International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Personalized Ontology Model for Web Information Gathering

Nandkishor Borse[1], Suraj Patil[2], Nihit Agrawal[3]

[1, 2, 3]SVKM'S NMIMS University, Mukesh Patel School of Technology Management and Engineering,
Department of Computer Science & Engineering, Shirpur, Maharashtra, India

Abstract: ***In personalized web information gathering, for the knowledge description ontology term is use. Mainly ontology used for acquires knowledge, share, reuse and increase relations description of knowledge. Paper shows different problems and searching techniques also related work shows working of different authors on ontology. Main work of ontology is to gather web information based on keywords that may be local repository or global repository. Initialization of information gathering is beginning according to user profile. Also section covers basic architecture of ontology which focuses on overall information gathering. Learning concept extract the information in structured format for unstructured input.***

Keywords: Ontology, Personalization, Information gathering, local profile, learning.

1. Introduction

In a recent, Ontology is an interested concept in a computer science. Ontology which is used for representing the knowledge as a concept [1]. Firstly, we are capable of collecting maximum degree of data from various web sites. Those data is set as a concept in a domain and the relationship assigned between those Concepts. We get it understood that ontology is a structural framework for arranging information which is used in different engineering fields like System Engineering, Software Engineering, Semantic Web, Artificial Intelligence[2].

Ontology firstly was a philosophical concept, but afterwards it was used in computer science. Ontology used for acquire knowledge share, reuse and increase relations description of knowledge. Presently, CAS (Chinese Academy of Science) developed a large commonsense knowledge base "Pangu system", [4] which uses ontology for understanding relation between concepts. Ontology is representing knowledge in the form of general description and specification. It provides a relevant data to be communicated between users and system. By using ontology, system can able to find out meanings of words and phrases and able to differentiate information by using concepts based techniques rather than keywords based techniques.

Ontology is defined by web intelligence community as important techniques which are used for web information gathering [3]. Many researchers now days started concentrating on semantic web for realizing the concept of "knowledge representation"[5]. Intelligent agents use Semantic web to performing complex actions for the users. Ontology plays an initial role in semantic web which share relevant data to the users. By definition, an "Ontology is An explicit and formal specification of a conceptualization" [5].Collecting large amount of useful data from the web is become the one of the challenge for the users. Users are satisfied from web according their interest. So, we need create detail information of user's background knowledge. Getting information individually depends on their knowledge description and formalization which is called as ontology based searching. If we define the Main object of this Web is Increasing information Gathering Performance to collect information from the web which called ontological user profiles. By using Secrete Algorithm Which take information from both global and local base repositories to show user profiles. Secrete Algorithm is created by getting feedback from users on the base of their interest [2].

1.1 General Problems

We observed Information Mismatching and overloading these are the general problems in ontology model for web information gathering:

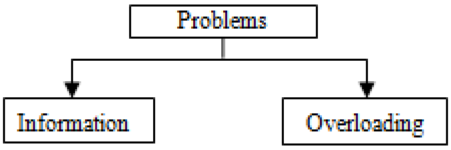


Figure 1: General Problems in Ontology

- **Information Mismatching:** Users are not satisfy from current web information gathering systems because most of systems based on keyword matching mechanism so it missed valuable information at time of information gathering. Which occurred the same topics but represented different syntactic.
- **Overloading:** the same word represent different meaning so if we search specific word we get lot of meaning of that word. E.g. Apple which use for fruit and iMac computer [6].

1.2 Searching Techniques

Keyword based and concept based techniques are used in ontology model for searching information.

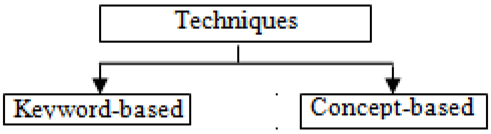


Figure 2: Searching Techniques in Ontology

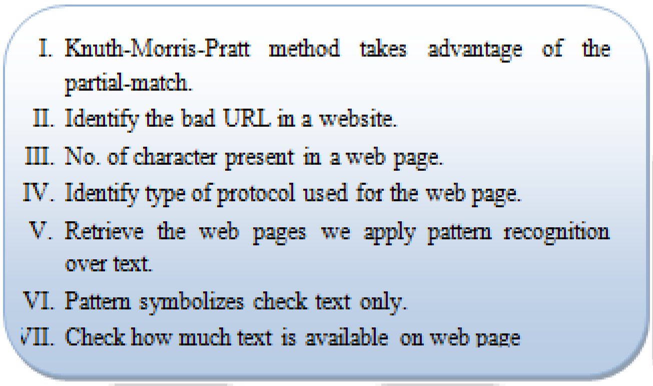
- **Keyword based technique:** It compares the feature vector of documents and Queries if both features match with each other then user satisfy his need.

