International Journal of Web Portals
Volume 11 • Issue 1 • January-June 2019

Semantics Based Web Ranking Using a Robust Weight Scheme

R. Vishnu Priya, VIT Vellore Institute of Technology Chennai Campus, Tamil Nadu, India
V. Vijayakumar, VIT Vellore Institute of Technology Chennai Campus, Tamil Nadu, India
Longzhi Yang, Northumbria University, Newcastle upon Tyne, UK

ABSTRACT

In this paper, HTML tags and attributes are used to determine different structural position of text in a web page. Tags- attributes based models are used to assign a weight to a text that exist in different structural position of web page. Genetic algorithms (GAs), harmony search (HS), and particle swarm optimization (PSO) algorithms are used to select the informative terms using a novel tags-attributes and term frequency weighting scheme. These informative terms with heuristic weight give emphasis to important terms, qualifying how well they semantically explain a webpage and distinguish them from each other. The proposed approach is developed by customizing Terrier and tested over the Clueweb09B, WT10g, .GOV2 and uncontrolled data collections. The performance of the proposed approach is found to be encouraging against five baseline ranking models. The percentage gain of approach achieved is 75-90%, 70-83% and 43-60% in P@5, P@10 and MAP, respectively.

KEYWORDS

Attributes, Feature Selection, HTML Tags, Image, Ranking, Weighting Model

1. INTRODUCTION

Information Retrieval (IR) is a field of study that helps to extract relevant information from a large collection of text documents[1] in the web. Web search is a most important application of IR and its challenges to retrieve the high quality web pages to the user query. While user using popular search engines (SEs) such as Google and Yahoo, they glance many web pages before finding a required one or they cannot find all the relevant information they are looking for (Fan et al., 2009).

Reasons can be explained from a number of perspectives. The existing search engines are learning the surfing habits of users through Analytics and Adsense code, which embedded on web pages for tracking the interest of users. This information is sold to companies for their development or used for targeted advertising, which allows businesses to advertise by popular keywords and advertise on particular sites. Both these Adsense and AdWords are increasing the revenue of search engines and create a profitable business for advertisers. This way of earnings will probably display the business and marketing web pages as top-k results, which is irrelevant to the users.

DOI: 10.4018/IJWP.2019010104

Due to the proprietary reasons, the accurate algorithms used by commercial SEs are not recognized. However, the correlation study has been made recently to conjecture the working nature of SEs algorithms and provide the factors that affecting the ranking. Those are mainly based on the backlinks, social networking websites, on page technical and on page content. While considering backlinks, various strategies are followed in increasing the rank of particular web sites, such as, links should point to an inner page, links should come from their country, remove the offending links, analysis links to keep and to get rid of and so on. Usually, webmaster and search engine optimizers (SEOs) carefully assess the above factors and rank pages higher than they deserve. Hence, SEOs earn money for better placement of websites in the search list. Due to all above listed facts, all commercial, social networking, fake, personal, advocacy web pages are ranked as top-k web pages. Therefore, users are frequently navigating for interesting pages result to increase in browsing time. A recent study has shown that the common characteristics of top ranked pages on the web are social network pages. However, social network pages are used by the certain aged people. As the result, the personal and advocacy web pages are present at the top-k of retrieved results. Further, in the TREC competition (Fan et al., 2009), it was stated that using the link information alone doesn't provide much help in performance improvement as compared to using content information. The ranking functions based on content alone are still very successful.

Meanwhile, on page technical and on page content are shown special attention on the texts associated with the web pages. Typically, web pages are developed using two characteristics: content of the document and the structure of the document. The ranking of web pages in content based ranking is done using various lexical/syntactical statistics of the words in web pages. Structure based considers the structural properties of web pages and a weight is assigned to each word existing in different structural position. This weighting heuristic improves the ranking performance of the web. Each web page is an HTML document where Tags are used for designing purpose. The HTML Tags gives idea for understanding a web page, says, the <title> TAG represents the title of the document. SEs understand the web page using few Tags related to the head and body sections and assign equal priority which lead to less progression. In addition to the text information, image, audio, video contents are also embedded into the pages and that can also be used in understanding the web content thereby the retrieval accuracy is improved. While considering the text content of images, it is assured that they are similar to the text annotations and closer to the semantic interpretation of the image. Recent works have been failed to exactly interpret the image using texts. A novel approach will need to propose which improves the top-k of retrieved results.

The main objective of this paper is to effectively capture the semantics of web pages that will rank the user interest web pages on top of the search result. This can be done through a novel TAGs-based models. The proposed approach captures the semantics of web page using the features of web pages say HTML Tags. These semantics are captured through assigning different weights for all HTML Tags. For this, the terms and its associated Tags extract from the web pages and each term is weighted based on the property of the Tags associated. Now, Genetic algorithm (GA), harmony search (HS) algorithm, and particle swarm optimization (PSO) algorithm are used to select the informative terms in web page using a novel Tags weighting scheme and term frequency. These informative terms with heuristic weight give emphases to important terms, qualifying how well they semantically explain webpage and distinguish them from each other. Further, the image Tag and its related attributes are retrieved, where the values of the attributes are being considered as terms and the weight is assigned to those terms based on the property of the associated attribute. Now, these weights are summed to assign a score for a web page and based on the scores the web pages are ranked. The rest of the paper is organized as follows. Section 2 presents the related work and the proposed technique is explained in Section 3. In Section 4, the experimental results are given and we conclude the paper in the last section.

2. RELATED WORKS

In this section, the related works are presented. Initially, the text retrieval from the documents is discussed in section 2.1. It is followed by various approaches to discover the knowledge from images in section 2.2.

2.1 Text Retrieval in IR

Initially, the research on web text retrieval was focused for developing techniques to make use of information obtained from hyperlinks. Query-independent and Query-dependent are the two categories of connectivity-based algorithms. Page Rank and On-line Page Importance Computation- OPIC (Abiteboul et al, 2003, kadry et al., 2013) were query-independent ranking algorithms. Kleinberg proposed Hyperlink-Induced Topic Search (1999) was a query-dependent ranking algorithm. Even though, these algorithms part of the current search engine with many issues. Say for example, the damping factors for all pages are used in a ranking model to remove sinking effect that will change the quality of the pages. In addition, these algorithms are suffering from rich-get richer problem. A recursive connectivity algorithm (Bidoke & Yazdani, 2008) based on reinforcement learning was proposed to handle rich-get richer problem. This approach considers the distance between pages as punishment, called 'Distance Rank' to compute ranks of web pages. Reinforcement learning algorithm favors only the odd pages and those high quality pages have been added recently without many inbound links were ignored. Hence, everyone has concentrated on raising the page rank by buying high inbound links.

