International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-4 Number-2 Issue-15 June-2014

Web Search Results Summarization Using Similarity Assessment

Sawant V.V.[1] , Takale S.A.[2]

Abstract

Now day's internet has become part of our life, the WWW is most important service of internet because it allows presenting information such as document, imaging etc. The WWW grows rapidly and caters to a diversified levels and categories of users. For user specified results web search results are extracted. Millions of information pouring online, users has no time to surf the contents completely .Moreover the information available is repeated or duplicated in nature. This issue has created the necessity to restructure the search results that could yield results summarized. The proposed approach comprises of different feature extraction of web pages. Web page visual similarity assessment has been employed to address the problems in different fields including phishing, web archiving, web search engine etc. In this approach, initially by enters user query the number of search results get stored. The Earth Mover's Distance is used to assessment of web page visual similarity, in this technique take the web page as a low resolution image, create signature of that web page image with color and co-ordinate features .Calculate the distance between web pages by applying EMD method. Compute the Layout Similarity value by using tag comparison algorithm and template comparison algorithm. Textual similarity is computed by using cosine similarity, and hyperlink analysis is performed to compute outward links. The final similarity value is calculated by fusion of layout, text, hyperlink and EMD value. Once the similarity matrix is found clustering is employed with the help of connected component. Finally group of similar web pages i.e. summarized results get displayed to user. Experiment conducted to demonstrate the effectiveness of four methods to generate summarized result on different web pages and user queries also.

Manuscript received May 20, 2014.

Sawant V.V., M.E. second year, Computer Department, Pune University/VPCOE, Baramati, India.

Takale S.A., HOD**,** IT Department, Pune University/ VPCOE, Baramati, India.

Keywords

Web mining, Layout Similarity, Visual Similarity, EMD Link Analysis, Summarization, Web Similarity, Text Similarity, WWW.

1. Introduction

Web page is the main part on the World Wide Web. Commercial search engines are those, which retrieve pages based on the user request. Number of popular search engines exists like Google, Alta Vista and others. Search engines are that crawl the web and gives the results in some indexed order based on some criteria. Finally results are displayed to the end user through the browser People wants the information should be get in few amount of time, rather than to check the number of search results urls. Web search engines helps in locating information content and normally provide thousands of results for a query. Users still have to spend lot of time to scan through the contents of this result set to locate the required information. It is not feasible for the user to open each link in the result set to find out its relevance. Numbers of urls or web pages are found duplicated. So how the true result set is produced for the given query in summarized format is explained in this approach.

Similarity of Web pages is very useful for Web content analysis. Some similarity computation methods have been used to compare Web pages. However, only text based similarity computation methods are not sufficient for Web page comparison, because Web page consists of not only text but also multimedia contents, such as, audio, video, image, hyperlink structure and so on. This paper proposes a new approach to evaluate visual similarity of Web pages considering some of the contents on them. It can make Web page similarity computation exactly and bring benefits for Web analysis. The size, dynamic nature and diversity of content of this information is necessary in the development of effective search tools .Web search engines are also today one of the most frequently used tools for retrieving information from the web. So, similarity between web pages becomes important to research problem. The subjects like web communities,

filtering, and cluster based search. This paper proposes the important technique which is used to restructure the search results for entered query by calculate the web page visual similarity, i.e. to measure the distance between two web pages. First the layout similarity of two web pages is calculated considering their associated DOM-Tree representation by using Templates Computation Algorithm and Simple Tags Comparison Algorithm. Then, Link Analysis, it performed by analyzing outward links, Text similarity is computed by applying the cosine similarity. To compute the image based EMD(Earth Mover's Distance) first we have to convert the web pages into the normalized web page images and then represent their image signatures with features composed of pixel color and its corresponding centroid coordinate to calculate the visual similarity of two Web pages. The linear programming approach of EMD is applied to visual similarity computation of the two signatures.

The similarity matrix is constructed and from that cluster is formed by connected component, then final result of similar urls is displayed. The system designed for web search results restructuring is described in section (3).The architecture for the system is given in figure 1.Initially query is entered through bing search engine. The retrieved results are stored in the form of webpage image and webpage. Then the webpage comparison methods are applied on that to generate the similarity matrix discussed in (3) section. On that connected component is applied to form similar webpage groups. From these groups similar web page images is display to user. Experiment and result discussion is described in section (4) and (5) respectively.

