Journal of Theoretical and Applied Information Technology
20th March 2015. Vol.73 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FRAMEWORK FOR ONTOLOGY BASED LINK ANALYSIS FOR WEB MINING

[1]S.MONISHA,[2]S.VIGNESHWARI

[1]Department of Computer Science and Engineering, Sathyabama University

[2]Faculty of Computing, Sathyabama University

E-mail: s.monisha.me@gmail.com, vikiraju@gmail.com

ABSTRACT

User defined keyword search have limitation in their results and deliverance based on first indications. They face a challenge against social media analysis. Searches based on user defined keyword results in poor quality and consumes more time. All relevant and irrelevant information present in a link becomes difficult for the users to extract. So, ontology is used for effective monitoring analysis. The main challenge is to create domain-specific ontology in social media analysis. The earlier heuristic algorithm, HITS(Hyperlink Induced Topic Search) was proposed for analyzing authority and hubs of pages, but since each site had extra hyperlinks like navigation panels, HITS failed in providing good precision.To overcome this difficulty, LAMIS(Link Analysis on Mining Web Informative Structures). LAMIS uses information entropy to give a higher precision from 133% to 232% and recall gets improved between 0.5 and 1 while retrieving the documents form online social media.

Keywords: *Ontology, Leads,Web- mining,LAMIS, Customer Relationship Management.*

1. INTRODUCTION

The growth of internet users increases rapidly as the number of active users for Facebook and Twitter increases worldwide. Some of the tools were used by social media intelligence to collect information about the company and its product. But complaints regarding the new features were unpredictable as they were based on user defined keyword search. A product may be identified with a specific name, whereas the username which is same as that of product name does not have both context and specific names.

ERP database contains information which is of valuable information,which explains domain explicit conceits. Evocation of ERP is compassed and values are stored in dictionary which can be retrieved at any time. Ontologies are selected, which describes inter-connection . Mining the text includes data mining and information retrieval filters the relevant information from ontology and social media.The LAMIS when compared with other heuristic methods gives higher precision and recall of a particular link.

The advantage of the proposed system is that the accuracy and efficiency can be increased using LAMIS. Both offline mode and online mode can be made more effective using LAMIS and Ontology**.** Time consumption can be reduced**.** Precision and Recall can be improved using LAMIS**.**

2. REVIEW OF RELATED WORK

Azad et al [1] discussed about accuracy and relevancy of information, being extracted from web. A combination of semantic-synaptic approach was used as a solution for this issue. This approach could be used for entire website but not for a particular link, nor hyperlinks.

Kosala et al [2] surveyed about the research area of web mining,regarding web mining categorization, web structure mining and web usage mining.Connections are done among these categories. Preprocessing transformations are done like stemming. Machine learning modes are recycled. Knowledge retrieval retrieves both relevant and irrelevant information.Information extraction extracts relevant information whereas information retrieval selects documents which are relevant.

Nithya et al [3] had done a novel research on web usage mining. In this paper noisiness is cleared. The advantage is that, it makes web searching easier.The drawback of this approach is,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

it can be used only for web search and not for security.

Panchal et al [4] proposed the prediction of popular web page and user navigation.
Using this technique, next location can also be predicted.The drawback is that, accurately predicting the location can minimize user précised latency. Implementation has been done by hiding the last page.

Kewen [5] found users access models automatically and quickly. Preprocessing details were analyzed in detail. The drawback in this mechanism is that, contrasting techniques should be used with reference to different characters, therefore it is complex.

Yadav et al [6]describes customer's behavior and the relationship with e-commerce. Web mining of customer's behavior includes methods of preprocessing, discovery in arrangement and behavioral inquiry. Using this behavior, customer requirements can be known. K-means algorithm, clustering of data could be done. But the drawback is, clusters calculation could be affected if any mistake in data occurs.

Reddy et al [7] uses effective data preprocessing methods used to improve the performance of data pre-processing technique. Meaningful patterns and relationship among stream of users are discovered.Failure in data collection occurs. Accuracy metric of user identification and session identification has not been implemented.

Rainer et al [8] increased the efficiency of defining ontologies by automatically extracting knowledge from existing enterprise application systems. Using existing systems in the field of ontology, interoperability and individualization are obtained.Drawback in this method is that, suitable criteria in the field of ontology cannot be selected. Reliability cannot be accessed using this methodology.

Halevi et al [9] discussed about emails, since most of us spend our time in emails, search is done whether semantic web can be used for email. Database query is included in email message. Light weight data collection and task manipulation is performed. Huge number of people can access the email which is not present in today's email process. But the drawback is that,limited constraint knowledge is adopted.

