https://doi.org/10.26438/ijcse/v6si1.3742

JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Special Issue-1, Feb 2018 E-ISSN: 2347-2693

On Applying Document Similarity Measures for Template based Clustering of Web Documents

T.I. Bagban[1*], P. J. Kulkarni[2]

[1*]Department of Information Technology, D.K.T.E'S T.E.I, Ichalkaranji, India
[2] Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India

[*]*Corresponding Author: tbagban@gmail.com, Tel.: +91-9860980990*

Available online at: www.ijcseonline.org

Abstract**—** World Wide Web is the useful and easy way to get the source of information on the Internet. In order to reduce the content generation and publishing time, templates are used to populate the contents in web documents. Template provides easy access to the web document contents through their layout and structures. However, for search engines, due to its irrelevant terms, the templates degrade search engines accuracy and performance. Also the templates are used by wrapper induction tools used in information extractor to extract and integrate information from various E-commerce sites. Thus it has received a lot of attention to improve the search engines performance and content integration. In this paper we have discussed how heterogeneous web documents i.e. web documents generated from different templates, can be clustered. We have applied document similarity measures to cluster the heterogeneous web documents generated from templates. Our experimental results on real data sets show that cosine distance similarity measure is more suitable for template based clustering of heterogeneous web documents.

Keywords**—**Template, Clustering, Cosine, Jaccard, Agglomerative Hierarchical Clustering

I. INTRODUCTION

World Wide Web (WWW) is widely popular in publishing information on the Web. However designing the rich and informative web documents is time consuming task. Nowadays Content Management Systems (CMS) are used to design web documents. Accordingly, large number of websites are designed using open source CMS tools such as Joomla, blog environments tools such WordPress.CMS systems make use of templates. Templates are document layout structures with empty slots, which are filled with contents from databases. Template generated web documents share common structure and is uniform in look and feel as shown in Fig.1. Templates help readers in easy assessing of information through their consistent structures but for search engines, due to its irrelevant information, templates are harmful as they degrade their accuracy and performance. Templates have received lot of attention in field of web content mining and information extraction. Its removal improves the performance of search engines, it is used in data integration in price comparison sites and so on [10],[1], [11], [12], [6],[13]. Wrappers use templates to extract the contents from web documents through the following steps 1) First all the web documents are clustered based on templates i.e webpages in a cluster are deemed to be generated from same templates. 2) Wrapper induction tools learn the underlying template structure from documents in the cluster. 3) From the web documents in a cluster except template part, relevant contents are extracted.

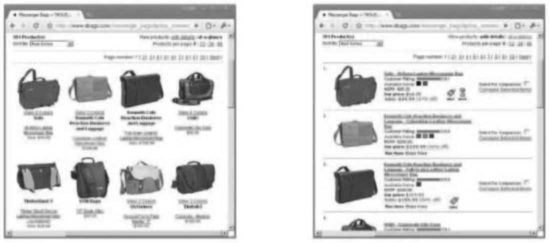
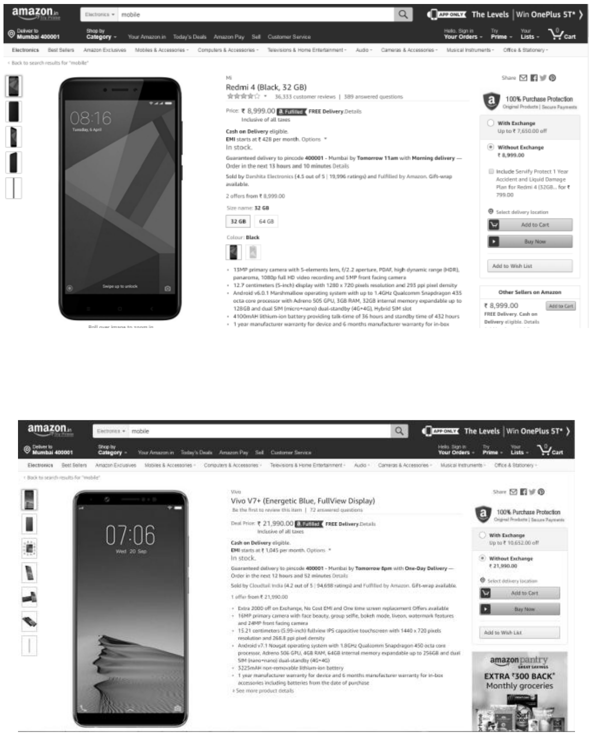


Fig. 2. Two different Template with similar URL

The problem of template based web documents clustering has been studied [10], [14], [10], [15]. Methods [11], [16] suggest that web pages having common URL constituents are generated from same template. However, from Fig. 2, it is observed that given the two web documents look different. In this case, their URLs are identical but the value of a layout parameter is different. If only URLs are used to cluster web documents then these pages will be included in the same cluster though they are from the different templates.



