Segmentation Based, Personalized Web Page Summarization Model

K.S.Kuppusamy
Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India
Email: kskuppu@gmail.com

G.Aghila
Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India
Email: aghilaa@yahoo.com

***Abstract*—The process of web page summarization differs from the traditional text summarization due to the inherent features in the structure of web pages comparing with normal documents. This paper proposes a model for web page summarization based on the segmentation approach. The proposed model performs an "inclusive summarization" by representing entities from different portions of the web page resulting in the miniature of the original page, termed as "Micro-page". With the incorporation of personalization in the summarization process, the micro-page can be tailored based on the user preferences. The empirical validation of the proposed model is carried out with the help of prototype implementation which depicts encouraging results.**

***Index Terms*— information retrieval, segmentation, web page summarization, personalization**

I. INTRODUCTION

The grasping of a lengthier document can be made simpler and faster by the summarization process. The web pages are a special kind of documents with some inherent additional features like hyperlinks, visual markups and "meta" tags etc. The summarization approaches followed for the traditional text documents need to be enriched with additional components so that they would harness the features present in the web pages, resulting in an enhanced output.

This paper proposes a model for web page summarization based on the web page segmentation. The segmentation process splits a web page into various distinct portions. The summary would be more effective if it includes representative items from these portions. The proposed model encapsulates this benefit by creating the summary as a bottom-up process from the segment level to the page level.

With the incorporation of personalization in the summarization stage makes it possible to render user specific results. Two different users looking at the same web page might expect a different summary based on various factors which includes their area of interest.

The proposed model captures the user interest in the form of profile-keywords. These profile-keywords would provide a valuable input during the summary creation process.

The objectives of this research work include the following:

- Proposing a model for web page summarization based on the segmentation approach.
- Enhancing the proposed model with the inclusion of personalization.
- Validation of the proposed segmentation based personalized web page summarization model with the help of prototype implementation.

The remainder of this paper is organized as follows: Section II would provide the related works carried out in this domain, which formed the basic motivation to propose this model. Section III would illustrate the proposed model and the algorithms. Section IV is about the experiments conducted on the prototype implementation and the result analysis. Section V would illustrate the conclusions and future directions for this work.

II. RELATED WORKS

This section would highlight the related works carried out in this domain. The proposed model includes three major active research domains which are as listed below:

- Web Page Segmentation
- Personalization
- Web page summarization

A. Web Page Segmentation

Web page segmentation is an active research topic in the information retrieval domain in which a wide range of experiments are being conducted. Web page segmentation is the process of dividing a web page into smaller units based on various criteria. The following are four basic types of web page segmentation methods [1]:

- Fixed length page segmentation
- DOM based page segmentation
- Vision based page segmentation
- Combined / Hybrid method

doi:10.4304/jait.3.3.155-161

A comparative study among all these four types of segmentation is illustrated in [1]. Each of above mentioned segmentation methods have been studied in detail in the literature. Fixed length page segmentation is simple and less complex in terms of implementation but the major problem with this approach is that it doesn't consider any semantics of the page while segmenting. In DOM base page segmentation, the HTML tag tree's Document Object Model would be used while segmenting. An arbitrary passages based approach is given in [2]. Vision based page segmentation (VIPS) is in parallel lines with the way how human views a page. VIPS [3] is a popular segmentation algorithm which segments a page based on various visual features.

Apart from the above mentioned segmentation methods a few novel approaches have been evolved during the last few years. An image processing based segmentation approach is illustrated in [4]. The segmentation process based text density of the contents is explained in [5]. The graph theory based approach to segmentation is presented in [6].

B. Personalization

Personalization is the process of customizing based on the user requirements and preferences. There exist many research works to personalize based on user feedbacks. The work presented in [7], proposes a method which utilizes the experiences of the earlier usage. Generally, the personalized result rendering is based upon the "feedback" from the end-users. There exist two types of feedbacks. They are as listed below:

- Explicit feedback
- Implicit feedback

In the explicit feedback mechanism user has to explicitly indicate the relevant and non-relevant items. In the case of implicit feedback it would be gathered automatically based on the actions performed by the user. Here the user is not required to explicitly mark it as relevant or irrelevant. Both these types of feedbacks are discussed in [8], [9], and [10].