Gribble et al [10] proposes semantic web which is not widespread. For ordinary people, it's so difficult. So, MANGROVE is used to make it ease. Cost is reduced and benefit is increased using MANGROVE. For design and implementation MANGROVE prototype is used. The main aim of this technique is to make ease for non-technical users. Complete semantic life-cycle is being supported. HTML turns as tangible benefit. But the drawback is that, it's not for peer data management system.

Mohammad Mustafa Taye [11] uses ontologyfor the transmission of logics from people to application systems. They are represented by web vocabulary. Characerization logic plays the basis role. Different types of description logics are described. SHOIN (D),SHIQ(D),SHIF(D) are the languages used.

Alper et al [12] explains the scientific workflows used by scientists for the purpose of big data analytics. Workflows are the data generated ones. Any changes in workflows like reports, methods and data tools, there will be a change in provenance. Reduction primitives are elaborated and used for different encoding techniques. Complexity is generated from the derivation. Complexity occurs when any changes in reports/methods, the entire process change, which is a drawback.

Gil et al [13] explains reuse of knowledge, sharing and modeling which plays the major role. Henceforth, ontology is used as a appraisal. The term 'PLANET' signifies plans. Applications can be rebuilt. Three various types of real world domains can be expressed. Proposed system in this case is that, devising a new format for each and every plan. Finally, all the plans are represented using Planet.The difficulty is that,representing a new format for each and every plan is time consuming

Valente et al [14] shares knowledge across systems, reduction in cost is benefitted, on the other side,sharing becomes uneasy.Recent research has made attempt in making that easy. They found that if two knowledge base share a common ontology, sharing is made easy. The system describes the relationship between ontology and the process that takes place.

Blythe et al [15] proposes the development of knowledge base which becomes impractical nowadays, so the better solution is, usage of reusable components. This paper explains the process of reusability.it briefly explains how to describe knowledge regarding problem solving. But it covers only 75% of the criteria. The remaining may fail at any time, may be of any domain.

Xing Jiang and Ah-Hwee Tan [16] use CRCTOL in the place of traditional text mining. CRCTOL makes the content concise and rich in semantic. Precision and recall can be increased using this CRCTOL.so many advantages has been

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

swamped.This technique works only to certain extent. Precision and recall is time consuming.

Vigneshwari and Aramudhan [17] describes OWL(Web Ontology Language) to represent knowledge using web. Relationships can be expressed using this tool. Concepts which are similar and with difference has been represented.they are not applied for multiple users where documents not indexed which leads to unformed ontology interrelationship.

Gowri et al [18] decribes clustering algorithm.Without implementing looping,the corresponding window can be searched by moving forward. The shift window occurs when in passive state.Data loss occurs since only stemming is done.

Vanitha et al[19] describes where multiple users run clusters and describes the spreading of data centres.multiple users share detached platform. Data security is also not achieved.

Isabella et al[20] describes IDF which is Information Document Frequency. This IDF identifies the word importance of every document. Retrieval of information is done. Mary and Jyothi[21] discussed about the feature clustering techniques and applied it on the XML benchmark datasets. XML forms the basis for building an ontology. The proposed technique aims at building an integrated ontological profile considering the document features. Vigneshwari and Aramudhan[22], suggested the XML based approach for designing standard ontologies.

3. MATERIALS AND METHODS

Earlier mechanisms did not work because of lack in query terms. So, to overcome, we propose LAMIS (Link Analysis on Mileadsning Informative Structures), on the basis of weight given to link related with the amount of information. LAMIS mines the structure in efficient and precise manner. Starting with the name of a person/company and discovering without manual work, user's requirements are met.LAMIS highlights the amount of information present in the link. Precision and recall of earlier methods like HITS, is inferior to LAMIS. The usage of LAMIS is highlighting only the relevant information which is needed for the user while browsing website.For ex. In a news website, banners, advertisements, etc. appears. It's difficult for user to extract. So, based on the count of number of visitors, the menus will be indexed. The similar websites with the highest rank is taken into account. Because of crawlers and information agents, precision increases. The one with lower rank will be blurred. Training data and related knowledge must be extracted first.