2. Literature Survey

***A. Search Results** Optimization & Summarization:*
Dragomir et al. [1] in their work have presented an open domain multi-document summarization in the context of web search. Thomas [2] in his work has developed a quality based web search engine based on human judgments. He analyzed the features for characterization of the web using machine-learning approaches. The author has developed a meta search service called AQUAINT where all result pages are evaluated according to their quality and re-ranked accordingly. Yitong Wang et al. [3] have proposed a new approach to cluster search results returned from Web search engine using link analysis. Delort et al. [4] in their work addresses the issue of web document summarization. The authors have considered the context of a web document by the textual content of all the documents linking to it.

B. Web Page Comparison:
Layout similarity method is based on Antiphish[5], which is also known as DOMAntiphish.The key disadvantage of AntiPhish is that user manual interaction is required to specify the information on a web site that is considered sensitive. Web pages consist of features like layout content, textual content, and visual content. Textual content is defined as the terms or words that appear in a given web page [6], except for the stop words (a set of common words like "a," "the," "this," etc.). We first separate the main text content from HTML tags and apply stemming [13] to each word. Distance measures have been proposed and widely applied, such as cosine similarity and the Jaccard correlation coefficient, Metric, Euclidean distance. Meanwhile, similarity is often conceived in terms of dissimilarity or distance as well [8]. Kleinberg has introduced the concepts of "authorities" and "hubs". This paper is perhaps one of the most widely cited papers in the areas of hyperlinked environments. He did utilize his algorithm to solve the "similarity queries", but he did not present a measure. There are other systems that find "similarity pages", online, like Google with its "Similar Pages" feature [9], Netscape with its "what's related?" option [10].Nevertheless, none of them present the concepts and techniques used in semantic link analysis. In this paper, we propose an effective approach for comparing two Web pages, which employs the Earth Mover's Distance (EMD) [11] to calculate the visual similarity of Web pages. DOM based [12] assessment technique is introduced and then phishing detection a 'visual based approach' is first introduced. EMD is a method to evaluate the distance (dissimilarity) between two web page signatures. A signature is consisting of features and their corresponding weights. The method arises from the well-known producer consumer problem.

3. Proposed Work

The system architecture is given in Fig. 1:
It shows the flow of the project. User enter the query for retrieve some information to the search engine, here bing search engine is used. So, search engine provides no. of search results urls. Many of the urls are repeated or duplicated and so, the large amount of result set is obtained. It is difficult and time consuming for users to check all the urls. As an

solution to reveal the users problem, system has introduced. It produces the appropriate and limited set of search results with rank so user can get the exact information in less time.

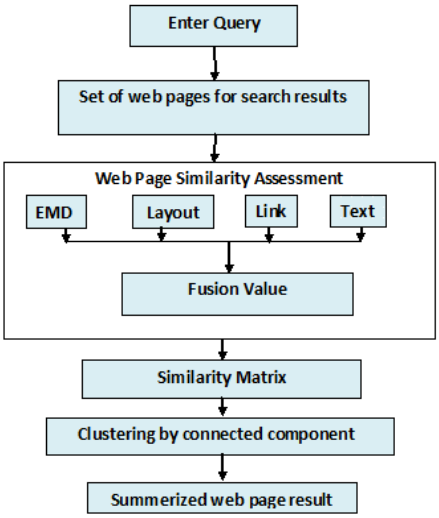

Figure 1: System Architecture

A .Webpage Visual Similarity Assessment
In this section, how to compare two webpages is described. There are no. of methods to compare the two webpages, here 4 methods are considered.

