SUBJECTIVE CONTENT ACCESSIBILITY USING DATABASE APPROACH FOR DIGITAL LIBRARY

Sachin yele
Rajiv Gandhi Prodhyogiki Vishwavidyala, Bhopal
ShriSatya Sai Institute of Technology, Sehore

Surendra Mishra
Rajiv Gandhi Prodhyogiki Vishwavidyala, Bhopal
ShriSatya Sai Institute of Technology, Sehore

Kamlesh Patidar
Rajiv Gandhi Prodhyogiki Vishwavidyala, Bhopal
Medicaps Institute of Technology & Management, Indore

Pankaj Dalal
Rajasthan Technical University,
Shrinathji Institute of Technology, Nathdwara

Dr. Suresh Jain
Institute of engineering & technology
Teaching Department of university
Devi Ahilaya Univercity, Indore

Abstract

Today's digital library is a massive collection of various types and categories of documents. The existing search engines do not provide subjective search from the collection, as no information about context is stored. The existing search engine mostly uses the agent based search then the database based search. The database search is simpler easier but static verses dynamic Web. The work shows how database become dynamic, subjective and search query becomes simpler. The subjective and context based search is necessity of searching in Digital Library. The user who may be researcher, students, and even common person expect subject or context and need content accessibility precise and subject specific. This paper presents the topic-word specific subjective search using the database approach in digital library, by data mining technique in warehouse.

Key-words: Warehouse & data mining, Database, Metadata, Accessibility, Content accessibility, data mart - lib mart.

"1. 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 web content mining is vast field as web has huge, diverse applications and unstructured resources. The approach for web content mining varies from application to application.

The Web content mining has two major activities, first to identify the relevant object document and second is to access the content of it.

The identifying (searching) the document on web has two distinct approaches. The **Agent Based** approaches use the caching, indexing, gatherer, broker or scatter techniques to generate web agents. There are three categories of such agents. They are Intelligent Search Agents, Information Filtering categorization and Personalized Web agents.

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

The **Database** approach focuses on techniques for organizing data into more structured way and using standard DB querying mechanisms & data mining techniques to analyze it. The categories in this approach are Multileveled Database (Multilayered) and Web Query System. This approach makes more organized, fast and precise query.

The digital library can be treated as warehouse of documents (information) which makes the subjective content accessibility vast, diverse and dynamic.

The content of the search item must be easily available – content accessibility is ultimate aim of a person. Accessibility has been defined by W3C as need to create web content that is perceivable, operable and understandable by the broadest range of users and robust enough to work with current and feature technologies [4]. The content accessibility requires a clear identification of few core contents that synthetically convey the information of the entire application, and then repeated use of few, well designed access patterns, to give users the impression of mastering the process of retrieval and navigation [4].

Modelling the content is most important aspect of data - intensive search. The proposed database approach model access the content subject domain specific.

Users have always want 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 or even one single webpage. The collection of digital library can be classified in different resource types of digital documents.

The content is an information-warehouse of verity of documents of different resource types. It is referred as resource objects, and need the process of classification of resource objects. The process of subjective extraction from warehouse makes data-mart subjective, say as lib-mart.

The classifications of the domain have subjective words or phrase of the domain. This search is subject domain specific – **Domain Semantic**.

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.

"2. Methodology"

2.1 Concept

The search item is one word or multiple words. The search item is mapped with metadata modelled as database. The data modelling of metadata provide the classification as lib-mart and word resource warehouse, where word is mapped to digital object like word index of book.

The concept is using multilayered database approach for searching rather then agent based. The model data are stored in multiple layers.

The first layer is domain semantic - specifies the domain of document. Now the search on subject specific is possible. The documents stored in digital libraries 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, which** points out page number on which content is available.

The second layer has classification by resource type from database view. The digital libraries have digital resources like e-books, research paper, white paper, reports, thesis, dissertation etc

The third layer is details of objects. The whole digital collection of documents needs to organised documents attributes-wise so that search is also possible by attribute of documents like title, author etc.

The digital resources in library are organised and maintains by using the **metadata** - data about data. The metadata are organized in tables of databases.

Here we model over this metadata database to visualize the Digital Library as warehouse of words and words are classified into subjective datamart – webmart-libmart.