Alternatively, several keyword-based search algorithms identify the structural relationships of pages for ranking. Multiway-SLCA (Sun et al., 2007) offers a search paradigm beyond the traditional AND semantics in support of keyword search, including both AND and OR Boolean operators. The semantics of exclusive lowest common answer to improve result quality is developed by Xu and Papakonstantinou (2008). According to the entities in data that are relevant to the search, keywords are explicitly inferred to generate return nodes (Liu & Chen, 2007). Li et al. (2008) presented a ranking mechanism to improve search effectiveness by combining structural compactness and textual relevance to rank the answers. Graupmann et al. (2005) proposed the Sphere search engine to provide unified ranked retrieval of heterogeneous web data. Li et al. (2008) proposed a unified keyword search framework, namely SAILER for unstructured and semi-structured data for answering keyword queries. EASE (Li et al., 2008) is an efficient and adaptive keyword search method designed for adaptively answering keyword search over heterogeneous data by summarizing the graphs. Li et al. (2008) demonstrated keyword search over heterogeneous data. However, graphs with a larger diameter were not meaningful frustrated the users by large and complex graphs. Recently, Li et al. (2011) have proposed keyword-based document retrieval, where information from multiple interrelated pages was integrated to provide meaningful results and for this they used graph model. The r-radius Steiner graphs were extracted using the set by removing the non-Steiner nodes from the corresponding r-radius graphs and ranked the result to return the top-K answers. Cai et al. (2016) rank using double hypergraph. However, it is prohibitively expensive to discover the rich structural relationships graph, adjacency matrix and Steiner graph, as each page consists of many internal links from other pages. In turn, links from many other websites will be pointing to them. Kaushik et al. (2017) and Zhan et al. (2017) have addressed the reranking algorithm for a specific domain. Thus, this structured relation search approach was suitable for a specific domain and may not be suitable for WWW.

Park et al. (2005) proposed an algorithm suitable for named page finding task using vector space model with baseline system. The Retrieval Status Value (RSV) for each document was retrieved using several stages. It assigns scores to query terms: in the title, the same sentence apart from the adjacent sentence and in the document being pointed by anchor tag. All the above generated scores are summed and RSV value of a document was calculated. The ranked documents were finally undergone stratifying and re-ranking stage. However, it is suitable only for named page finding task and do not

work well for the short queries. The value for sentence-query similarity for a document varies once there is a change in query terms with time consumption. The titles defined within the bodies of web pages were more relevant compared to the content in title field as they are more noticeable to the readers. Addressing this idea, Xue et al. (2007), Bhardwaj et al. (2014) and Akhtar et al. (2014) have proposed an automatic extraction of the title from the bodies of web pages using DOM-tree and vision based methods. Neha Sharma et al. (2014) have captured sematic of web using decision tree. However, this approach considers only title of the document, the information present in other Tags was ignored and thus result in low precision. Akritidis et al. (2011) have developed a new Metasearch engine for running simultaneously user query across multiple component search engines, retrieve the generated results and aggregate them. It is known that this engine doesn't maintain own document index and new rank aggregation method is used to rank the content in a single list. The author proposed zone scoring by assigning weight to each zone and based on the zone the weight varies. Weight factor rewards the documents containing the query terms in its title, snippet or URL as many times as possible. However, wasnt considered the importance of all the zones for scoring a document. Fan et al. (2009) developed a framework for few HTML tags and used those tags in genetic programming techniques. The content of the HTML web page have been identified by the framework (Wu et al., 2016; Serrano-Guerrero et al., 2015; Uzan et al., 2014 and Mohammad zadeh et al., 2013). Ling et al. (2006 & 2007) proposed the idea for designing a web page and the above framework and design standard ideas were support to identify the essential content in the web page and the same is used in our approach. However, these algorithms consider only few HTML tags for capturing the semantics. Thus, it is imperative that a comprehensive approach is required to make use of keywords in tags of all categories for capturing the semantics of HTML pages.

2.2. Image Retrieval in IR

The first experimental approach for image retrieval is Annotated based information retrieval (ABIR), where each image is manually annotated using textual data and the standard database management systems (Chang and Fu, 1980; Chang and Kunii, 1981) were employed. In the early 90's, manually annotation approach for vast images collection became inoperable. As an outcome, CBIR approach (Rui et al., 1999) proposed and those extracted features were stored in the database for retrieval purpose (Aslandogan and Yu, 1999). Various CBIR techniques with different extraction and storing features (Kwae and Kabuka, 2000; Ogle and Stonebraker, 1995; Wu, 1997) and on the image searching approach (Cox et al., 2000; Flickner et al., 1995; Santini and Jain, 2000) were suggested. The research direction for image retrieval was changed due to the expansion of WWW (Benitez et al., 1998; Benitez et al., 1997; Sclaroff et al., 1997). Generally, the images were identified with image filenames, html tags and surrounding text in the web. In the course of time, text was combined with image feature to develop the multi-modal systems for improving image search results (Wang et al., 2001; Chen et al., 2001; Hu et al., 2000; Smith and Chang, 1997; Wu et al; 2001). Finally, it was hard to extract low-level features from the web images (or) manual annotation is hard. Hence, ABIR was reasonable and Kılınc et al., (2000) recently suggested this approach, where the image retrieval system was introduced for Wikipedia pages using textual data around images. The document and queries expansion techniques was suggested using WordNet (Miller, 1990), Word Sense Disambiguation and similarity functions. Normally, this expansion lead to high recall with low precision and the low-level re-ranking approach is introduced to increase precision. However, selection of appropriate sentences to describe an image are an important and valid problem.

The studies to analyse the textual information from monochromatic were suited better to describe the images. In addition, text regions extracted from mixed text/ graphic compound document images, the most of the textual regions show distinctive texture features that were unlike other non-text background regions (Hasan and Karam, 2000). A three level rule based model was developed in (Niyogi and Srihari, 1996; Levine and Nazif, 1985; Lee et al., 2000) for the image analysis domain. Chen et al. (2009) extract text-lines from mixed compound with different illumination levels, sizes

and font styles. This systems decompose the document image into distinct object planes using multi-plane segmentation technique for extracting and separating homogenous objects. Therefore, (Chen et al., 2012) come up with a new algorithm that easily handles text-line that overlap with pictorial objects and background. Knowledge based text extraction and identification procedure were applied on each plane with various characteristics to detect, extract and identify text-lines. The geometrical and statistical features of text-lines were encoded using knowledge rules which establish two rule sets. However, rules generation and extraction of feasible test from mixed component is a consumes time.

