Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 5, Issue 1; January-March, 2018, pp. 26-30

http://www.krishisanskriti.org/Publication.html

KNOWLEDGE ACQUISITION AFTER FORMALIZING WEBPAGES THROUGH KM FRAMEWORK

Shri Rajesh Doss

Assistant Professor in Computer Science, National Defence Academy, Khadakwasla, Pune, Maharashtra, India

Abstract—*Learning phase leads to the acquisition of knowledge through Listening, Embarking, Acquiring, Reading and narrating the observed patterns. These observed patterns vary from one individual to another. The innovation of many new technologies for acquiring knowledge or learning patterns has superseded the traditional method of teaching-learning processes. The addictive usage of IoT has made the learning generation to be so quick such that every query of a learner gets solved instantly. Today's web technology provides a standard space for sharing information but does not help us for analyzing unstructured web data such that the users instantly can use the acquired learning patterns. Unknown facts from unknown or known sources with unethical trustiness will mislead the learners towards their decision-making system and also will blame the user who reciprocated the information. As a way out to the observed problem, this paper enhances to derive a formalized learning pattern after its recovery from the webpage with the help of more specific datasets of social networking. This paper uses knowledge management principles to explore the created SNKA framework and also compares the data mining algorithms suited for finding the similarity analysis of the web content.*

Keywords: *Knowledge Management, Data Mining, Social Networking.*

1. INTRODUCTION

The Impact on the usage of technology towards the learning process has become more user-friendly while comparing the past technological development. Learning starts from the observation of anything from anywhere, or from anyone and also by many other means. Today, Internet (IoT) provides the fastest mode of transactions like an exchange of things, money or knowledge concerned through the unstructured web. The unstructured web has also introduced a network for sharing their ideas and interests which are in common among the known or unknown members of the group formed [1]. We call them as Social Networks, which is technically defined as a social structure of people, related one to another through a common relation or interest [2]. A few popular social networking websites in the world are Facebook, Myspace, LinkedIn, Twitter, YouTube, BlogSpot, Myspace & Flicker [2].

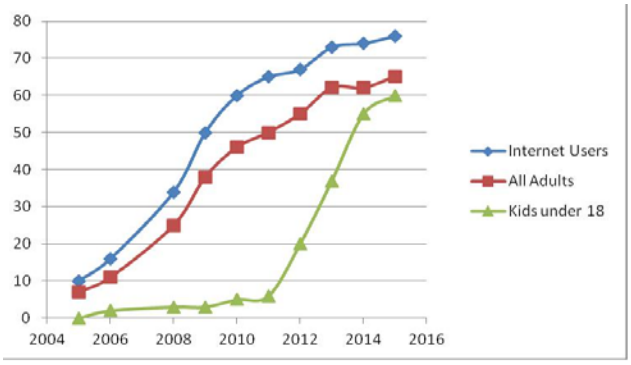


Fig.1 - Impact of Social Networking Usage of Indian Learners (Questionaire Based)

The primary characteristics of these social networks as defined by Rajan Dutu, 2007 Research of Rice University areUser-Based, Interactive, Community Driven, Relationship, and Emotional control. The Usage of these social websites has expanded broadly from 2004 onwards. To query more on the study, a Questionnaire was generated. The impact of learning patterns was acquired from many web users through unstructured webpage with more specific to Social networking web pages of Facebook and LinkedIn.

Fig1. depicts the impact on usage of social websites among Indian learners. The Blue colored graphical representation refers the overall usage of internet among Indian users irrespective of their ages. The red colored graphical representation refers to the adults lesser in age of 18 years. The green colored graph points out to the drastic growth of social network users of <18 age. This targets the learners >18 years of age and also who are in the habit of frequent usage of social networks. These social networking sites are the part of the conceptual umbrella of Web2.0. Web 2.0 provides a uniformity to create a standard data format to upload and download the content where there is no predefined structure to either verify the originality or to check the quality of the information or message shared in those environments. Hence, this paper theoretically tries to define a KM framework for

streamlining web data and also applies data mining algorithm to create and implement a formalized knowledge extraction tool from social Web pages.