The subject-wise words are gathered in datamart serve as index. And the document- metadata serve as object resource warehouse like **book**

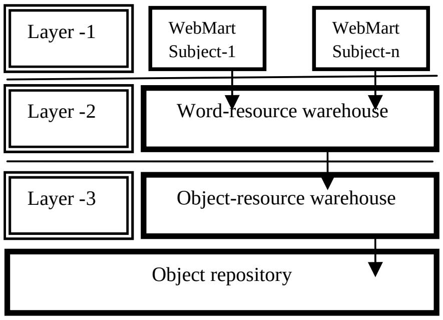


Figure 1 – Block diagram of multilayered database using libmart

The model treats the documents as bucket of words. The words are classified subject-wise. The classified words are kept in many buckets i.e. libmart. The words in the libmart do the job of domain classification – Domain Semantic. The Unique ID identifies the words in libmarts, which points out the resource. However, the digital objects are also classified by resource type like e-Book, Research Paper, etc. The words in first layer are mapped with the resource type of digital object. The accessibility of contents is made easy by the page number of digital object. This is second layer word-resource warehouse operates. The third layer is metadata of digital objects operates as an object resource warehouse as shown in figure 1.

The object resource warehouse is descriptive metadata of object of physical repository. The descriptive metadata are modified Dublin descriptive metadata element set [5]. The Dublin metadata element set keeps details of entire range of objects.

The model helps to kept all digital storage in compressed form, which saves the storage space, and it uncompressed it at the time of content delivery.

2.2 Data Modelling

The normalise Data model block diagram is shown in figure 2

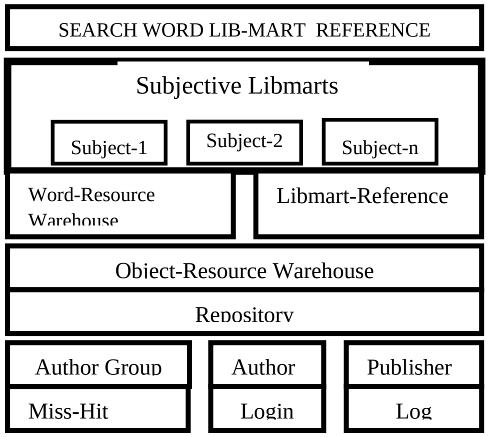


Figure 2 – Block Diagram of Data Model in normalised form

2.2.1 SEARCH WORD LIB- MART_REFERENCE

The two attribute Word and Lib-mart_ID, gives Libmart_IDs in which the word is located. This provides the search the word in multiple libmarts. The word can be in more then one libmart as belongs to more then one domain i.e. subjects

2.2.2 LIB-MART_CLASSIFICATION (No of Lib-marts)

This shows the lib-mart is of which subject domain and its path where it is stored. The three attributes, one shows ID, second shows the subject domain as classification (Engineering, Astrology,….), the last keeps the physical name & location of lib-

2.2.3 SUBJECTIVE LIB-MART

The three attributes maps search word with resource type and the structure metadata of object as key Resource_word_ID which classifies the digital objects by resource type. The word can be found in multiple digital objects, and user can select one of own interest.

There are various libmarts as per subject-classifications. The word may found in multiple libmarts i.e. multiple domain word, can be distinctly shown to user, and can select one domain and avoided unwanted results. The word 'model' is used by almost all sub-domains of engineering & technology. This way datamart concept makes the search results narrowed to specific context – Subject.

The words are repeated in the table but with resource_word_ID it makes unique key.

The classification of content may be like

A. Professional & Technical
    a. Engineering
        1. Aerospace
        ………………
        9. Computer Science & Engineering
            i. Data structure
            …. ……
            vi. Neural Network
            ….…
        99. Environmental
        100. Information Technology
        ………………

2.2.4 WORD RESOURCE WAREHOUSE

The search word finally maps with page number where the word appears in specified resource type of digital object. There may be multiple digital object of same resource type. Moreover, same time the word is on multiple pages of one digital object.

The user may choose one page number or can choose one after other. The unique object_id of the digital object of word, identified here by resource_word_ID, is used to get various metadata from Object_resource_warehouse table.

This layer acts as the warehouse of words with two classifications, one by word and other by resource type. The resource_word_ID with resource type and page number is unique key.

2.2.5 OBJECT RESOURCE WAREHOUSE

The complete Descriptive, Technical, Administrative, and structural metadata of digital object, irrespective of subject domain or resource type are stored in this table. The Dublin metadata element set [5] is modified to use in this model to represent verity of objects.

The Object_ID is the unique key to identify a digital object. The model allows keeping metadata of object from conventional physical library and remote digital objects.

The metadata facilitated application to search on Title, Author, and other metadata.

The various following other tables are out comes of normalisation in data model. The tables are:

1. **Libmart Referance –** The libmarts are created subject-wise, and word may be in more then one libmart.

2. **Author_Group Table –** The digital objects have one are more authors. The author_group table have author_group_ID and Author_ID.

3. **Author Table –** The author table maintains author details

4. **Publisher –** The table maintains publisher details.

5. **Miss-Hit –** The table is added to facilitate the Machine Learning. The words which not found in our libmarts are stored in this table.

6. **Log –** The log table is maintain to understand the requirements of the user as well as trends of search. The document rank concept is implementing through log analysis.

7. **Book-self** – The details of user marked objects for book-self are kept. The attributes are user id, word, object id, page number.

8. **Login –** The table is used to keep the track of authorised user.

2.3 Other Models

The Making of America II project (MOA2) of University of California at Berkeley uses the Metadata Encoding and Transmission Standard (METS). MOA2 providing an encoding format for descriptive, administrative, and structural metadata for textual and image-based works. The U.C. Berkeley has three component based modular object management environment called GenDL Generic Digital Library, Discover, Display & Navigate. The model is designed to support Distributed object middleware environment [5].

The Digital Library Federation initiative, uses MOA2 and provide an XML document format for encoding metadata for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and user).

The Harvard University Digital Library has different multimedia repositories in one layer and next layer has content source. The strength lies with the multiple catalogues With Access system [5].

The Standford University Digital library uses the metadata modelled to discovery and delivery of content.

The Fedora – A project of university of Varginia, metadata is modelled to manage the objects.

The many more Digital Libraries uses the metadata only for preserving, managing the digital object but the presented model uses also for classification of object, it index and maps the content to the searched- word.

2.4 Machine Learning

The Miss-hit table and miss-hit crawler add the machine-learning concept.

The missed-hit word or phrase will keep in separate miss-hit table. The table stores the domain of word and the missed word or phrase. The miss-hit crawler searches in repository and update metadata to appropriate tables in different layers of proposed model. The other case may be that domain does not exist. The miss-hit crawler creates a new domain libmart and update metadata to appropriate tables in different layers of proposed model.

The machine learning can be extended to keep documents uncompressed for most frequently access documents, also controls the growth of individual libmarts as well as words, and object warehouse.

2.5 Advantages & Disadvantages

Advantages

The first advantage is that whole concept models metadata to provide the context.

The second advantage is multiple classification of same content to different contexts. The repository has digital object on the Tajmahal. This document can be classified in **history, architecture** and **Wonders of world** lib-mart.

The third major advantage is that classification is dynamic. A new lib-mart can be added with another classification at any point of time. The lib-mart **architecture** is further sub-classified and an **ancient-architecture** new lib-mart can be added.

The fourth advantage is a new libmart may be added not only on the subject classification basis, but on the need basis also, say for some project, or combining few subjects and generating a new libmart. This helps in fast accessibility of content for the user.

The fifth advantage is that a powerful search engine can be used to search on word, title, author, publisher, other attributes and combination of these.

The sixth advantage is that indexing required for metadata tables only. The digital object metadata are only indexed on object_ID not on other attributes like keywords.

The seventh advantage is that digital objects can be stored in compressed form to save storage space and decompressed when it is required.

The eighth advantage is that machine learning. The miss-hit crawler, which finds miss-hit word in the documents of repository and update metadata details in respective tables.

The ninth advantage is that the concept of user book-self. The user can search the desired document and mark for his book-self so next time when user logins can directly access from book-self.

The tenth advantage is that the layering of metadata made the SQL query simpler.

Disadvantages

The major disadvantage is static in nature, i.e. the metadata is added to database manually, though it is only one time job. This disadvantage can be handled at large by automatic extraction of metadata from digital object through lib-crawlers for few of metadata rest metadata can be manually added to table.

The second disadvantage is need of human intervention to provide context i.e. classification.

The third disadvantage is copy write law for remote documents

The fourth disadvantage is word redundancy in two tables of different layers. The issue of integrity may arise.

"3. Application Architecture"

3.1 Three Module Architecture

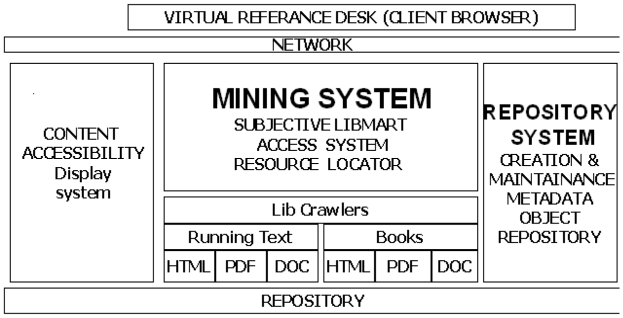


Figure 3 – Application Architecture & components

The model is only useful if it supports application architecture. The client-server architecture, web architecture can implement proposed model. Here discussing one general-purpose application architecture to support the proposed model of 3 modules as shown in figure 3.

The **first module** covers creation and maintenance of metadata. The details of objects are kept in database tables which classifies the repository in subjective topics.

There are two ways to update metadata, one is manual for offline entry, second is lib-crawlers for each storage type of digital objects

The lib-crawlers, which makes the application dynamic from static nature of database approach. The lib-crawlers will read metadata from the digital objects and update. The digital objects can be broadly classified into three HTML, running text and books. Further sub classified by storage formats as PDF, DOC, etc..The crawlers for running text have separate approach then books and HTML. The key words are available in the index of books where keywords are not available in running text. The user gives key words in case of running text. The hyper text plays the role of keywords in HTML document.

The digital library has stored objects in the local repository. It can further enhance by adding two more database table containing metadata for remote repository on web referring to other sites where object is available and paper repository (traditional library) or all in one metadata table by modifying the metadata element set. The all required metadata are not embedded with documents. This leads to enter remaining metadata manually.

The **second module** is mining system i.e. search the item from three layered database. The first item is searched in the word-lib table, which shows that word is associated with one or multiple lib-marts. The mining module picks-up the resource_IDs from each libmart. These resource_IDs are searched in the Word_resource_warehouse table, which maps the Object_id, resource_Type and page numbers in objects. This may have multiple object resource. This details are used in two ways, one ask user to choose one or can deliver one by one.

The **third module** is displays/delivery of content. The module diagnoses the storage format of item and invokes appropriate application like acrobat reader etc.

The proposed architecture can be implemented in two-tier or three-tier client-server architecture for digital library on Intranet or Internet using layered database approach which can provide SQL query simpler, access security and makes whole digital library as classified information warehouse by using the lib-mart (data mart) to serve the purpose of context to content for context based information retrieval.

“4. Experimental Result”

The model has three-layered approach in figure 4.

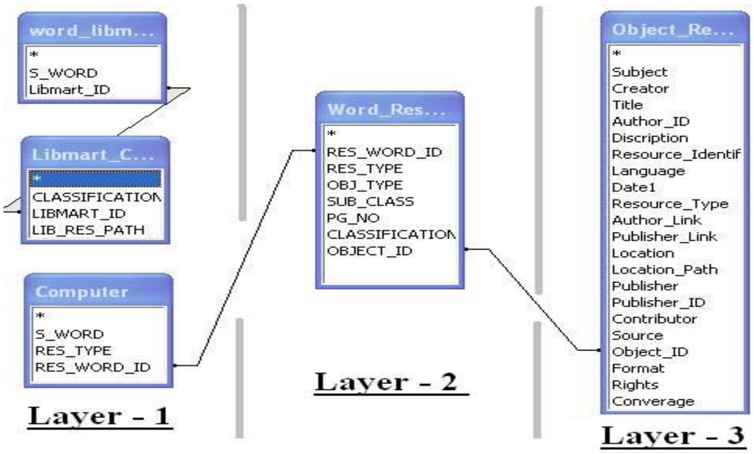
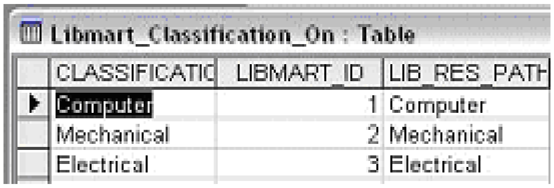


Figure 4 – Database tables layers-wise

The **layer one** does the domain analysis. It has three tables, which identify the libmarts. The word may have more then one record as it is referred in many domain i.e. subjective libmart.

The value of classification is the name of subjective libmart. It is shown in Database table 2 Libmart_Classification in layer 1 of figure 4. The attribute ‘Classification’ has value “Computer” as the name of libmart and lib_Res_path gives physical path as show In figure 5.