An automatic personalization system based on usage mining is depicted in [11]. Aggregate usage profile based web personalization is explored in [12].

C. Web page summarization

The web page summarization is a sub-domain of the text summarization which is also an active research area. The process of summarization can be broadly sub divided in to two types. They are as listed below:

- Extractive summarization
- Abstractive summarization

In the case of extractive summarization the candidate sentences are chosen from the original text to form the summary. In the abstractive approach novel sentences are created based on the semantics. This approach is more complicated and employs various Natural Language Processing (NLP) techniques [13].

The research work explained in [14] falls under the extractive summarization technique. In the case of extractive summarization the candidate sentences are chosen based on their ranks. Sentences with higher ranks would be chosen as part of the summary based on the compression ratio.

There exist certain additional features associated with the web pages comparing the normal documents. So the summarizer for web pages needs to exploit these features to provide a better summary. An approach based on the usage of click through data while summarizing web pages is provided in [15].

The approach illustrated in [16] utilizes the hyperlinks in the web pages to enhance the summarization process.

III. THE MODEL

This section would illustrate the proposed model for web page summarization using segmentation. The Fig.1 illustrates the proposed model with various components in it.

The proposed model receives the source web page as input. This source web page needs to be segmented to carry out the summarization, as shown in (1).

$$P = \{S_1, S_2, S_3 \ldots S_n\} \tag{1}$$

The segmentation is carried out so that the segments cover the entire page and there exist no overlap among the segments. This is illustrated in (2), (3) and (4) as two criterion.

Criteria 1: During segmentation the components are selected such that they are non-overlapping.

$$\forall(s_i, s_j): s_i \cap s_j = \text{NULL}?\ i, j = (1, k) \tag{2}$$

Criteria 2: Segmentation incorporates all parts of the web page.

$$s_1 \cup s_2 \cup s_3 \cup \ldots . s_k = P_i \tag{3}$$

$$P_i = \bigcup_{j=1}^{k} S_k \tag{4}$$

The above two criteria ensures that all portions of the web page is covered and there is no overlap among them.

The summarizer has to take these segments as input. The summarization process is carried out on these segments individually. The summarization task on each of these segments is carried by incorporating four critical factors. They are as listed below:

- Segment Weight
- Luhn's Significance Factor
- Profile Keywords
- Compression ratio

The summarization on each of the segments would be based on this quadruple as shown in (5)

$$\Sigma = (\varepsilon, \phi, \gamma, \eta) \tag{5}$$

The four parameters specified in (5) represent the above specified four factors.