2. KNOWLEDGE MANAGEMENT

Knowledge Management is explored as a process to search, select, record and store information to disseminate knowledge and then apply the acquired knowledge throughout the organization or individual for finishing faster transactions with reduced rework [Sanjay Mohapatra, Macmillan].Various other clarifications dealing with knowledge Management are defined by other popular authors are Davenport (1998], Alavi al(2001), Allee (1997), Holm(2001).Managing knowledge is one of the primary action taken either towards individual growth or the organization's development. The Construction of Knowledge Management includes four major phases like Gathering, Organizing, Generic flow for easier understanding and reapplying the knowledge. These phases help us to disseminate the tacit knowledge from experienced brains, to define a process for storing the acquired knowledge from potential loss, to reuse the effective gained knowledge.

In knowledge management, the Sharing of knowledge gets accomplished only after the created knowledge gets shared further or reused. There are various levels in constructing a productive knowledge management model. The different levels behind KM model are individual wise, Group wise and the organization wise. The individual level identifies the knowledge and discusses with peers/COP. The Group level reviews the knowledge, categorizes and codes the metadata. A technology is looked up for knowledge sharing. The organization level finally refines the knowledge and creates the learning culture. The common model for knowledge acquiring process is analyzed for the web-based environment such that a knowledge acquisition tool is created with the base process of knowledge Acquisition cycle. The knowledge acquisition process can be seen itself as a lifecycle:

- It must be ***created*** or ***searched*** either within a trusted source inside or outside the organization. This is typically comprised of repetitive tacit and explicit looping process until the knowledge is ready for distribution to those outside the creating group.
- For reusing the observed knowledge, it can be ***stored*** at some defined locations, either transparently or in a hidden manner so that it is accessible for other members of the group or organization can find and use.
- Those who need the specific knowledge must then ***find*** out where it is, when they need it, bysearching in the right places and/or asking the right people.
- Once the knowledge source is found, the user will then go through the act of actually ***acquiring*** it. This will be knowledge gaining phase irrespective of their sources whether acquired from other fellow humans or documented sources.
- Once acquired, the knowledge can be put to ***use*** towards some productive purpose.

after using respectively, the user understands to learn what worked perfectly well and what went against the expected as a result of applying the knowledge gained. These steps can be further be taken as a successful continuous process for all the input given or taken for further consecutive looping steps to achieve the knowledge creation and knowledge sharing processes.

3. SOCIAL NETWORK DATA ANALYSIS

Analyzing the architecture of the Social group of networks is known as Social Network analysis. The increase in the use of the web has interested more researchers from 2004. Various tools for social network analysis are Graph Characterization Toolkit, Tweet Hood, Meerkat, thriller, Hits/ISAC Social Network Analysis Tool, D-Dupe and X-RIME, a cloud generated library or large sized social network analysis. The social network analysis is broadly done only by two ways either by generating User Graph analysis nor by creating questionnaires or by generating interview questions among users. The case study was taken up as the secondary analysis method. This paper was the outcome of an online questionnaire which was circulated through google blogs where many online learners have responded. In order to study the corporate view certain interview questions among the leading corporate users were recorded.

The main cause of this study to achieve the following concerns about analyzing the users:

i) Behavior in the case of sharing?
ii) Searching Capacity for an online content?
iii) Testing the trustiness among users?

Table 1: Impact on various Learning Platforms

Hypothesis-1	Impact
Social Networking as Learning Environment	66.5
Internet Based Learning	24.2
Traditional Learning Environment	9.3
Hypothesis-2	Impact
Share Data based on User's Trust	25.3
Share Data after Verifying or Confirmation	7.7
Blind Sharing or Simply Forward	67

Table 2: Impact on sharing Interest

Hypothesis-3	Impact
Use Popular Searching Browsers	58
Use Guided Source for searching data	12
Apply various Filters or by specific algorithms	30
Table 3 Impact on Searching Patterns	

The above result helps us to study & work upon to derive a Knowledge Acquisition Framework or cycle to provide an efficient learning environment among the online users.

3.1 SNKA CYCLE – PROPOSED

The Social Networking based Knowledge Acquisition (SNKA) Cycle involves the series of steps in finding the absoluteness in knowledge building or knowledge gaining procedures. It provides a faster learning and structured acquisition patterns. These conversion process was further applied to the unstructured web data into structured web data that solves many deficiencies in learning patterns.

It provides a cyclic process where all social websites work under the same umbrella with a finite medication in typical human behavior. The Figure1 above describes the Social Networking based Knowledge Acquisition Cycle as:

Sender's / Receiver's information: Information is sent or received from various domains of social webpages.