Another approach retrieves image and video on the WWW is presented in (Smith and Chang, 1997), where the image was described using the words from a URL and ALT tag with a predefined category. Lu and Willam (1999) proposed the technique of weighting the extracted term based on their location in the HTML document. A normalized sum of these weights were combined with a color comparison measure by adding both values. The similar approaches with some refinement were introduced in (Chen et al., 2001) and (Hu et al., 2000). In (Wu et al., 2001), a self-organizing neural network was used through combining image features and textual information. However, there is no clear information on all these works such as the part selected from the HTML document for image retrieval. Sanderson and Dulop (1997) used a combination of texts from associated HTML pages with one-several links away for modeling image contents. Shen et al. (2000) used the more summarized information on the related web page and construct a simple model to chain associated terms. The model for the image (Benitez et al., 1997; Feng et al., 2004) was constructed using the combination of text, link information surrounding text of images. Chen and Xiaohua, 2010 modelled the visual content by including features such as color and region. However, the features were too low level and inadequate to model image contents (Chang and Fu, 1980; Yang 2006) that result to the ineffectiveness of retrieval. Images are classified into various categories (Chen et al., 2009) to handle this issue by using domain specific features. Recently, keywords related approaches were spreading over unlabeled images to capture the semantic concept. Users implicitly labelled more images with the relevance feedback mechanism was used to regularly update this model. However, the relevance feedback mechanism will takes large to converge the learning knowledge. Xu and Zang (2007) suggested an integrated patch model and generative model for image categorization based on feature selection. The feature selection method was divided into three steps for acquiring representative characteristic and also used to eliminate the noise characteristic. Recently, the textual keywords appeared in the web pages was used for identifying unsuitable, offensive and pornographic web sites (Hu et al., 2007). In this work, the decision tree was used to divide the text pages into continuous, discrete and image using respective classifiers. Pornographic nature of continuous text in web pages is identified using statistical and semantic features. Similarly, pornographic content of discrete and image web pages were identified using Bayesian classifier and object's contour of images respectively. Recently, the gap between the extracted features of the systems and the user's query were reduced using another method (Barnard and Forsyth et al., 2001; Yang et al., 2008; Zhao and Grosky, 2002). The semantic-based image retrieval using self-organizing maps (Yang et al., 2008) discovered the semantics of the image from web pages and those semantics are described based on the text surrounding the images. The text mining procedure used by adopting Self Organizing Map learning method as a kernel. After this process, some of the implicit semantic information was also discovered and the semantic relevance measure was used for retrieval. Recently, both the keyword and low-level features of images in HTML pages were extracted for retrieval (Vadivel et al., 2008 and 2009). Several low-level features were integrated with high-level for retrieving relevant image retrieval. However, the association or the importance of the keywords with the HTML page or with the image, is not measured. It was noticed from the above discussion that none of the approaches dynamically measure the association of textual keywords with image information for capturing the semantics retrieval (Lienhart and Hartmann, 2002; Liu and Chen, 2007). Based on the above discussion, it was noticed that the information available in HTML documents should be used effectively. The structural position of the texts were helped to capture the semantics of the web page and rank the desired web pages on top of the retrieved results.

3. PROPOSED WORK

World Wide Web inventor developed the semantic web to use intelligently for serving the users need. This can be done through formulating users need in a form that can be understood by the retrieval mechanism. Meanwhile, the contents of document should be described in a form that make the retrieval mechanism to recognize the highly relevant documents. Recently, these two phases are inherently unsatisfied and the solution for one of the phase is discussed as follows.

The web is a collection of interrelated web pages without strict and identical data structures or schemas that follow. As a result, there is increase in need to better deal with the unstructured nature of web pages to capture the knowledge. The web pages in web are designed in a well-defined HTML format, that format contains some preliminary web data structures, say Tags and it can primarily help to design the appearance quality of web pages. For e.g., the designer wants to place the important text at the center with highlighted form, the texts are placed between the respective Tags. This appearance provides an importance of text that helps us to capture the semantic nature of web pages. In this paper, the information in web pages is uniquely identified using Tags with suitably assigned weight.

3.1. TAGs-Based Model for Texts

Let us consider that a web contains a large number of web pages, say, $WP = \{wp_1, wp_2, \ldots wp_{|WP|}\}$. Each wp_i is designed using various Tags $TG = \left\{tg_1, tg_2, \ldots tg_{|TG|}\right\}$ and $T = \{t_1, t_2, \ldots t_{|T|}\}$ are the set of texts occurred between TG. In our earlier work (Vishnu Priya et al., 2012), the semantics of the web page is captured and weight is assigned using TAGs-based model. The weight is calculated for each text using Equation (1). However, the text can occur several times in a page and it failed to consider the reoccurrence. Ihis issue is addressed in this work to increase the efficiency of retrieval rate:

$$DWGT = \sum_{L=1}^{n}\left(\frac{(N-(T*L))}{(2*L)}\right) \tag{1}$$

The texts and its associated Tags are extracted from each web page. The weight is then assigned to the Tag based on the TAGs-based model. Each text has both frequency and Tag weight, such that, $f : T \times WP \rightarrow \{1,0\}$:

$$f\left(t_i, wp_j\right) = \begin{cases} 1 & if\ wp_j\ contains\ t_i \\ 0 & otherwise \end{cases}, where 1 \le i \le \left|T\right| and \le j \le \left|WP\right| \tag{2.1}$$

f(t_i,wp_j) of text t_i in web page wp_j is defined as the frequency of t_i=1, when t_i appears in wp_j. Similarly:

$$Twgt : T \times WP \rightarrow N$$

$$Twgt\left(t_i, wp_j\right) = Twgt_{lw}\ \ if\ \ Level\text{-}lw\ \ contains\ \ tg_k,\ \ where\ 1 \le i \le \left|T\right|,\ 1 \le j \le \left|WP\right|\ \ and\ 1 \le k \le \left|TG\right| \tag{2.2}$$

*Twgt(t_i,wp_j)*of text t_i in web page wp_j is defined as the Tag weight of t_i= $Twgt_{lw}$, where tg_k is an associated TAG of t_i belongs to the *Level-lw* and *lw*= 1 to 4.