Mi

Redmi 4 (Black, 32 GB)

36,333 customer reviews | 389 answered questions

Price: ₹ 8,999.00 Fulfilled FREE Delivery. Details

Inclusive of all taxes

Cash on Delivery eligible.

EMI starts at ₹ 428 per month. Options

In stock.

Guaranteed delivery to pincode 400001 - Mumbai by Tomorrow 11am with Morning delivery — Order in the next 13 hours and 10 minutes Details

Sold by Darshita Electronics (4.5 out of 5 | 19,996 ratings) and Fulfilled by Amazon. Gift-wrap available.

2 offers from ₹ 8,999.00

Size name: 32 GB

32 GB 64 GB

Colour: Black

- 13MP primary camera with 5-elements lens, f/2.2 aperture, PDAF, high dynamic range (HDR), panaroma, 1080p full HD video recording and 5MP front facing camera
- 12.7 centimeters (5-inch) display with 1280 x 720 pixels resolution and 293 ppi pixel density
- Android v6.0.1 Marshmallow operating system with up to 1.4GHz Qualcomm Snapdragon 435 octa core processor with Adreno 505 GPU, 3GB RAM, 32GB internal memory expandable up to 128GB and dual SIM (micro+nano) dual-standby (4G+4G), Hybrid SIM slot
- 4100mAH lithium-ion battery providing talk-time of 36 hours and standby time of 432 hours
- 1 year manufacturer warranty for device and 6 months manufacturer warranty for in-box

○ With Exchange Up to ₹ 7,650.00 off

● Without Exchange ₹ 8,999.00

□ Include Servify Protect 1 Year Accident and Liquid Damage Plan for Redmi 4 (32GB... for ₹ 799.00

₹ 8,999.00 FREE Delivery. Cash on Delivery eligible. Details

Vivo

Vivo V7+ (Energetic Blue, FullView Display)

Be the first to review this item | 72 answered questions

Deal Price: ₹ 21,990.00 Fulfilled FREE Delivery. Details

Inclusive of all taxes

Cash on Delivery eligible.

EMI starts at ₹ 1,045 per month. Options

In stock.

Guaranteed delivery to pincode 400001 - Mumbai by Tomorrow 8pm with One-Day Delivery — Order in the next 12 hours and 52 minutes Details

Sold by Cloudtail India (4.2 out of 5 | 94,698 ratings) and Fulfilled by Amazon. Gift-wrap available.

1 offer from ₹ 21,990.00

- Extra 2000 off on Exchange, No Cost EMI and One time screen replacement Offers available
- 16MP primary camera with face beauty, group selfie, bokeh mode, liveon, watermark features and 24MP front facing camera
- 15.21 centimeters (5.99-inch) fullview IPS capacitive touchscreen with 1440 x 720 pixels resolution and 268.8 ppi pixel density
- Android v7.1 Nougat operating system with 1.8GHz Qualcomm Snapdragon 450 octa core processor, Adreno 506 GPU, 4GB RAM, 64GB internal memory expandable up to 256GB and dual SIM (nano+nano) dual-standby (4G+4G)
- 3225mAH non-removable lithium-ion battery
- 1 year manufacturer warranty for device and 6 months manufacturer warranty for in-box accessories including batteries from the date of purchase

› See more product details

○ With Exchange Up to ₹ 10,652.00 off

● Without Exchange ₹ 21,990.00

Fig. 1. Two Webpages with same Template

Lot of work in template detection and extraction through clustering has been done but they address the problem of detection and extraction for only web documents belonging to same website or same webpage. This is termed as homogeneous case. Our solution addresses the issue of identifying template class of heterogeneous WebPages through clustering of heterogeneous web documents i.e. web documents belonging to different websites. Our system clusters heterogeneous web documents in such a way that web documents sharing same template becomes a part of same cluster.

In this paper we investigate the problem of template based clustering of heterogeneous web documents and show how structural similarity based clustering techniques can be used to cluster the heterogeneous web documents generated from different templates. In this paper we discuss different distance measures that can be used to reflect template based similarities. The distance measures are applied and evaluated from computational costs and accuracy parameters. The evaluation is based on a 40 large web documents taken from 4 different templates. The system under consideration is a part of research where we plan to make the current system more scalable and efficient in running time using Locality Sensitive Hashing Technique.

In Section II we present an overview of related works in the field of structural similarity based web documents clustering. In Section III we describe various document similarity measuring techniques and Agglomerative Hierarchical Clustering Algorithm (AHCA), In Section IV Experimental Results are presented. Section V concludes our research findings.

II. RELATED WORK

Problem of template based clustering of web documents has been explored by many researchers for more than a decade. Bar-Yossef and Rajagopalan[1] were the first to address the problem of template identification in 2002. They proposed template detection algorithm by detecting recurrent webpage segments which are sub-document elements in a set of webpages. The recurrent webpage segments are extracted via DOM tree segmentation. Lin and Ho[2] who developed Infodiscoverer, based their work on the fact that in web documents frequency of template generated contents is more than main contents. They used the concept of block entropy to filter frequent and redundant DOM blocks. Debnath et al. also used redundant blocks concept in ContentExtractor[3] along with words and text. It also considers other features like images. Yi, Liu and Li[4]adapted Site Style Tree approach by focusing more on visual aspects and concludes that identically formatted DOM sub-trees are template generated. Reis et al[5] used the top down tree mapping algorithm (RTDM) to calculate a tree edit distance between two DOM trees. The RTDM tree edit distance is used to find clusters of different templates. Gibson et al. [6] used a site-level template detection algorithm. Cruz et al. [7] describe several distance measures based on tag vectors, parametric functions or tree edit distances for clustering of web documents. Buttler[8] proved tree edit distances to be the best but most expensive similarity measures in web page clustering. Hence Buttler proposed path shingling approach which makes use of the shingling technique suggested by Broder et al.[9], Chulyun Kim et al[17] proposed automatic template extraction based on heterogeneous web documents. They used information theoretic approach and agglomerative hierarchical clustering in combination with Extended MinHash technique. Since they have used co-matrix that does clustering and extraction at same time it has got limited scalability.

III. METHODOLOGY

System Architecture

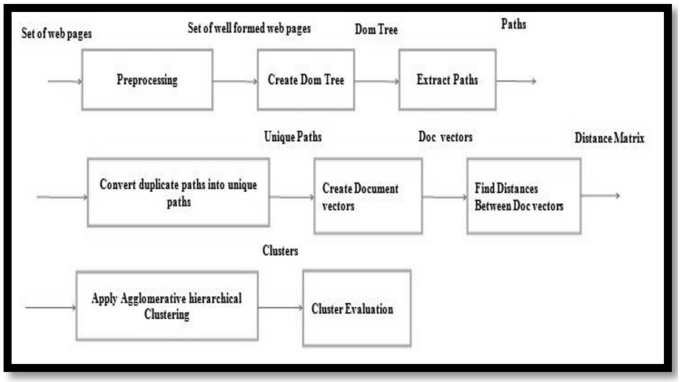


Fig. 3. System Architecture

The proposed system architecture given in Fig.3 consists of following Modules:

3.1 Preprocessing Module

This module converts all the web documents in a directory to well-formed web documents suitable for Dom tree construction.

3.2 Create DOM Tree Module

This module creates DOM tree for each webpage using JTidy package. Input to this module is well formed web pages and output is Dom Tree for each page as shown in Fig. 4.

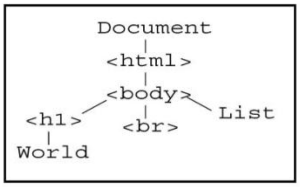


Fig. 4. DOM Tree

3.3 Extract Paths Module

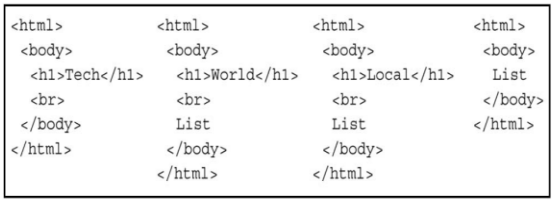
For each webpage Extract Path Module extracts paths from root to elements from the DOM Tree. For above DOM Tree Extracted paths are as shown below
\html\body,\html\body\h1,
\html\body\h1\World,\html\body\br, \html\body\List

3.4 Convert Duplicate Paths into Unique Paths Module

For each webpage Extract Path Module extract paths from root to elements from the DOM Tree. If there exist repetition of extracted paths then those are converted to unique paths by appending extra numeric information.