1) Layout similarity:
To calculate the layout similarity of web page we considered the DOM tree representation of the web page. We assume that identical layout is generated if two web pages having same DOM tree representation. It is possible that by having the different DOM tree representation we can generate the identical layout. Given two DOM-Trees, we compare their similarity in two different ways: 1) comparing the tags of the two web pages; 2) Extracting regular sub graphs from the trees. Templates denotes particular sub-graph of the original graph with at least two tags. The layout similarity of the two webpage is defined as the ratio of the weighted number of matched vertices of the DOM-Trees to the number of total vertices in the web page [12], as shown in Equation 1.

$$\alpha = \sum_{n=0}^{n=v} Wt(Vn)/Vn \qquad (1)$$

Where, Wt is a function that assigns a similarity weight between 0 and 1 to each vertex of the DOM-Tree, while V_n represents the n-th vertex of the DOM-Tree. If the layout similarity value α, as defined in (1) exceeds a certain threshold δ then two pages are similar one.

2) Text similarity:
A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity and the Jaccard correlation coefficient. Meanwhile, similarity is often conceived in terms of dissimilarity or distance as well [7] In this approach text data is extracted from webpage The bag of word model is widely used in text mining [12] and information retrieval. Words are counted in the bag, which differs from the mathematical definition of set. We use the frequency of each word as its weight, which means words that appear more frequently are more important and descriptive for the document. Let D = *d*1,..., *dn* be a set of documents and W = *w*1, ...,*wn* the set of distinct terms occurring in D. We discuss more precisely what we mean by "terms" below: for the moment just assume they are words. A document is then represented as a n dimensional vector $\overrightarrow{t_d}$. Let tf (d, w) denote the frequency of term $w \epsilon W$ in document $d \epsilon D$. Then the vector representation of a document d is: td = (tf(d, w1), . . tf(d, wn). words like a, the, an are probably the most frequent words that appear in English text, but neither are neither descriptive nor so important for the document's subject. Documents presented as vectors, we measure the degree of similarity of two documents as the correlation between their corresponding vectors, which can be further quantified as the cosine of the angle between the two vectors. Terms are basically words. But we applied standard transformations on the basic term vector to represent in keyword vector. First, we removed stop words. There are words that are non-descriptive for the topic of a document, such as a ,and, are and do words were stemmed using Porter's suffix-stripping algorithm [13], so that words with different endings will be mapped into a single word.

Cosine similarity:
When documents are represented as keyword vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors that is, called as cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval

applications .Given two documents $\overrightarrow{t_x}$ and $\overrightarrow{t_y}$ their cosine similarity is;

$$Sim(\overrightarrow{t_x}, \overrightarrow{t_y}) = \frac{\overrightarrow{t_x} \cdot \overrightarrow{t_y}}{|\overrightarrow{t_x}| \times |\overrightarrow{t_y}|} \qquad (2)$$

Where $\overrightarrow{t_x}$ and $\overrightarrow{t_y}$ are n-dimensional keyword vectors over the term set T = *w*1, ...,*wm*. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is nonnegative and bounded between [0, 1].

3) Link similarity:

Hyperlinks inside HTML pages contain a wealth of information about the relationships among webpages. Kleinberg introduced the concepts of "authorities" and "hubs". His article presented on an analysis of the link structure which states that "a link recognizes authority of the other document". The main statement is that those conferring the recognition are called "hubs" and those receiving the recognition are called "authorities".[14] Our definition of core is similar to the idea of a hub. In this section we are primarily interested in the similarity based on the hyperlink structure among the pages.

Successor Check: If we have two web pages linking to the same web page, we may also consider these two pages are similar. This leads us to the second definition.

Definition: Given a set of web pages W and two pages
x and y in W, the similarity of these two pages is:

$$\sigma_1(x, y) = \frac{|suc(x) \cap suc(y)|}{|suc(x) \cup suc(y)|} \qquad (3)$$

If denominator != 0. Otherwise, σ 1(*x, y*) = 0.

4) Earth Mover's Distance:

In [15] have given, EMD is method to evaluate the dissimilarity or distance between two signatures. A signature is consisting of set of features and their corresponding weights. This method comes from the well-known transportation/consumer producer problem it has been practically proved that EMD is advantageous in representing problems involving multifeatured signatures.EMD allows for partial matches in a very natural way and is especially fit for cognitive distance evaluation. To calculate EMD, give input as a two Url. It consists of following tasks.

a) Page Processing and Signature Generation:

The task of our Web page preprocessing approach contains three procedures: i) obtain the image of a Web page from its URL, ii) perform normalization, and iii)represent the Web page image into a Web page visual signature(consists of color and coordinate features), which is used to evaluate the visual similarity of a pair of Web pages.