Let us consider the text 't_i' is represented as <wp_j, f_i, $Txtwgt_i$>, it means t_i appear f_i times in the web page wp_j that related to the Tag weight of $Txtwgt_i$. It is calculated by summing up all the frequencies (f_x) with its associated Tags weights ($Twgt_x$). It is noticed that, a text can physically present more than once in wp_j. Then, $Txtwgt_i$ of t_i in wp_j is calculated using Equation (3), where 'n' is the total number of times 't_i' present in wp_j:

$$Txtwgt_i = \left(\sum_{x=1}^{n} f_x\right) + \left(\frac{\sum_{x=1}^{n} Twgt_x}{n}\right) \tag{3}$$

The average Tags weight is calculated in Equation (3) due to the reason as follows. Let us assume, my name appears in two web pages, say, my home page and my supervisor's journal publication page. In the first case, it appears near <TITLE> Tag and its weight is 49 (i.e. 1+48). In another case, it appears 5 times near <p> Tag and its weight is 15 (i.e. 5+ (5*10)/5). If the total weight is taken, my supervisor's journal publication page will be at the top. Henc, it is necessary to consider the total contribution of the same texts in a web page. This concludes the average weight helps to effectively assign weight to each text and rank those web pages in an effective manner.

There is a possibility that the same texts could occur among different Tags from different Levels, in such case, we make the first four Levels into one and the last two into another groups. Since, the last two with very low weights, which will drop the total Tag weight of the text. Therefore, the Tag weight of the text that appears among different TAGs is calculated based on Equation (4). For instance, t_2 appears four times with Tags weights $Twgt_1$, $Twgt_2$, $Gwgt_1$ and $Gwgt_2$ in wp_1, where $Twgt$ represents the weight of the Tags from Level-1 to Level-4 and $Gwgt$ denotes the Tags from Level-5. The Tag weight of t_2 ($Txtwgt_2$) is defined as:

$$Txtwgt_g = \sum\left(\sum_{x=1}^{n} f_x\right)\left(\frac{\sum_{y=1}^{m} Twgt_y}{m}\right)\left(\frac{\sum_{z=1}^{l} Gwgt_z}{l}\right) \tag{4}$$

where:

$$Gwgt = \begin{cases} 1 & if\ \ t_g \in TL_5 \\ 0 & otherwise \end{cases}$$

and g, l, m = 2 and n = 4.

Before to construct the inverted index, GA, HS and PSO algorithms are used to select the informative terms from the web page based on the Tags and term frequency weight. Inverted index is constructed for those terms. For each text, there is a posting list (PL_T) that contains <*WP, Txtf, Txtwgt*>, where *WP* is a web page id and <*Txtf, Txtwgt*> is the frequency and Tag weight of the text.

3.2. ATTRIBUTEs-Based Model for Images

The precision of retrieval of the proposed approach can be improved by considering the texts associated with the images. An image related information can extract from <img> Tag, which contains both attributes and texts. Typically, the designer uses 24 attributes to design an image and is divided into 4 Levels based on its characteristic. This division has been done as same as the earlier work, called

ATTRIBUTEs-based model. Each level contains the set of attributes of the same characteristics with the weight.

Let *I* be a set of images exist in *WP* and its <img> Tags says $I = \left\{i_1, i_2, ...i_{|I|}\right\}$, which contains both attributes $IA = \left\{ia_1, ia_2, ... ia_{|IA|}\right\}$ and texts $T = \{t_1, t_2, ... t_{|T|}\}$. *AL* = $\left\{al_1, al_2, al_3, al_4\right\}$ are four different levels and *WT= {wt₁, wt₂, wt₃, wt₄}* are weights related to each level. The subset I' given in Equation (5), extracts all images contain texts:

$$I' = \left\{\left(i_y, wp_j\right) \middle/ \forall i_y \in I \;\; and \; f\left(t_x, i_y\right) = 1\right\}, where \; 1 \le x \le \left|T\right|, 1 \le y \le \left|I\right|, 1 \le j \le WP \tag{5}$$

where the frequency of each text is defined as:

$$f : T \times I \rightarrow \left\{1, 0\right\}$$

$$f\left(t_x, i_y\right) = \begin{cases} 1 & if \; i_y \; contains \; t_x \\ 0 & otherwise \end{cases}, where \; 1 \le x \le \left|T\right| \;\; and \;\; 1 \le y \le \left|I\right| \tag{6}$$

A set of texts with its attribute are available at T' from the subset I' is shown in Equation (7):

$$T' = \left\{\left(t_x, ia_z, wp_j\right) \middle/ \forall I' \in I \; \exists \; t_x \;\; s.t \; for \; each \; t_x \in T \; \exists \; some \; ia_z \in IA\right\},$$
$$where \; 1 \le x \le \left|T\right|, 1 \le z \le \left|IA\right| \;\; and \; 1 \le j \le \left|WP\right| \tag{7}$$

A text with its associated frequency, weight and document id is shown in Equation (8), where weight assigned based on ATTRIBUTEs-based model:

$$g : T \times IA \rightarrow N$$

$$TR = \left\{\begin{matrix} \left(t_x, \; f\left(t_x, \; I'\right), \; wt_{lw}, wp_j\right) \middle/ for \; each \;\; ia_z \;\; contain \;\; in \;\; T' \; \exists \;\; al_{lw} \in AL \\ s.t \;\; g(t_x, ia_z) = wt_{lw} \;\; if \;\; al_{lw} \;\; contains \; ia_z \;\; and \; f(t_x, I') \;\; = 1 \end{matrix}\right\},$$
$$where \;\;\; lw = 1 \; to \; 4, \; 1 \le x \le \left|TA\right|, 1 \le y \le \left|I\right|, 1 \le z \le \left|IA\right| \;\; and \; 1 \le j \le \left|WP\right| \tag{8}$$

TR consists of interested information of each term and can be defined as the relationship of $<t_x, f_x, wt_x, wp_j>$ as follows. Actually, a term from different images of same document can repeat many times in *TR*. Therefore, the total frequency and attribute weight of each term for a document can be calculated as:

$$Imgf : TR \times TR \rightarrow N$$
$$Imgwgt : TR \times TR \rightarrow N$$

$$Imgf\left(t_x, wp_j\right) = \sum_{x=1}^{m} f_x \tag{9}$$

$$Imgwgt\left(t_x, wp_j\right) = \sum\left(\sum_{x=1}^{m} f_x \sum_{x=1}^{m} wt_x\right), where\ 1 \le j \le \left|WP\right|\ \ and\ 1 \le x \le \left|RA\right| \tag{10}$$

where *m* is the total number of same terms in wp_j. While comparing Equation (10) with 4, the total weight has been taken for attribute weight calculation of a term. This is due to the attributes are obtained from the same <img> Tag. Now, another posting list (PL_I) for each text from the images is stored in the inverted index (*II*), which represents as <*WP,Imgf,Imgwgt*>. A newly proposed weight in posting lists of each term as <<PL_I>, <PL_I>>. The user query process these posting lists to capture semantic of web pages. Retrieval task for a given query is discussed as follows.

