Taxonomy based Metadata Classifier

Asiya Abdus Salam Qureshi
Department of Computer Science and Software Engineering
University of Hail, Saudi Arabia

Syed Muhammad Khalid Jamal
Department of Computer Science
University of Karachi, Pakistan

ABSTRACT

This paper proposes new approach based on the breakdown of metadata repository into taxonomy based metadata classifiers to classify the information. Due to the raising quality issues, it results in avoiding metadata from being processed correctly. The inconsistent metadata makes it difficult to locate relevant information. In the multitier architecture of data warehousing, there is a need to break metadata repository to handle the information.

From warehouse, information like data names and definitions of that warehouse is marked by metadata. The reason for construction of metadata is also discussed. Information like data warehouse structure, operational metadata, algorithms, mapping, system performance related data and business metadata are contained by the repository. This storage of information and management should be persistent. This approach will split the heavily populated data warehouse into data marts to control and manage data in immensity which results in controlling of time consuming and slow working. New method is introduced here based on dividing the metadata repository into data marts.

This paper is discussed as follows. First part is the introduction of the metadata and taxonomies. In the second part, need of breaking metadata repository into data marts is discussed. Statistical framework from metadata repository's point of view is delineated in third part of this paper. How data warehouse is managed and its components are conversed in next section. Methodology is conferred as steps in implementing a data mart. After the related work, conclusion and future directions are given at the last part of the paper.

Keywords

Data warehouse, Data marts, Metadata repository, Taxonomy based data marts, Taxonomy based metadata classifier.

1. AN INTRODUCTION TO METADATA AND TAXONOMIES

As data warehouse[1] collect information from multiple sources, store under unified schemas so this enormous area should be more organized. Metadata is considered to be the key for assurance about the existence of resources and to be handy in future. As warehouse objects are to be defined, metadata is used.

Metadata and taxonomies play vital role in defining strategies for quality management system or web content management system that support the goals of the group for categorizing and linking both resources and contents.

Metadata can be taken as information about an quality instead of just the basic filename. Metadata would be any kind of element or component that helps to classify or portray a meticulous illustration, file, presentation or spreadsheet.

Taxonomy is a Greek word "taxis" which means array or division and "nomos" means law. It can be defined as the science of categorization based on a pre-determined scheme with the consequential catalog that provides an abstract for argument, investigation, or content recovery.[2]

There is difference between taxonomies and metadata. Taxonomy helps to systematize content and resources into hierarchical relationships. In case of indefinite searching, classifying content and quality in taxonomy then go for computerized quality organization or web based content management. [2]

There are many benefits in defining and using taxonomy. With the help of taxonomy, system users can categorize subject and information by using a proscribed terminology. This restricted lexis can be an integration reference point.

Taxonomy should have a particular use and organized in a logical hierarchy. Users can easily understand about the levels in a form of different sections or units. Also unnecessary taxonomy cannot be defined for other metadata. This makes the provided metadata information easily searchable and defined to user and computer.

Meta data can be structural, administrative or descriptive. In structural type, metadata elements are collected based on a mix of organizational and system needs.

The word taxonomy has different contents. The Information Technology world uses taxonomy to any sort of formation that organizes information. Instead of taxonomies, "controlled vocabularies" is used by information science people. The major objectives are to describe a content component to create some level of steadiness and manage the information used, and illuminate relationships between them.

Merely, taxonomy arranges and control information and metadata portrays it. Metadata store the information for the taxonomy to be able to organize it. It all works together to make the content searchable, identifiable, and functional.

2. NEED OF BREAKING METADATA REPOSITORY INTO DATA MARTS

Metadata grant structural design or agenda relating user's data contained by a data background. This architecture ought to supply specific, consistent and rational structure that "paints a picture" of the records. It also ought to depict how the internal environmental data is interfaced to the external world. [1]

Repositories allow for a centralized metadata location. They can systematize and hold various forms of meta data into different modules to allow suppleness between systems. This modularity allows dissimilar applications to accept and decide modules needed to incorporate these systems.

Among organizations assimilation information assets is not an effortless task. Frequently end users find themselves as a victim to unapproachable or contrary metadata protected for a specific application's tool set. In the area of information substitution, metadata has become the number one mixing problem for the following reasons given below.

For dispensation of information assets, there are multiple tools. Usually each tool has its own way of utilizing and configuring metadata. The same metadata is entered multiple times in different formats to integrate systems.

