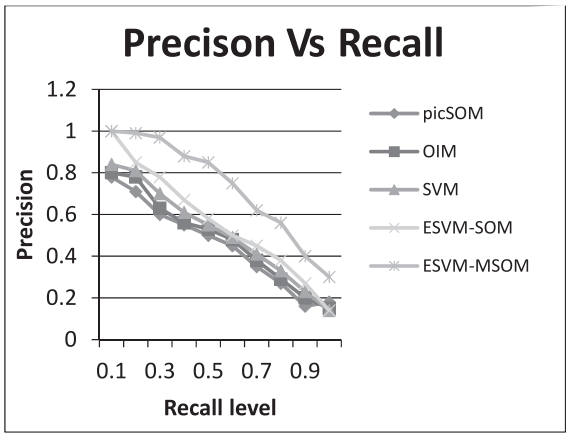

Figure 3: Average of Precision vs recall for all images

Table 1: Average of precision versus recall curve over all images

Recall	*picSOM*	*OIM*	*SVM*	*ESVM-SOM*	*ESVM-MSOM*
0.1	0.78	0.8	0.84	1	1
0.2	0.71	0.78	0.81	0.85	0.99
0.3	0.6	0.63	0.70	0.78	0.97
0.4	0.55	0.56	0.61	0.67	0.88
0.5	0.5	0.53	0.55	0.58	0.85
0.6	0.45	0.48	0.49	0.5	0.75
0.7	0.35	0.38	0.41	0.45	0.62
0.8	0.27	0.29	0.33	0.38	0.56
0.9	0.16	0.2	0.23	0.27	0.40
1.0	0.18	0.15	0.14	0.14	0.30

Accuracy Comparison

The Retrieval system proposed employing the Web crawler & DOM tree with ESVM-MSOM yields a better accuracy rate as illustrated in Figure 4having a much better accuracy results compared to the other available image retrieval techniques. When the number of images is increased, the accuracy of the result is also increased. This approach yields a high accuracy rate in comparison with the other available system as illustrated in Table 2.

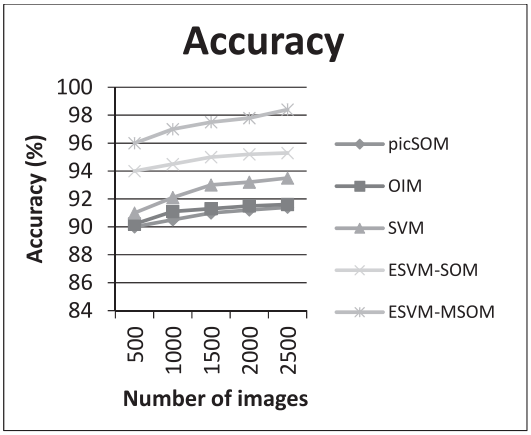

Figure 4: Accuracy Comparison

Table 2
Accuracy (%) comparison over all images

No. of images	*picSOM*	*OIM*	*SVM*	*ESVM-SOM*	*ESVM-MSOM*
500	90	90.2	91	94	96
1000	90.5	91.1	92.1	94.5	97
1500	91	91.3	93	95	97.5
2000	91.2	91.5	93.2	95.2	97.8
2500	91.4	91.6	93.5	95.3	98.4

F1-measure Comparison

F1-measure is defined to be the harmonic mean of precision and recall. A good classifier is supposed to to have a high F1-measure that shows that the classifier performs better with regard to both precision (P) and recall (R). F1-measure $= \frac{2 \times PR}{P \times R}$. The retrieval system proposed employing Web crawler & DOM tree with ESVM-MSOM yields a greater F1-measure as illustrated in Figure 5, which is greater compared to the available image retrieval techniques. With the increase in the number of images, the F1-measure of the result also increases. This approach yields a good F1-measure rate while comparing with the existing system. Table 3 illustrates the comparison between the retrieval systems with regard to F1 measure.