- *Web Page Rendering Process:* The process of displaying a Web page in a Web browser on the screen from HTML and accessory files (including images, flash movies, activeX plugins, java Applets, etc.)is the Web page rendering process. We use webscreencapture to get Web page images (in png format).
- *Perform Normalization:* The images of the original sizes are processed into images with normalized size (e.g.10*10) The Lanczos algorithm is used to calculate the resized image because the Lanczos algorithm has very strong antialiasing properties in Fourier domain, and it is also easy to be computed in spatial domain. Lanczos algorithm is used to calculate resized image. Sharp images can be generated with the Lanczos algorithm as intuitively, the sharp images could provide better signature for identification than others. We store the normalized images to present the signature of each Web page.
- *Signature Generation:* A signature of an image is a feature vector which can effectively represent the web page image. The signature of an image in our approach is comprised of features and their corresponding weights. A feature is comprised of a color and the centroid of its position distribution in the image. The feature-weight tuples in Si are ranked in the descending order of their weights.

The color of each pixel in the resized images is represented using the ARGB (alpha, red, green, and blue) scheme with 32 bits. A color can be represented with a 4-tuple < *A,R, G,B* > . However, this is a huge color space, which includes 2^{32}=4, 294, 967, 296 colors. In practice, use a degraded color space to represent the signature of an image .Define the Color Degrading Factor (DF) to be the scale of each color

component making a change. Thus, we have $(2^8/DF)^4$ colors in our degraded color space. A degraded color can be represented as:
<A-(A Mod DF), B-(B Mod DF),C-(C Mod DF),D-(D Mod DF)>
For example, when DF = 32, we have 4,096 colors in the degraded color space. The centroid of each degraded color is calculated using

$$Cen = \sum_{i=1}^{N} \frac{Cen,i}{Ndc} \qquad (4)$$

Cen is the centroid of degraded color dc, Cen,i is the coordinates of the i[th] pixel that has degraded color dc, and *Ndc* is the total number of pixels that have degraded color dc, i.e., the frequency of dc. A feature F, which has degraded color dc, can be represented with dc and Cen, *F =<dc,Cen >* .The weight corresponding to this feature is the colors frequency Ndc. A complete signature Sc is represented as:
*Sc =<< Fdc*1,*Ndc*1 *>,< Fdc*2,*Ndc*2 *>, ...,<Fdcn,Ndcn >>*
Where N is the total number of degraded colors. The feature-weight tuples in Sc are ranked in the descending order of their weights, i.e., for. In our approach, we do not use all of the features. We choose the first Nsi most frequent colors in Sc to be the signature, where Nsi is less or equal to N, and we denote it as Ssi. When N is less than Nsi, Sc is chosen to be exactly Ssi.

b) Computing Visual Similarity From EMD:
The distance matrix *Ds* = [*dmij*] $1 \le i \le m$ *and* $1 \le j \le n$ is defined in advance using a straightforward way. First calculate the normalized Euclidean distance of the degraded ARGB colors, and then calculate the normalized Euclidean distance of centroids. The two distances are added up with weights a and b, respectively, to form the feature distance, where *a + b* = 1.Suppose we have feature α_i =< *dci,Cdi* > where
dci =< dAi, dRi, dGi, dBi > feature α_j =< *dcj ,Cdj* >
Where *dcj =< dAj, dRj, dGj, dBj* > ,
MDcolr =||<MaxA − 0,MaxR − 0,MaxG − 0,MaxB − 0 >|| where MaxA, MaxR, MaxG, and MaxB are the maximum
numbers of the four components of ARGB, respectively, in the specified color space, and the maximum centroid distance, $\sqrt{w^2 + h^2}$ Where, w and h are the width and height of the resized images, respectively. The normalized color distance *NDcolr*(*dci*,*dcj*) is defined as

$$ND_{colr(dci,dcj)} = \frac{\sqrt{(dci - dcj) * (dci - dcj)^T}}{MD_{colr}} \qquad (5)$$

