

Managing Unstructured Data Using Agent Technology

Amit Kumar Goel¹, Ritu Sindhu², Monica Mehrotra³, G.N. Purohit⁴
(goelkarnit@rediffmail.com, ritu.sindhu2628@gmail.com, monica_mehrotra2000@yahoo.com,
gn_purohitjaipur@yahoo.co.in)

ABSTRACT

In today's scenario, all the companies are storing their data electronically, due to reducing cost of Media, Servers etc. So now a big problem arises to organize unstructured data properly. Mainly source of unstructured data are documents, spreadsheets and Emails. Most of the information companies generate—more than 70 percent, according to experts—won't fit into the cells of a traditional relational database. So the problem is how to handle unstructured data. This paper raises the issues related to Information Lifecycle Management (ILM) and gives the powerful approach to manage such type of data. This paper includes issues surrounding Information Lifecycle Management (ILM) and resolves the problem of Information Retrieval through agents.

Keywords: Information Lifecycle Management (ILM), Extraction Agent, Retrieval Agent, Categorization Agent, Customer Relationship Management, Ontology

1.0 INTRODUCTION

Corporate information stored on file servers and network attached storage (NAS) devices is in danger of compromise because IT governance policies and access rules in many companies are incapable of dealing with a massive growth of unstructured data.

A Ponemon survey of 870 IT professionals found that only 23% believe unstructured data stored by their companies is properly secured and protected.

A wide majority - 84% -- of respondents said that too many workers at their companies can access critical corporate unstructured data. About 76% said their companies have no process in place to control which employees can access specific unstructured data. Such unchecked access could expose internal security gaps and increase the potential for misuse of data, the study notes [5][14].

Larry Ponemon, chairman of the Traverse City, Mich.-based research firm, noted that IT managers say that it's difficult to find automated access control processes that can determine the importance of information the moment it's created.

About 61% of respondents said they cannot keep track of which user's access specific unstructured data, and 91% said their organizations lack the ability to determine data ownership because of faulty governance policies and a lack of available storage tools that can remedy the problem.

While IT managers continue to spend significant sums of money on storage technology to hold rapidly increasing amounts of structured data, many admit that the complexity of unstructured data still makes it difficult to secure it.

"What we find is not that they won't spend money on it, but they really don't know how to resolve the issue because of the complexity; it's a knowledge issue," The respondents

said that without adequate controls for unstructured data, the top potential problems are insider negligence and deliberate misuse or theft of information from within an organization.

Unstructured data is defined as electronic information residing on file servers and NAS devices that is not stored in a database or in a document/content management system. He said it can include: e-mail, instant messages, Microsoft Word documents; PowerPoint files; electronic spreadsheets; and source code.

2.0 ILM

ILM is used to manage data from the beginning of its creation to end. ILM is comprised of the policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost-effective. Since many organizations have no formal Records Management Policies that have been transferred to electronic content this may mean that final disposition is never reached. Other organizations have even worse Electronic Records Management Policies that are based on questionable analysis of existing rules and regulations, and call for the destruction of possibly valuable data by fiat directive. E.g. All e-mails older than 90 days will be deleted from company systems.

ILM uses a number of technologies and business methodologies, including the following:

- Assessment
- Socialization
- Classification

- Automation
- Review

In the assessment phase of ILM, storage administrators can take advantage of storage resource management (SRM) technologies. SRM solutions help IT administrators figure out what data resides on the storage assets in their environment.

Most SRM tools can generate reports for IT that outline data usage patterns. Once the IT department understands what data it has and where this data lives, it can begin the next steps of the ILM process: generating reports from the SRM tools, presenting them to the company's department heads and explaining the breakdown of storage asset utilization and the costs involved. This process is known as the socialization phase of ILM [14] [15].

Once IT meets with the department heads, and the groups collaborate to understand data usage patterns, department heads must determine how this data is used and how critical it is to the business at any given point in time. The ability to prioritize data based on business requirements (that is, mission-critical, business-sensitive and departmental) will allow IT to determine where data should live through its lifecycle and assist in creating policies to migrate data to the proper storage "class" over time.