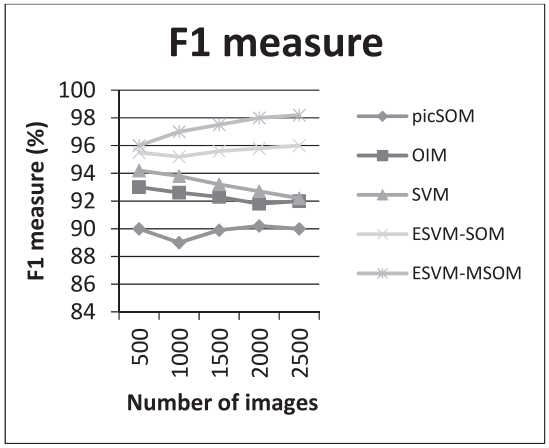

Figure 5: F1-Measure Comparison

Table 3
F1-Measure comparison over all images

No. of images	*picSOM*	*OIM*	*SVM*	*ESVM-SOM*	*ESVM-MSOM*
500	90	93	94.2	95.5	96
1000	89	92.6	93.8	95.2	97
1500	89.9	92.3	93.2	95.6	97.5
2000	90.2	91.8	92.7	95.8	98
2500	90	92	92.2	96	98.2

Mean Absolute Error

Statistically, the mean absolute error (MAE) is a quantity employed for measuring how nears the forecasts or predictions are to the eventual outcomes. The mean absolute error is expressed by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

As from the name, the mean absolute error is measured as an average of the absolute errors $|e_i| = |f_i - y_i|$, where f_i refers to the prediction and y_i indicates the true value.

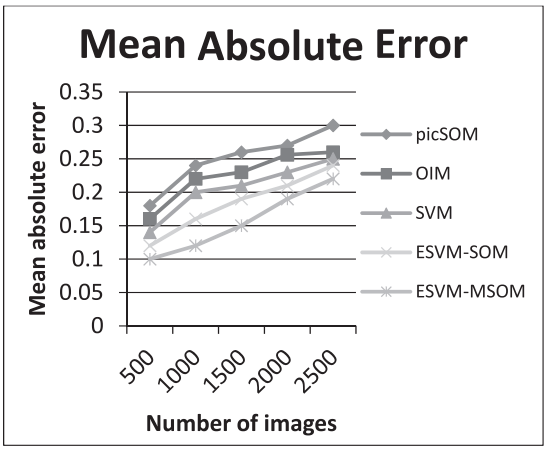

Figure 6: Mean Absolute Error Comparison

Table 4
Mean Absolute Error comparison over all images

No. of images	*picSOM*	*OIM*	*SVM*	*ESVM-SOM*	*ESVM-MSOM*
500	0.18	0.16	0.14	0.12	0.10
1000	0.24	0.22	0.20	0.16	0.12
1500	0.26	0.23	0.21	0.19	0.15
2000	0.27	0.256	0.23	0.21	0.19
2500	0.3	0.26	0.25	0.24	0.22

Figure 6 illustrates the comparison made between the mean absolute errors of the retrieval systems. Thus the best results of mean absolute error are received for the proposed retrieval system employing Web crawler & DOM tree with ESVM-MSOM compared to the other available techniques.

Execution Time

The time taken by the schemes for completing the web data retrieval determines the efficiency of the scheme and the amount of complexity involved.

Table 5
Execution time (seconds) comparison over all images

No. of images	*picSOM*	*OIM*	*SVM*	*ESVM-SOM*	*ESVM-MSOM*
500	88	80	75	66	60
1000	110	98	88	80	74
1500	130	115	102	92	80
2000	160	142	130	110	96
2500	180	165	148	130	112

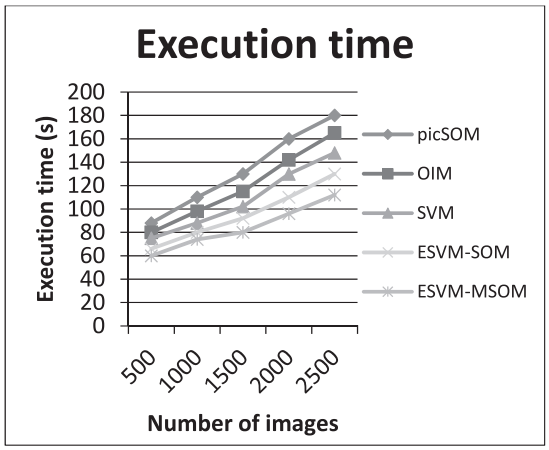

Figure 7: Execution time Comparison

Figure 7 illustrates the comparison made between the execution time of the retrieval systems. It proves that the proposed scheme using ESVM-MSOM takes less time for execution. This way, the model proposed does an accurate retrieval of the images and texts with large efficiency and also considerably minimizes the computational complexity in the biomedical information retrieval.