3.3. Retrieval Phase

Once the user enters the query $Q' = \{q_1, q_2, \ldots q_{|Q'|}\}$, such that $q_\alpha \in Q'$, the mapping will be happened between q_α and it_α, an indexed term in *II*. Each indexed term holds posting list, say, PL_I, which contain $< wp'_\beta, w_{\alpha,\beta}, f_{\alpha,\beta} >$. Here, the weight of text exists in TAG represents as $w_{\alpha,\beta} \ge 0$. If doesn't exist, the weight will be $w_{\alpha,\beta} = 0$. Finally, the ranking is done using the weight and it is computed for each document as below:

$$S_\beta = \frac{\sum_{\alpha=1}^{n(Q')} \frac{w_{\alpha,\beta}}{\max_{\chi \in wp'}\left(w_{\alpha,\chi}\right)}}{n\left(Q'\right)},\ where\ 1 \le \beta \le n\left(wp'\right)\ and\ t_\alpha \equiv q_\alpha \tag{11}$$

Let $n(wp')$ be the number of documents containing it_α in either <TAGs> or <img>. Different weights of the same text contain in different documents are given as χ. S_β is a score of individual document and its range vary between:

$$\left[0, n\left(\frac{Q'}{2}\right)\right]$$

4. EXPERIMENTS

The experiments are run using Terrier to validate the proposed text-weighting method based on TAGs and ATTRIBUTEs weights. We chose Terrier because it is renowned open source search engine that has been validated (ladh Ounis et al., 2005). The impact of the proposed approach is shown by modifying each module in Terrier which has been described as follows.

4.1. Experimental Setup

The terrier is a highly efficient, effective and flexible web search that readily deployable on large-scale collections of documents. It consists of two main facets namely indexing and retrieving. Source codes of these facets are modified to incorporate our needs.

4.1.1. Indexing and Retrieval Phases

Initially, the set of web pages is extracted from the collections. The customization has been done in the indexing phase of Terrier to extract the TAGs along with the raw texts in the web

page. Texts are embedded within lots of TAGs, among them; the TAG which exists in the higher level of TAGs-based model is considered. Then the texts and its related TAG are returned to the Term Pipeline. Texts passed through the Term Pipeline are stemmed using Porter's English stemmer along with the removal of stopwords. Each term has three fundamental properties, say, the position of the text occurs in the web page, the frequency of occurrences, TAG in which the text occurs. The TAG weight of the text is applied according to the Equation (1). Further, the same text in a webpage could be re-occurred many times, whose TAG weight is calculated based on Equation (4). Eventually, the texts contained in <img> tag are weighted as in Equation (10). Using the texts and related info, the inverted index is built and it writes to the appropriate data structure. In retrieval phase, the queries are read from the topics file and it is processed. All the retrieval parameters are used to determine the web pages match a specific query and the web pages are scored with respect to a query. Weighting model is employed to assign a score to each of the query terms in a web page. It represents the retrieval model that is given in Equation (11) to weight the terms of a web page. The matching module uses a weighting model to assign a weight to each query term in a web page.

4.2. Datasets and Queries Used

Four collections are used for the experimental results, .GOV2, WT10G, ClueWeb09 and our own crawled web pages. The.Gov2 (topics 701-850) collection crawled from.Gov domain, which contains 25 million documents with 426GB in size. It has been employed in the TREC 2004, 2005 and 2006 TB tracks. The WT10G is a medium size crawl of web document collections that is used in the TREC9 and 10 web tracks. The topics regard this data set is 451-550. The ClueWeb09 is a very large TREC test collection, and is currently the largest crawl of the Web. The category B of ClueWeb09 is used, that contains about 50 million English Web pages. Its associated topics are constructed from three topic fields, namely, title, descripitive and narrative in TREC 2009 web topics. The fifty topics released from the adhoc task of TREC 2009 on the ClueWeb09 corpus are randomly taken for evaluation purpose. Finally, the uncontrolled dataset is generated using Google and the queries related to various domains such as colleges, Universities, Institutes, research center, flower, famous leaders, newspapers, sports, cine field, tourism, etc., are provided as a query into Google and for each query, the top-500 links are collected. This collection contains at least one qualified image on each web page and those links are given as input to the crawler for fetching those particular pages. The web pages associated with each query are saved in a repository. The aim is to evaluate and investigate retrieval approaches using this heterogeneous collection of web pages that are browsed by users with various information categories. The total number of web pages in the dataset is 2, 59, 818 that cover various topics of interest and each page includes the mixed and overlapping of text/ images. The queries for this heterogeneous database are taken from n-grams data (2, 3, 4, 5-word sequences, with their frequency) and it is available at http://www.ngrams.info/.

4.3. Performance Measures

A proposed IR model designed using the Terrier platform is evaluated with IR measures. It includes Precision @ 5, Precision @ 10 and Mean Average Precision (MAP). These measures could be essential for usual web search tasks because the user is willing to see only the top few web pages and would expect as many relevant documents as possible in the top-k retrieved results. To ascertain the outcome of the output, we have formed a group containing graduate, undergraduate and research students from different disciplines who give feedback about the proposed model for the queries in topics file. Users are asked to examine the results pages by opening the page and identify the most relevant, partially relevant and irrelevant web pages. Based on the feedback, the results are validated and detailed as below.

4.4. Results

The results of experiments done for the proposed approach are discussed in this section. Initially, the queries of varies collections are processed in the batch mode of the topic file. It reads into the Terrier, where the indexed weighted documents of the four collections are used by the retrieval model to assign a score for the web pages. According to the degree of relevance, the web pages are sorted and ranked to the users. The user examines this ranked list from the top. The main objective of the proposed approach to help the web users to find the most relevant web pages with less effort. Based on the objectives of the proposed approach, P@5, P@10 and MAP are considered as retrieval performance evaluation. Terrier supplies two well-known ranking functions which include TF-IDF and BM25 in order to facilitate the cross-comparison between the developed models. Therefore, we used five different ranking systems for the comparison purpose. The first is the proposed work and the other four are renowned functions in IR literature and they have used in the TREC evaluation study. The functions are TF-IDF, BM25, Title extraction approach (Xue et al., 2007) and GA-based approach (Fan et al., 2009), Decision Tree Induction (Neha et al., 2014), Page Content Rank (kadry et al., 2013 and kaushik et al., 2017) and web document reranking (Zhao et al. 2017) will be considered as a retrieval baseline function for comparisons.