Updating tool sets is crucial for the success as tools for integrated systems change. As more efficient tools come along, the need to interface today's saving application with the accessible tools interfaces and metadata can be devastating and at times impractical.

For possible review organizations may wish to follow metadata for online transactions processing or data warehousing. This often requires removing metadata from the individual mechanism and then guiding them together to show the combined arrangement. Due to the different ways the tools represent the data, this adhoc approach makes the metadata may be unnecessary and appear conflicting.

Increasing and preserving a complete set of metadata for an organization may results in unaffordable cost. Large businesses may have several organizations for developing and maintaining metadata for their individual use. Thus they are missing out on the benefit of being able to share and reuse the metadata across the various departments.

This new approach will introduce to overcome all the issues discussed as shown in figure 1 taxonomy based meta data classifier.

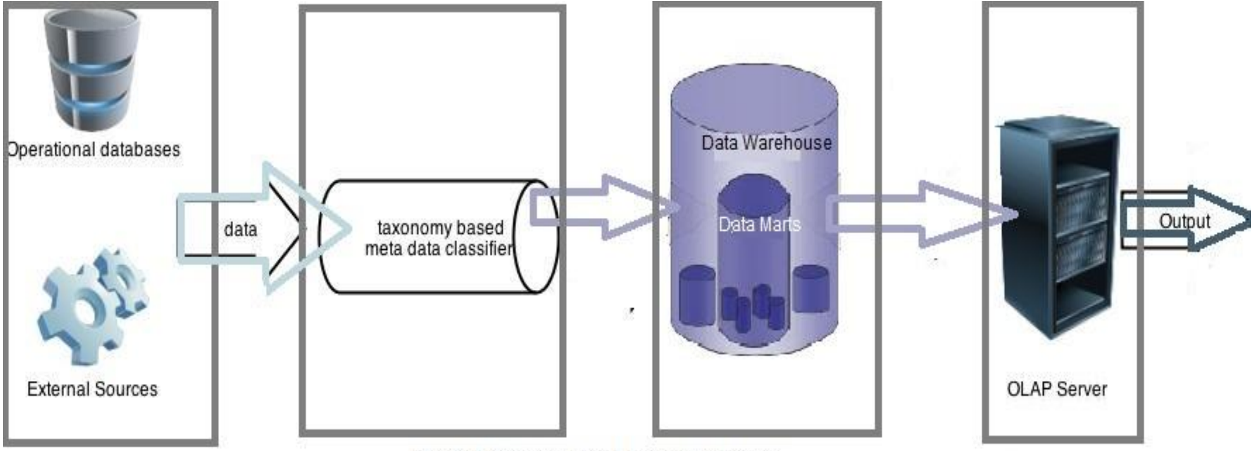

Figure 1 Taxonomy Based Metadata classifier

3. STATISTICAL META DATA REPOSITORY

Rationally a statistical metadata repository is a fundamental statistical metadata warehouse that allows for the query, controlling, and organizing of metadata. It is a system that provides looking up information about statistical field as well as their plan, expansion, and analysis. [4]

Frequently metadata is sprinkled, shortened or omitted. Many times the only basis for some information is from focused authority. The useful and proficient organization of statistical metadata greatly increases the effectiveness of statistical data. Because metadata is data or statistics, it can be stored and retrieved in a warehouse similar to the data stored and retrieved in a database. Statistical metadata repositories are designed for many functions mainly it is a customary means for researchers and analysts to find information and descriptions of surveys. The sort of information that is directly available in such a repository comprised of statistical dictionaries, record layouts, feedback forms, model designs, and standard errors. Evaluation of designs of different surveys is done by users and through different survey finding for common information collected.

The exposure of the organization of statistical metadata is the fabrication of metadata storage area with global network system, access to retain Internet records allocation to automatic analysis plan and dispensation tools. The warehouse manages data reciting the plan, dispensation, investigation, and records. It will be potential to trace all the accessible information describing a single survey and locate explicit types of information for any set of assessment. The availability of metadata is for lifetime of a review.

Metadata is gathered mechanically as a purpose of programmed assessment plan and processing tools. Metadata relating surveys, variables, data sets, products, and other items will be obtained through a range of thorough competences. Comprehensive terminologies (e.g. thesauri or taxonomies) will aid to explore the unambiguous objects or matter of concern.