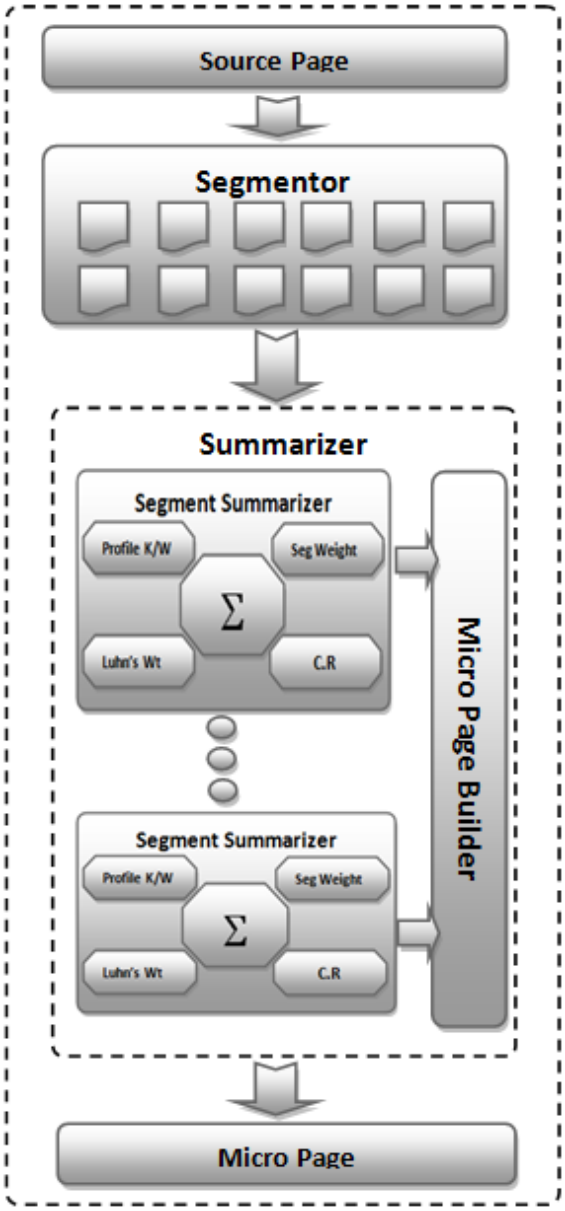

Figure 1.The Proposed Web Page Summarizer Model

A. *Segment Weight*

To calculate the segment weight, a customized version of our earlier model [17] is used. The segment weight for each of the segments is calculated as a sum of four different weights as shown in (6).

$$\omega(s_i) = (L,T,V,M) \quad (6)$$

Where

- L represents the link weight
- T represents the Theme weight
- V represents the Visual Feature weight
- M represents the Image weight

To calculate the above specified four weight factors the following steps are followed:

Step 1: Remove the stop words from the web page

$$P = P - \{w_1, w_2, ..w_n\} \quad (7)$$

Step 2: Sort the words in the descending order based on the word occurrences. The word occurrence is indicated by the | | operator in (8).

$$P = \{t_1, t_2, ...t_n\}, \forall i, i-1 \,|\, t_i \,|\lhd\, |\, t_{i-1} \,|\, i = 1..n-1 \quad (8)$$

Step 3: Consider the top N terms from the above list which is termed as page seed array, β as shown in (9). The value of N can be selected so that it reflects the top ten percentages of the terms extracted.

$$\beta = \{t_1, t_2, ...t_j\} \quad (9)$$

The number of terms matching between the specified component and the page seed array terms β is used to calculate the four weights specified in (6).

In addition to the terms selected from the content of the page, the keywords in the "meta" tag would also be added to the page seed array β. The addition of this component is carried out because of the fact that the "meta" keywords are a good indicator for theme of the document. So inclusion of this "meta" keyword component would enrich the quality of summarization process.

$$\beta = \beta \cup \{\lambda_1, \lambda_2 ... \lambda_n\} \quad (10)$$

After the construction of the page seed array, the remaining weight components can be calculated based on this page seed array. The link weight calculation is done as shown in (11).

$$L = \{|\, l_i \cap \beta \,| + (|\, syn(l_i) \cap syn(\beta) / 2)\} \quad (11)$$

Where l_i represents the terms in the individual links anchor tag. The "syn" indicates the synonym operation. This process would assign a weight for each link tag. The top n links with maximum weight would be chosen.

The image weight calculation is done as shown in (12).

$$M = \{|\, m_i \cap \beta \,| + (|\, syn(m_i) \cap syn(\beta) \,| / 2)\} \quad (12)$$

Where m_i represent the terms in the "alt" attribute of the image present in that segment. The image with maximum weight can be chosen for the summary.

The "visual feature weight" calculation is done as shown in (13).

$$V = \{|\, e_i \cap \beta \,| * |\, vf \,| + ((|\, syn(e_i) \cap syn(\beta) \,| * |\, vf \,|) / 2)\} \quad (13)$$

Where e_i represents the html elements present in the segment and |vf| represents the weight associated with that visual element. This visual weight feature is calculated to give more weight to elements that have been given additional visual emphasized. For example the text

appears inside the bold tag should be given more weight than the normal ones.

The title of a page is a good indicator of theme of the page. So the contents in the segment which are matching the title should get more importance to be included in the summary. This is depicted in (14).

$$T = \{|e_i \cap t| + |(syn(e_i) \cap syn(t)|)/2)\} \quad (14)$$

In (14), t represents set of terms in the title of the page and e_i represents the elements in the segment.

B. Profile Keywords

The profile keyword is a set that would hold the keywords that represent the user's area of interest. The inclusion of profile keywords in the summarization process makes the summary to be tailor made according to the preferences of the user.

The elements in the segment which contains the terms from both the page seed array and profile keywords should be given more weight in the summary creation process.

$$\gamma = \{|e_i \cap [K \cup \beta]| + (|syn(e_i) \cap [syn(\beta) \cup syn(K)]|/2)\} \quad (15)$$

In (15), K represent the terms from the profile keywords.

C. Luhn's Significance Factor

The Luhn's algorithm [18] for auto summarization is a well known statistics based summarization method. The Luhn's formula is utilized to calculate the importance of sentences present in a document based on the distance measure of important words in that document.

The proposed model utilizes the Luhn's significance factor to select important sentences from the segment.

$$\phi = \{LS(S_i)\} \quad (16)$$

In (16) LS represents the Luhn's significance factor. The set ϕ would hold the Luhn's significance factor of each of the sentences in that segment.

D. Compression Ratio

The compression ratio is an important factor that decides the final contents of the micro-page. In all the above steps, different sets have been derived with weights associated with their elements. The compression ratio would decide the number of entities to be selected from the derived sets.

The final summary would be formed by selecting the top ranked items whose count would be decided by the compression ratio.

$$\Sigma = \left\{\left\lceil \frac{\eta|L|}{100} \right\rceil L \cup \left\lceil \frac{\eta|T|}{100} \right\rceil T \cup \left\lceil \frac{\eta|V|}{100} \right\rceil V \cup \left\lceil \frac{\eta|M|}{100} \right\rceil M \right\} \cup \left\lceil \frac{\eta|\gamma|}{100} \right\rceil \gamma \cup \left\lceil \frac{\eta|\phi|}{100} \right\rceil \phi \quad (17)$$

The compression ratio η is multiplied by the number of items in that set and divided by hundred. The resultant value of the above calculation is used as the threshold value to select the top "n" items in individual set.

The algorithm for the above specified model is as given below:

```
Algorithm SegmentSummarize
Input  : Page P, Segment Si , Profile Keywords K,
         Compression ratio η , Page Seed Array β
Output : Segment Summary

Begin
  For each link li in the segment
    begin
      Linkweight[li] = |li ∩ β| +(| syn(li) ∩ syn(β) / 2)
    end
  For each image mi in the segment
    begin
      Imageweight[mi] = |mi ∩ β| +(| syn(mi) ∩ syn(β)| /2)
    end
  //calculate the visual feature weight
  For each element ei in the segment
    begin
      visualweight[ei]=
                |ei ∩ β|*|vf| +((| syn(ei) ∩ syn(β)|*|vf|) / 2)
        end
  For each element ei in the segment
    begin
      themeweight[ei]= |ei ∩ t| +|(syn(ei) ∩ syn(t)|) / 2)
    end
  //calculate the Lunh's significance factor for
  //segment sentences
  for each sentence in segment
    begin
      φ [sni]= LS(sentence)
    end
  //identify the elements to be present in the summary
  links_count  = η * |L| /100
  image_count  = η  * |M| /100
  vf_count     = η  * |V| /100
  theme_count  = η * |T| /100

  profile_count = η * | γ | /100
  luhn's count  = η  * | φ | /100

  segsum = segsum +  topn(links[links_count])
  segsum  = segsum +topn(images[image_count])
  segsum  = segsum + topn(V[vf_count])
  segsum=segsum+topn(T[theme_count])
  segsum  = segsum + topn( γ [profile_count])
  segsum = segsum + topn( φ [luhn's count])

return segsum

End
```

The SegmentSummarize algorithm uses the page seed Array as an input. The algorithm BuildPageSeedArray is used to construct the seed elements.

Algorithm BuildPageSeedArray
Input : Page-Url PU

Output : Page Seed Array

Begin
//Fetch the Page Contents for PU.
P = fetch_contents(PU)

//Extract the keywords from meta tag.
M = $\{\lambda_1, \lambda_2 ... \lambda_n\}$

//remove stop words from P
$P = P - \{w_1, w_2, .. w_n\}$

//calculate the frequency of occurrence of each word
Freq_array = |occurrence(P)|

//sort the array in descending order
Freq_array = sort_descending(freq_array)

//fetch the top 10% of items from freq_array

n = count(freq_array)

for index = 0 to n * 0.1
top(index) = freq_array(index)

top= freq_array

// merge the array top with meta tag keywords array

$\beta = top \cup M$

return β

End

The "micro page builder" component would receive the "segsum" from above algorithm as input and build the target summary page.