3.1 Algorithm

Information from the business database can be extracted by using either ODBC or JDBC. Appropriate algorithms are used in the transformation of database connectivity to ontology. The algorithms are used for manually created ontologies but not of specific. The developers should be expert in domain knowledge.

```
Step 1:  G := set of pages
         for each page p in G do
Step 2:  p.auth = 1 //p. auth is the authority score of the page p
Step 3:  p.hub =1 //p. hub is the hub score of the page p
Step 4:  function Hubs and Authorities(G)
Step 5:  for step from 1 to k do // run the algorithm for k steps
Step 6:  norm = 0
Step 7:  for each page p in G do // update all authority values first
Step 8:  p.auth = 0
Step 9:  for each page q in p.incomingNeighbours do //
         p.incomingNeighbours is the set of pages that links to p //
Step 10: p.auth += q.hub
Step 11: norm += square (p.auth) //calculate the sum of the squared auth
         values to normalize //
Step 12: norm = sqrt(norm)
Step 13: for each page p in G do  //then update all hub values
Step 14: p.hub =0
Step 15: for each page r in p. outgoingNeighbours do //
         outgoingNeighbours is the set of pages that p links to //

Step 16: p.hub += r.auth
Step 17: norm += square(p. hub) // calculate the sum of the squared hub
         values to normalize //

Step 18: norm = sqrt(norm)
Step 19: for each page p in G do //then update all hub values
Step 20: p. hub = p. hub / norm      // normalize the hub values
```

3.2 Proposed System

Fig 1. Represents the proposed framwork for ontology based link anlaysis. The accuracy and relevancy of information is achieved by semantic-synaptic web mining which is the existing system but the drawback is that complete details not retrieved. Data loss occurs. Another issue is that, it

Journal of Theoretical and Applied Information Technology
20th March 2015. Vol.73 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is possible only for entire website not for a particular link. In the proposed system, an ontology based link analysis mechanism to mine informative structures with high precision is being used. They can be used for interlinks. Using ontology, database contents are integrated to online links, so leads and security can be improved. Queries are written for Netflix website. Information for queries is extracted from database. The information from Netflix is imported as training dataset. Computation can be done offline using the formula,

$$E(p) = -\sum_{j=1}^{n} u \, log x_{y^n}(u) \quad (1)$$

In equation 1, u is the weight given to pages.In near line, the information regarding TOC (Topic of Contents) gets stored. This information is stored in cache memory, SQL and database. It is crosschecked against ontology for interrelationship. The information is given according to user's needs and stores the information for computation.

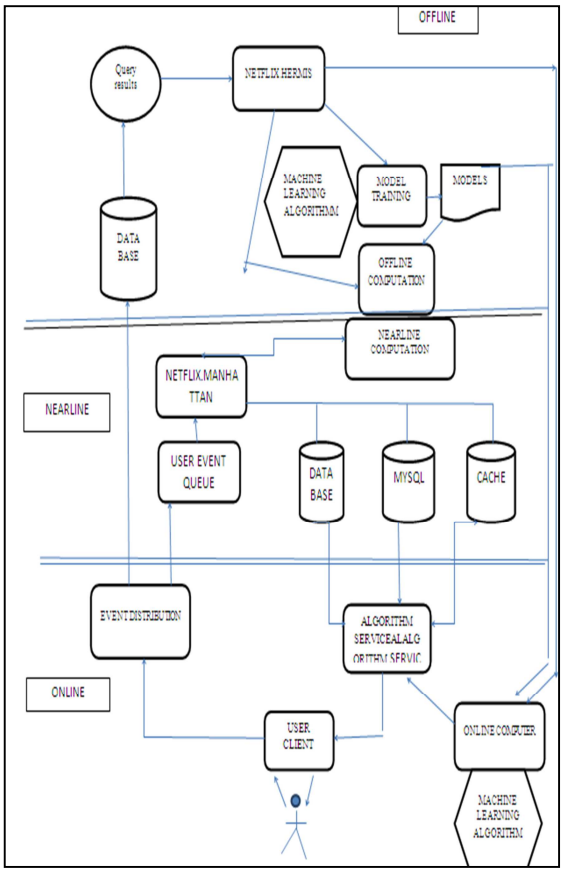

Figure 1: Proposed Framework

3.3 Experimental Setup

Table 1: Training Data Set

E-Commerce sites	No.of pages	Index	Sections	Blocks
Flipkart	1224	26	3628	3291
Snapdeal	987	20	3522	2519
Ebay	1587	29	8628	7295
Amazon	2127	33	10,222	9655
Redbus	567	16	1514	1555
Yepme	255	8	1819	1319
pinterest	323	11	566	546

Table 1 comprises of the training datasets which are obtained from a variety of E-Commerce sites. In online module, information from database,cache memory and SQL is integrated to database through offline mode. The user takes his' own requirements and the TOC also get increased which describes the LEADS process. The administrator can view the integrated database and retrieves information, since the online computation follows machine learning skills. When each profile is created by user, it gets stored in administrator's database, which can be retrieved for future referral. The admin's database automatically stores the user's activities, like commenting on feedback. So, here, leads provide security, wherein only the admin can view the database. Leads are generated from web. LAMIS is augmented to improve precision and recall in the terms of ontology.