In absence of ground truth value, the performance is evaluated using Precision which is a standard measure in information retrieval. Precision is defined in Equation (12) and is measured by identifying the top-5 and10 results for every query and calculate the average:

$$Precision = \frac{No.\ of\ relevant\ documents\ retrieved}{Total\ no.\ of\ documents\ retrieved} \tag{12}$$

In Tables 1 and 2, the average precision of top-5 and 10 web pages is shown for all the collection and it is noticed the proposed approach outperform the comparative approaches. The baseline functions TFIDF and BM25 are failed to weight each text based on the appears and its outcome degrades. The performance by Xue et al. and Fan et al. are more or less similar and however lower than the proposed approach. The average precision of the proposed and Fan et al. approach are ranged from 75-90% and 40%-60% respectively for the top-5 results. It shows that the top-5 web pages displayed in the proposed approach for the given queries are found to be relevant. Even the maximum achieved an average precision for Fan et al. and Xue et al. are 60% and 45% respectively for the top-5 web pages. This is due to the fact that these algorithms typically take only few structural information such as, anchor, title, abstract and body. Moreover, equal weight is assigned to the texts appear in the different structural position in the document.

Besides having high value of precision@5 and @10, it is also important to estimate Mean Average precision (MAP). It is an average precision at relevant document exist. It determines precision at each point while a new relevant web page retrieved and is shown in Equation (13):

$$MAP = \frac{\sum_{q=1}^{Q} Ave\ P(q)}{Q} \tag{13}$$

where Q is the total number of queries. It is observed that MAP with the proposed approach is out performing all other comparative approaches. It performance varies within the range of 45 -60% for the proposed approach and for recent comparative approach it varies between 20-40%. This is due to the fact that most of the texts appear in different structural position on the web pages are weighted depending on their nature. This discriminate nature of each text with others will effectively capture

the semantics of each web page in the web. Further, the proposed weighting model is used to assign a valid score to the web pages that place the relevant pages in the top of retrieval set. In addition to the entire above outcome on controlled data collections, it is also essential to estimate the performance of the proposed approach over uncontrolled data collection. It is depicted in the Tables 1-3 that the outcome values of the proposed approach for uncontrolled collection are higher compared to the other collections. The reason is the crawled web pages contain at least more than two images. Hence, both the tags and attributes weights of the text give a higher score to web page. It is in a similar range of the other comparative works. It is concluded the performance rate of the proposed approach is encouraging compared to the recent baseline ranking functions.

5. CONCLUSION

Usually, retrieval system returns a large number of web pages for a user query. While the users visit the web pages that are ranked higher as it assumed that they are more relevant to the query. It is found to be tedious to display the most relevant web pages in top results. This is due to the ranking mechanism used by the conventional system is ineffective. Typically, the retrieval system considers various techniques for ranking, such as link based, connectivity based and keyword based techniques. We found the keyword based technique effectively retrieves the relevant web pages. In this work, an open source Terrier is customized to extract Tag-text pairs of the web page and weight is assigned

Table 1. Performance metrics for Precision@5

Coll.	Topics	TFIDF	BM25	Xue et al	Fan et al	Neha et al	Kadry et al	Kaushik et al	Zhao et al	TAGs_ATTR
GOV2	701-850	0.281	0.414	0.450	0.582	0.378	0.459	0.381	0.541	0.8534
WT10G	451-550	0.187	0.314	0.413	0.579	0.328	0.514	0.45	0.521	0.7865
ClueWeb09B	50 Topics	0.181	0.289	0.303	0.462	0.426	0.443	0.34	0.511	0.7402
UnCont. DB	5-grams	0.150	0.231	0.259	0.383	0.335	0.371	0.47	0.569	0.8973

Table 2. Performance metrics for Precision@10

Coll.	Topics	TFIDF	BM25	Xue et al	Fan et al	Neha et al	Kadry et al	Kaushik et al	Zhao et al	TAGs_ATTRIBUTEs
GOV2	701-850	0.272	0.326	0.338	0.514	0.313	0.211	0.220	0.341	0.7980
WT10G	451-550	0.171	0.311	0.317	0.434	0.351	0.129	0.231	0.412	0.7514
ClueWeb09B	50 Topics	0.173	0.225	0.245	0.274	0.213	0.113	0.242	0.472	0.7062
UnCont. DB	5-grams	0.134	0.173	0.218	0.211	0.251	0.218	0.440	0.519	0.8259

Table 3. Performance metrics for mean average precision

Coll.	Topics	TFIDF	BM25	Xue et al	Fan et al	Neha et al	Kadry et al	Kaushik et al	Zhao et al	TAGs_ATTRIBUTEs
GOV2	701-850	0.244	0.305	0.359	0.383	0.212	0.121	0.152	0.311	0.5794
WT10G	451-550	0.186	0.210	0.243	0.340	0.238	0.142	0.161	0.281	0.4648
ClueWeb09B	50 Topics	0.174	0.206	0.208	0.329	0.251	0.173	0.154	0.300	0.4316
UnCont. DB	5-grams	0.153	0.183	0.121	0.202	0.254	0.154	0.148	0.292	0.5925

to the text based on the characteristic of Tag with which it is associated. Eventually, the weight is assigned to the Attribute-text pair in <img> tag according to attribute-based model. Finally, all the weights related to a text in the web page are summed using the proposed weighting formula. The GA, HS, and PSO algorithms are used to select the informative terms based on weight and indexed. While user enters the query, the matching between query and document texts are made and relevant documents are ranked depand on the score. The performance of the proposed work was compared against five baseline function using renowned benchmark collections. It found that the proposed approach achieves maximum 90% precision for the batch queries.

REFERENCES

Abiteboul, S., Preda, M., & Cobena, G. (2003). Adaptive on-line page importance computation. In *Proceedings of the twelfth international conference on World Wide Web* (pp. 280–290). doi:10.1145/775152.775192

Akhtar, N., Siddique, B., & Afroz, R. (2014). Visual and Textual Summarization of Webpages. In *International Conference on Data Mining and Intelligent Computing (ICDMIC)*. IEEE.

Akritidis, L., Katsaros, D., & Bozanis, P. (2011). Effective rank aggregation for metasearching. *Journal of Systems and Software*, *84*(1), 130–143. doi:10.1016/j.jss.2010.09.001

Alp Aslandogan, Y., & Clement, T. (1999). Techniques and Systems for Image and Video Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *11*(1), 56–63. doi:10.1109/69.755615

Barnard, K., & Forsyth, D. A. (2001). Learning the semantics of words and pictures. In *IEEE International Conference on Computer Vision* (Vol. 2, p. 408).

Bhardwaj, A. (2014). *A Novel Approach for Content Extraction from Web Pages. In Proceedings of 2014 RAECS UIET, Panjab University, Chandigarh*. IEEE.