- **Concept based technique:** This technique match the semantic concept of documents with those Queries [6].

2. Related Work

In [7] author proposed an ontology model for gathering web information which representing user background knowledge by constructing local as well as global search from local instance repository and world knowledge based respectively. Also they compared proposed model with existing baseline model and it showed that proposed model work better due to combination of local and global search. But still there is a issues related investigate the method that generate user local repositories to match the global base.

Xiaohui Tao, Yuefeng Li, and Ning Zhong [11] also provide solution for global and local knowledge in single computational model is available by proposed ontology model. Here main method is to design of web information gathering system. Functions provided by this model are Information Retrieval, web Intelligence, Recommendation Systems, and Information Systems. So here problem comes for content based descriptors for large volume of documents. Here author focus on method that generate local instance repositories to match the representation of global knowledge base. And this method uses strategies like ontology mapping and text classification. Hence author's ontology model sets the experimental set up and used TREC-11 filtering track to evaluate method of persistent user profiles. Focused crawler is developed to extract only the relevant web pages of interested topic from the Internet [12]. For retrieving web pages semantic search technique is used and to finding the web page content KMP algorithm is referred. Here parameters considered are Speed, Query Processing time, accuracy, multidimensional mining, parallel processing, and efficiency. Author's model provides solution to emphasizing global and local knowledge in a single computational model.



I. Knuth-Morris-Pratt method takes advantage of the partial-match.
II. Identify the bad URL in a website.
III. No. of character present in a web page.
IV. Identify type of protocol used for the web page.
V. Retrieve the web pages we apply pattern recognition over text.
VI. Pattern symbolizes check text only.
VII. Check how much text is available on web page

Figure 3: Steps under KMP Algorithm [12]

In [4] Jing Wang, Jianpei Zhang, Ying Wang Proposed new extended ontology model is finding known as Extension Knowledge Ontology Model to achieve communication between knowledge and make computer stronger in associative ability. The action of ontology is to organize knowledge horizontally, and make computer associate the other concepts which belong to the same ontology with it by one concept, thereby achieve associative reasoning of intelligent system.

For e-Government ontology model has three parts: Process meta-model, Organ meta-model, Document meta-model. Process meta-model use of different objects such as flow, function, router and control object [13].

Author Satya Bhanu Jonnalagadda, A. Sravani, Prof. S. V. Achutha Rao implemented semantic relationships ontology model algorithm which is proposed for representing user background knowledge for gathering for web information gathering. Personalized ontologism is constructed by extracting world knowledge and discovering background knowledge from local instance repositories work [2].

3. Related Work

Ontology learning is also called as ontology extraction and it is related with a subtask of information extraction. The main aim of information extraction is to automatically extract structured information from unstructured and / or semi structured documents. Ontology learning is to semi-automatically extract relevant concepts and relation between them in a group which create Ontology [7].

3.1 Local Profiles

We need to create user profiles by collecting personal data of each user which is used in web information gathering. It represents the identity of users. User profiles include description of the characteristics of person [7]. In next section we will go through concepts of some models which are necessary to create user profiles in ontology.

User profiles are divided into three groups:

1) Interviewing
2) Semi-interviewing
3) Non-interviewing

3.2 Models

3.2.1 Ontology Model

In this model, the input was a topic and output was display the data based on user interest. The global and local database it can be linked by using an id. For example in local database if person interested in cricket but he like Indian team so related to Indian team he create his own profile. When that person searches in global database he can get accurate information about Indian team instead of cricket as common search. In this proposed model query execution time and navigation cost is reduced [7].

3.2.2Category Model

Category model is demonstrated by using non-interviewing user profiles. It creates user profile without involving user but it creates user profile by using a set of weighted subjects learned from user's browsing history. In this model there were no relations used like is-a, part-of , related-to and no ontology mining performed .In OBIWAN Model ,When an OBIWAN agent receives the search results for a given topic, it filters and re-ranks the results based on their semantic similarity with the subjects. The similar documents are awarded and re-ranked higher on the result list [7].

3.2.3Golden Model: TREC Model

Interviewing user profiles is to demonstrate by TREC Model which reflects on user concept. TREC Model is used to read relevant information to the given user request. TREC Model looks for the interested topic of the web users and judged each as relevant or non-relevant to the topic. Relevant judgment is provided. Users profile gives an idea about user's personal interests; TREC Model uses set of documents and user profiles for making judgment [9].