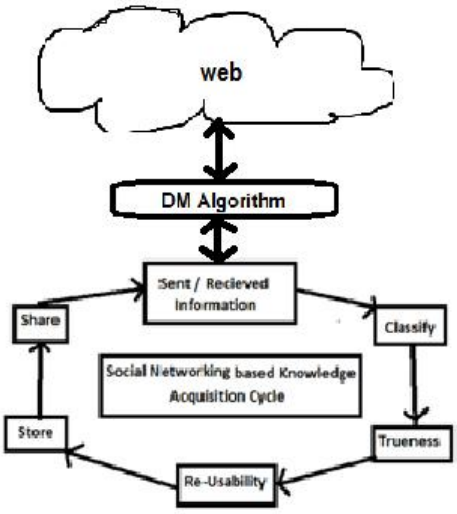


Fig.2. Proposed SNKA CYCLE

Classification and categorization: Information received or observed is to be classified as smoother usages.

Truth: The validity of the received information is to tested through search engines or through trusted experts of the domain. This testing helps in finding the exactness in web data.

Re-creation: The received information is involved in the verification process that defines the status of reusability nature.

Storage process: The observed or received information is stored for future usage in the exact form of knowledge after satisfying the verification and validation process.

Sharing: After the successful completion of the entire cycle, the acquired knowledge could be shared with others and hence this process surely leads to quality in the decision making process.

4. DATA MINING IMPLICATION

Data Mining helps us to explore the processes in extracting hidden and predictive information observed from large datasets. The efficient and powerful technology that supports individuals and companies to focus on the information retrieval in their data warehouses.

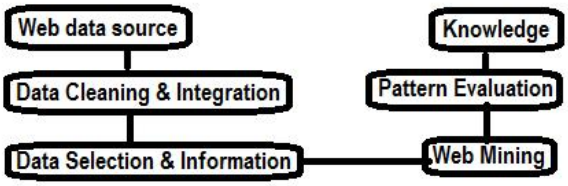


Fig.3 Basic Web Mining Steps

The different data mining methodologies help us to predict the apparent behaviors of data. This is more supportive for researchers, scientists and corporate to make productive knowledge-based decisions. Most corporate sectors have already the collection and refining the large volume of data. The broad areas of data mining are found in Jiawei (2003) literatures as corroborated that includes medical treatment prediction, disease identification based on symptoms, retail management industry, tele call recording patterns, DNA pattern sequences, natural disaster, weblog click stream, financial data analysis, music selection, content based. e-mail processing systems, analyses of data from specific experiments conducted over time, analysis of nation's census database, and so on. The Data mining supports all types of operating system software and hardware platforms to improve the information resources and can be easily integrated into the new products and systems supportive for online users. The users of data mining are categorized into three groups as Application users, Designers, and Theorists. It is usually common that the theorists based on some principal assumptions usually formulate new ideas.

Therefore, some users are primarily interested in this group. Those who work with the application of DM such as knowledge Managers are referred as 'DM researcher /designer'. The most adequate techniques behind data mining methods are:

1. Artificial Neural Networks (ANN) belongs to a predictive model that understands more through training process and resembles the biological neural networks.
2. Decision trees represent a tree-shaped pattern that provokes different sets of decisions. These decisions create rules for the classifying and categorization of a dataset.
3. Genetic Algorithms: They are optimization techniques that use a process such as genetic combination, mutation, and

natural selection in a design based on concepts of evolution.

It reciprocates the way exactly as the nature works.

4. Rule Induction in data mining leads to the extraction based on if then rules to acquire useful information from data observed from statistical intelligence.
5. Regression analysis act a source to identify the best linear pattern so as to predict the value of one's behavior and relation to another
6. Semantic Web simply refers to online or web data sources that supports in smooth exchange of information in an easy understandable environment.

Zhou et.al (2008) in his innovative research has explained the sources of applying the learning methods through statistics on semantic web data. the friend-of-a-friend is the ontology-based schema that integrates Social Networking web pages.

Tushar et.al (2008) explains the usage of Semantic Web technology to detect the associations between multiple domains in a Social Network. Opuszko and Ruhland (2012) introduced a novel approach of using semantic similarity measure based on pre-defined ontologies for classifying social network data. Ostrowski (2012) has developed an algorithm to retrieve information in social networks to identify trends. The Algorithm has use semantics todetermine the relevancy of networks using unstructured data. The algorithm was tested on twitter messages.