Collecting the descriptive statistical information to inhabit the warehouse is not simple. Assessment forecasters and designers frequently construct metadata only for themselves and occasionally as a postscript. If wanted to know about the significance of metadata, the designers and analysts shows the importance of data but they don’t have the point or resources to enter it into a repository. Valuable, preset breakdown map and dispensation tools collect the metadata without considerable additional attempt. When the users perceive the repository as an essential part of the work success is achieved.

Different types of metadata warehouse include substances like compilation, listing, crosswalk, upholding and query. Here describe exertion to put up computerized analysis that supplies data collection as a consequence of their functions. Listing guarantees and follows the quality of metadata.

Data warehouses, data marts and other large storage network, are greatly vital to worldwide associations. Managing these databases begins with a superior life-cycle procedure that contemplates the functional aspects of the organization which is based on the accessibility and presentation of the structure. The functional supervision of a data depot should ideally focus on speedy aspects. Placing the controlled part of the database as a primary storage, for instance Oracle, declare the RDBMS vendor and the vendor’s own administration. Working database, along with other data storage archives, provides utmost simplicity of management best possible execution and accessibility.

The previous ingenious support method is also available that can identify hot spots and eliminate them on the fly. The decision support needs of the data warehouse can be ‘localized’ to the user group. File-system level of duplication services can be used to provide enhanced disaster recovery, remote backup and multiple remote sites adding further flexibility and providing finest ‘read’ access. Frequently available, emergent, high presentation data warehouses provide the facilities like elevated accessibility and sophisticated organizing information.

4. MANAGING DATA WAREHOUSE

Data to data warehouse is approached from an ample variety of sources, and is accessible through a logical record system. Data ware houses tend to be very large in size, used on a large percentage and can be reachable transversely to the world. The supposed value of a data warehouse is that executives and others can gain accurate result by having immediate access to relevant shared information. In fact, leading organizations now comprehend that information has become the most effective asset in today’s demanding markets.

A data warehouse contain divergent things, ranging from the predictable monetary, development, sorting and patron data, through manuscript, authorized and project data, to the daring new world of internet and local internet including market facts, press, multi-media, and links to web sites depending on the business having a common distinctiveness like absolute dimension, interconnected data from various sources, and use by workforce.

Care must be taken to ensure that the information contained within essential introductory data in such a particular central spot, its performance meets the demanding needs of the modern user and is obtainable greatly, managed effortlessly, recoverable in time, protected and backed up as shown in figure 2.

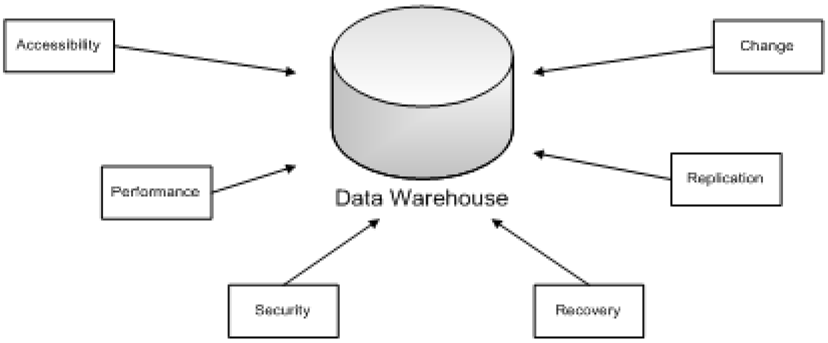

Figure 2 Management of Data warehouse

4.1 Data Warehouse Components

The data warehouse is fashioned by integration of allied data from many different sources into a single database which is more managed. Archives and data sources may be accessed by associations from the original source, across partner sites, unique intranet sites or out to internet. [5] More intricate systems copy associated files that may be better kept outside the database for diagrams, illustrations, statement dispensation documents, images, resonance etc.

More structured data, that is summarized and coherent information, is often a blend of current, instantaneous and unfiltered data. It might relate to a quarterly period or a snapshot of the business. To promote faster, drive profits and service levels and manage change, the target is to enhance information and made quick access to enable information.[9]

In general, a data warehouse is comprised of three parts – a mass administration, data repository, and a query organization module as shown in the figure given below. This diagram is taken as a reference for discussing the need of data warehouse.

Load or mass administration is the compilation of data from contrasting in-house or outer sources. The loading system consists of summarizing, manipulating and changing the data structure into a format that provides itself to logical processing. Genuine data should be kept within the data warehouse thus enabling the new construction and different representations.

