ISSN:0975-9646

A Novel Approach for Subjective Content Accessibility Mining using Lib Mart & Web Mart Classification having Multilayered Architecture

Sachin Yele[1], Ravindra Gupta[2]
[1]Shri Satya Sai Institute of Science & Technology, Sehore
[1]Rajiv Gandhi Prodhyougoki Vishwavidhyalaya, Bhopal, India
[2]Shri Satya Sai Institute of Science & Technology, Sehore
[2]Rajiv Gandhi Prodhyougoki Vishwavidhyalaya, Bhopal, India

***Abstract*-- Users have always wanted to find their desired topic in documents, whether a document is such as a book, an article, a paragraph, a section, a chapter, a research paper and even one single webpage, which can be seen as a kind of information material and information quickly and easily from a set of latest documents on the Web. This means that the user has three main goals. Get the content of desired domain, Get the details of the document set in which search topic present, content accessibility of specified topics present in the document set. The above goals of user are the issues for searching content on web. The content searching on web is dealt by web content mining. The approach for web content mining varies from application to application. The Web content mining has two major activities Identify the object i.e. find the relevant document and access the content from identified document this paper give the concept of subjective accessibility test using multilayered architecture using Libmart & Web Mart Classification.**

***Keyword*--** Web content Mining, Libmart, Web mart, Multilayered architecture

I. INTRODUCTION

The user has three main goals for subject-topics accessibility. First is domain specific, second is all those document set in which search topic is present, and third is Content accessibility of specified topics. The above goals of user are the issues for **web content mining**. The identifying (searching) the document on web has two distinct approaches.

1. **Agent Based**: These types of approaches uses the caching, indexing, gatherer, broker or scatter techniques to generate web agents. These are three categories of such agents.
   a) Intelligent Search Agents
   b) Information Filtering categorization &
   c) Personalized Web agents.

There are many agents such as Harvest, FAQFinder, HyPursuit, Bookmark Organizer, WebWatchers

2) **Database**: This approach focused on techniques for organizing data on the web into more structured collections of resources and using standard DB querying mechanisms & data mining techniques to analyze it.

There are following categories in this approach:
   a) Multileveled Database (warehouse)
   b) Web Query System

This approach makes more organized, fast and precise query.

The web can be treated as warehouse of documents (information) which makes the subjective content accessibility vast, diverse and dynamic. The vast web can be narrow-down by segmenting it application-wise.

The scope of work is limited to one of the web application. The precise subjective or context based content accessibility needs to classified the documents subject-wise i.e. **Domain-Semantic** – specifies the domain of document. The concept of Domain-Semantic is implemented by the Warehouse – datamart concept. The datamart on web application called webmart. The web-mart is collection of words of specific domain. The first context to the information is the Domain of document. The user generally searches by the **word or by phrase**. The other possibility is search by **attribute of documents** like title, author etc. or combination of these attributes. The whole collection of documents is organised in such a fashion that the search may be on word or document attribute would be simple query. The document needs to farther classify by word-wise and attribute-wise. This leads to **metadata** solution. The accessibility of contents and search of word is easy by the one of document attribute **resource type** in domain. The resource types are Physical Book, E-Book, Research Paper, White Paper etc. The scope of work is restricted to textual content. This further can be extended for all type of data. This paper deals about the topic-specific, subjective-search using the **Lib Mart & Web Mart classification** by data mining techniques. The Multi layered Architecture approach is implemented using metadata. The metadata-database approach needs automated extraction of metadata from the documents and stored into metadata-database which makes database approach dynamic. The **crawlers** are used for automated extraction of metadata. The metadata is the key to locate, use, and preserve digital content. The structure data - metadata about digital objects and collections are of three types, all ensures the usability and preservation successfully over a period. The Descriptive Metadata describes the digital object at fullest of its verity. The Structural Metadata describes the relation, association within among the objects. The Administrative Metadata helps to access, manage, and preserve the digital collection.

II. METHODOLOGY

A. Concept

The concept is using multilayered Architecture approach for searching rather then agent based. The model data are stored in multiple layers. The first layer is domain semantic. The subject-wise classifications of the documents make subjective groups of the documents. This subjective group is referred as Domain of document. Now the search on subject specific is possible. The subjective classification of the documents is Domain-Semantic – specifies the domain of document. The documents stored on web are treated as warehouse of unstructured, semi-structured and structured documents. The mining always tries to infer the structure, unstructured or semi-structured data which needs to convert it into structured manner. The structured of the document can not be changed as to maintain originality. The user searches information by the word or by the phrase similar to searching word in the index of book**.** The index of book also contain one more information i.e. page number on which content is available. The second layer is classification by resource type. The digital libraries have digital resources like e-books, research paper, white paper, reports, thesis, dissertation etc. This resource type is another classification from database view. The third layer is details of objects. The whole digital collection of documents needs to organised document attributes-wise so that search is possible by attribute of documents like title, author etc.