IT must work with department heads to set up a classification schema for the company. Data can be classified in the following ways:

- Data type
- Data "Organization"
- Data age
- Data "Value"

IT will use all data collected at this point to establish policies to automate the data's migration through the environment, with a minimum amount of hands-on data management.

SRM solutions should be employed throughout the ILM process, not merely for an initial assessment. SRM technology can monitor the storage environment constantly, revealing where excess capacity, duplicate files, "unnecessary files" or aged files exist. This information is very important in the ILM process, for it is essential in understanding which data should be migrated, archived or purged [5].

2.1 AUTOMATION PROCESS

There are various elements of most ILM systems.

- Tiered Infrastructure
- Data Management Layer
- Application Specific Interfaces

Tiered infrastructure is essentially the same as HSM (Hierarchical Storage Management) without using Data Management Layer. The idea behind this is to use different storage solution, moving from expensive Tier I to Tier IV.

Tier I: fast access, high-performance primary disk.

Tier II: low-cost disk such as SATA [serial advanced technology architecture] disk.

Tier III: tape technology where data must be retained but is unlikely to be referenced again.

Tier IV: offline tape in a secure facility, possibly offsite, which can be manually reintroduced into a tape library in the very unlikely event that it needs to be recalled."

The emergence of SATA disks has done much to boost ILM efforts, he says. SATA arrays can store data at a fraction of the cost of high-performance disk. When shifting point-in-time copies of data, for example, SATA is often the obvious choice.

That is not to say, however, that tape technology is becoming redundant. "Despite advances in disk, there are still huge advantages to tape technology. Tape technology that can hold around 1.5 terabytes on an [pounds sterling] 80 tape. The costs involved in storing large volumes of data that will probably not be accessed again on tape are now staggeringly low, and disk - even low-cost disk -- still cannot match them."

Example: One customer is used ILM to balance availability and cost by automating payroll data management and migration. Payroll processing is a mission-critical application, so it made sense to store the data on high-performance disk during the processing cycle and replicate it every two hours.

Once the pay cycle is complete, the automated management system now moves payroll data to mid-range SATA disk arrays. At this stage, users can access payroll data from the company's web site for a period of three months.

After three months, the data is written to a tape library, which is on the same campus as the data archive. For disaster recovery protection, the data is replicated to a remote location, where it is stored on a back-up tape library.

ILM is an ongoing process - data storage administrators will need to continually maintain a balance between data performance needs and storage options. "The struggle is to get the client to realise that getting benefit out of ILM is only 20% about technology and 80% about business processes. It's that kind of housework that drives the biggest savings," he says.

The Data Management Layers are the tools responsible for performing the "aging" process. Rudimentary retention policies can be applied to the data as it passes through the tier storage layers.

Application Specific Interface: Most systems do not provide direct interfaces into Specific Applications Interfaces, without having some sort of helper application. They may have an Application Programming Interface (API) for addressing their specific methods of applying retention policies.

2.2 FUNCTIONALITY OF ILM

There are five phases identified as part of ILM.

- Creation and Receipt
- Distribution

- Use
- Maintenance
- Disposition
- Exception

Creation and Receipt: deals with records from their point of origination. This could include their creation by a member of an organization at varying levels or receipt of information from an external source. It includes correspondence, forms, reports, drawings, computer input/output, or other sources.

Distribution is the process of managing the information once it has been created or received. This includes both internal and external distribution, as information that leaves an organization becomes a record of a transaction with others.

Use takes place after information is distributed internally, and can generate business decisions, document further actions, or serve other purposes.

Maintenance is the management of information. This can include processes such as filing, retrieval and transfers. While the connotation of 'filing' presumes the placing of information in a prescribed container and leaving it there, there is much more involved. Filing is actually the process of arranging information in a predetermined sequence and creating a system to manage it for its useful existence within an organization. Failure to establish a sound method for filing information makes its retrieval and use nearly impossible. Transferring information refers to the process of responding to requests, retrieval from files and providing access to users authorized by the organization to have access to the information. While removed from the files, the information is tracked by the use of various processes to ensure it is returned and/or available to others who may need access to it.