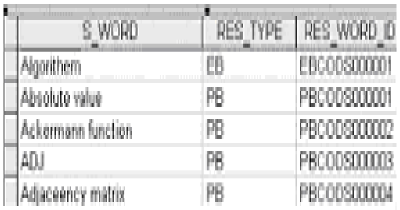


Libmart_Classification_On : Table

CLASSIFICATIO	LIBMART_ID	LIB_RES_PATH
Computer	1	Computer
Mechanical	2	Mechanical
Electrical	3	Electrical

Figure 5 – Libmart Classification Table

The figure 6 shows mapping of word in Subjective Libmart ‘Computer’

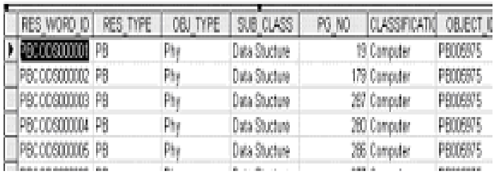


S_WORD	RES_TYPE	RES_WORD_ID
Algorithem	EB	EBCODS000001
Absolute value	PB	PBCODS000001
Ackermann function	PB	PBCODS000002
ADJ	PB	PBCODS000003
Adjaceency matrix	PB	PBCODS000004

Figure 6 – Subjective Libmart ‘Computer’

The **layer Two** identifies objects of various resource types, which are having the word from table, which contains EB - eBook, PB - Paper Book , PR - Physical Research Paper, PW - Physical White Paper, EM – eMagezine, PM - Paper Magazine, ER - Digital Research Paper, EW - Digital White Paper, EA - Digital Articles, EJ - Digital Journals, RR - Remote Research Paper, RW - Remote White Paper.Here we take Physical for Paper Library, Digital for Digital Library, Remote for Link to other WWW

The res_word_ID is consist of first pair of letters is resource type, second pair of letters is classification, third pair of letters is short name derived from title to make more unique and six digits unique number.

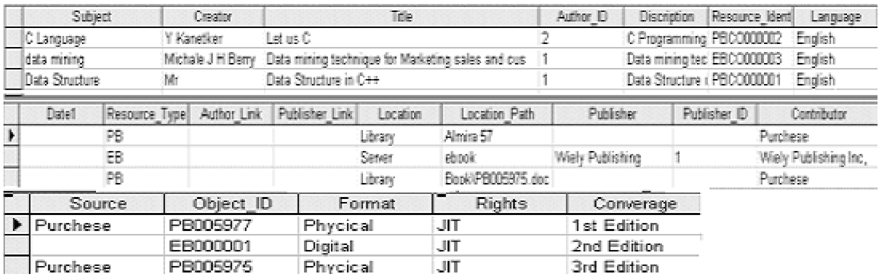


RES_WORD_ID	RES_TYPE	OBJ_TYPE	SUB_CLASS	PG_NO	CLASSIFICATIO	OBJECT_I
PBCODS000001	PB	Phy	Data Stucture	19	Computer	PB006975
PBCODS000002	PB	Phy	Data Stucture	179	Computer	PB006975
PBCODS000003	PB	Phy	Data Stucture	287	Computer	PB006975
PBCODS000004	PB	Phy	Data Stucture	280	Computer	PB006975
PBCODS000006	PB	Phy	Data Stucture	286	Computer	PB006975

Figure 7 – Word Resource table

The table word_resource_warehouse contain the words, page numbers that are associated with resource_ID and object_ID of the resource. The table is having two records of resource_word_ID PBCODS000008, but page numbers associated with it are different.

The **layer three** furnishes the metadata of document objects in table Object resource metadata as shown in figure 8. The details of the selected object resource from Word Resource table are gathered and used for content accessibility.



Subject	Creator	Title	Author_ID	Discription	Resource_Ident	Language
C Language	Y Kanetker	Let us C	2	C Programming	PBCO000002	English
data mining	Michale J H Berry	Data mining technique for Marketing sales and cus	1	Data mining tec	EBCO000003	English
Data Structure	Mr	Data Structure in C++	1	Data Structure i	PBCO000001	English

Date1	Resource_Type	Author_Link	Publisher_Link	Location	Location_Path	Publisher	Publisher_ID	Contributor
	PB			Library	Almira 57			Purchese
	EB			Server	ebook	Wiely Publishing	1	Wiely Publishing Inc,
	PB			Library	Book\PB005975.doc			Purchese

Source	Object_ID	Format	Rights	Converage
Purchese	PB005977	Phycical	JIT	1st Edition
	EB000001	Digital	JIT	2nd Edition
Purchese	PB005975	Phycical	JIT	3rd Edition