The digital resources are organising by using the metadata - data about data. The metadata are organize in tables of databases to visualize the Digital Library as warehouse of words and words are classified into subjective data mart – webmart-libmart. The subject-wise words are gathered in data mart serve as index. And the document- metadata serve as object resource warehouse like book. The words, page number and attributes from the unstructured document or from semi-structured documents are extracted and kept with data mart and resource & object warehouses. The concept in this way converts logically unstructured documents to structure documents.The search item is one word or phrase (multiple words). The data modelling of metadata provide the database tables as data-mart for words and word resource warehouse for resource type classification. And object resource warehouse maps word with digital object like word from index maps to content of page in the book. The data mart in web application refers as **Web mart – Lib mart classification**.

B. Metadata

The literature survey about metadata made the following definition list:

a) 'Data about data' (Boehm 1999; Butterfield 1995; Daniel and Lagoze 1997; Henze and Schefczik 1997; Lynch 1998; Rust 1998; Weibel 1997)

b) 'A (usually brief) characterization of the individual Information Objects in the collection of a library' (Smith 1996)

c) 'Classifying the content of Web objects' (Marchiori 1998)

d) 'Metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics' (Dempsey and Heery 1998)

e) 'The Internet-age term for structured data about data' (EU-NSF 1999)

f) 'A small summary of characteristics of each available resource' (Wood 1999)

g) '[A] relationship that someone claims to exist between two entities' (Rust and Bide 2000)

The “data - structured or not - that describes something about another data resource” is mention in almost all definitions. The term Metadata is used when this data resource resides on the Internet.

Some definitions seem to use quite a different angle.

- 'There is no clear line between content and meta-content' (Guha 1996/1997b)
- 'The lifeblood of commerce systems' (Erickson 1998)

This Paper uses the first definition in the list for metadata**:** data about data. Metadata is structured information about the characteristics of a physical or digital object. Metadata serves the same function as a label. The metadata will describe a digital information resource, and also used to describe physical and remote resource. The metadata itself is also digital data. The Web resource lacks essential metadata, or if the metadata is inaccurate or incorrect, search results affects quality and consistency. The good metadata helps to find the information easy and fast, but searching by uncontrolled keywords may give tens of thousands of results, in which majority of are usually irrelevant to the user. The structure of metadata allows searching by elements of document e.g. title, subject, etc. and search results are less and more relevant.The metadata are necessary for successful management and organization of digital objects. The metadata are also used for accessibility of digital objects. The metadata in both case are more extensive and different from the metadata used for only managing collections of digital or printed works and other physical materials.

The Architecture needs to store descriptive metadata regarding a book in its collection. The **structural metadata** regarding the book's organization are stored to dissolve the book into a series of unconnected pages. The library also need to record structural metadata regarding the book's organization, user can be able to evaluate the book's worth. The page image or text files comprising the digital work are of little use, without structural metadata and user may be unsure about the originality of the digital version provides. The user can not access the content without technical metadata regarding the digitization process. The library must have access to appropriate technical metadata in order to periodically refresh, migrate the data, and ensuring the durability of valuable resources for internal management purposes.

The use of metadata can briefly describe as below:

- Metadata organizes information
- Metadata makes things accessible
- Metadata makes things discoverable
- Metadata makes things endure
- Metadata return better performance needs to be used in an organized way.

Librarian perspective: Only think of information that is easily retrieved by end-users

Publisher perspective: Only think of realized by information that furthers the organizational view.

C. Web Mart Concept

The metadata are organize in tables of databases to visualize the Digital Library as warehouse of words and classifies as data mart – web mart referred here as lib mart.The documents on the web can be treated as bucket of words. The words are classified subject-wise. The classified words are kept in many bucket i.e. web mart. The words in the web mart do the job of domain classification –Domain Semantic. The words in web marts are identified by unique ID which points out the resource. The documents are resources of words. The words of web marts are mapped to resource i.e. digital objects (document). But the digital objects are also classified by resource type like e-Book, Research Paper, etc. The words of first layer are mapped with the resource type of digital object and page number. The accessibility of contents is made easy by the page number of digital object in which word is presents. This is second layer word-resource warehouse operates. The third layer is metadata of digital objects operates as an object resource warehouse.