Disposition is the practice of handling information that is less frequently accessed or has met its assigned retention periods. Less frequently accessed records may be considered for relocation to an 'inactive records facility' until they have met their assigned retention period. Retention periods are based on the creation of an organization-specific retention schedule, based on research of the regulatory, statutory and legal requirements for management of information for the industry in which the organization operates. Additional items to consider when establishing a retention period are any business needs that may exceed those requirements and consideration of the potential historic, intrinsic or enduring value of the information. If the information has met all of these needs and is no longer considered to be valuable, it should be disposed of by means appropriate for the content. This may include ensuring that others cannot obtain access to outdated or obsolete information as well as measures for protection privacy and confidentiality.

Long-term records are those that are identified to have a continuing value to an organization. Based on the period assigned in the retention schedule, these may be held for periods of 25 years or longer, or may even be assigned a

retention period of "indefinite" or "permanent". The term "permanent" is used much less frequently outside of the Federal Government, as it is impossible to establish a requirement for such a retention period. There is a need to ensure records of a continuing value are managed using methods that ensure they remain persistently accessible for length of the time they are retained. While this is relatively easy to accomplishing with paper or microfilm based records by providing appropriate environmental conditions and adequate protection from potential hazards, it is less simple for electronic format records. There are unique concerns related to ensuring the format they are generated/captured in remains viable and the media they are stored on remains accessible. Media is subject to both degradation and obsolescence over its lifespan, and therefore, policies and procedures must be established for the periodic conversion and migration of information stored electronically to ensure it remains accessible for its required retention periods.

Exceptions occur with non-recurring issues outside the normal day to day operations. One example of this is a legal hold, litigation hold or legal freeze is requested by an attorney. What follows is that the records manager will place a legal hold inside the records management application which will stop the files from being in queued for disposition.

3.0 AN AGENT APPROACH TO MANAGE UNSTRUCTURED DATA

Three Agents can be formed to manage unstructured data.

- 3.1.1. Extraction Agent
- 3.1.2. Categorization Agent
- 3.1.3. Retrieval Agent

Extraction Agent: This agent is used for examining the semantics of document. This agent extract document before categorizing them.

Categorization Agent: This agent is responsible to categorize document & consider the way in which document is subdivided.

Retrieval Agent: This agent is responsible for retrieving information from the collection of documents efficiently and effectively. Before applying information retrieval technique the document should be categorized

To determine whether or a document is pertinent to a particular retrieval process for retrieval agent. In Artificial Intelligence ontologies are developed by humans as models. Ontology serves as a representation vocabulary that provides a set of terms with which to describe the facts in some domain. Concepts represented by an ontology can usually be clearly depicted through natural language because the ontology and natural language function similiarly.. Depending on the construction of the ontology, the meaning of each

world could remain the same as in natural language. In a computer system; context may be represented and constrained by ontology. In other words ontology provides a context for the vocabulary it contains [1][2].

Categorization agent is responsible to categorize data because manually categorize information is highly inefficient and often impractical.

Once awareness of the issue is raised, the next step is to identify the unstructured data in the organization. In content-management systems, such as those from Interwoven, Web pages are typically considered unstructured data even though essentially all Web pages are defined by the HTML markup language, which has a rich structure. This is because Web pages also contain links and references to external, often unstructured content such as images, XML files, animations and databases (see Figure 1).