3.2.4Baseline Model: Web Model

The web model was the implementation of typical semi interviewing user profiles [9]. In this model user profile is acquired from web by employing search engine.

3.2.5Global Knowledge Representation

In this model user background knowledge is extracted from a world knowledge base which is encoded from the Library of Congress Subject Headings (LCSH). LCSH is used for retrieval of information. World knowledge mainly focuses on the accurateness and effectiveness for representation of set of facts within knowledge domain [7].

4. Methodology

We have used two methodologies:

1) TF-IDF
2) Cosine Similarity

4.1 TF-IDF

In information retrieval, the term frequency-Inverse document frequency also called TF-IDF, is used for calculating how important is a word in a document .TF-IDF is convert textual representation of information into a vector space model.

It is composed by four terms:

Step 1: Calculate term frequency, which is the number of times a word occurs in a document.
Step 2: The normalized term frequency, which is the number of times a word, appears in a document, divided by the total number of words in that documents. For example: In document 1 the term t_i occurs 2 times and the total number of terms in the document 1 is 20.Hence the normalized term frequency is 2/20=0.1 for term t_i in document 1.
Step 3: Then the inverse document frequency, which is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the term t_i appears.
Step 4: Finally compute TF*IDF Weight.

4.2 Cosine Similarity

The cosine similarity is used for measuring similarity between two documents. It is depends on envisioning user preferences as points in space and user interest as points in an N-dimensional space. Now imagine two lines from the origin, or point (0,0,…,0), to each of these two points. When two users are similar, they'll have similar ratings, and so will be relatively close in space at least; they will be in roughly the same direction from the origin. The angle formed between these two lines will be relatively small. In contrast, when the two users are dissimilar, their points will be distant, and likely in different directions from the origin, forming a wide angle. This angle can be used as the basis for a similarity metric in the same way that the Euclidean distance was used to form a similarity metric. In this case, the cosine of the angle leads to a similarity value. If we are rusty on trigonometry, we need to remember to understand this is that the cosine value is always between –1 and 1: the cosine of a small angle is near 1, and the cosine of a large angle near 180 degrees is close to –1. This is good, because small angles should map to high similarity, near 1, and large angles should map to near –1.

Cosine Similarity is same as TF-IDF but only one extra step perform in cosine similarity lather than TF-IDF Algorithm. Extra step measure similarity between two documents.

Cosine Similarity Algorithm given below:

Step 1: Calculate TF.
Step 2: Normalized TF.
Step3: Calculate IDF.
Step4: Calculate TF*IDF.
Step5: perform Cosine Similarity on Documents.

How to calculate cosine similarity between two documents given below:

The set of documents in a collection then is viewed as a set of vectors in a vector space. Each term will have its own axis. Using the formula given below we can find out the similarity between any two documents.

Cosine Similarity (d1, d2) =Dot product (d1, d2) / ||d1|| * ||d2||
Dot product (d1, d2) = d1 [0] * d2 [0] + d1 [1] * d2 [1] * … * d1 [n] * d2 [n]

||d1|| = square root (d1 [0]2 + d1 [1]2 + ... + d1 [n] 2)
||d2|| = square root (d2 [0]2 + d2 [1]2 + ... + d2 [n] 2)

5. System Architecture

We have implement two mechanism for constructing ontology model. In this system, we can retrieve user search query from both database i.e. World Knowledge Database and Local Database by using any one mechanism.

Implemented system architecture given below:

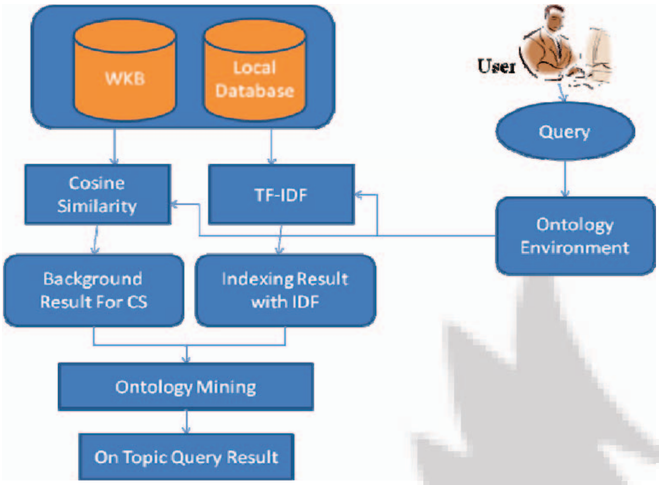


Figure 4: System Architecture

The world knowledge and local user repositories are used in the implemented system. World knowledge is commencing knowledge acquired by people from experience and education. an local instance users repository users personal collection of information items from a world knowledge base here we constructed personalized ontologisms by adopting user feedback on interesting knowledge, a multi-dimensional ontology mining method TF-IDF and Cosine Similarity is introduced in the implemented model for analyzing concepts specified in the ontologisms the users local instance repositories are then used to background knowledge and to populate the personalized ontologism.

6. Results

We have analyzed performance of our implemented techniques such as TF-IDF and Cosine Similarity by considering two factors for performance analysis i.e. Precision-Recall and Execution Time. In TF-IDF, it simply compute weight for search query in each documents and display all documents which is related search query. But, in cosine Similarity, We have used Dot Product which estimates similarity between two documents.

6.1 Precision-Recall

After performing the implementation the result computed is as shown in the following figures. We have used the different user request for both mechanisms one by one and depending on that the result calculated with the precision-recall has been shown in figure 5. In fig. 5 shows, X-axis denotes Recall and Y-axis denotes Precision. We have find out that the precision of cosine similarity mechanism is better than the TF-IDF mechanism.

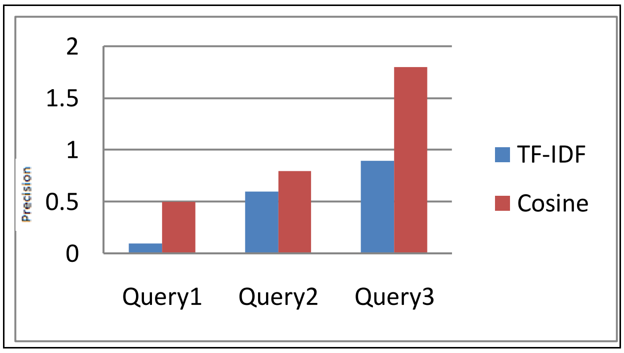


Figure 5: Precision-Recall

6.2 Execution Time

After performing the implementation the result computed is as shown in the following figures. We have used the different user request for both mechanisms one by one and depending on that the result calculated with the Execution Time has been shown in figure 6. In this graph, X-axis denotes number of documents and Y-axis denotes execution time. We have find out that the execution time of cosine similarity mechanism is improved than the TF-IDF mechanism.

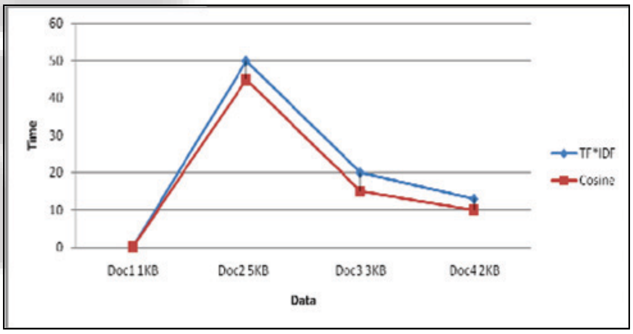


Figure 6: Execution Time

6.3 Compare TF-IDF and Cosine Similarity Mechanism

Here, we have compared to major techniques such as Cosine similarity and TF-IDF. Generally cosine similarity mechanism having ability to provide relevant documents with the help of dot product. So that they can provide exact information which related to the user request.

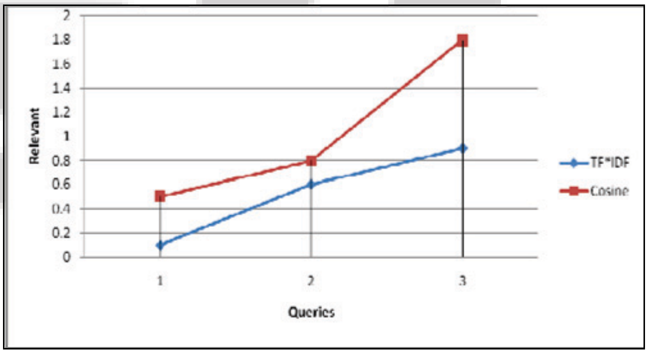


Figure 7: Compare TF-IDF and Cosine Similarity

7. Conclusion and Future Scope

An Ontology Model is implemented for representing user background knowledge for personalized web information gathering. The model constructs user personalized ontologies by extracting world knowledge from global database and discovering user background knowledge from user local instance repositories. Ontology mining methods, TF-IDF and Cosine Similarity, is introduced for user background knowledge discovery. And finally conclude that Cosine Similarity is better than TF-IDF mechanism by using different performance parameters. The future work can be implementation of improved Cosine Similarity mechanism for multi domain approach. The future work can be analyzing the performance of improved mechanism with difference set of performance parameters like efficiency, throughput and consistency, etc.