3.5 Create Document Vectors Module

For each webpage, it removes those paths which occur only in one document or in all documents and then obtain the dictionary of paths with numeric indexes. By removing paths, dimensionality reduction is done. For given set of WebPages in Fig. 5, document vectors are given below:



```
<html>           <html>            <html>             <html>
 <body>           <body>            <body>             <body>
  <h1>Tech</h1>    <h1>World</h1>    <h1>Local</h1>     List
  <br>             <br>              <br>              </body>
 </body>           List              List             </html>
</html>           </body>           </body>
                 </html>           </html>
```

Fig. 5. Sample HTML code for set of web Pages

Doc1-012345678, Doc2-012345910117, Doc3-0123451278, Doc4-0123413

3.6 Find distances between Webpages using various distance measures

In this module documents similarities are calculated using various distance measures. The distance matrix is used to indicate how similar the documents are. The distance 0 between two documents in distance matrix indicates the documents are identical.

Following are various distance measures considered

a) Document Similarity using Levenshtein Similarity

The Levenshtein distance between two words is the minimum number of single character changes like insertions, deletions or substitutions required to convert word in one form to the other. We have implemented it using Wagner–Fischer algorithm which is based on dynamic programming design principles. Its space complexity is O (m) and its time complexity is O (mn) where m and n are number of characters in two words. Here we have calculated distance matrix using Levenshtein distance metric.

b) Document Similarity using Cosine Similarity

In this shingling technique is used to convert document vectors into shingles of size 2 or 3 each. Dictionary of shingles is created and binary document vectors are obtained based on presence or absence of shingles in dictionary. Shingles set of size two is given for document vectors A and B as:

A-012345678 Shingle set of A={01,12,23,34,45,56,67,78}
B-0123413 Shingle set of B={01,12,23,34,41,13}
Vector A-{1111111100}
Vector B-{1111000011}

Cosine similarity of A and B =A.B/|A|.|B|
For each document vector determine its cosine similarity with other document vector. Here we have calculated distance matrix using cosine distance metric

c) Document Similarity using Jaccard Similarity

In this, shingling technique is used to convert document vectors into shingles of size 2 or 3 each. Dictionary of shingles is created and binary document vectors are obtained based on presence or absence of shingles in dictionary. Shingles set of size two is given for document vectors A and B as:

A-012345678 Shingle set of A= {01,12,23,34,45,56,67,78}
B-0123413 Shingle set of B= {01,12,23,34,41,13}
Vector A-{1111111100}
Vector B-{1111000011}

The Jaccard index measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the two documents. The Jaccard similarity coefficient is used for comparing the similarity and diversity of sample sets.

$$\mathsf{JS}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

For each document, vector determines its Jaccard similarity with other document vector. Here we have calculated distance matrix using Jaccard similarity metric. Cost of calculating Jaccard similarity metric is given by O(m+n)

3.7 Applying Agglomerative Hierarchical Clustering(AHC) Module

Once the distance matrix is obtained then AHC technique is used to cluster the various web documents based on distance matrix. Other Clustering technique that can be used is K-Median Clustering. We have considered AHC technique to cluster web documents because hierarchical clustering methods do not require fix number of clusters as a parameter. It starts with each document as a single cluster. The clusters are iteratively merged until only one cluster is obtained. The information about merging of clusters, location of merging is represented in the form of a tree structure called as dendogram.

Agglomerative Hierarchical Clustering Algorithm
Input: Distance Matrix, Approach
Output: Set of Clusters in the form of Dendogram
Procedure

1. Initially each item $x_1, . . . ,x_n$ is in its own cluster $C_1, . . . , C_n$.
2. Repeat the following process until only one cluster is left:
3. Merge the nearest clusters, say C_i and C as per given approach
4. Display Dendogram

There are many ways to decide which clusters are to be merged.

a) The single linkage approach

It merges those two clusters for which the distance between two of the web documents from the two clusters is minimum over all inter-cluster distances.
Given C_i, $C_{j,}$ $x \in C_i$, $x' \in C_j$.
$d(C_i , C_j) = \min_{x \in Ci,\, x' \in Cj} d(x, x')$.

b) The complete linkage approach

It merges those two clusters for which the distance between two of the web documents from the two clusters is maximum over all inter-cluster distances.
Given C_i , $C_{j,}$ $x \in Ci$, $x' \in Cj$.
$d(C_i , C_j) = \max_{x \in Ci, x' \in Cj} d(x, x')$.