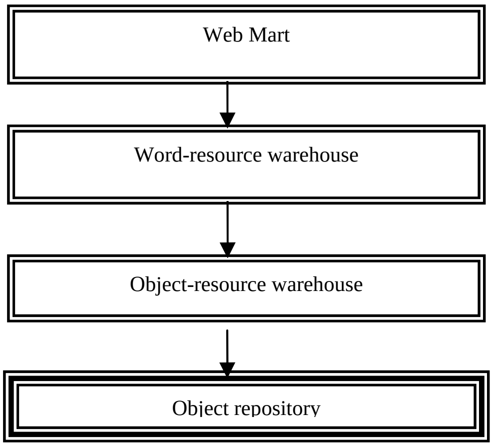

Fig.1 Block diagram of multilayered database using lib mart

D. Proposed Data model

The model data is modelled in three layers. The layer maps Search-word and classification. The Search-word is mapped with subjective lib-mart. The Search-word can be referred in more than one subject domain so is in lib-mart.

Let the database schema of layer-1 contain two relations, Word and domain-lib mart.

1. Word (word, lib mart number) The word may be in more then one lib mart.
2. lib mart (lib mart number, lib mart path, classification)

The second layer is word resource, which maps the Search-word with the resource type and page number. The resources are e-book, research paper, white paper, etc. The Search-word can be in more then one resource type.

Let the database schema of layer-2 contain three relations, Word, Type of document (resource type) and page number.

- word (word, resource type) The word may be in more then one resource type and may be in number of documents of same resource type.
- resource type (resource_id, Resource type, page number, object_id)
- page number (the page number may be more then one)

The third layer is object resource warehouse, which maps Search-word with object resource. The Search-word can be in more than one object of same resource type.Let the database schema of layer-3 contain two relations.

Resource type The word may be in more then one resource type and may be in number of documents of same resource type.

- object id (Resource type, object_id)

The object resource warehouse keeps metadata of object which is in physical repository.

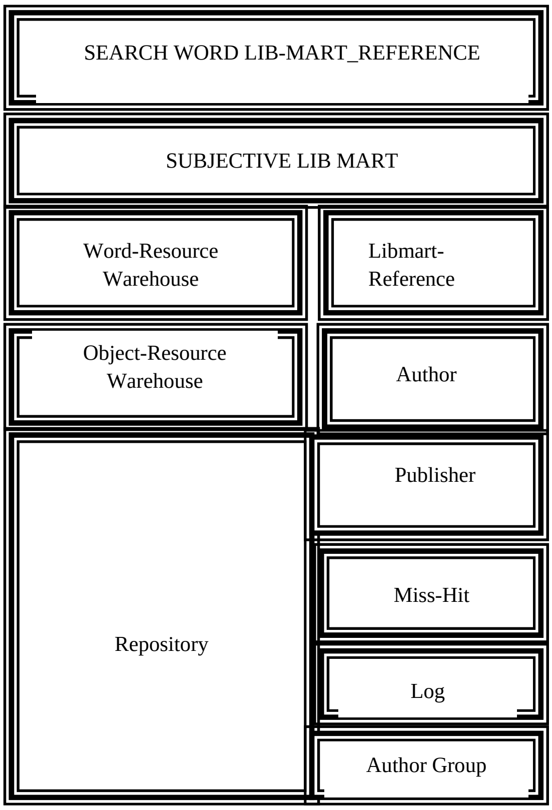

Fig.2 Block Diagram of Layered Architecture in normalized form

III. MINING ALGORITHM

The search algorithm is as below:

```
Start
Fetch the word
Search the word in [Search_word_lib-mart_referace]
Display Classification /* Engineering ,Astrology,…*/
Get classification from user
Read lib-mart Details
Search the word in [Subjective lib-mart]
Display Resource type available where word is referred
Get Resource Type from user
Read word Details
Search the word in [Word Resource Warehouse]
Display all object details of selected Resource type
/* may more document of same resource type /*
Get Object from user
Read Object Details
Search the Object in [Object Resource Warehouse]
Display the content of selected word from selected
Resource type of selected subject domain
END
```

This mining algorithm apply for searching the word for the exact searching and subjective mining of the word this also include the machine learning concept The search word is not found in any of existing lib mart then the search phrase will kept in separate table miss-hit. The table stores the domain of word and the missed word or phrase. The miss-hit crawler picks the word from the table and finds in stored Digital objects of that domain irrespective of resource type, with repository either find out from the web and that word added in the database.