The normalized centroid distance *NDcen*(*Ceni*,*Cenj*) is defined as

$$ND_{cen(ceni,cenj)} = \frac{\sqrt{(ceni - cenj) * (ceni - cenj)^T}}{MD_{cen}} \qquad (6)$$

The normalized feature distance between α_i and α_j is defined as $ND_{feature(\alpha i,\alpha j)}$

$$ND_{feature(\alpha i,\alpha j)} = a.ND_{colr}(dci, dcj) + b.ND_{Cen}(Ceni, Cenj) \qquad (7)$$

So far, *Ds = dmij*, where *dmij* = $ND_{feature(\alpha i,\alpha j)}$ can be calculated before performing EMD calculation. Suppose we have signature *Ss,x* and signature *Ss,y* where *Ss,x* has m features and *Ss,y* has n features. The flow matrix *Fxy* = [*fmij*] $1 \le i \le m$ *and* $1 \le j \le n$ can be calculated through linear programming and the EMD between *Ss.x* and *Ss.y* can be calculated as:

$$EMD(S_{s.x}, S_{s.y}, D) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} fm_{ij}.dm_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} fm_{ij}} \qquad (8)$$

βϵ(0,+1) is the amplifier visual similarity. Use β to make visual similarity to be better distributed in (0, 1) rather than too dense at either side without affecting the ranking relationship of the visual similarity values of Web pages. After performing the web page similarity assessment we get four similarity values for each method.So, to get the exact similarity value for every pair, fusion takes place, in which weight is equally distributed for each method. The similarity matrix is formed from that similarity values. Group of similar web pages are get by clustering with maximal subgraph component. From that group only webpage is selected by most visited site count or hit count. Number of groups found with threshold value, but from every group only webpage is taken in final result set.

4. Experiment

For the experimental performance of the given system some queries are given to the system. Then search results are get retrieved by using bing search engine. The first 10 urls are taken as input to get the 10*10 fusion similarity matrix. This matrix is obtained by web page similarity assessment methods i.e. fusion of layout, text, link and EMD value each having equivalent weight. For the given experiment it is necessary to store the webpage image and webpage so for that web screen capture API is used. All the methods similarity value must be taken between 0 and 1(0 i.e. similar 1 i.e. non similar).If fusion value is 0 then two web pages are similar or if it's 1 then it is non-similar. Clustering is performed by using connected component by using threshold value. Component having more than 1 element is displayed to user. Component having more than 1element means set of similar web pages.

5. Result Discussion

There are two main part of the system first is web page visual similarity assessment, i.e. used for web page comparison and get the similarity value. This is compared with two popular tools viz. SEO tool and copyscape. This is shown in Table 1.

Table 1: Result Analysis (a)

URL1	URL2	Layout	Text	Link	EMD	Fusion	SEO Tool	copyscape
gmail.com	gmail.com	1	1	1	0	1	1	1
pune.olx.in	Baramati.olx.in	0.15	0.66	0	0.43	0.13	0.64	0.62
google.co.in	google.co.au	0.65	0.81	0.08	0.39	0.53	0.77	0.75

The first two columns are for urls and layout, text, link, EMD are for their similarity. If EMD value is 0 means distance between two web page is 0 and they are similar, and for others it's non similar. The seo tools compare the web pages based on text and link similarity, while copyscape is based on text similarity. Fusion column represent the similarity of proposed work. Proposed system considered all the parameters like text, hyperlink, layout, image distance.

In second part after assessment of web pages we obtain the similarity matrix. Let's consider the query search sites retrieved results as shown in Fig.2.The obtained result shows there are redundant urls occurred. So after setting the threshold value 0.25 we obtain the connected component elements i.e. 1, 3, 4 and 5, 6, 10. Which mean component related web pages are 25% similar or more than it.

Fetching https://www.bing.com/search?q=search sites...
URLs to be Featch Are
1)http://www.google.com/
2)http://mail.google.com/mail/?hl=en&tab=wm
3)https://www.google.com/search
4)http://www.google.co.in/
5)http://en.wikipedia.org/wiki/List_of_search_engines
6)http://www.makeuseof.com/tag/4-ways-monitor-search-engine-health-site/
7)http://www.thisismoney.co.uk/money/bills/article-2611313/How-energy-switching-sites-hide-deals-wont-make money.html?ns_mchannel=rss&ns_campaign=1490
8)http://www.bgr.in/search/myntra/
9)http://in.search.yahoo.com/
10)http://www.freefind.com/

Figure 2: Result Analysis (b)

So, the final summarized result for the given query contains two groups, the groups i.e. the web pages images are get displayed to user.