III. Experiments and Results

This section would highlight the experimental setup used for the validation of above mentioned model and algorithms. The prototype implementation is done with the software stack including Linux, Apache, MySql and PHP. For client side scripting JavaScript is used. With respect to the hardware, a dual processor system with 3 GHz of speed and 4 GB of RAM is used. The internet connection used in the experimental setup is a 64 Mbps leased line.

The Fig.2 shows a sample page to be summarized. The screenshot in Fig.3 shows the user interface of the proposed system with a text box to get the url from the user; a combo box with compression ratio values listed in the order of 10; a command button to initiate the summarization. The micro-page is displayed in the same screen once the server finishes the task and sends the output back to the client.

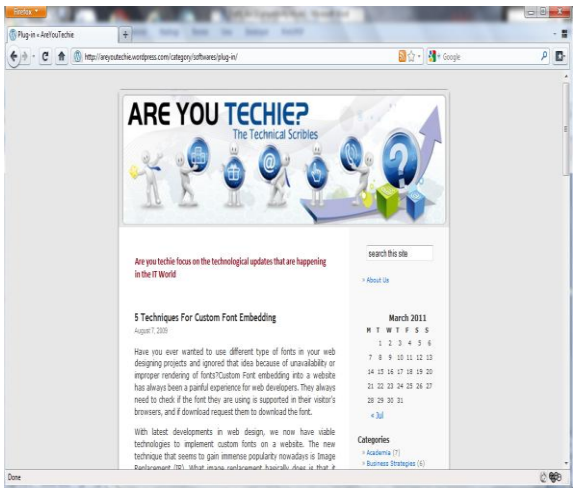

Figure 2. The Source Web Page

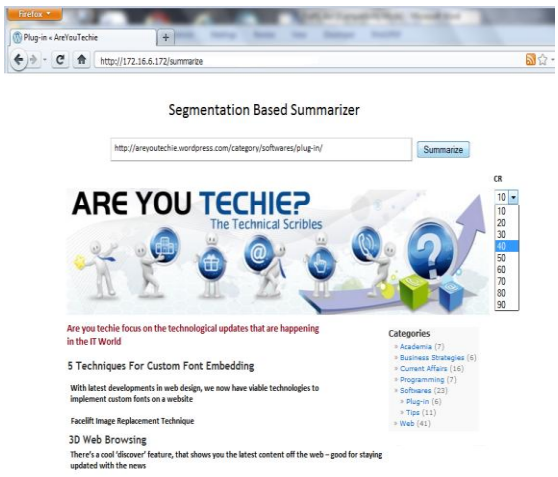

Figure 3. The Micro Page

To validate the proposed model a set of experiments were conducted. The values listed in Table. I correspond to twenty sample experiments of user I. The column SSP indicates Segments in Source Page, SMP indicates Segments in Micro-Page, ISP indicates Images in Source Page, IMP indicates Images in Micro-Page, LSP indicates Links in Source Page and LMP indicates Links in Micro-Page.

TABLE I. EXPERIMENTAL RESULTS USER I

Page	SSP	SMP	ISP	IMP	LSP	LMP
1	25	20	4	2	14	7
2	27	22	3	1	12	5
3	15	14	6	4	14	6
4	10	8	5	5	12	5
5	5	5	2	0	14	6
6	17	16	6	3	15	11
7	12	8	4	1	25	12
8	8	6	3	3	12	8
9	11	9	1	1	10	6
10	12	11	0	0	11	8
11	24	20	3	2	14	7
12	25	22	2	1	12	5
13	13	10	6	5	13	6
14	8	6	4	4	11	4
15	6	5	2	1	12	5
16	17	16	5	3	1	1
17	11	8	4	2	25	17
18	7	6	2	2	11	8
19	11	9	1	1	11	7
20	11	11	4	3	12	8

The experimental result analysis for User I is charted in Fig.2. Hence personalization is incorporated in to the summarization process; users with different interest would get different summaries. The purpose of the experiments is to check the fact that the micro page represents elements from various segments of the source page. The mean of the values for SSP and SMP establishes the fact that relevant portions from majority of segments are carried out in to the micro-page.