5. CONCLUSION

Since the image retrieval system which is based on SVM uses PCA for feature dimensionality reduction, it is affected by accuracy challenges in the image retrieval stage of the medical analysis. Therefore in this work, a retrieval system is introduced that uses a web crawler and DOM tree based Parser module for the purpose of pre-processing. The proposed technique for biomedical query processing uses feature extraction strategies for the extraction of the color, texture and shape features. Thereafter the images are clustered employing the PSO based clustering whereas the texts are grouped employing the kNN algorithm. Then MSOM training with labeling process is carried out which is followed by the matching process employing ESVM. This way, the images and text are accurately retrieved. The experimental results also show that the model proposed yields a greater efficiency in information retrieval along with a reduction in complexity in retrieving the medical information related to the input medical keyword.

REFERENCES

[1] J. Eakins, M. Graham, Content-based image retrieval, Technical Report, University of Northumbria at Newcastle, 1999.

[2] I.K. Sethi, I.L. Coman, Mining association rules between low-level image features and high-level concepts, Proceedings of the SPIE Data Mining and Knowledge Discovery, Vol. III, 2001, pp. 279-290.

[3] F. Long, H.J. Zhang, D.D. Feng, Fundamentals of content-based image retrieval, in: D. Feng (Ed.), Multimedia Information Retrieval and Management, Springer, Berlin, 2003.

[4] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (9) (2001) 947-963.

[5] Kekre, H.B. and Sudeep D. Thepade, (2008). Creating the Color Panoramic View using Medley of Grayscale and Color Partial Images, WASET International Journal of Electrical Computer and System Engineering (IJECSE), 2: 3.

[6] Kekre, H.B. and Sudeep D. Thepade, (2008). Color Traits Transfer to Gray scale Images, In Proc of IEEE First International Conference on Emerging Trends in Engineering & Technology.

[7] Kekre, H.B. and Sudeep D. Thepade, (2008). Scaling Invariant Fusion of Image Pieces in Panorama Making and Novel Image Blending Technique, International Journal on Imaging (IJI), 1(A08): 31-46.

[8] Müller, H., N. Michoux, D. Bandon and A. Geissbuhler, (2004). A review of content-based image Retrieval systems in medical applications – clinical benefits and future directions, International Journal Med. Inf., 73(1): 1-23.

[9] Lehmann, T., B. Wein, J. Dahmen, J. Bredno, F. Vogelsang and M. Kohnen, (2000). Content-Based Image Retrieval in Medical Applications: A Novel Multi-Step Approach, International Society for Optical Engineering, 3972(32): 312-320.

[10] Kekre, H.B., Sudeep D. Thepade and Akshay Maloo, (2010). Performance Comparison for Face Recognition using PCA, DCT & Walsh Transform of Row Mean and Column Mean, ICGST International Journal on Graphics, Vision and Image Processing (GVIP), 10: II.

[11] Jarrah, K., M. Kyan, S. Krishnan and L. Guan, (2006). Computational Intelligence Techniques and Their Applications in Content-Based Image Retrieval, IEEE Int. Conf. on Multimedia & Expo, pp: 33-36, Toronto, Canada.

[12] Antani, S., R. Kasturi and R. Jain, (2002). A survey on the use of pattern recognition methods for Abstraction, indexing and retrieval of images and video, Pattern Recognition, 35(4): 945-965.

[13] Vailaya, A., M.A. T. Figueiredo, A.K. Jain and H.J. Zhang, (2001). Image classification for content based indexing, IEEE Trans. Image Processing, 10(1): 117-130.

[14] Kekre, H.B. and Sudeep D. Thepade, (2009). Improving the Performance of Image Retrieval using Partial Coefficients of Transformed Image, International Journal of Information Retrieval, Serials Publications, 2(1).

[15] Sinha, U., A. Ton, A. Yaghmai, R.K. Taira and H. Kangarloo, (2001). Image Content Extraction: Application to MR Images of the Brain, Radio Graphics, 21(2): 535-547.

[16] Rahman M.M., Bhattacharya M.P. and Desai B.C. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Trans. Inform. Technol. Biomed., Vol. 11, No. 1, pp. 58-69, 2007.

[17] HuiHui Wang, Dzulkifli Mohamad & N.A. Ismail. Semantic Gap in CBIR: Automatic Objects Spatial Relationships Semantic Extraction and Representation. International Journal of Image Processing (IJIP), Vol. 4, Issue 3, pp. 192-204, 2010.

[18] Lars Jacob Hove, (2004). Extending Image Retrieval Systems with a Thesaurus for Shapes, Masters Thesis.

[19] Wilson, C., B. Srinivasan and M. Indrawan, (2000), A General Inference Network Based Architecture For Multimedia Retrieval, In Proceedings of IEEE International Conference on Multimedia pp: 347-350, New York City, USA, 2000.