6. Conclusion and Future Work

The system has been carried out to improve the search results efficiency. It is very novel approach for web search results summarization. It helps the users to retrieve the exact and accurate result in short time rather to check all results produced. It compares web pages with text data, layout, hyperlink, pixel level. This system also removes the redundant results, which is also applicable in phishing detection, copyright checker like systems. More number of retrieved result consideration, and testing on queries of different areas is the future work of the system to get improved accurate result.

References

[1] Radev.D.R and Weiguo Fan (2000). "Automatic summarization of search engine hit lists", Proceedings of the ACL-2000 workshop on

Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11, pp 99-109.

[2] Thomas Mandl,(2006). "Implementation and Evaluation of a Quality-Based Search Engine", ACM-HT'06, August 22–25.

[3] Yitong Wang and Masaru Kitsuregawa. (2001) "Clustering of Web Search Results with Link Analysis, Second International Conference on Advances in Web-Age Information Management (WAIM 2001).

[4] J.Y.Delort, B. BouchonMeunier and M. Rifqi. (2003). "Enhanced Web Document Summarization Using Hyperlinks" ACM HT'03, August 26–30.

[5] E. Kirda and C. Kruegel. Protecting Users against Phishing Attacks. The Computer Journal, 2006.

[6] Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu, Senior Member, IEEE, "Textual and Visual Content-Based Anti-Phishing:A Bayesian approach", IEEE Transactions on Neural Networks, VOL. 22, NO. 10, October 2011.

[7] G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.

[8] Nn Haveliwala, T., H., Glonis, A., Klein, D., Indyk, P. Evaluating strategies for similarity on the web. In Proceedings of WWW11, 2002.

[9] Netscape Communications Corporation. 'What's Related FAQ' web page. http://home.netscape.com/escapes/related/faq.html.

[10] E. Levina and P. Bickel,"The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics," Proc. IEEE Int'l Conf. Computer Vision, vol. 2, 2001.

[11] L. Wood, Document Object Model Level 1 Specification, http:// www.w3.org, 2005.

[12] Angelo P. E. Rosiello, Engin Kirda, Christopher Kruegel, and Fabrizio Ferrandi,"A Layout-Similarity- Based Approach for Detecting Phishing Pages", 2008.

[13] M. F. Porter "An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130âAS137, 1980.

[14] Shou-Hsuan Stephen Huang, Jesus Ubaldo Quevedo Torrero, Carlos Humberto Molina-Rodriguez ,Mario Francisco Fonseca-Lozada," Exploring Similarity among Web Pages Using the Hyperlink Structure", International Conference on Information Technology, 2005.

[15] Anthony Y. Fu, Liu Wenyin, Senior Member, IEEE, and Xiaotie Deng, Senior Member, IEEE," Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)",IEEE Transactions on Dependable And Secure Computing, VOL. 3, NO. 4, October-December 2006.

Vanita V. Sawant received the B.E.(CSE) degree from Solapur University,India in 2008.She is now pursuing M.E.(computer) degree from VPCOE,Baramati, Pune University, India. Her research interests include Web Mining and Web page similarity

Sheetal Ajay Takale has obtained her B.E. in 1998 and M.E. in 2005 in Computer Science and Engineering from Walchand College of Engineering, Sangli. She also qualified GATE 2011. She is now pursuing her Ph.D. in Computer Science and Engineering from Shivaji University Kolhapur. Sheetal has worked as a lecturer in Computer Engineering from 1998 to 2002 at M.I. T. Pune and from 2002 to 2007 at VPCOE, Baramati. Since 2007 to till date Sheetal is Assistant Professor and HOD at Vidya Pratishthan's college of Engineering, Baramati, Her research interests includes span Information Retrieval Web Mining and Artificial Intelligence. She is working on various topics related to above fields including similarity and summarization of image and text documents.