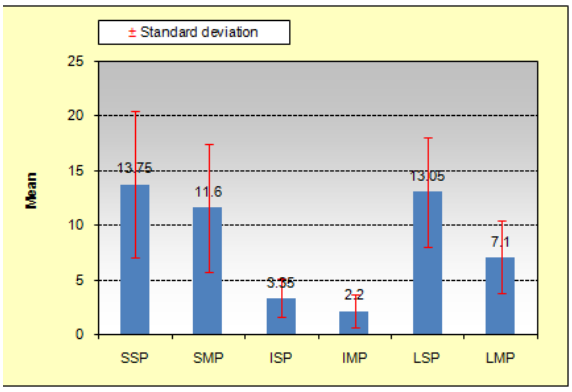

Figure 2. The Mean Analysis for Users I

In Fig.2 the values listed on top of SSP and SMP indicates the mean after clustering. The data set is clustered in to various groups. This clustering is done to illustrate the fact that the proposed model works fine for pages with less number of segments and large number segments as well. With respect to ISP and IMP, only the images satisfying the filtering criteria have become part of the micro-page. A similar criterion is applied to LSP and LMP as well.

For User II and III the results are charted out in Table II and Table II respectively.

TABLE II. EXPERIMENTAL RESULTS USER II

Page	SSP	SMP	ISP	IMP	LSP	LMP
1	25	21	4	4	14	10
2	27	25	3	2	12	6
3	15	13	6	5	14	7
4	10	9	5	5	12	6
5	5	5	2	1	14	8
6	17	13	6	4	15	12
7	12	10	4	2	25	13
8	8	7	3	2	12	8
9	11	9	1	0	10	7
10	12	11	0	0	11	8
11	24	21	3	1	14	13
12	25	23	2	1	12	11
13	13	11	6	4	13	8
14	8	7	4	3	11	4
15	6	4	2	2	12	11
16	17	15	5	5	1	1
17	11	10	4	3	25	10
18	7	5	2	1	11	6
19	11	10	1	0	11	4
20	11	9	4	2	12	9

TABLE III. EXPERIMENTAL RESULTS USER III

Page	SSP	SMP	ISP	IMP	LSP	LMP
1	25	20	4	3	14	12
2	27	24	3	1	12	6
3	15	11	6	4	14	8
4	10	8	5	4	12	10
5	5	4	2	0	14	11
6	17	16	6	5	15	13
7	12	11	4	3	25	13
8	8	6	3	1	12	8
9	11	10	1	0	10	6
10	12	10	0	0	11	8
11	24	22	3	2	14	12
12	25	22	2	1	12	10
13	13	11	6	5	13	6
14	8	6	4	2	11	10
15	6	5	2	1	12	8
16	17	15	5	4	1	1
17	11	8	4	3	25	14
18	7	6	2	1	11	5
19	11	8	1	0	11	7
20	11	9	4	2	12	10

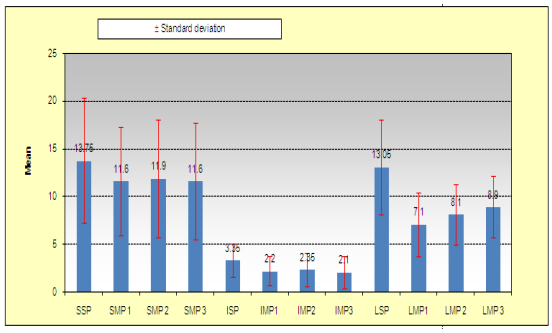

Figure 3. The Mean Analysis for all three users

From the values listed in all the above three tables, Table I, Table II and Table III, it can observed the values of SMP, LMP, IMP varies across the users. The user profile-keywords play an important role in the proposed model in the summary creation process.

V. Conclusions and Future Directions

The proposed model summarizes a page incorporating both segmentation as well as personalization. The derived conclusions are as listed below:

- The web page summarization process carried out by associating segmentation creates a representative micro page which incorporates items from various portions of the web page.
- The summaries generated can be tailor made to suit the needs and preferences of the user.