Figure 8 – Object resource metadata table

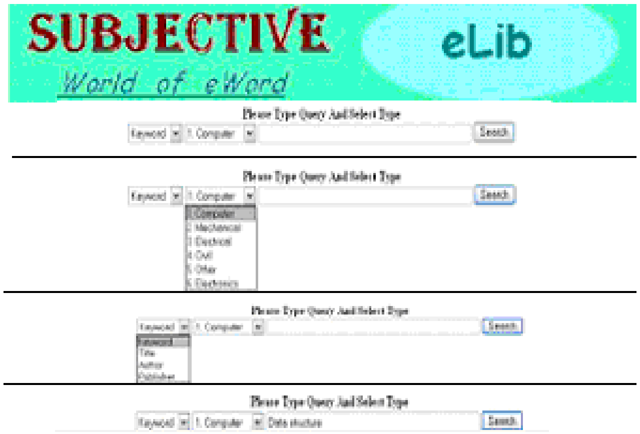
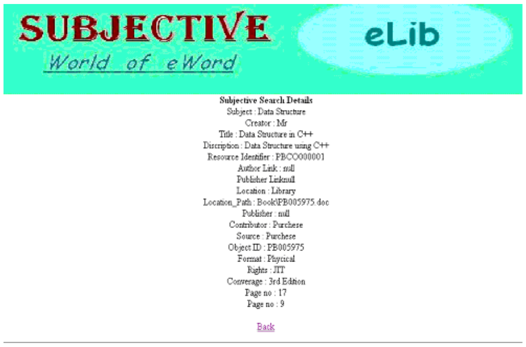


Figure 9 – GUI for asking search details

The application GUI shown in figure 9 asks three things for search. The first the search on attributes of documents title, author, etc. second domain and third phrase to search. The figure shows search on 'keyword' of domain 'computer' and word 'Data structure'

Figure 10 -Query results if success



Subjective Search Details
Subject : Data Structure
Creator : Mr
Title : Data Structure in C++
Discription : Data Structure using C++
Resource Identifier : PBCO000001
Author Link : null
Publisher Linknull
Location : Library
Location_Path : Book\PB005975.doc
Publisher : null
Contributor : Purchese
Source : Purchese
Object ID : PB005975
Format : Phycical
Rights : JIT
Converage : 3rd Edition
Page no : 17
Page no : 9

Back

Figure 11 -Detail of Query results

The Figure 10 shows resource type available for the ‘Data Structure’ phrase like ‘Electronic Book’, ‘Physical Book’ and ‘Electronic Research Paper’out of which The ‘Electronic Book’ is selected. The figure 11 shows the details of the same

“5. Conclusion”

The Different from the existing digital library, a new approach, called multiple layered database (MLDB) approach using webmart, has been proposed and investigated for resource discovery and content accessibility. The approach is to construct subjective digital library in which accessibility of content is as simple as accessing content of word from index of a book.

The major strength of the MLDB approach provides a tight integration of database and data mining with resource discovery and content accessibility from repository of objects.

With the dynamically growing digital library the performance will remain almost constant.

The study shows that the subjective accessibility of content is simple with MLDB 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 subject classification can finely granule to deepest branch of main subject.

The study presents a general application architecture which shows that the MLDB approach for resource discovery, simplifies the SQL query.

The MLDB approach with libmart provides the dynamic structure to digital library. The growth of digital object doesn’t affect the search time and structure of digital library.

The model facilitates **Machine learning** and My **Book-self** concept makes digital library dynamic and reduce the additional load of searching same for same user.

The digital library can also contain the metadata of physical library and remote link of objects in other digital library or web server. The globalization of digital library can achieve by sharing metadata level.

“6. 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] www.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] 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. May 2000.

[14] M.Spiliopoulou. Data mining for the Web. In Proceedings of Principles of Data Mining and Knowledge Discovery. Third European conference, PKDD'99, p588-589.

[15] J.Borges and M.Levene. Data mining of user navigation patterns. In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999. San Diego, CA, USA, pages 31-39, 1999

[16] R. Kosala, H. Blockeel. Web mining Research: A Survey.

[17] O. Zaiane, J. Han, Z. Li, S.H. Chee, J.Y. Chiang. MultiMediaMiner: A system prototype for MultiMedia Data Mining

[18] Metadata for E-commerce **by** Tom Worthington FACS This document is Version 4.1 16 April 2002: www.tomw.net.au/2001/metadata.html