We have presented single linkage approach only because it proved to be effective in our case.

IV. Results and Discussion

Experimentation was carried on a set of samples. 40 large size web documents were selected from the data set consisting of web documents from four different areas, namely 1)List of artist from website www.amazon.com site 2)List of Cars from different car manufacture from website www.amazon.com 3) Product information from website www.buy.com 4) Category wise list of product description from website www.buy.com. All four categories of web documents are generated from different templates.

4.1 Running Time Performance

While computing the distance matrices for the 40 large web documents, we measured the time needed to calculate the distance matrix for an increasing number of documents. The graph in figure shows the time in seconds required for calculating distance matrices based on the number of documents involved and using the different distance measures. Based on the graph, we observed that time required to calculate cosine distance matrix is less as compared to Jaccard Index. The time required to calculate distance matrix using Levenshtein distance is exponential to the number of documents as shown in Fig. 6.

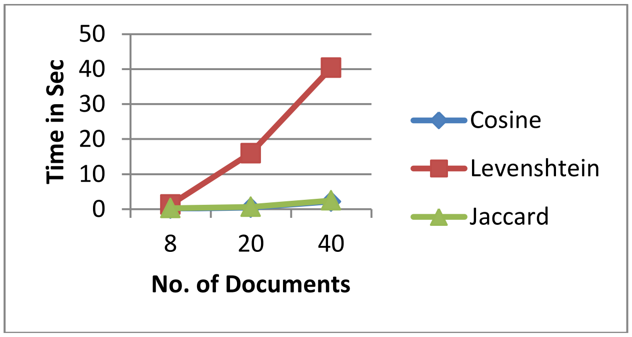


Fig. 6.Time required to calculate the distance matrix with the different similaity measures,with varing number of documents

4.2 Cluster Quality Evaluation

There are different Cluster Quality Evaluation measures like Internal Cluster Quality evaluation and External Cluster Quality evaluation. We have used External Cluster Quality Evaluation measure. Following are the measures we have used.

Here we have considered standard classes of 4 clusters

a)Precision Recall Based Cluster Quality Evaluation

Here we have considered pair wise approach to calculate Precision and Recall.

Precision=TP/(Tp+Fp)
Recall=Tp/(Tp+Fn)

Tp is pair of documents belonging to same class and also belongs to same cluster

Fp is pair of documents belonging to different class but belongs to same cluster

Tn is pair of documents belonging to different class and also belongs to different cluster

Fn is pair of documents belonging to same class but belongs to different cluster

Ideally both precision and recall should be 1.

b) Purity based Cluster Evaluation

Purity is calculated per cluster. It is the measure number of web pages generated from same template that belongs to same cluster. Average purity gives overall purity of clustering technique using average of purity of all clusters.

Given a cluster C_x and $n^{(i)}$ the number of documents in cluster x which according to the ground truth actually belong to cluster i, the purity is:

$$P(C_x) =(1/(\sum_i n_x^i)).\mathrm{Max}_i\, n^i_x$$

Average purity is expected to be 1.

The following table shows quality of clusters obtained by applying ARC algorithm on distance matrix. This is obtained using document Similarity measures like Levenshtein, Jaccard Index and Cosine on sample 40 large size web pages by considering similarity threshold as 0.7 and this is applied to dendogram output.

From the table it is clear that cosine measure performs better than other two measures.

Table 1.Evaluation of clusters obtained with given distance measures and by applying AHC with single linkage for given threshold

Distance Measure	Levenshtein	Jaccard	Cosine
Threshold	0.7	0.7	0.7
No. of Clusters	13	11	10
Avg. Purity	1	1	1
Precision	1	1	1
Recall	0.55	0.80	0.89

V. Conclusion and Future Scope

We compared different distance measures by applying ARC clustering method for clustering web documents which are generated from same template. Levenshtein distance measures perform poor along all parameters, while Cosine distance measure performs better than Jaccard in terms of run time as well as quality of clusters.

We intend to make our system more scalable and accurate by applying it to large number of heterogeneous web documents with more accuracy.

References

[1] Bar-Yossef, Z., Rajagopalan, S,*"Template detection via data mining and its applications"*,WWW '02: Proceedings of the 11th International Conference on World Wide Web, New York, NY, USA, ACM Press 580–591, 2002.

[2] Lin, S.H., Ho, J.M,*"Discovering informative content blocks from web documents"*, KDD '02: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM Press 588–593, 2002.