Bidoki, A. M. Z., & Yazdani, N. (2008). DistanceRank: An intelligent ranking algorithm for web pages. *Int. Journal of Information Processing and Management*, *44*(2), 877–892. doi:10.1016/j.ipm.2007.06.004

Cai, D., He, X., Ma, W. Y., Wen, J. R., & Zhang, H. (2004). Organizing WWW Images based on the Analysis of Page Layout and Web Link Structure. In *Proc. of International Conference on Multimedia Expo* (p. 113).

Chang, E., Goh, K., Sychay, G., & Wu, G. (2003). CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology, 13(1), 26-38.

Chen, Y.-L., Hong, Z.-W., & Chuang, C.-H. (2012). A knowledge-based system for extracting text-lines from mixed and overlapping text/graphics compound document images. *Expert Systems with Applications: An International Journal*, *39*(1), 494–507. doi:10.1016/j.eswa.2011.07.040

Chen, Y.-L., & Wu, B.-F. (2009). A multi-plane approach for text segmentation of complex document images. *Pattern Recognition*, *42*(7), 1419–1444. doi:10.1016/j.patcog.2008.10.032

Chen, Z., Liu, W., Zhang, F., & Li, M. (2001). Web mining for web image retrieval. *Journal of the American Society for Information Science and Technology*, *52*(10), 831–839. doi:10.1002/asi.1132

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., & Yianilos, P. N. (2000). The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, *9*(1), 20–37. doi:10.1109/83.817596 PMID:18255370

Essam, A. (2000). Efficient content-based indexing of large image databases. *ACM Transactions on Information Systems*, *18*(2), 171–210. doi:10.1145/348751.348762

Feng, H., Shi, R., & Chua, T. S. (2004, October). A bootstrapping framework for annotating and retrieving WWW images. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 960-967). ACM.

Forsati, R., & Meybodi, M. R. (2010). Effective page recommendation algorithms based on distributed learning automata and weighted association rules. Int. *J. Expert Systems with Applications*, *37*(2), 1316–1330. doi:10.1016/j.eswa.2009.06.010

Long, F., Zhang, H., & Feng, D. D. (2003). Fundamentals of content-based image retrieval with high-level semantics. *Pattern Recognition*, *40*(1), 262. doi:10.1016/j.patcog.2006.04.045

Graupmann, J., Schenkel, R., & Weikum, G. (2005). he sphere search engine for unified ranked retrieval of heterogeneous XML and web documents. In *Int. conference on Very large data bases* (pp. 529-540).

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Journal of Data Mining and Knowledge Discovery*, *8*(1), 53–87. doi:10.1023/B:DAMI.0000005258.31418.83

Hauptmann, A., Baron, R. V., Chen, M. Y., Christel, M., Duygulu, P., Huang, C., ... & Moraveji, N. (2004). Informedia at TRECVID 2003: Analyzing and searching broadcast news video. Carnegie-Mellon Univ Pittsburgh PA School of Computer Science. Retrieved from http://www-nlpir.nist.gov/projects/tv.pubs.org

Heng, T. S., Ooi, B. C., & Tan, K.-L. (2000). Giving meanings to WWW images. In *Proceedings of the eighth ACM international conference on Multimedia*, Marina del Rey, CA (pp. 39-47).

Hu, W., Wu, O., Chen, Z., Fu, Z., & Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1019–1034. doi:10.1109/TPAMI.2007.1133 PMID:17431300

Vadivel, A., Sural, S., & Majumdar, A. K. (2009). Image Retrieval from Web using Multiple Features. *Online Information Review*, *33*(6), 1169.

Jing, F., Li, M., Zhang, H. J., & Zhang, B. (2005). A unified framework for image retrieval using keyword and visual features. IEEE Transactions on Image Processing, 14(7), 979-989. doi:10.1109/TIP.2005.847289 PMID:16028561

Kaushik, S. & Yadav, B. (2017). Improved Page Ranking in Search Engine Using Web Content Mining and Web Structure Mining. *International Journal of Current Trends in Engineering & Research*, *3*(5), 191-194.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, *46*(5), 604–632. doi:10.1145/324133.324140

Li, G., Feng, J., Wang, J., & Zhou, L. (2008). RACE: finding and ranking compact connected trees for keyword proximity search over XML documents. In *Int. conference on World Wide Web* (pp. 1045-1046). doi:10.1145/1367497.1367648

Li, G., Feng, J., Wang, J., & Zhou, L. (2008). An effective and versatile keyword search engine on heterogeneous data sources. VLDB, 1(2), 1452-1455.

Li, G., Feng, J., Wang, J., & Zhou, L. (2008). SAILER: an effective search engine for unified retrieval of heterogeneous XML and web documents. In International conference on World Wide Web (pp. 1061-1062).

Li, G., Feng, J., Wang, J., & Zhou, L. (2010). RACE: Finding and ranking compact connected trees for keyword proximity search over XML documents. *Information Systems Journal*, *35*(2), 186–203.

Li, G., Ooi, B. C., Feng, J., Wang, J., & Zhou, L. (2008, June). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 903-914). ACM.

Lienhart, R., & Hartmann, A. (2002). Classifying images on the web automatically. *Journal of Electronic Imaging*, *11*(4), 1. doi:10.1117/1.1502259

Ling, J., & Schaik, P. V. (2006). The influence of font type and line length on visual search and information retrieval in web pages. *Int. J. Human-Computer Studies*, *64*(5), 395–404. doi:10.1016/j.ijhcs.2005.08.015

Ling, J., & Schaik, P. V. (2007). The influence of line spacing and text alignment on visual search of web pages. *Int. J. Displays*, *28*(2), 60–67. doi:10.1016/j.displa.2007.04.003

Liu, Y., Zhang, D., & Lu, G. (2008). Region-Based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, *41*(8), p2554. doi:10.1016/j.patcog.2007.12.003

Liu, Z., & Chen, Y. (2007). Identifying return information for XML keyword search. In ACM SIGMOD international conference on management of data (pp. 329-340).

Liu, Z., & Chen, Y. (2007). Identifying return information for XML keyword search. In *Proc. of ACM SIGMOD International conference on management of data* (p. 329). doi:10.1145/1247480.1247518

Miliaraki, I., Berberich, K., Gemulla, R., & Zoupanos, S. (2013, June). Mind the gap: Large-scale frequent sequence mining. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 797-808). ACM. doi:10.1145/2463676.2465285

Mohammadzadeh, H., Gottron, T., Schweiggert, F., & Nakhaeizadeh, G. (2013). Extracting the main content of Web documents based on character encoding and a naïve smoothing method. *Communications in Computer and Information Science*, *303*, 17–236. doi:10.1007/978-3-642-36177-7_14

Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., & Taubin, G. et al. (1993). *'The QBIC project: querying images by content using colour, texture and shape* (pp. 173–187). Soc. Opt. Eng., in Storage and Retrieval for Image and Video Databases.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web (Technical report)*. Stanford University.

Park, E.-K., Ra, D.-Y., & Jang, M.-G. (2005). 'Techniques for improving web retrieval effectiveness', Int. *Journal of Information Processing and Management*, *41*(5), 1207–1223. doi:10.1016/j.ipm.2004.08.002

Sanderson, H. M., & Dunlop, M. D. (1997). Image retrieval by hypertext links. In *Proc. of ACM SIGIR* (p. 296).

Santini, S., & Jain, R. (2000). Integrated Browsing and Querying for Image Databases. *IEEE MultiMedia*, *7*(3), 26–39. doi:10.1109/93.879766

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Inf. Sci.*, *311*, 18–38. doi:10.1016/j.ins.2015.03.040

Sharma, N. (2014),. Semantic Based Web Prefetching Using Decision Tree Induction. In *5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*. IEEE.

Shen, H. T., Ooi, B. C. Tan, K. L. (2000). Giving meaning to WWW images. In Proc. Of ACM Multimedia (p. 39).

Smith, J. R., & Chang, S. F. (1997, February). VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia* (pp. 87-98). ACM.

Smith, J. R., & Chang, S. F. (1997, February). VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia* (pp. 87-98). ACM.

Sun, C., Chan, C. Y., & Goenka, A. K. (2007). Multiway SLCA-based keyword search inXMLdata. In *Int. conference on World Wide Web* (pp. 1043-1052).

Uzun, E., Serdar Güner, E., Kılıçaslan, Y., Yerlikaya, T., & Agun, H. V. (2014). An effective and efficient web content extract or for optimizing the crawling process. *Software, Practice & Experience*, *44*(10), 1181–1199. doi:10.1002/spe.2195

Vadivel, A., Sural, S., & Majumdar, A. K. (2008). Robust Histogram Generation from the HSV Color Space based on Visual Perception. *International Journal on Signals and Imaging Systems Engineering*, *1*(3/4), 245. doi:10.1504/IJSISE.2008.026796

Vishnu Priya, R., & Vadivel, A. (2012). Capturing Semantics of Web Page using Weighted Tag-tree for Information Retrieval. *International Journal of Asian Business and Information Management*, *3*(4), 7-24.

Wu, S., Er, M. J., & Gao, Y. (2001). A fast approach for automatic generation of fuzzy rules by generalized dynamic fuzzy neural networks. *IEEE Transactions on Fuzzy Systems*, *9*(4), 578–594. doi:10.1109/91.940970

Wu, Y.-C. (2016). Language independent web news extraction system based on text detection framework. *Information Sciences*, *342*, 132–149. doi:10.1016/j.ins.2015.12.025

Xu, F., & Zhang, Y. J. (2007). Integrated patch model: A Generative Model for Image Categorization based on Feature Selection. *Pattern Recognition Letters*, *28*(12), 1581–1591. doi:10.1016/j.patrec.2007.03.016

Xu, Y., & Papakonstantinou, Y. (2008, March). Efficient LCA based keyword search in XML data. In Proceedings of the 11th international conference on Extending database technology: Advances in database technology (pp. 535-546). ACM. doi:10.1145/1353343.1353408

Xue, Y., Hu, Y., Xin, G., Song, R., Shi, S., Cao, Y., & Li, H. et al. (2007). Web page title extraction and its application. *Information Processing & Management*, *43*(5), 1332–1347. doi:10.1016/j.ipm.2006.11.007

Yanai, K. (2003, November). Generic image classification using visual knowledge on the web. In Proceedings of the eleventh ACM international conference on Multimedia (pp. 167-176). ACM. doi:10.1145/957013.957047

Yang, H. C., & Lee, C. H. (2008). Image Semantics Discovery from Web Pages for Semantic-based Image Retrieval using Self-organizing maps. *Expert Systems with Applications*, *34*(1), 266–279. doi:10.1016/j.eswa.2006.09.016

Zhao, R., & Grosky, W. I. (2002). Narrowing the semantic gap—improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, *4*(2), 189–200. doi:10.1109/TMM.2002.1017733

Zhao, G., & Zhang, X. (2017, July). A Domain-Specific Web Document Re-ranking Algorithm. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 385-390). IEEE.

R. Vishnu Priya received her post graduate under the Faculty of Information and Communication ngineering in College of Engineering Guindy, Anna University Main Campus, Chennai, India, in 2009. She received her PhD from National Institute of Technology, Tiruchirappalli, India. She worked under the DST project entitled "Internet Multimedia Search Engine for Information Retrieval in Distributed Environment" of DST/TSG/ICT/2009/27 dated 3rd September 2010, from July 2009 – December 2012. Currently she is working as an Associate Professor in School of Computing and Engineering, VIT University Chennai Campus. Her research interest includes Big Data, Information retrieval and Sentimental analysis.

Vijayakumar Varadarajan is currently a Professor and an Associate Dean for School of Computing Science and Engineering at VIT University, Chennai, India. He has more than 16 years of experience including industrial and institutional. He also served as a Team Lead in industries like Satyam, Mahindra Satyam and Tech Mahindra for several years. He has completed Diploma with First Class Honors. has completed BE CSE and MBA HRD with First Class. He has also completed ME CSE with First Rank Award. He has completed his PhD from Anna University in 2012. He has published many articles in national and international level journals/conferences/books. He is a reviewer in IEEE Transactions, Inderscience and Springer Journals. He has initiated a number of international research collaborations with universities in Europe, Australia, Africa and North America including University of Missouri. He had also initiated joint research collaboration between VIT University and industries including FSS. He is also the Guest Editor for few journals in Inderscience, Springer and IGI Global. He also organized several international conferences and special sessions in USA, Vietnam, Africa and India including IEEE, ISBCC etc. His research interests include computational areas covering grid computing, cloud computing, computer networks and big data. He received his university-level Best Faculty Award for 2015–2016. He is also a member of several national and international professional bodies including ISTE, IAENG, CSTA, etc.

Longzhi Yang (MBCs, SMIEEE, FHEA) is a senior lecturer in computer science, and the programmer cluster leader of computer networks, cyber security and digital forensics programmes at Northumbria University. He is also the chair of IEEE SIG of Big Data for Cyber Security and Privacy. His research interest include machine learning, intelligent control systems and the application of such techniques in real-world uncertain environments. He was the recipient of two best paper awards and a number of IEEE CIS travel grants.