IV. IMPLEMENTATION PROTOTYPE

For the above mining algorithm and proposed model prepare the implementation prototype and this prototype give the detail of the work implementation and the scope of the prototype is limited and the following sequence show the client server communication of the proposed work and finally it shows the detail of the searching method or mining method.

```
Client
    Prog-cl0
Send request for opening page (Browser)
Server
    Prog-se0
Read request
Open session
Send response – opening_page
Client
    Prog-cl1
 Ask Search word
 Search Domain   ( Computer,Electr..)
 Submit
  Server
      Prog-se1
 Receive (Search Word, Search Domain)
Get domain libmart address from
libmat_classification.DB
Open libmart of specific domain libmart (subjective
Libmart)
Search the word in libmart
If   found
Read all resource_word_id & resource_type of search
word from  subjective_libmart.DB (there may more then
one resource type for seach word)
Send response for further selection of resource type
Else
Send response "Not available" – Hit fail
Save in miss-hit table
Save in log table
Terminate the session
Request to open_page progcl0
                End

    Client
        Prog-cl2
                Get response
 Display response
 If  no_Error response
 Read selection for resource type
 Sent response
  Else
  Save in log table
  Terminate the session
  Request for opening_page progcl0
  End

      Server
          Prog-se2
Receive  Resource type  from response
Open word_resource_warehouse.DB
Search Resource_ID in word_resource_warehouse.DB
Read (Resource location, Object_ID, Object type, page
no)
If more then one page no
While end of list
Open Object_resource_warehouse.DB
Search for object_ID in
Object_resource_warehouse.DB
 Read Locatin_path, Locaton,
Invoke the editor of object_type at client side & send
content of page_no
End while
Else
Open Object_resource_warehouse.DB
Search for object_ID in
Object_resource_warehouse.DB
Read Locatin_path, Locaton,
Invoke the editor of object_type at client side & send
content of page_no
Endif
Write log
Terminate session
 Request for opening_page progcl0
```

V. CONCLUSION

The major strength of the Multilayered Architecture approach provides a tight integration of database and data mining with resource discovery and content accessibility from repository of objects. a new approach, called multilayered Architecture approach using web mart, has been proposed and investigated for resource discovery and content accessibility. The study shows that the subjective accessibility of content is simple with Multilayered Architecture approach. The creation of metadata can be constructed automatically and updated by integration of data analysis and data mining techniques. The search performance depends upon how efficiently and effectively classification of resource is done in such a multiple layered database. The model facilitate **Machine learning** and My **Book-self** concept makes database dynamic and reduce the additional load of searching same for same user.

REFERENCES

[1] R Cooler, B Mobser & J Shrivastava, "Web mining : Information & Pattern Discovery on WWW"

[2] Margaret H Dunhum "Data Mining : Introductory and Advance Topics" LPE – Pearson Education Publising

[3] Arun K Pujari "Data Mining Techniques " Universities Press P Ltd.

[4] Stefano Ceri, Marristela Matra, Francisca Rizzo, and Vera Demalde , "Designing Data-Intensive Web Applications for Content Accessibility Using Webmarts", Communication of the ACM , Vol 50 No. 4 April 2007.

[5] Sun Micro system INC. "Digital Library Technology Trand"

[6] http://archive.cabinetoffice.gov.uk/egovernment/resources/handbook/html /htmlindex.html) 2002

[7] METS: metadata encoding and transmission standard: primer and reference manual -An Overview & Tutorial Version 1.6 September 2007 http://www.loc.gov/standards/METS

[8] Ronald Snijder **"**Metadata Standards and Information Analysis" A Survey of Current Metadata Standards and the Underlying Models 2001(Metadata Standards and Information Analysis1.doc)

[9] http://www.cyberartsweb.org/cpace/ht/lanman/wsm1.htm

[10] Frawley**,**W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992, pgs 213-228, 1992

[11] Oren Etzioni. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11), 65-68, 1996.

[12] S.K.Madria, S.S.Rhowmich, W.K.Ng, and F.P.Lim. Research issues in Web data mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference. DaWaK'99, pages 303-312, 1999.

[13] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota