A MULTIMODAL APPROACH TO INCREMENTAL USER PROFILE BUILDING

K.S.Kuppusamy[1] and G.Aghila[2]
[1]Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India
kskuppu@gmail.com
[2]Department of Computer Science, School of Engineering and Technology, Pondicherry University, Pondicherry, India
aghilaa@yahoo.com

ABSTRACT

The World Wide Web is the largest distributed information source which is accessed by billions of people all across the world. A unique content source on the web can be accessed by various users for different purposes. Hence it becomes mandatory to capture specific information requirements of each user. This paper proposes a multimodal approach to build profile of users in an incremental manner. The approach proposed in this paper achieves the goal of building user profiles using a hybrid approach. The profile building process is further enriched with the incorporation of web page segmentation technique involving page trees and densitometry. The proposed model extracts the requirement context of the user by utilizing both local and global sources during the profile building process. The experiments conducted on the proposed model confirm the effectiveness of the approach and highlights the channels which carries an edge over others in capturing the user interest.

KEYWORDS
User profile building, incremental profiles, web page segmentation, FOAF

1. INTRODUCTION

The World Wide Web has become the colossal source of information which provides solution to all the informational needs of billions of people across the globe. The web hosts the largest collection of information in a distributed manner. The information retrieval domain has shifted its focus from the problem of "information scarcity" to "information overload". In the early days of web, it was difficult for the users to find the information that they need, but these days the users has to filter out the required information from a large collection of unrelated set.
The web surfing sessions starting with typing a keyword in the search engine to locate the required page rather than directly typing the address in the browser, has become a common task these days. The same keyword supplied by two different users shall mean different things. For example the keyword "Cricket" generally points to the sport Cricket but when supplied by an entomologist it has a higher probability of representing the insect cricket. In order to differentiate between these representations, the search keywords need to be supported with the user profile information. These user profile information would be a critical information source to properly locate the relevant document in alignment with the user's current information requirement context.

This paper proposes a model for representing the user profile which would assist in better disambiguation of the user supplied query there by presenting the relevant information to the user's requirements. The objectives of the proposed research work are as listed below:

DOI : 10.5121/ijwest.2012.3405

- Proposing a multimodal approach to represent the user profiles.
- Incorporating the incremental approach in profile building using local and global context data.
- Supporting the profile building process with the help of the web page segment evaluation techniques.

The remainder of this paper is organized as follows: In Section 2, some of the related works carried out in this domain are explored. Section 3 deals with the proposed model's mathematical representation and algorithms. Section 4 is about prototype implementation and experiments. Section 5 focuses on the conclusions and future directions for this research work.

2. RELATED WORKS

This section explores the related works which have been carried out in this domain. The proposed model incorporates the following two major fields of study:

- Contextual Information Retrieval
- Web Page Segmentation

2.1 Contextual Information Retrieval

The Contextual Information Retrieval is the process of harnessing the knowledge about the user in retrieving relevant information [1]. The retrieval using a collaborative approach is explained in [2]. The interest of the user can be gathered through feedbacks. The feedbacks can be of either implicit or explicit forms. The systems harnessing both these types of feedbacks are explored in [3][4][5]. The implicit feedback based systems are illustrated in [6][7]. Both explicit and implicit approach has its own merits and demerits. In the implicit approach the user need not provide much information. The system gathers the information through the user actions. In the case of explicit approach the users need to provide the information themselves. The proposed approach follows a hybrid method where a combination of both implicit and explicit approaches is utilized.

The user interests are represented using techniques like keyword vectors and concept hierarchies etc [8][9][10]. An effort to represent the profile of the user using Ontology is given in [11]. The context search by reasoning the user context through recent activities performed is explored in [12]. A combinatorial approach on long term and short term interest is explored in [13].

The approach followed in this paper is to utilize the Friend Of A Friend (FOAF) and the Open Directory Project (ODP) Ontology [14][15]. The FOAF is considered to be one of the popular tools in the semantic web domain. It has the capability to represent the user in a machine readable format.

Another advantage is that it can provide a network of users by linking users among themselves through the FOAF files. These FOAF files can be consumed both by humans and by the programs like web crawlers and agents etc.

The proposed model utilizes the fields given in the FOAF specification for representing the initial profile of the user.

2.2 Web Page Segmentation

Web page segmentation is an active research topic in the information retrieval domain in which a wide range of experiments are conducted. Web page segmentation is the process of dividing a

web page into smaller units based on various criteria. Web page segmentation can be performed using various types of techniques ranging from simple fixed length segmentation to advanced semantic segmentation. Segmentation of web pages has applications in various domains like mobile rendering, information retrieval etc. The following are four basic types of web page segmentation methods. They are fixed length page segmentation, DOM based page segmentation, vision based page segmentation and Combined / Hybrid method.

A comparative study among all these four types of segmentation is illustrated in [16]. 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 [17]. Vision based page segmentation (VIPS) is in parallel lines with the way, humans views a page. VIPS [18] 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 [19]. The segmentation process based text density of the contents is explained in [20]. The graph theory based approach to segmentation is presented in [21].

The proposed multimodal approach to incremental profile building utilizes the hybrid segmentation approach. A combination of Page Tree (DOM) and Densitometry based approaches are utilized. The initial block finding is done through the page tree and further sub dividing the inner blocks are performed through the densitometry based approach. The densitometry based approach measures the amount of text present in a fixed area and considers the change in density of the text while performing the segmentation.

3. THE MODEL

This section illustrates the approach used in the proposed multimodal profile building approach which is based on the web page segmentation process. The proposed model involves two different scenarios in building the profile, as listed below:

- Creating the profile of the user for the first time where no action log data is available to further enrich the profile.
- The incremental enrichment of the user profile based on the global and local context data gathered.

The block diagram of the proposed model is as shown in Fig.1. The model involves various components to build the profile. The top layer in the block diagram indicates the FOAF layer. It shows the different fields which would be utilized in fetching the profile keywords. The Segmentor and scorer modules are responsible for performing the hybrid web page segmentation of the source pages indicated by the FOAF fields. The session parser and the activity logger track the user actions and collects the data associated with those user actions. The ODP fetcher gathers the ODP categories for the keywords and maps them. The profile pool stores the user profile keywords.
The initial user credentials are gathered as a Friend Of A Friend (FOAF) ontology file. The FOAF is used in representing the user preferences and links among the users.

The FOAF specification involves various fields. Among these fields the relevant items to this research work are chosen. The important fields which would be utilized in fetching the keywords

related to the users are foaf:topic_interst, foaf:interest, foaf:weblog, foaf:workplaceHomepage. Their usages are as shown in Table 1.

Table 1. FOAF Fields and their Description

Field	**Description**
foaf:topic_interest	The thing of interest to the user
foaf:interest	A page about a topic of interest to this person.
foaf:weblog	A weblog of some thing / person
foaf: workplaceHomepage	A workplace homepage of some person; the homepage of an organization they work for.

The model starts by gathering these fields from the FOAF file of the user. The initial step in the profile building process is to utilize these fields in generating the user specific data. Among the four fields specified in Table.1, the values of three fields are URLs. The foaf:topic_interest would direct point to a thing which is interested to the user.

3.1 Mathematical Model

In (1) α denotes the foaf:topic_interest, β denotes the foaf:interest, δ denotes the foaf:weblog and ε denotes th foaf:workplaceHomepage.

$$\Omega = \begin{Bmatrix} \alpha \\ \beta \\ \delta \\ \varepsilon \end{Bmatrix} \tag{1}$$

The foaf:topic_interest is searched in the Open Directory Projects and categories that are retrieved would give a broader picture on what are all the fields in which the user is interested in as shown in (2).

$$\Omega' = \begin{Bmatrix} \Gamma(\alpha) = \{\lambda_1, \lambda_2 \ldots \lambda_n\} \\ \beta \\ \delta \\ \varepsilon \end{Bmatrix} \tag{2}$$

