Unstructured Content Analysis & Classification System for the IRS

R.Palson Kennedy,
Prof. ,Dept of IT,RajarajeswariEngg College
Research Scholar ,Anna University

ABSTRACT
Creating ontological approaches to personalizing queries of unstructured data requires intensive use of XML-based tables and schema. From the legacy design efforts for CSDL to the myriad of approaches to XML schema development including the development of XIRQL , Hybrid XML retrieval and XML queries , the adoption of advanced techniques for unstructured content management is progressing rapidly. Paralleling these research advances is pervasive adoption of Cloud Computing platforms including Software-as-a-Service (SaaS), driven by the growth of the Amazon Web Services platform in addition to others. The intent of this thesis proposal is to define an XML schema that can aggregate unstructured content that when combined based on the individualized taxonomies and ontological preferences of system users, delivers highly relevant and timely data. The proposed XML Schema Model for Unstructured Content Personalization shown in Figure 1. This model is further supported by the development of and continual fine-tuning of Quantum Information Algorithms to define approximate taxonomies and approaches to creating role-based query is used as the basis of creating personalization pathways in the data. Quantum Information Theory also makes it possible to create enterprise-wide networks of knowledge management systems that can effectively "learn" over time through the use of latent semantic indexing (LSI) to create linguistic models of representation of the data. Quantum Information Theory provides the basis for creating an entire network of systems that can in essence learn over time, continually fueling new insights into the knowledgebase of the complex of systems themselves.

Keywords: *Unstructured data, XML schema, LSI, Cloud Computing, Personalization,*
knowledge management systems.

INTRODUCTION

Information storage and retrieval

Information storage and retrieval, the systematic process of collecting and cataloging data so that they can be located and displayed on request. Computer device capable of performing a series of arithmetic or logical operations. A computer is distinguished from a calculating machine, such as an electronic calculator, by being able to store a computer program (so that it can repeat its operations and data processing techniques have made possible the high-speed, selective retrieval of large amounts of information for government, commercial, and academic purposes. There are several basic types of information-storage-and-retrieval systems.

Document-retrieval systems store entire documents, which are usually retrieved by title or by key words associated with the document. In some systems, the text of documents is stored as data. This permits full text searching, enabling retrieval on the basis of any words in the document. In others, a digitized image of the document is stored, usually on a write-once optical disc. *Database systems* store the information as a series of discrete records that are, in turn, divided into discrete fields (e.g., name, address, and phone number); records can be searched and retrieved on the basis of the content of the fields (e.g., all people who have a particular telephone area code). The data are stored within the computer, either in main storage or auxiliary storage, for ready access. *Reference-retrieval systems* store references to documents rather than the documents themselves. Such systems, in response to a search request, provide the titles of relevant documents and frequently their physical locations. Such systems are efficient when large amounts of different types of printed data must be stored. They have proven extremely effective in libraries, where material is constantly changing.

The volume of information has been rapidly increasing in the past few decades. While computer technology has played a significant role in encouraging the information growth, the latter has also had a great impact on the evolution of computer technology in processing data throughout the years. Historically, many different kinds of databases have been developed to handle information, including the early hierarchical and network models, the relational model, as well as the latest object-oriented and deductive databases. However, no matter how much these databases have improved, they still have their deficiencies. Much information is in textual format. This unstructured style of data, in contrast to the old structured record format data, cannot be managed properly by the traditional database models. Furthermore, since so much information is available, storage and indexing are not the only problems. We need to ensure that relevant information can be obtained upon querying the database.

1.MOTIVATION

Ontological classification of unstructured data is critically important in the managing of initiatives, programs and strategies. When the specialized requirements of the Internal Revenue Service (IRS) and their requirements to stay in compliance to government regulations are included, the complexity of their roles and the need for accuracy, auditability and transparency are critically important. The IRS has long been challenged by a lack of transparency and in creating an ontologically based model that can take into account role-based requirements; the proposed XML Schema Model(1) for Unstructured Content Personalization integrates unstructured and structured content into role-based taxonomies. With regard to structured data integrated, XSLT style sheets and XML integration into the taxonomies and ontological frameworks is defined. For unstructured data,

XML data Integration and a Latent Semantic Indexing (LSI) filter that classifies and organizes the content into ontologically-defined roles is used. These two XML integration workflows from structured data and unstructured data are also used for creating knowledge management structures, systems and processes. In evaluating how the XML Schema Model for Unstructured Content Personalization would accomplish this, the recursive nature of its workflow needs to be seen as a factor driving the accumulation of knowledge as a result of velocity of data transactions and fluidity of communication.