Data repository is the management of the data depot regularly. The warehouse management tasks includes ensuring its availability, the efficient backup of its stuffing, and its fortification.

Query organization[11] communicates the stipulation of the warehouse contents. It may include the splitting of information into unrelated areas with different privileges to different users. Access may be provided through ad hoc query tools or custom-built purposes.

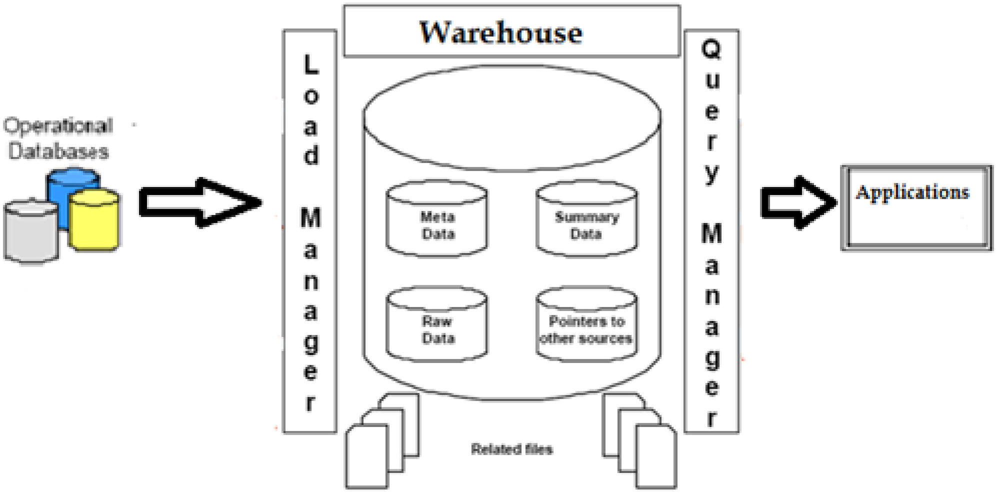

Figure 3 Data warehouse Components

5. STEPS IN IMPLEMENTING A DATA MART

In order to implement a data mart the main steps involved are designing schema, constructing substantial storage, occupying data mart with data from source systems, accessing data to make conversant assessments and managing it ultimately. [8]

In data mart process, scheming is the first step which is entirely responsible for commencing the data mart demand through collection of information about the necessities regarding data and mounting rational and substantial plan of the data marts. Designing phase involves gathering the commerce and technical requirement, classifying the major data supply, selecting the suitable detachment of data and scheming the physical and rational structure of data mart.

Second step in data mart process is the construction of storage which includes making of physical database and rational structures linked with data mart. The advantage of this step is to grant fast and efficient access to the data. Main tasks involves in this step are generating physical database and storage structures. Also includes creating scheme substances defined in design step such as tables and indexes. It also determine how best to set up the tables and access structures.

Populating is the next step that includes tasks related to receiving the data from the supply, cleaning it up and moving the modified data which issue right format and detailed levels to the data mart. Mapping of data sources is also done in this step to mark data structures. Data is extracted and cleaned which is transformed afterwards to load into data mart, producing and storing meta data leads to populating phase.

Accessing step includes applicability of data that results in data utilization, data querying, exploring it, forming reports, visual aid and grids and distributing them. These are the application for end users. To submit queries for storage and show result of these, end users use graphical front end tool. The main tasks performed here involves setting up an intermediate layer for the front end tool to use. Business interfaces are maintained and managed. Setting up and managed database structures execute quickly and efficiently that help queries submitted through front end tool.

The last step in data mart process is the managing data marts over time period. It provide secure access to data and manage data growth. For better performance, the system is optimized and ensure the accessibility of data at failures.

The aim of data mart is to grant access to data that is explicit to a particular department or serviceable area. For analysis that the end users want to perform, data should be at a significant level of detail. Data should be existing for the business that end user understand[10]. More informed business decisions are result of the analysis of data in data mart.

Technical requirements should be identified to get information about data that is input to the data mart. Operational systems are the primary sources of data for data marts. They handle daily transactional activities. They are Online Transaction Processing System. Data marts are fed from more than one such source. Data cannot transfer from operational system into data mart without intermediate processing. Also information is required on the frequency of data refreshed and need of data cleaning with the identification of data source.

For designing data mart, physical and logical structure is needed to define specific data content, relationship within and between groups of data, supporting environment for data mart, transformation on data and data refreshed frequency. With the help of logical and physical design relations among data objects can be made and objects can be store and retrieve effectively[13].