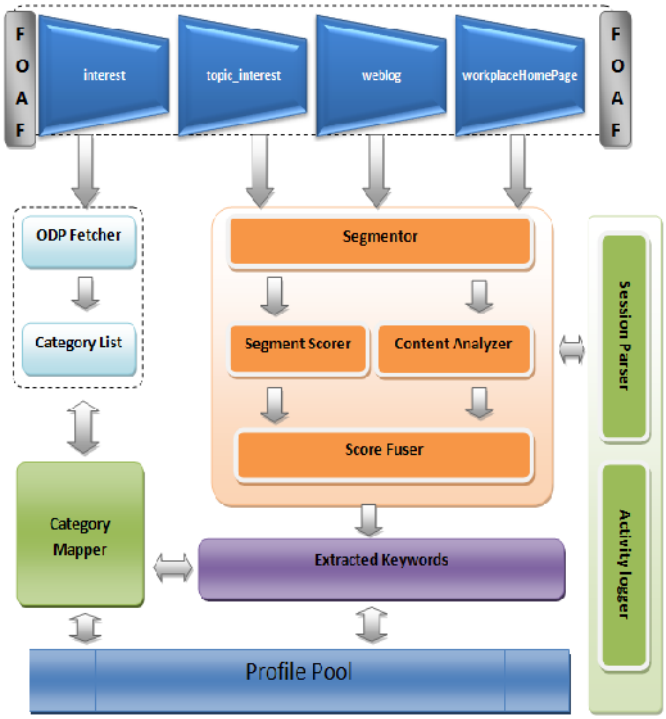

Figure 1. Block Diagram of the Proposed Model

In (2), $\Gamma(\alpha) = \{\lambda_1, \lambda_2 ... \lambda_n\}$ represent the category fetching from the ODP. Each λ_i represent a category fetched from ODP. For the topic_interest, the page pointed is extracted and segmented as shown in (3).

$$\Omega'' = \left\{ \begin{array}{c} \Gamma(\alpha) = \{\lambda_1, \lambda_2 ... \lambda_n\} \\ \Psi(\beta) = \{\omega_1, \omega_2 ... \omega_n\} \\ \delta \\ \varepsilon \end{array} \right\} \quad (3)$$

The $\Psi(\beta) = \{\omega_1, \omega_2 ... \omega_n\}$ represent the segments generated from the source page β. For each of the segments in (3) the score is calculated using a variation of the MUSEUM (Mutli-dimensional Segment Evaluation Model) [22].

$$\Psi(\beta) = \{\forall_{i=1..n} \psi(\omega_i) \oplus \vartheta(\omega_i)\} \quad (4)$$

In (4), $\psi(\omega_i)$ represent the segment score calculated and $\vartheta(\omega_i)$ represent the Content Analysis score. The content analysis would output a weighted array with extracted terms and their weight. There exist many content analysis services like Yahoo! Content Analysis Service [23]. The proposed model utilizes the content analysis service from Yahoo! In extracting the conceptual terms present in the supplied unstructured information source. The $\oplus$ operator indicates the fusion of scores generated by both the components. This approach makes the model to extract personalized keyword from the page identified by topic_interest.

The other two components are also evaluated using the same procedure. The representation for weblog is as shown in (5).

$$\Omega''' = \left\{ \begin{array}{c} \Gamma(\alpha) = \{\lambda_1, \lambda_2 \ldots \lambda_n\} \\ \Psi(\beta) = \{\omega_1, \omega_2 \ldots \omega_n\} \\ \Theta(\delta) = \{\nu_1, \nu_2 \ldots \nu_n\} \\ \varepsilon \end{array} \right\} \quad (5)$$

The representation for "workplaceHomePage" is as shown in (6).

$$\vec{\Omega} = \left\{ \begin{array}{l} \Gamma(\alpha) = \{\lambda_1, \lambda_2 \ldots \lambda_n\} \\ \Psi(\beta) = \{\omega_1, \omega_2 \ldots \omega_n\} \\ \Theta(\delta) = \{\nu_1, \nu_2 \ldots \nu_n\} \\ \Phi(\varepsilon) = \{\kappa_1, \kappa_2 \ldots \kappa_n\} \end{array} \right\} \quad (6)$$

The initial profile after the evaluation all four components are represented by $\vec{\Omega}$. The proposed model incorporates incremental profile enhancement. The initial profile of the user would be further enhanced by logging the user activities.

The locally available context information like browser bookmark is utilized in enhancing the profile at regular intervals.

$$\vec{\vec{\Omega}} = \vec{\Omega} \bigcup \{\forall_{i=1..n} score(\tau_i)\} \quad (7)$$

In (7), each bookmark is specified as τ_i. As each bookmark τ_i is pointing to a web page the procedure stated in (3), (4) can be utilized in scoring.

Apart from these browser bookmarks, the user's interactions with pages can be a critical indicator in understanding the user preferences. The proposed model considers the following parameters in evaluation:

- The pages in which user has spent more than a threshold time limit. This indicates user is interested in topic of that page.
- The pages which got saved by the user to their local hard-disk. This indicates that user is interested in referring the contents of the page.
- The pages which got printed by the user. The user printing a page confirms the fact that user is interested in contents of that page.

$$\vec{\vec{\Omega}} = \vec{\Omega} \bigcup \left\{ \begin{array}{l} \forall_{i=1..n} score(p_i) * k_1 \\ \forall_{i=1..n} score(q_i) * k_2 \\ \forall_{i=1..n} score(r_i) * k_3 \end{array} \right\} \quad (8)$$

In (8), the set of all pages in which user has spent more than the threshold time limit is indicated by "p". The score is calculated for those pages and keywords are extracted to enrich the profile. The pages which got saved by the user to their hard-disk are indicated by "q" and pages which got printed by the user are represented as "r".

It can be observed from (8) that, the scores are multiplied by a scalar weight "k". The weightage for actions performed by the users can be altered. Due to this three different constants are used in (8).

3.2 The Algorithm

The algorithmic representation of the user profile building model is depicted in this section. The algorithm "BuildProfile" illustrates the steps involved in the profile building process based on segmentation.

In the above algorithm the CA service represent the service used for content analysis. In the implementation level the Yahoo! Content Analysis service shall be used.

The algorithm "ReBuildProfile" is used to enhance the profile by monitoring user activities.

Algorithm BuildProfile

Input: FOAF file for user ;

Output : weighted profile terms with categories

Begin

1. Fetch the FOAF file and Parse it.
2. Extract the fields
   - 2.1 Extract "foaf:interest" to α
   - 2.2 Extract "foaf:topic_interest" to β
   - 2.3 Extract "foaf:weblog" to δ
   - 2.4 Extract "foaf:workplaceHomePage" to ε
3. Search α in ODP
4. Fetch the categories for α
5. Fetch terms from β
   - 5.1 Segment the page β
   - 5.2 for each segment in β
     - 5.2.1 compute the segment score
     - 5.2.2 extract terms from segment using CA service
     - 5.2.3 fuse the scores and filter the terms
6. Repeat step 5 with δ
7. Repeat step 5 with ε
8. Merge all terms from step 5, 6, 7
9. For each term t
   - 9.1 Search in ODP
   - 9.2 Fetch the categories
10. Combine all terms, categories
11. Return the weighted term, category vector

End

```
Algorithm ReBuildProfile
Input: Activity Log L;
Output : Profile
Begin
        For each page in L
            fetch the time spend t on page p
            if (t > T) then
               call score (p, k1)
            if page is saved then
               call score(p, k2)
            if page p is printed then
               call score (p, k3)
       end for
      for each bookmark page p
          call score(p)
      update the profile with these terms
end
function score(p, k)
            Segment the page p
             for each segment in p
                  compute the segment score
                  extract terms from segment using CA service
                  multiply the score by k
                  fuse the scores and filter the terms
                  return terms with score s;
end function
```

4. EXPERIMENTS AND RESULT ANALYSIS

This section explores the experimentation and results associated with the proposed model for profile building. The prototype implementation is done with the software stack including Ubuntu Linux , Apache, MySql and PHP. For client side scripting JavaScript is used. With respect to the hardware, an Intel Quad core processor system with ~3 GHz of speed, 8 GB of RAM is used. The internet connection used in the experimental setup is a 128 Mbps leased line. The experiments were conducted with various groups of users covering a diverse range.

The experimental results are tabulated in Table 2. The MTIT stands for mean of terms extracted from item interest, MWBT refers to mean of terms extracted from weblog and MWHT indicates mean of terms extracted from workplaceHomepage.

Table 2. Mean Term Count extracted through various sources

Group ID	MTIT	MWBT	MWHT
1	13.12	17.32	8.45
2	16.45	18.21	9.47
3	18.45	19.53	11.76
4	14.32	14.38	12.45
5	12.76	16.54	15.35
6	11.28	12.67	12.12
7	22.34	25.23	18.43
8	11.78	23.54	14.34
9	16.43	18.12	16.12
10	14.12	15.13	12.12
11	18.75	19.24	13.13
12	14.68	16.32	14.78
13	13.65	18.43	14.21
14	12.79	16.51	15.31
15	11.33	14.71	14.12

The comparative chart is as shown in Fig. 2. It can be observed from the results that the weblog has an edge over other components in retrieving the profile terms. The mean of terms extracted using weblog is 17.72.

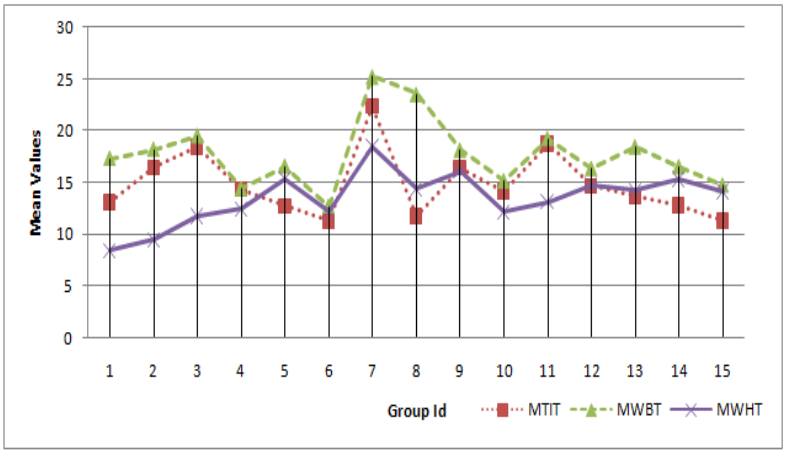

Figure 2. The profile terms chart from three components

The reason for weblog having an edge over other channels is due to the fact that the contents of the weblog are generally written by the users themselves. Hence the content is created by the user himself / herself it has an edge over other sources which are simply pointed by user but not authored themselves.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The following conclusions are derived from the proposed model for user profile building

- The profile building process can be effectively carried out by make it into an incremental process by adapting a multimodal approach.

- The recursive utilization of the segment scoring model can be incorporated in to the user profile building to effectively score the pages and retrieve the keywords.
- The action-log mining data plays a critical role in the keeping the profile of the user updated.

The future directions for this research work include the following:

- The model can be further enriched by introducing concept/ category mapping techniques.
- The number of fields utilized from the FOAF specification can be extended so that it covers a wider spectrum.
- The profile representation can be extended to incorporate the Long Term Interest (LTI) and Short Term Interest (STI) data.
- The collaborative profile building can be introduced by harnessing some of the fields from the FOAF specifications among a group of users.

REFERENCES

[1] Allan, J., al.: Challenges in information retrieval and langage modelling. In: Work-shop held at the center for intelligent information retrieval, Septembre (2002)

[2] Barry Smyth, Evelyn Balfe, "Anonymous Personalization in Collaborative Web Search", Information Retrieval, (2006) 9: 165–190.

[3] 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.

[4] Jung, S., Herlocker, J.L, and Webster, J. "Click Data as Implicit Relevance Feedback in Web Search" in Information Processing and Management, 43, 791-807, 2007.

[5] 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, 23(2), 147-168, 2005.

[6] Mc Gowan, J.P.: A multiple model approach to personalised information access. In:Master Thesis in computer science, Faculty of science, University College Dublin,February (2003)

[7] Vallet, D., Fernndez, M., Castells, P., Mylonas, Ph., Avrithis, Y.: Personalized Information Retrieval in Context. In: 3rd International Workshop on Modeling and Retrieval of Context, Boston USA 16-17 July (2006)

[8] Tamine, L., Boughanem, M., Zemirli, W.N.: Inferring the user's interests using the search history. In:Workshop on information retrieval, Learning, Knowledge and Adaptatbility (LWA 2006), ildesheim Germany november 9 - 11 (2006)108{110

[9] Kim, H. R., Chan, P. K.: Learning implicit user interest hierarchy for context in personalization. In: Proceedings of the 8th international Conference on intelligent User interfaces IUI '03, Miami Florida USA January 12 - 15 (2003)

[10] Liu,F., Yu, C., Meng, W.: Personalized Web Search For Improving Retrieval Effectivenss. In: IEEE Transactions on Knowledge and Data Engineering,Vol. 16(1).,January (2004)

[11] Sieg, A., Mobasher, B., Burke, R.,Prabu, G., Lytinen, S.: representing user information context with ontologies.

[12] Challam, V., Gauch, S., Chandramouli, A.: Contextual Search Using Ontology Based User Profiles. In: Proceedings of RIAO 2007, Pittsburgh USA 30 may - 1 june (2007)

[13] Widyantoro, H, Ioerger, T & Yen J (2000). Learning User Interest Dynamics with a Three Descriptor Representation. Journal of the American Society for Information Science, 52(3),212--225.

[14] Friend Of A Friend (FOAF), http://www.foaf-project.org/

[15] The Open Directory Project, http://www.dmoz.org

[16] 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

[17] 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.

[18] D. Cai, S. Yu, J. Wen, and W.-Y. Ma, VIPS: A vision-based page segmentation algorithm, Tech. Rep. MSR-TR-2003-79, 2003.
[19] 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.
[20] 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.
[21] 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.
[22] K.S.Kuppusamy, G.Aghila, "Museum: Multidimensional Web page Segment Evaluation Model" Journal of Computing, Vol 3, Issue 3.pp.24-27, ISSN 2151-9617
[23] Yahoo! Content Analysis Service http://developer.yahoo.com/search/content/V2/contentAnalysis.htmlZX

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. He has 14 International publications in peer reviewed and indexed journals and conferences.

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 22 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 more than 60 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 includes Intelligent Information Management, artificial intelligence, text mining and semantic web technologies.