The future directions for this research work are as listed below:

- Making the personalization more effective by following ontology based data representation instead of using profile-key word approach.
- Extending this work to include languages other than English.

References

[1] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Block-based web search. In SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 456–463, New York, NY, USA, 2004. ACM

[2] Kaszkiel, M. and Zobel, J., Effective Ranking with Arbitrary Passages, Journal of the American Society for Information Science, Vol. 52, No. 4, 2001, pp. 344-364.

[3] D. Cai, S. Yu, J. Wen, and W.-Y. Ma, VIPS: A vision-based page segmentation algorithm, Tech. Rep. MSR-TR-2003-79, 2003.

[4] Cao, Jiuxin , Mao, Bo and Luo, Junzhou, 'A segmentation method for web page analysis using shrinking and dividing', International Journal of Parallel, Emergent and Distributed Systems, 25: 2, 93 — 104, 2010.

[5] Kohlschütter, C. and Nejdl, W. A densitometric approach to web page segmentation. In Proceeding of the 17th ACM Conference on information and Knowledge Management (Napa Valley, California, USA, October 26 - 30, 2008). CIKM '08. ACM, New York, NY, 1173-1182, 2008.

[6] Deepayan Chakrabarti , Ravi Kumar , Kunal Punera, A graph-theoretic approach to webpage segmentation, Proceeding of the 17th international conference on World Wide Web, April 21-25, Beijing, China, 2008.

[7] Barry Smyth, Evelyn Balfe, "Anonymous personalization in collaborative web search", Information Retrieval (2006) 9: 165–190.

[8] Rocchio, J. "Relevance feedback in information retrieval" in G. Salton (Ed.), The SMART retrieval system: Experiments in automatic document processing (pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall, 1971.

[9] Fox, S., Kamawat, K., Mydland, M., Dumais, S., and White, T."Evaluating implicit measures to improve the search experiences" in ACM Transactions on Information Systems, vol. 23(2), 147-168, 2005.

[10] Jung, S., Herlocker, J.L, and Webster, J. "Click data as implicit relevance feedback in web search" in Information Processing and Management vol. 43, 791-807, 2007.

[11] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. 2000. Automatic personalization based on Web usage mining. Commun. ACM 43, 8 (August 2000), 142-151. DOI=10.1145/345124.345169 http://doi.acm.org/10.1145/345124.345169I.

[12] Mobasher Bamshad., Luo Tao., Nakagawa Miki., "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization" in "Data Mining and Knowledge Discovery", 61-82, 2002.

[13] Yeh Ye et al., 2007. "Document concept lattice for text understanding and summarization". Information Processing & Management 43 (6), 1643–1662.

[14] Nomoto, T., Matsumoto, Y., 2001. A new approach to unsupervised text summarization. In: Proceedings of the 24th ACM SIGIR, pp.26-34.

[15] Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). ACM, New York, NY, USA, 194-201. http://doi.acm.org/10.1145/1076034.1076070

[16] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. "Enhanced web document summarization using hyperlinks" In Proceedings of the 14th ACM conference on Hypertext and hypermedia, pages 208-215, New York, NY, USA, 2003. ACM Press.

[17] K.S.Kuppusamy, G.Aghila, "Museum:Multidimensional Segment Evaluation Model", Journal of Computing, Vol 3, Issue 3, March 2011 (Accepted Paper)

[18] H. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159--165, 1958.

K.S. Kuppusamy is an Assistant Professor at Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India. He has obtained his Masters degree in Computer Science and Information Technology from Madurai Kamaraj University. He is currently pursuing his Ph.D in the field of Intelligent Information Management. His research interest includes Web Search Engines, Semantic Web.

G. Aghila is a Professor at Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India. She has got a total of 20 years of teaching experience. She has received her M.E (Computer Science and Engineering) and Ph.D. from Anna University, Chennai, India.

She has published nearly 40 research papers in web crawlers, ontology based information retrieval. She is currently a supervisor guiding 8 Ph.D. scholars. She was in receipt of Schrneiger award. She is an expert in ontology development. Her area of interest include Intelligent Information Management, artificial intelligence, text mining and semantic web technologies.