4. RESULTS AND DISCUSSIONS

For various websites, precision and recall is calculated. Table(2) denotes precision and recall without LAMIS for the following websites, Flipkart,Snapdeal,ebay,Amazon,Redbus.

Table 2: Absence of LAMIS

WEBSITES	No. of documents retrieved	PRECISION	RECALL
Flipkart	145	0.45	0.40
Snapdeal	155	0.86	0.79
ebay	147	0.4	0.36
Amazon	152	0.2	0.17
Redbus	167	0.62	0.58

Precision and recall is calculated for these websites without using LAMIS. Equation 2 represents the formula to calculate precision. Here PP is the Product Precision and recall is calculated

Journal of Theoretical and Applied Information Technology
20th March 2015. Vol.73 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for the product search using equation (3), where PR represents Product Recall.

$$PP = \frac{X \cap Y}{\sum_I Y_I} \quad (2)$$

$$PR = \frac{X \cap Y}{\sum_I X_I} \quad (3)$$

From equations (2) and equations (3), X is the number of relevant documents; Y is the number of retrieved documents.

Table 3: Presence of LAMIS

WEBSITES	No. of documents retrieved	PRECISION	RECALL
Flipkart	135	0.5	0.52
Snapdeal	211	0.9	1
ebay	134	0.6	0.56
Amazon	218	1	0.82
Redbus	136	0.8	0.78

Table (3) represents the precision-recall rate upton using LAMIS.From these calculations, it is observed that LAMIS improves the precision which is of range between 133 to 232% and recall of range between 0.5 and 1. The combined observations of LAMIS-ontology are the best in comparison with others.

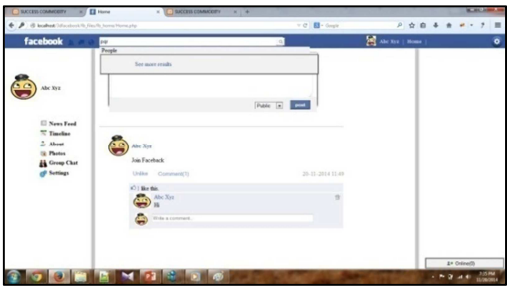

Figure 2: Before the use of LAMIS and Ontology.

The figure (2) describes the absence of LAMISwhere it consumes more time for user to search.When a word 'P' is typed, it shows the entire names related to that product begining with the letter, 'P'. Search is complicated and consumes more time.

Figure (3) clearly describes the usage of LAMIS. When the same letter 'P' is typed, the names related to 'P' is displayed but with fewer links. It displays search results, based on the count of TOC.Based on the higher rank with reference to ontology, search becomes easier.

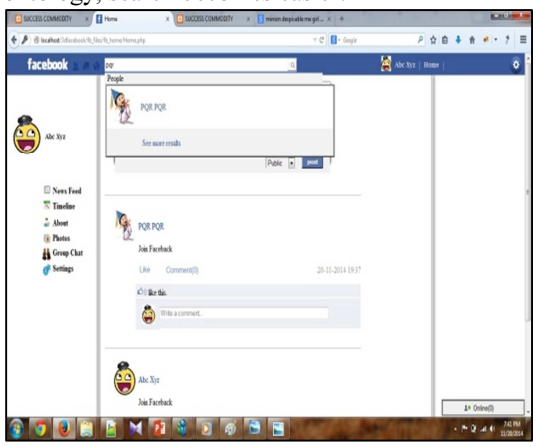

Figure 3: After the use of LAMIS and Ontology

5. CONCLUSION AND FUTURE WORK

Therefore by using the proposed system, complete details are retrieved. Data can be retrieved from different media channels. Ontologies are derived from ERP database. The suitable criteria can be selected in the determination of social media. LAMIS is incorporated to improve the precision and recall. Accuracy and reliability is achieved using semantic-synaptic approaches.

The LAMIS can be augmented to improve the precision and recall still more. Enhancing the LAMIS will be useful practically for mining real websites.

REFERENCES

[1] Hiteshwar Kumar Azad and Kumar Abhishek,"Semantic-Synaptic Web Mining: A Novel Model for improving the Web Mining", Fourth International Conference on Communication Systems and Network Technologies 978-1-4799-3070-8(2014).