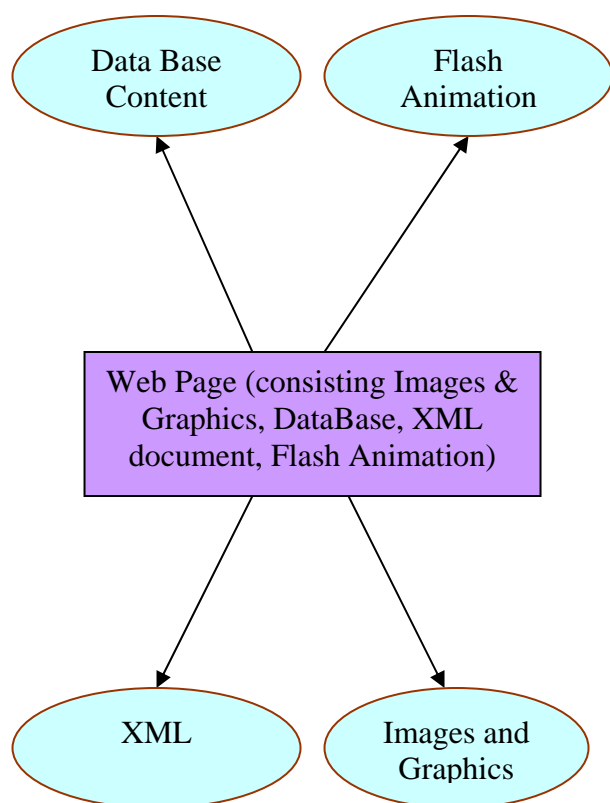


Figure1: Web Page Extraction

Unstructured data is also prevalent in customer relationship management (CRM) systems, specifically when customer-service representatives and call-center staff create notes. However, once again the verbatim text in call-center and customer-service notes is embedded within a form that is

both highly structured and easily represented in a database format. In sum, unstructured data nearly always occurs within documents. Even though many documents follow a defined format, they may also contain unstructured parts. This is another reason why it's more accurate to talk about the problem of semi-structured documents. A basic requirement for semi-structured documents is that they be searchable. Prior to the emergence of the Web, full-text and other text-search techniques were widely implemented within library, document- management and database management systems. However, with the growth of the Internet, the Web browser quickly became the standard tool for information searching. Indeed, office workers now spend an average of 9.5 hours each week searching, gathering and analyzing information, according to market-research firm Outsell Inc.; and nearly 60 percent of that time, or 5.5 hours a week, is spent on the Internet, at an average cost of \$13,182 per worker per year.

Is all this searching efficient? Not really. Current Web search engines operate similarly to traditional information-retrieval systems: They create indexes of keywords within documents and then return a ranked list of documents in response to a user query. Several studies have shown that the average length of search terms used on the public Web is only 1.5 to 2.5 words and that the average search contains efficient Boolean operators (such as and, or and not) fewer than 10 percent of the time. With such short queries and so little use of advanced search techniques, the results are predictably poor. In fact, a performance assessment of the top five Web search engines, conducted by the U.S. National Institute of Standards and Technology, showed that when 2.5 search words are used, only 23 to 30 percent of the first 20 documents returned are actually relevant to the query.

In recognition of the weakness of basic, keyword search, the search-engine vendors have continued to improve their technology. For example, Verity has added techniques such as stemming and spelling correction to its K2 arsenal, while newcomer phrase employs natural language processing.

Information Retrieval through Agents: Due to the popularity of www, has created bulk of unstructured data in the form of documents, spreadsheets, Emails and PDF. So a great issue is to extract information from online documents. There is a information retrieval agent (IR agent) and information extraction agent (IE agent) for same [4] [6].

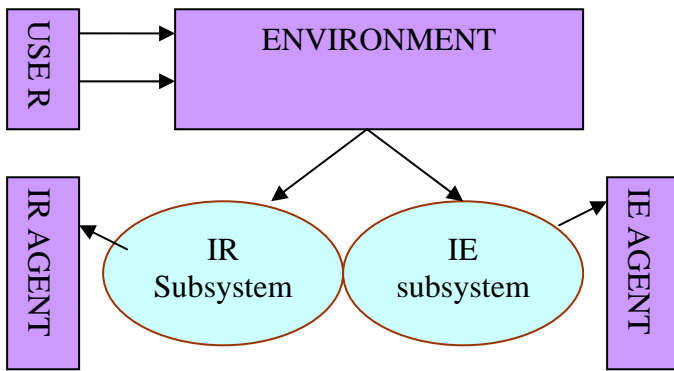


Fig 2: An Agent Overview for Information Extraction

Agents are intelligent because they can adapt their behavior according to the user's instructions and the feedback they get from their environments. In other words, they are learning agents the user, an intelligent agent that use neural networks to store and modify their knowledge [8][9][11].