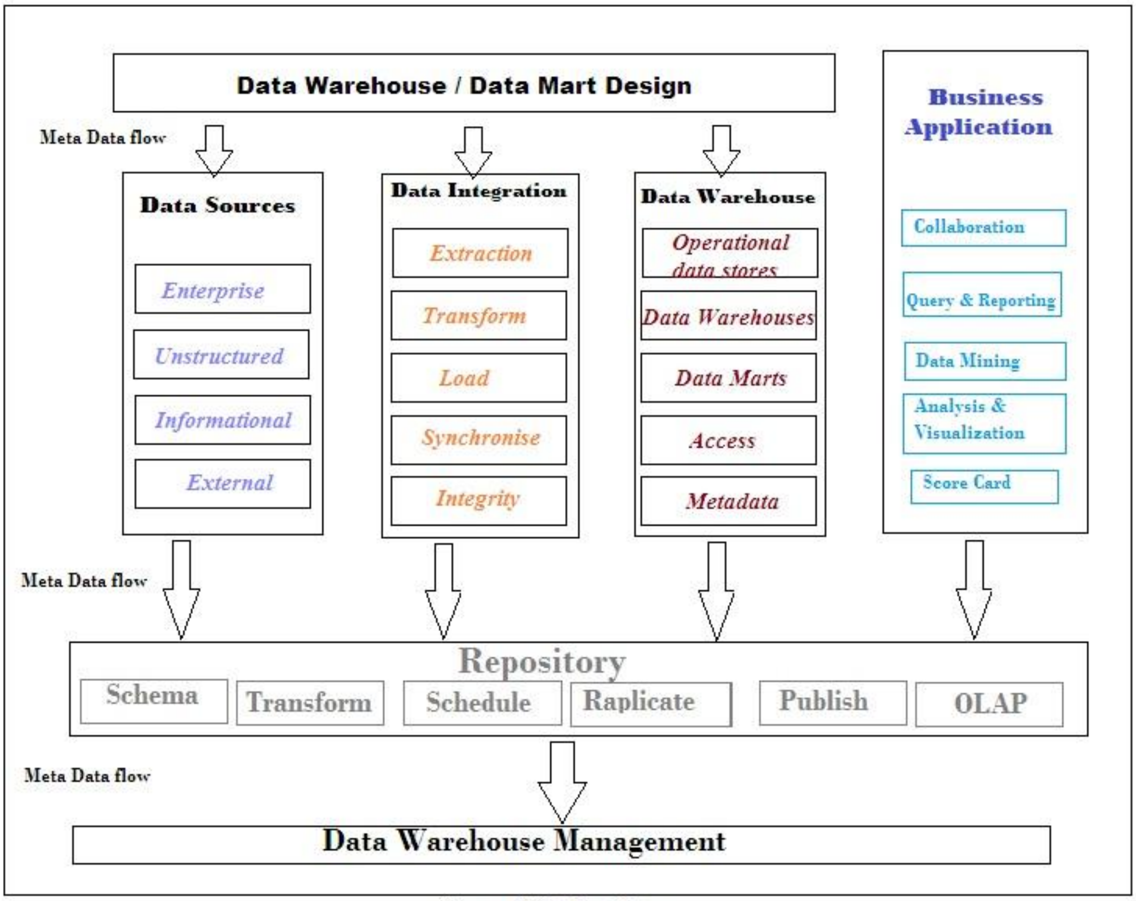

Figure 4 Meta Data Flow

6. RELATED WORK

Related work has been done formerly on the data warehouse three tier architecture in which relational database system holds warehouse database server where backend tools and utilities input information. Performance of data includes data extraction, cleaning and transformation. Loading and refreshing functions update the data warehouse.

There is a meta data repository to store information regarding warehouse and its contents. OLAP server is at middle tier and finally front end tools are at the top tier containing query, reporting , analysis and data mining tools.

Another related concept, taxonomy based data marts guide this paper through which classification of queries are done using data warehouse. New algorithm is introduced for time saving and specific results and the use of data marts are discussed.

Usually data warehouse implements an architecture that is based on the layers and query classification techniques and approaches are determined.

Lot of work is already done on meta data and its repository. Role of meta data is discussed by various researchers team. Meta data used here is regarding a directory that helps the assessment maintenance system forecaster that define the elements of the data depot.

From the prepared environment of the data depot surroundings, malformed data is mapped. This leads to guide for the summarized algorithms used between the existing data with detail and the lightly as well as highly summarized data. Persistently, meta data should be stored and managed.

This research paper splits meta data repository into data marts for efficient and easy access. Internal to the data marts, the populated data is also classified within data marts like tree structure where selection of only one specific path is done which help in creating internally divided data warehouse, ultimately yielding the goal of selected search. [14]

7. CONCLUSION AND FUTURE DIRECTION

This paper concludes that meta data repository breaks into taxonomy based metadata classifiers for categorizing the information more proficiently. Metadata repository is considered to be incompatible and unpredictable.

Metadata repository creates problem in locating or searching the significant information. The deeply inhabited and occupied data warehouses need to be handled and controlled. Therefore new approach called taxonomy based metadata classifiers is delineated.

More work can be done in this respect for handling of information. New techniques can be defined to classify meta data. Moreover query diversification based on statistical framework can be added through profound research.

Relational schemas can be redesigned in terms of query classifiers. Check for disaster recovery and online storage or offline storage can be one of the future task.

8. REFERENCES

[1] Asiya Abdus Salam Qureshi and Syed Mohammad Khalid jamal. 2012.Taxonomy based data marts. In International Journal of Computer Application., December 2012.

[2] Asiya Abdus Salam Qureshi and Syed Mohammad Khalid jamal. 2012.Web supported query taxonomy classifier. In International Journal of Computer Application., August 2012.

[3] Christian Platzer, Clemens Kolbitsch and Manuel Egele. 2011. Removing web spam links from search engine results. In Journal in Computer Virology, Volume 7 Issue 1, February 2011, Pages 51-62, Springer-Verlag New York, Inc. Secaucus, NJ, USA.

[4] Alessandro Marchetto, Filippo Ricca and Paolo Tonella. 2009. An empirical validation of a web fault taxonomy and its usage for web testing. In Journal of Web Engineering, Volume 8 Issue 4, December 2009, Pages 316-345, Rinton Press, Incorporated.

[5] Fabrizio Silvestri. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. In Journal of Foundations and Trends in Information Retrieval, Volume 4 Issue 1—2, January 2010 , Pages 1-174, Hanover, MA, USA.

[6] Jihie Kim, Peter Will, S. Ri Ling and Bob Neches. 2003. Knowledge-rich catalog services for engineering design. In Journal of Artificial Intelligence for Engineering Design, Analysis and Manufacturing, Volume 17 Issue 4, September 2003, Pages 349 – 366, Cambridge University Press New York, NY, USA.

[7] Ying Li, Zijian Zhen and Honghua Dai. 2005. KDD CUP-2005 Report: Facing a Great Challenge. In SIGKDD Explorations Volume 7.

[8] Utkasrsh Srivastava, Kamesh Munagala, Jennifer Widom and Rajeev Motwani. 2006. Query Optimization over Web Services. In VLDB'06 September 12-15, 2006, Seol, Korea, ACM.

[9] Pu-Jeng Cheng, Ching-Hsiang Tsai and Chen-Ming Hung. 2006. Query Taxonomy Generation for Web Search. In CIKM'06, November 5-11, 2006 Arlington, Virginia, USA, ACM.

[10] Joseph M. Hellerstein, Jeffrey F. Naughton. 1996 Query Execution Techniques for Caching Expensive Methods. In SIGMOD'96 6/96 Montreal, Canada, ACM.

[11] Evgeniy Gabrilovich, Andrei broder, Marcus Fontoura, Amruta Joshi and Vanja Jasifovski. 2007. Classifying Search Queries Using the Web as a Source of Knowledge. In ACM international Conference on Research and Development in Information Retrieval (SIGIR) Amsterdam, Netherlands.

[12] S. Chaudhuri,U. Dayal and T. Yan. 1995. Join queries with external text sources: Execution and optimization techbiques. In Proc. of the ACM SIGMOD Intl Conference on Management of Data,San Jose, California.

[13] Nikos Kirtsis and Sofia Stamou. 2011. Query Reformulation for Task Oriented web searches. In Proc. of IEEE/WIC/ACM Intl conference on Web Intelligence and Intelligent Agent Technology.

[14] Shuai Ding, Josh Attenberg, Ricardo Baeza and Torsten Suel. 2011. Batch Query Processing for web search engines. In proc. of the fourth ACM intl conference on web search and data mining