[2] R.Kosala and H.Blockeel, "Web mining research: A Survey"SIGKDD: *SIGKDD Explorations: newsletter of SIG on knowledge discovery and data mining*", ACM, vol-2.Pp1-15(2000).

[3] P.Nithya, and Dr.P.Sumathi," An Enhanced Pre-Processing Technique for Web Log Mining by Removing Web Robots", IEEE 978-1-4673-1344-5(2012).

[4] Priyanka S. Panchal and Prof. Urmi D. Agravat," Hybrid Technique for User's Web Page Access Prediction based on Markov Model", IEEE -31661(2013).Applied Sciences VOL. 6, NO. 10, OCTOBER 2011

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] Liu Kewen," Analysis of Preprocessing Methods for Web Usage Data",*1ntemational Conference on Measurement, Information and Control (MIC)*, 978-1-4577-1604-11(2012)

[6] Mahendra Pratap Yadav, MHD Feeroz and Vinod Kumar Yadav," Mining the customer behavior using web usage mining in e-commerce", *IEEE*-201S0 (2012).

[7] Sudheer Reddy, M. Kantha Reddy, V. Sitaramulu," An effective Data Preprocessing method for Web Usage Mining", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009

[8] Alt, Rainer, Wittwer, Matthias ," Towards an ontology based approach for social media analysis",5th *International Conference on Computational Intelligence and Communication Networks* 978-0-7695-5069-5/13 $26.00 © 2013 IEEE (2013)

[9] Oren Etzioni, Alon Halevy, Henry Levy, andLuke McDowell, "Semantic Email: Adding Lightweight Data Manipulation Capabilities to the Email Habitat",*International Journal of Information, Technology & Decision Making* Vol. 7, No. 4, 2008

[10] Luke MC Dowell, Oren Etzioni,Steven D. Gribble,Alon Halevy,Henry Levy, William Pentney,Deepak Verma ,"MANGROVE: Enticing ordinary people onto the semantic web via Instant Gratification",*National Conference on Computing and Communication Systems (NCCCS)*,978-1-4673-1953-9/12/$31.00 ©2012 IEEE

[11] Mohammad Mustafa Taye, "Web-Based Ontology Languages and its Based Description Logics", T he Research Bulletin of Jordan ACM , ISSN : 2078 - 7952 , Volume II (II) Page 1

[12] Alper, Belhajjame, Goble C.; Karagoz, "Small Is Beautiful: Summarizing Scientific Workflows Using Semantic Annotations", Big Data (Big Data Congress), IEEE International Congress, 978-0-7695-5006-0, June 27 2013-July 2 2013.

[13] Yolanda Gil and Jim Blythe. "PLANET: A Shareable and Reusable Ontology for Representing Plans". In AAAI 2000 workshop on Representational Issues for Real-world Planning Systems, Proceedings of the *International Conference on Intelligent User Interfaces,* 2001.

[14] Andre Valente, Tom Russ, Robert MacGregor, and William Swartout. "Building, Using and Reusing Ontology of Air Campaign Planning". IEEE Intelligent Systems, special issue on Ontologies, 14(1), January 1999.

[15] Jim Blythe and Yolanda Gil. "A Problem-Solving Method for Plan Evaluation and Critiquing". Proceedings of the Twelfth Banff Knowledge Acquisition for Knowledge-Based Systems Workshop.

[16] Xing Jiang and Ah-Hwee Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents", *International Journal of Human-Computer Studies*, 1998. *International Journal of Human-Computer Studies*, 1998

[17] **S.**Vigneshwari**,**Dr.M.Aramudhan,"A technique to user profiling ontology mining and relationship ranking", *Journal of Theoretical and Applied Information Technology*,31st December 2013. Vol. 58 No.3

[18] S.Gowri,G.S.AnandhaMala,G.Divya, "Enhancing the digital data retrieval system using novel techniques", *Journal of Theoretical and Applied Information Technology,* 20th August 2014. Vol. 66 No.2

[19] Vanitha Muthuswamy,Kavitha.C, "Secured data deletion in cloud based multi-tenent databased architecture",*International Journal on information Sciences and Computing "*, Vol 6, No 2,July 2012.

[20] Isabella.J,Suresh.R.M,"Application of feature extraction technique to unstructured texts",*National Journal on advances in computing and management*,vol3,no 2,oct 2012.

[21] A. Mart Posonia and V. L. Jyothi, Context-based Classification of XML Documents in Feature Clustering, Indian Journal of Science and Technology, 2014, 7(9): 1355–1358

[22] S.Vigneshwari, Aramudhan M, An approach to personalize the web using XML based ontologies, 2012, 2012 World Congress on Information and Communication Technologies, 759-762.