7. Markov models: These Markov chains help us to predict the user's behavior based on the current visit of web pages. the mathematical notation in this algorithm undergoes the transitions undertaken from one to another or countable number of expected possible states. this process leads to next state but depends only on the current state. it does not concentrate on the previous sequence of events that preceded it. Implementation of Markov models can be applied in web mining to predict users' next action. The Social web pages can be mapped to a place where nodes will be projecting user's previous visits. But most of the researchers have tested this model only on static networks yet to be tried over dynamic behavior web pages.

4.1 WEB MINING ALGORITHMS

This generation learners are fully dependent over the digital communication sources where social network is an internet centric. the role of web mining for understanding frequent item sets may be studied as part of discovering the associations and correlations among items searched in large transactional or related data sets.

The derived SNKA is also dependent on the data mining algorithm for obtaining accuracy in the searched or queried data. The most famous algorithms identifying the frequency in data are:

4.1.1 Apriori algorithm

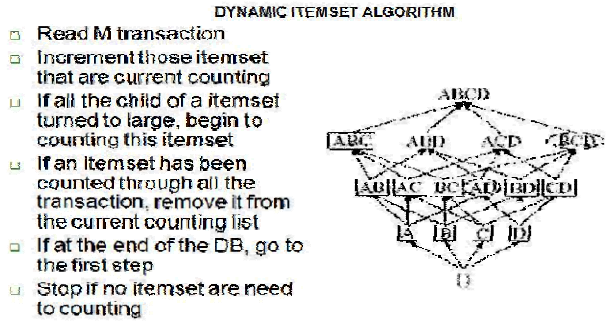
The algorithm (with or without candidate Key generation) Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikanth (1994) for applying data mining in finding frequent item sets for obtaining Boolean association rules.

The algorithm name supports the conceptual fact that algorithm uses *previous knowledge* of frequent itemset properties. Apriori employs an iterative approach known as a *level-wise* search, where *k*-item sets are used to explore (*k*+1)-item sets. As the first step, the group of frequently occurring 1-itemsets are identified by scanning the database to acquire the count for each item and gather those items that satisfy minimum support. The resulting set is indicated as *L*1. Next, *L* and later it is used to trace *L*2. the group of frequently occurring 2-itemsets are studied to find *L*3 and moves further until no other frequent *k*-item sets can be found. The finding of each *Lk* requires one full scan of the database. The objective of apriori is to study and increase the efficiency of the different generation of frequent item sets through an important property called the Apriori property. apriori mainly reduces the search environment.

Algorithm: **Apriori**. Find frequent itemsets using an iterative level-wise approach based on candidate generation.
Input:
D, a database of trans actions;
m in sup, the minimum support count threshold.
Output: *L*, frequent itemsets in *D*.
Method:
(1) *L*1 = find frequent 1-itemsets(D);
(2) for (*k* = 2;*Lk*? 1 6= f;*k*++) f
(3) *Ck* = apriori gen(*Lk*? 1);
(4) for each transaction *t* 2 *D* f // scan *D* for counts
(5) *Ct* = subset(*Ck*, *t*); // get the subsets of *t* that are candidates
(6) for each candidate *c* 2 *Ct*
(7) c.count++;
(8) g
(9) *Lk* = fc 2 *Ck*jc:*count* _ *m in sup*g
(10) g
(11) return *L* = [k*Lk*;
procedure apriori gen(*Lk*? 1:frequent (*k*? 1)-itemsets)
(1) for each itemset *l*1 2 *Lk*? 1
(2) for each itemset *l*2 2 *Lk*? 1
(3) if (*l*1[1] = *l*2[1])^(*l*1[2] = *l*2[2])^:::^(*l*1[k? 2] = *l*2[k? 2])^(*l*1[k? 1] < *l*2[k? 1]) then f
(4) *c* = *l*1 on *l*2; // join step: generate candidates
(5) if has infrequent subset(*c*, *Lk*? 1) then
(6) delete *c*; // prune step: remove unfruitful candidate
(7) else add *c* to *Ck*;
(8) g
(9) return *Ck*;
procedure has infrequent subset(*c*: candidate *k*- itemset;
Lk? 1: frequent (*k*? 1)- itemsets); // use prior knowledge
(1) for each (*k*? 1)-subset *s* of *c*
(2) if *s* 62 *Lk*? 1 then
(3) return TRUE;
(4) return FALSE;

4.1.2 Dynamic Item set Counting

A dynamic item set counting technique was proposed in which the database is partitioned into blocks marked by start points.