Further supporting this rationale and the coupling points throughout the Proposed XML Schema Model for Unstructured Content Personalization is the need for having support for information filters that can be modified to reflect specific process-based and role-based information needs throughout this network. The use of condensation filters that can aggregate data over time and provide a synopsis of the content in either a learned taxonomy or in the context of a learning hierarchy is critically important. In conjunction with latent semantic indexing there is also the need for creating contextual filters that can be configured in real-time to match the specific preferences of role-based taxonomies throughout the network of users and learning-based systems as well. The use of contextual analysis and categorization together can serve to create more effective tailoring of content, specifically to role-based categorizations as defined by taxonomies. There is also the need for creating calculation-based filters in conjunction with other filters, provide additional insight and intelligence into the data sets. In addition to all of these factors, the development of spatially-based databases through the integration of a knowledge state engine that would be able to coordinate and manage the continual learning of these processes would also be critical.

2.PROBLEMS IDENTIFIED AND FUTURE PROBLEMS OF UNSTRUCTURED DATA

The lack of consistency across frameworks, methodologies and taxonomies used in classification and organization of data is showing there is no single best approach to modeling and interpreting unstructured data. Instead of a single holistic standard, multiple ones are emerging, each with specific insights and value. Parsing of structured and unstructured data and its interpretation through XML modeling (10), Bayesian analysis (35), and Latent Semantic Indexing with the creation of context- and role-based taxonomies (26) are examples of this divergence. Replicating computing and process-centric platforms the use of Web Services and extraction agents (12) seeks to attain a holistic and scalable standard. Unstructured Data Management Systems (UDMS) continue to also be prevalently in use throughout commercial applications that have a high percentage of content inbound from customer service and website traffic (13). All of these factors are unified and made holistic at a theoretical level from the context of risk mitigation and minimization (35). Yet all lack a consistently scalable and functional approach to delivering consistency of linguistic modeling analysis and consistency of parsing. This lack of harmonization of standard is far from balkanizing the areas of parallel research; instead there is recognition that variation in parsing, taxonomy and codification processes are acting as a catalyst of unstructured content analysis growth (26).

3.PROPOSED METHODOLOGY

The natural tendency to apply reductionism to the area of unstructured content analysis needs to be countered with a holistic foundation within which the entire ecosystem of unstructured content within an organization can be defined. Ecosystems are by nature more oriented towards reciprocity and a continual rejuvenation of content and are therefore holistic in nature. Reductionism seeks to define the contributions of each component. Yet for there to be balanced in any unstructured content system there must be a holistically-based model as well.

Defining ontology-based approaches for the analysis, classification, personalization and retrieval of unstructured content that are compatible with role, process and personalized taxonomies have the potential to augment and enhance content analysis and classification systems. Taking these in the context of a holistic approach to defining an unstructured content analysis model(1), they can be more seen as ecosystems that must be structured to provide for balance of inbound, process and output process workflows. They are not discrete unto themselves but rather contributory aspects of the holistic ecosystem comprising unstructured content.

Initial efforts at applying linguistic analysis to unstructured content have contribute to latent semantic analysis (6), integration of linguistic structures to neural networks (14), and the use of knowledge management linguistic analysis techniques (22). Together these three ecosystems of latent semantic analysis, linguistic structures and linguistic analysis techniques combine to create an integrated ecosystem which makes it possible to create linguistic models based on a critical mass of unstructured content. These three process areas form the catalyst of unstructured content analysis ecosystems.

Once a specific ecosystem has been put into place, it specific contribution to an organization cab be defined. From a holistic standpoint, the use of unstructured content ecosystems is also useful for defining business process re-definition in the context of business intelligence workflows (11). Unstructured content analysis emanating from these areas of development lack shared structured data schema (12) which are critical for the development of a holistic ecosystem. Integration, data replication, XML parsing and reliance on XSLT style sheet definitions are providing taxonomy-based personalization for structured content yet is unproven for unstructured content use (18). These aspects of unstructured content analysis are used for creating specific reductionism models. They are as a result also critical for defining the integration between the components of an ecosystem as well.

The intent of this methodology is to validate the feasibility of the proposed model shown in Figure 1, Proposed Holistic Model for Unstructured Content Personalization. The methodology to validate this model has several prerequisites that will defined in the following section, followed by recommended series of testing and analysis phases. The key performance indicators (KPIs) defining the proposed models; level of accuracy, velocity and orthogonality of support for XML-based taxonomies to the role, process and personal levels are also defined. XML is used as the integration technology to validate or refute the accuracy, integration precision and velocity of the proposed Holistic Model.

3.1 Methodology Objectives and Hypotheses

Methodology Objectives

1. To validate the data accuracy, replication and validity of XML as a transport mechanism between unstructured content and structured content data structures in the context of a holistic model of unstructured content personalization. XMLs' adoption as a data transport layer of models has been verified for structured content and retrieval (19).

2. Ascertain that data repositories based on XML-extensible schema can be queried using taxonomy-based workflows (5).
3. Evaluate the compatibility, extensibility and scalability of XML query languages when used in conjunction with latent semantic indexing (LSI) to ascertain the reliability of this specific area of the proposed model (20).
4. To evaluate and validate if content integration scenarios that are process-based with cloud-based databases as the main repositories are scalable, secure, and capable of supporting LSI-based integration for taxonomy support (28) .

Hypotheses

Null Hypothesis 1: There is no significant increase in accuracy, speed and taxonomy-based parsing of data based on a Cloud-based platform.

Null Hypothesis 2: The level of transaction velocity is attributable to the structured data speed of parsing and validation in conjunction with constraint-based taxonomies being defined through role-based and process-based definitions.

3.2 Methodology Prerequisites

The following perquisites apply to the development of this research effort. First, there is the need for defining which cloud platform will be used for completing the analysis. The Amazon Web Services (AWS) platform in addition to its development services including Mechanical Turk is preferred as the hosting and cloud-based platform for the validation of the proposed model.

The unstructured content can reside in open source Linux databases which are available at no charge. The recent offer from Oracle and Microsoft of free downloads of their databases would also be potentially useful. Registering as a member of the AWS development team under educational access is highly advisable for guidance on creating the test cloud platform.

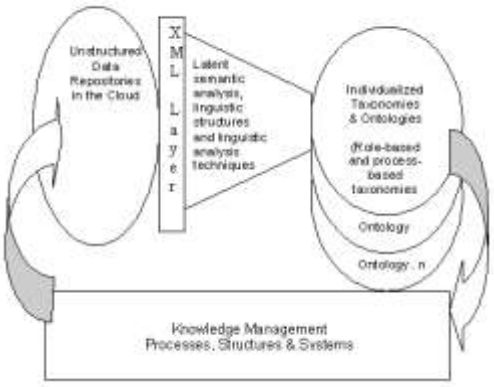
Additional prerequisites include the development of XML and XSLT style sheets that can be structured to support LIS-based constraint logic on the inbound connection and role, process and personal based taxonomies on the outbound side. Allowances for testing variations in CSDL(33), Hybrid XML (31)and XIRQL(15) also need to be defined for the methodology to also be effective in capturing the range of functionality XML for role-based taxonomy development. In addition to these prerequisites there is also the need to create an analytics layer that can capture specific performance data and represent it in a balanced scorecard. Each of the KPIs in this scorecard need to also include XML and XSLT measurements of performance for each supported taxonomy and role-based exception to the data structures represented.

3.3 Methodology

With AWS test region defined and the Linux-based databases defined with unstructured content replicating the IRS data sets, XML and XSLT links will be assigned using a randomized traffic algorithm to ensure representativeness of access is achieved. Each of these XML and XSLT connections will be in turn load balanced with heavy queries by role, process and person-based taxonomies. From these queries a performance table will be devised by each type of XML variant technology used. The load balancing could potentially make this a cube or three dimensional matrixes, yet additional tables for each XML-based technology could also be used. The table consists of role, process &personal taxonomies is tested with High, Midrange & low bandwidth unstructured query will be shown ,in the table as XML & XSLT Parsing Performance.

This matrix and additional versions for CSDL, Hybrid XML and XIRQL would be used for validating the hypotheses and objectives of the study and also validating the main precept of this study, which is cloud-based unstructured data query and structured data query can be efficient in the context of a single system.

Figure 1: Proposed Holistic Model for Unstructured Content Personalization



4.PRELIMINARY FINDINGS

Initial findings based on the research undertaken to evaluate the hypothesis is discussed in this section.

The Hypotheses of this study are:NH1 & NH2 stated earlier in methodology.

Amazon Web Services (AWS) in Education program was used to produce the following table, XML and XSLT Parsing Performance.

Table 1:XML & XSLT Parsing Performance

Access Type	High Bandwidth test of unstructured query	Midrange Bandwidth test of unstructured query	Low Bandwidth test of unstructured query
Role-based Taxonomy (Director)	0.34 ms	0.54 ms	0.89 ms
Role-based Taxonomy (Manager)	2.47 ms	3.76 ms	4. 35 ms
Role-based Taxonomy (Associate)	6.75 ms	10.43 ms	14.43ms
Process-based Taxonomy Query (interval-based scaling)	0.24ms	1.26ms	2.12ms

High bandwidth was executed at 1GB/second, Mid-range bandwidth at 100MB/second and low at 10MB/second. It does not relate directly to the metrics shown, as the speed results shown in the table are based on the performance of each taxonomy purely over the network. In other words the figures are performance levels, in ms, not specific benchmarks.

Based on assessment of these hypotheses using a broad base of previous empirical studies the following findings are reported. With the first methodology objective to validate the data accuracy, replication and validity of XML as a transport

mechanism between unstructured content and structured content data structures the findings of Maletic & Collard (26) indicate that XML is used as the basis of parsing logic is effective in taxonomy creation. Their analysis is predicated on the development of XML-derived data structures specific to each document analyzed (26). With the goal of creating a replicable algorithm and methodology for evaluating the accuracy of extraction of lexical, structural, syntactical, and documentary information from documents (i.e., source code files) Maletic and Collard (10) present arguments for the development of parsing standards that can be interpolated into a syntax tree and scaled across a wide variety of content sources. Comparable to the work of Maletic and Collard (26) are the initial results from a team of researchers at the University of Wisconsin, Madison who sought to parse, categorize unstructured content based on linguistic models, creating context-based taxonomies in the process (13)

The research competed by Doan. et.al. (2008) also supports the second methodology objective of data repositories based on XML-extensible schema capable of being queried using taxonomy-based workflows (5,20). XML-based query engines (26) and context-based query engines that seek to define taxonomies (Doan, et.al. 2008) through empirical research have also proven this objective of the methodology can be attained.

The third methodology objective is to evaluate the compatibility, extensibility and scalability of XML query languages when used in conjunction with latent semantic indexing (LSI) to ascertain the reliability of this specific area of the proposed model (28). The use of XML as the basis of a parsing and query engine development and integration to Latent Semantic Indexing (LSI) platforms has been achieved across a representative number of unstructured content sources (20). LSI-based technology integration and the development of XML-based taxonomies through the use of Web Services to function as content extraction and interpolation agents (13) through the use of AJAX programming standards shows significant performance potential yet cannot scale to a holistic platform over time. This is because AJAX as a programming standard has functionality in parallel to J2EE and C# yet does not have replication, transparency and scalability across all commands. From the research completed, the progression of Web Services from a process standpoint to the development of unstructured data management systems (UDMS) to address the need for interpreting unstructured content originating from disparate Web content management systems and inbound content over the Web (Doan, et.al., 2008) it is apparent that a single holistic approach to creating an information retrieval system has not yet been attained through theoretical or empirical studies.

Instead there is the development of linguistic modeling based on XML-based parsing engines (10), Bayesian analysis (36), and Web Services designed as extraction engines (12). The third objective of the methodology is to evaluate and validate if content integration scenarios that are process-based with cloud-based databases as the main repositories are scalable, secure, and capable of supporting LSI-based integration (28). Clearly this objective has a basis for validation and continual testing once AWS for Education is approved by Amazon. XML parsing engines that are cloud resident have comparable performance to server-based performance metrics (26) and have scalability factors comparable to Web Services (13). From the standpoint of supporting the creation and continual updating of taxonomies and data structures that are entirely cloud-based, the development and refinement of XML to LSI integration as the front-end process from studies completed (35,36) indicate this level of integration and parsing can be defined and achieved.

4.1 Key Performance Indicators

As part of the prerequisites it was stated that an open source analytics application would be used for capturing the performance of the XML and XSLT performance levels, in addition to defining KPIs for the unstructured database performance based on the use of LSI-based algorithms for data query. Based on these requirements the following KPIs are recommended:

XML Variant Performance Ratio – Defines the level of performance depending on the high, medium or low level of load from the unstructured databases.

XML-to-Role based Taxonomy Accuracy – This metric will measure how effective the XML and XSLT style sheets are in translating LSI-based queries into taxonomies defined by role, process and customization (personalization).

Unstructured Content Latency - This metric would measure the relative level of performance of randomized queries of the cloud-based unstructured content databases. Over time this data would indicate if the hypotheses were accurate or not.

5.CONCLUSION

The divergent paths unstructured content analysis concepts, frameworks, methodologies and technologies are progressing on indicates that holistic attainment of an unstructured content platform will be elusive. XML is the integration standard and the basis of LSI parsing engines, yet will compete with J2EE and C# over the long-term. The development of Web Services in AJAX for performance introduces a competing standard. The validation of role-based taxonomy, process-based taxonomy and personally defined taxonomies across high, medium and low bandwidths within a virtualized server environment (cloud) based on XML and XSLT integration is proven from a compatibility standpoint from previous research (35) (11).

6.FUTURE RECOMMENDATIONS

Linguistic versus probabilistic modelling, XML-based approaches to taxonomy creation and the development of XML to LSI integration, and the use of Web Services and extraction agents (13) all illustrate the variation in approaches used. For a holistic ecosystem to be created the basis needs to be on process scalability not necessarily integration technology (XML) or constraint-and rules-based Web Services (13). The validation of process-based methodologies evaluated from an XML and XLST performance level is needed.

7.REFERENCES

[1]. Agosti, Maristella, Crestani, Fabio, Gradenigo, Girolamo. (1989). Towards Data Modelling in Information Retrieval. Journal of Information Science, 15(6), 307.

[2]. Alain Azagury, Michael E Facto, Yoelle S Maarek, Benny Mandler. (2002). A novel navigation paradigm for XML repositories. Journal of the American Society for Information Science and Technology, 53(6), 515-525.

[3]. Michael Benedikt, Christoph Koch. (2008). XPath leashed. ACM Computing Surveys, 41(1), 23.

[4]. Elisa Bertino, Giovanna Guerrini, Marco

Mesiti. (2008). Measuring the structural similarity among XML documents and DTDs. Journal of Intelligent Information Systems, 30(1), 55-92.

[5]. Angela Bonifati, Stefano Ceri, Stefano Paraboschi. (2002). Pushing reactive services to XML repositories using active rules. Computer Networks, 39(5), 645-660.

[6]. Falguni Bhuta. (2006, June). Put Unstructured Data In Its Place. Information Week, (1094), 21.

[7]. Yannis Charalabidis, Fenareti Lampathaki, Dimitris Askounis. (2008). Unified Data Modelling and Document Standardization Using Core Components Technical Specification for Electronic Government Applications. Journal of Theoretical and Applied Electronic Commerce Research, 3(3), 38-51.

[8]. Shu-Yao Chien, Vassilis J. Tsotras, Carlo Zaniolo, Donghui Zhang. (2006). Supporting complex queries on multiversion XML documents. ACM Transactions on Internet Technology, 6(1), 53.

[9]. Tae-Sun Chung, Hyoung-Joo Kim. (2002). A two phase optimization technique for XML queries with multiple regular path expressions. The Journal of Systems and Software, 64(3), 183-193.

[10]. Collard, M. L. and Maletic, J. I., (2004), "Document-Oriented Source Code Transformation using XML", in Proceedings of 1st International Workshop on Software Evolution Transformation (SET'04), Delft, The Netherlands, Nov. 9, pp. 11-14.

[11]. Samuel Robert Collins, Shamkant Navathe, Leo Mark. (2002). XML schema mappings for heterogeneous database access. Information and Software Technology, 44(4), 251-257.

[12]. Conlon, S., J. Hale, S. Lukose, and J. Strong. 2008. INFORMATION EXTRACTION AGENTS FOR SERVICE-ORIENTED ARCHITECTURE USING WEB SERVICE SYSTEMS: A FRAMEWORK. The Journal of Computer Information Systems 48, no. 3, (April 1): 74-83.

[13]. Doan, A.; Naughton, J. F.; Ramakrishnan, R.; Baid, A.; Chai, X.; 0002, F. C.; Chen, T.; Chu, E.; DeRose, P.; Gao, B. J.;W. & Vuong, B.-Q. (2008), 'Information extraction challenges in managing unstructured data.', SIGMOD Record 37 (4) , 14-20 .

[14]. Adam Fadlalla, Chien-Hua Lin. (2001). An analysis of the applications of neural networks in finance. Interfaces, 31(4), 112-122.

[15]. Norbert Fuhr, Kai Grojohann. (2004). XIRQL :An XML query language based on information retrieval concepts. ACM Transactions on Information Systems, 22(2),

[16]. Norbert Fuhr, Norbert Gövert. (2006). Retrieval quality vs. effectiveness of specificity-oriented search in XML collections. Information Retrieval, 9(1), 55-70.

[17]. J E Funderburk, S Malaika, B Reinwald. (2002). XML programming with SQL/XML and XQuery. IBM Systems Journal, 41(4), 642-665.

[18]. Sven Groppe, Jinghua Groppe, Stefan Böttcher, Thomas Wycisk, Le Gruenwald. (2009). Optimizing the execution of XSLT stylesheets for querying transformed XML data. Knowledge and Information Systems, 18(3), 331-391.

[19]. Norbert Gövert, Norbert Fuhr, Mounia Lalmas, Gabriella Kazai. (2006). Evaluating the effectiveness of content-oriented XML retrieval methods. Information Retrieval, 9(6), 699-722.

[20]. Jaap Kamps, Maarten Marx, Maarten de Rijke, Börkur Sigurbjörnsson. (2006). Articulating information needs in XML query languages. ACM Transactions on Information Systems, 24(4), 407-436.

[21]. Shmuel T Klein. (2008). Processing queries with metrical constraints in XML-based IR systems. Journal of the American Society for Information Science and Technology, 59(1), 86.

[22]. Judith Lamont. (2007, February). Semantic Web holds promise for KM. KM World, 16(2), 22,26.

[23]. Leah S Larkey, Margaret E Connell. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. Information Processing & Management, 41(3), 457-473.

[24]. William Laurent. (2008). Mining the Business Intelligence from Unstructured Information. DM Review, 18(4), 28.

[25]. Libby, Robert, Tan, Hun-Tong. (1994). Modeling the determinants of audit expertise. Accounting, Organizations and Society, 19(8), 701.

[26]. Maletic, J.I., Collard, M.L., 2005, Adding Structure to Unstructured Text Wright Center for Innovation/LexisNexis Conference on Using Metadata to Manage Unstructured Text Dayton, Ohio, October 7, 2005, 5 pages

[27]. S Liu, C A McMahon, S J Culley (2008). A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management. .Computers in Industry, 59(1), 3.

[28]. Robert M Losee. (2006). Browsing mixed structured and unstructured data. Information Processing & Management, 42(2), 440-452.

[29]. M Mercedes Martínez-González, Pablo de la Fuente. (2007). Introducing structure management in automatic reference resolution: An XML-based approach. Information Processing & Management, 43(6), 1808.

[30]. Young-Ho Park, Kyu-Young Whang, Byung Suk Lee, Wook-Shin Han. (2006). Efficient evaluation of linear path expressions on large-scale heterogeneous XML documents using information retrieval techniques. The Journal of Systems and Software, 79(2), 180-190.

[31]. Jovan Pehcevski, James A. Thom, Anne-Marie Vercoustre. (2005). Hybrid XML Retrieval: Combining Information Retrieval and a Native XML

Database. Information Retrieval, 8(4), 571-600.

[32]. Juan Manuel Pérez, Rafael Berlanga, María José Aramburu. (2009). A relevance model for a data warehouse contextualized with documents. Information Processing & Management, 45(3), 356.

[33]. Roussopoulos, Nicholas. (1979). CSDL: A Conceptual Schema Definition Language for the Design of Data Base Applications. IEEE Transactions on Software Engineering, 5(5), 481-496.

[34]. Kun-Woo Yang, Soon-Young Huh. (2007). Intelligent Search for Experts Using Fuzzy Abstraction Hierarchy in Knowledge Management Systems. Journal of Database Management, 18(3), 47-68.

[35]. ChengXiang Zhai, and John Lafferty. 2006. A risk minimization framework for information retrieval. Information Processing & Management 42, no. 1, (January 1): 31-55.

[36]. Jose Zubcoff, Jesús Pardillo, Juan Trujillo. (2009). A UML profile for the conceptual modelling of data-mining with time-series in data warehouses. Information and Software Technology, 51(6), 977.

[37]. CC Kane morekotte (2006) The importance of sibling for efficient clustering of XML documents IBM Systems Journal, 45(2), 321-334