[3] Debnath, S., Mitra, P., Giles, C.L,*"Automatic extraction of informative blocks from webpages"*, SAC '05: Proceedings of the 2005 ACM Symposium on Applied Computing, New York, NY, USA, ACM Press 1722–1726,2005.

[4] Yi, L., Liu, B., Li, X,*"Eliminating noisy information in web pages for data mining"*, KDD '03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM Press 296–305, 2003

[5] [5] Reis, D.C., Golgher, P.B., Silva, A.S., Laender, A.F,*"Automatic web news extraction using tree edit distance"*, WWW '04: Proceedings of the 13th International Conference on World Wide Web, New York, NY, USA, ACM Press 502–511,2004

[6] Gibson, D., Punera, K., Tomkins, A,*"The volume and evolution of web page templates"*,WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, New York, NY, USA, ACM Press ,830–839,2005

[7] Cruz, I.F., Borisov, S., Marks, M.A., Webbs, T.R,*"Measuring structural similarity among webdocuments: preliminary results"*, EP '98: Proceedings of the 7th international Conference on Electronic Publishing, Artistic Imaging, and Digital Typography,.513 – 524, 1998

[8] Buttler, D,*"A short survey of document structure similarity algorithms"*, IC '04: Proceedings of theInternational Conference on Internet Computing, CSREA Press 3–9, 2004

[9] Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G,*"Syntactic clustering of the web"*,ComputerNetworks 29(8-13) 1157–1166, 1997

[10] A. Arasu and H. Garcia-Molina,*"Extracting Structured Data from Web Pages"*, Proc. ACM SIGMOD, 2003.

[11] M. de Castro Reis, P.B. Golgher, A.S. da Silva, and A.H.F. Laender,*"Automatic Web News Extraction Using Tree Edit Distance"*, Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

[12] M.N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim,*"Xtract: A System for Extracting Document Type Descriptors from Xml Documents",* Proc. ACM SIGMOD, 2000.

[13] Y. Zhai and B. Liu,*"Web Data Extraction Based on Partial Tree Alignment",* Proc. 14th Int'l Conf. World Wide Web (WWW), 2005.

[14] V. Crescenzi, G. Mecca, and P. Merialdo,*"Roadrunner: Towards Automatic Data Extraction from Large Web Sites",* Proc. 27th Int'l Conf. Very Large Data Bases (VLDB), 2001.

[15] K. Vieira, A.S. da Silva, N. Pinto, E.S. de Moura, J.M.B. Cavalcanti, and J. Freire,*"A Fast and Robust Method for Web Page Template Detection and Removal",* Proc. 15th ACM Int'l Conf. Information andKnowledge Management (CIKM), 2006.

[16] S. Zheng, D. Wu, R. Song, and J.-R. Wen,*"Joint Optimization of Wrapper Generation and Template Detection",* Proc. ACMtiSIGKDD, 2007.

[17] Chulyun Kim and Kyuseok Shim,*"TEXT: Automatic Template Extraction from Heterogeneous Web Pages"*,IEEE Transaction on Knowledge and Data Engineering, 2011

Authors Profile

Mr.TanveerI.Bagban pursed Bachelor of Engineering from Shivaji University, Kolhapur in year 2001, Master of Engineering from Shivaji University, Kolhapur in year 2007. He is currently pursuing Ph.D. and currently working as Associate Professor in Department of Information Technology, D.K.T.E T.E.I. Ichalkaranji. He has published 7 papers in International Journal and 4 papers at international conferences. His research interest includes WebMining and Theory of Computation. He worked as expert faculty to teach course Operating System at Busitema University Uganda, Africa. He worked as Board of Studies Member for Rajarambapu Institute of Technology, Islampur. He is working as member of Departmental Academic Advisory Board for Annasaheb Dange college of Engg. and Tech. Ashta.

Prakash Jayant Kulkarni pursed Bachelor of Engineering from University of Poona in 1979, Master of Engineering in the subject Digital Signal Synthesis from Shivaji University, Kolhapur in 1986, and Ph.D. in Electronics in the subject Digital Image Processing from Shivaji University, Kolhapur in 1993. He is currently working as Professor in Computer Science and Engineering Dept., Walchand College of Engineering, Sangli.He has provided guidance to many PhD students in the areas of Electronics Engineering and Computer Science and Engineering. His research interest includes Computer Vision, Pattern recognition, Artificial Neural Networks, Data Mining, Web mining and Information retrieval. He is a recipient of Best Teacher Award of Maharashtra State Government for the year 2011–2012.