References

[1] B.Umamaheswari, Pramod Patil,"Personalized Ontology Model-Survey"International Conference on Hybrid Intelligent Systems(HIS),2012

[2] Satya Bhanu Jonnalagadda, A.Sravani, Prof.S.V.Achutha Rao," A Knowledge Based Model for Percularised Web Information Gathering using Ontologies" International Journal of Computer Trends and Technology(IJCTT), Vol.4 Issue 9-Sep 2013.

[3] Xiaohui Tao, Yuefeng Li, Ning Zhong, Richi Nayak," Ontology Mining for Personalized Web Information Gathering" IEEE/WIC/ACM International Conference on Web Intelligence,2007

[4] Jing Wang, Jianpei Zhang, Ying Wang," Research on Semantic Web-Oriented Ontology Model" International Conference on Internet Computing in Science and Engineering,2008

[5] Krishna Chandramouli, Craig Stewart, Tim Brailsford, Ebroul Izquierdo," CAE-L An Ontology Modelling Cultural Behaviour in Adaptive Education " Third International Workshop on Semantic Media Adaptation and Personalization, 2008

[6] Xiaohui Tao," Personalised Ontology Learning And Mining For Web Information Gathering" Queensland University of Technology Brisbane, Australia

[7] Surya Natarajan, Mr. J.Sethuraman," A Personalized Ontology Model for Web Information Gathering Using Local Instance Repository" Journal of Theoretical and Applied Information Technology,30th April 2012,Vol.38

[8] Ming Li," Ontology-Based Context Information Modeling for Smart Space"Proc.10th IEEE I.C. on Cognitive Informatics & Cognitive Computing

[9] Miss. Deshmukh Rupali R., Prof. Keole R.R," Ontology Mining for Personalized Web Information Gathering" International Journal of Engineering and Computer Science,Vol.1 Issue 3 Dec 2012

[10] Qiuyu Zhang, Fengman Miao, Zhanting Yuan, Qikun Zhang, Zhi Fan,"Construction of A Dynamic Trust Ontology Model" International Conference on Computational Intelligence and Security

[11] Xiaohui Tao, Yuefeng Li, Ning Zhong," A Personalized Ontology Model for Web Information Gathering" IEEE Transactions on Knowledge and Data Engineering, Vol.23

[12] Shubhangi Shindikar, M.V.Nimbalkar, Anand Deshpande,"A Personalized Ontology Model for Web Information Gathering By Domain Specific Search" International Journal of Scientific & Engineering Research,Vol.3, Issue 7,July 2012

[13] Liquan Han, Ming Li," Application And Research on Ontology In E-Government Workflow Model" International Conference on Computer, Mechatronics, Control and Electronic Engineering(CMCE),2010

[14] Richi Nayak Bryan Lee," Web Service Discovery With Additional Semantics and Clustering" IEEE/WIC/ACM International Conference on Web Intelligence,2007

[15] Leonidas Kallipolitis, Vassilis Karpis, Isambo Karali," Semantic Search in the World News domain using automatically extracted metadata files" Knowledge Based System,2012

[16] Mohd. Sadik Ahamad, S. Naga Raju," Web Personalization using Efficient Ontology Relations" International Journal of Computational Engineering Research,Vol.2

[17] FAN Jing, ZHANG Xin-pei, DONG Tian-yang," Research of Plant Domain Knowledge Model on Ontology" IEEE The 3rd International Conference on Innovative Computing Information and Control,2008

Author Profile

Mr. Nandkishor Borse received the bachelor's degree from North Maharashtra University, Jalgaon in 2011. Currently, He is a Student of Master of Technology from Department of Computer Engineering at Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, (Maharashtra) of SVKM's NMIMS (Deemed to be University). His current research focuses on A Personalized Ontology Model for Web Information Gathering.

Prof. Suraj Patil is an Assistant Professor in the Department of Computer Engineering at Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, Dist. Dhule (Maharashtra) of SVKM's NMIMS (Deemed to be University). He received the bachelor's degree from Bharatiy vidyapeath, Shivaji University, Kolhapur 2006 and Master's degree from North Maharashtra University, Jalgaon in 2013. His research interests include cloud computing and data mining.

Prof. Nihit Agrawal is an Assistant Professor in the Department of Computer Engineering at Mukesh Patel School of Technology Management and Engineering, Shirpur Campus, Dist. Dhule (Maharashtra) of SVKM's NMIMS (Deemed to be University).He received the Bachelors degree from Institute of Information Technology and Management, Gwalior 2010 and Master's degree from Shri Govindram Seksaria Institute of Technology and Science,Indore 2013. His research includes data mining.