This causes creation of new candidate item set that can be added at any initial point unlike in Apriori. This dynamic technique calculates the confidence count of all of the item sets that have been counted along with new candidate item sets if all subsets are estimated to be frequent. The resulting algorithm requires lesser time for database scanning than Apriori.

Comparison of Algorithms:

As part of finding the closest algorithm to find the frequency in shared blogs and web pages, a sample database of web links and web pages were stored and later both the algorithms were implied. DIC algorithm was faster in comparative and fig 4 depicts the efficiency of search results in the shortest span of time.

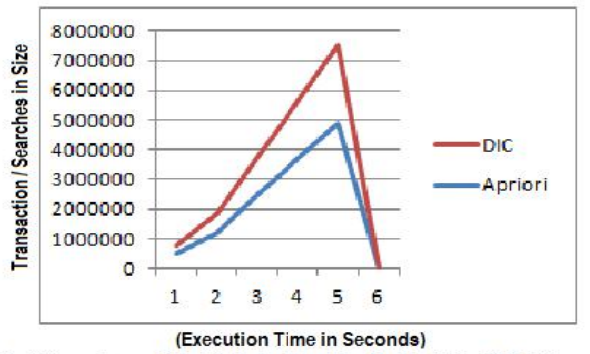


Fig 4 Comparison of DIC with Apriori Algorithm for identifying Web Patterns

The Questionnaire output was analyzed only based on Indian learners.

The algorithmic implementation is under process to be converted into a tool such that it becomes more effective for analyzing a content either forwarded content or downloaded content.

5. LIMITATIONS OF THE STUDY

The Questionnaire output was analyzed only based on Indian learners. The algorithmic implementation is under process to be converted into a tool such that it becomes more effective for analyzing a content either forwarded content or downloaded content.

6. CONCLUSION

The addictive usage of IoT has made the learning generation to be so quick such that every query of a learner gets solved instantly. The technology does not provide any universal platform for analyzing the unstructured web data and the instantly acquired learning patterns that will mislead the learners towards their decision making and also will blame the user who reciprocated the information. The unprecedented growth of the World Wide Web coupled with the recent advances in the telecommunication networks has made possible the transmission of large amounts of data in a short period of time resulting in the accumulation of data on the Internet.

This data is stored in files specially created for this purpose calledlog files, generated by servers showing a list of actions that occurred e.g. user's behavior at a particular organization's website. There are many data mining tools in existence to turn the raw data in the log files to useful information. the problem of accuracy creates more serious issue among general learners. Hence, this paper enumerates the different concepts applied in different approaches are compiled as a study in interlinking the concepts like Information Retrieval, searching methods and framing a knowledge management for deriving as a permanent solution for Correctness or trueness in the information acquired from the web.

REFERENCES

[1] Davenport, Thomas H. (1994). "Saving its Soul: Human Centered Information Management".
Harvard Business Review 72(2): 119–131.
[2] Bordia, P., Irmer, B. E., & Abusah, D. (2006). Differences in sharing knowledge interpersonally and via databases: The role of evaluation apprehension and perceived benefits. European Journal of Work and Organizational Psychology, 15(3), 262−280
[3] Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. Administrative Science Quarterly, 48(2), 240−267
[4] G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma "Elimination of Redundant Links in Web Pages" ,World Academy of Science, Engineering and Technology 52 2009
[5] Jaroslav Pokorny, Jozef Smizansky, "page content rank: an approach to the web content mining", Charles University, Faculty of Mathematics and Physics,Malostranskénám. 25, 118 00 Praha 1, Czech Republic
[6] Fergus Toolan "Web Mining" Intelligent Information Retrieval, Group University College, Dublin
[7] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to InformationRetrieval, Cambridge University Press, 2008
[8] W. Bruce Croft, Donald Metzler, and Trevor Strohman, Search Engines: Information Retrievalin Practice, Addison Wesley, 2009
[9] Peter Brusilovsky, Jae-wook Ahn and Edie Rasmussen "Teaching Information Retrieval with Web-based Interactive Visualization" 25 July 2010
[10] Prahaladrao.M "Knowledge Management", Defence Electronics research laboratory Hyderabad.