[20] Roland Kwitt, Peter Meerwald and Andreas Uhl, (2011). Efficient Texture Image Retrieval Using Copulas in a Bayesian Framework, IEEE transactions on image processing, pp: 20.

[21] Naphade, M.R. and T.S. Huang, (2002). "Extracting semantics from audiovisual content: The final Frontier in multimedia retrieval," IEEE Trans. Neural Network., 13(4): 793-810

[22] Yixin Chen, (2005). Member, IEEE, James Z. Wang, Member, IEEE and Robert Krovetz, CLUE: Cluster-Based Retrieval of Image by Unsupervised Learning, IEEE transactions on image Processing, 14.

[23] Long, F., H. Zhang and D.D. Feng, (2003). Fundamentals of content-based image retrieval, in Multimedia Information Retrieval and Management Technological Fundamentals and Applications. New York: Springer-Verlag.

[24] Sumathi, P., & Manickachezian, R. (2016). Semantic-Based Web Mining For Image Retrieval Using Enhanced Support Vector Machine. International Journal of Applied Engineering Research, 11(5), 3276-3281.

[25] Qi Zhou, Chong Wang, Miao Xiong, Haofen Wang, and Yong Yu. Spark: Adapting keyword query to semantic search. In Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, volume 4825 of LNCS, pages 687–700, Berlin, Heidelberg, November 2007. Springer Verlag.

[26] "Public Control Algorithm for a Multi Access Scenario comparsing GPRS and UMTS", at Department of Computer Science and Engineering, National Conference on "Intelligent computing With IoT on April 16 2016 in Dhirajlal Gandhi College of Technology.

[27] "Teleimersion" Research Journal of Pharmaceutical, Biological and Chemical Sciences on March – April 2016 issue.(Impact Factor 0.35 Indexed in Scopus). http://www.rjpbcs.com/pdf/2016_7(2)/%5b131%5d.pdf

[28] Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, and James Matfield. Information retrieval on the semantic web. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, pages 461-468, New York, NY, USA, 2002. ACM.

[29] John Davies and Richard Weeks. Quizrdf: Search technology for the semantic web. Hawaii International Conference on System Sciences, 4:40112+, 2004.

[30] Irene Celino, Emanuele Della Valle, Dario Cerizza, and Andrea Turati. Squiggle: An experience in model-driven development of real-world semantic search engines. In Luciano Baresi, Piero Fraternali, and Geert-Jan Houben, editors, ICWE, volume 4607 of Lecture Notes in Computer Science, pages 485-490. Springer, 2007.

[31] Justin Johnson, Ranjay Krishna, Michael Stark, LiJia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[32] H. Pourghassem and H. Ghassemian, "Content-based medical image classification using a new hierarchical merging scheme," Comput. Med. Imag. Graph., Vol. 32, No. 8, pp. 651-661, 2008.

[33] N. Mehta, R.S. Alomari, and V. Chaudhary, "Content based sub-image retrieval system for high resolution pathology images using salient interest points," Int. Conf. Proc IEEE EMBS, Vol. 1, pp. 3719-3722, 2009.

[34] A. Cruz Roa, J. Caicedo, and F. Gonzlez, "Visual pattern analysis in histopathology images using bag of features," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2009, pp. 521-528.

[35] L. Zheng, A. Wetzel, J. Gilbertson, and M. Becich, "Design and analysis of a content-based pathology image retrieval system," IEEE Trans. Inf. Technol. Biomed., Vol. 7, No. 4, pp. 249-255, Dec. 2003.

[36] L. Yang, O. Tuzel,W. Chen, P.Meer, G. Salaru, L. Goodell, and D. Foran, "Pathminer: Aweb-based tool for computer-assisted diagnostics in pathology," IEEE Trans. Inf. Technol. Biomed., Vol. 13, No. 3, pp. 291-299, May 2009.

[37] A.F.R. Rahman, H. Alam and R. Hartono. "Content Extraction from HTML Documents". Document Analysis and Recognition Team (DART). BCL Computers Inc. August 11, 2002.

[38] O. Buyukkokten, H. Garcia-Molina, A, Paepcke, "Text Summarization for Web Browsing on Handheld Devices", In Proc. of 10th Int. World-Wide Web Conf., 2001.

[39] E. Kaasinen, M. Aaltonen, J. Kolari, S. Melakoski and T. Laakko. "Two Approaches to Bringing Internet Services to WAP devices". In Proc. of 9th Int. World-Wide Web Conf. 2000. pp. 231-246.