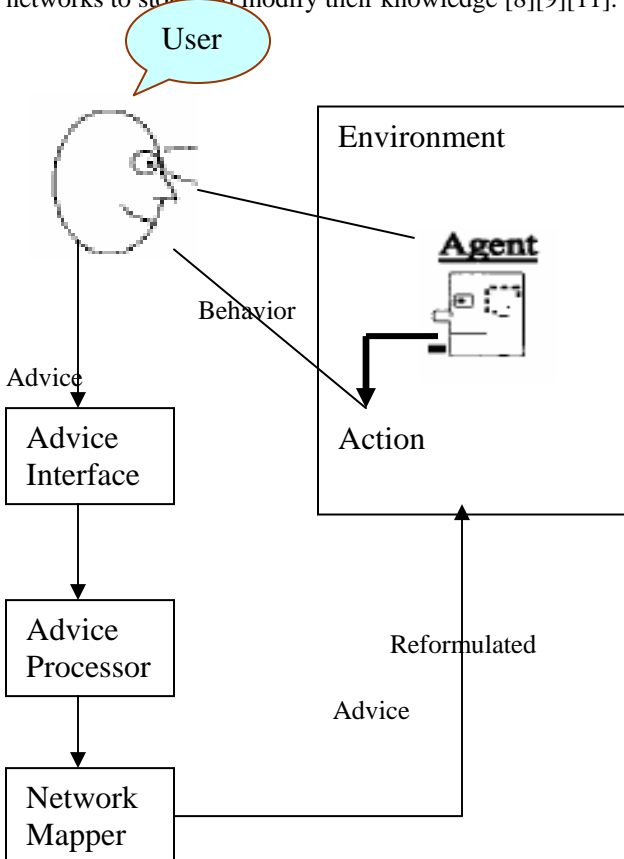


Figure 3: The interaction between a user, an intelligent agent, and Environment agent's

Figure 3 illustrates the interaction between the user, an intelligent agent, and the agent's environment. The user observes the agent's behavior, and provides helpful instructions to the agent. We refer to users instructions as advice, since this name emphasizes that the agent does not blindly follow the user-provided instructions, but instead refines the advice based on its experiences. The user inputs his/her advice into a user-friendly advice interface. The given advice is then processed and mapped into the agent's knowledge base (i.e., into two neural networks), where it gets refined based on the agent's experiences. Hence, the agent is able to represent the user model in its neural networks, which is used for effective learning

4.0 CONCLUSION

The Information Lifecycle Management is the complete automation of the entire unstructured data management process.

The ideal system will monitor the network, automatically enforcing policies on file naming and storage availability based on how valuable the content is. Intelligent analysis tools will suggest which files should be imported into structured data system, and which should be downgraded to low cost storage or deleted.

This article presented concept of linking unstructured information. Solution is presented by using agent oriented approach with emphasis on cooperation with business user while searching for information and exploiting navigational support [12][13].

We envision future research to focus in the area of integrating user's context when retrieving information from unstructured documents. The semantic web is one possible approach, in which pages can be given well defined meaning. Software agents can also assist web users by using this information to search, filter and prepare information in new ways. This approach allows better integration between machine and people and assists the evolution of human knowledge. In addition, future technologies must have the capability to automatically extract the meaning of unstructured documents with reference to the context of the users with minimal human intervention.

5.0 REFERENCES

- [1] Albers M, Jonker CM, Karami M, Treur J (2004) Agents models and different user ontology's for an electronic market place, Knowl Inf Syst 6 (1): 1:41.
- [2] Alexander Smirnov & Nikolay Shilov (2007), Ontology-driven intelligent service for

- configuration support in networked organization. Springer-Verlag London Limited.
- [3] Belew, R. K.: 2000, Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. New York, NY: Cambridge University Press.
 - [4] Croft, W., Turtle, H. and Lewis, D.: 1991, 'The use of phrases and structured queries in information retrieval' In: proceedings of the Fourteenth International ACM SIGIR Conference on R & D in Information Retrieval. Chicago, IL, pp. 32-45.
 - [5] Ching Kang Cheng and Xiao Shan Pan, Using Perception in Managing Unstructured Documents. ACM Student Magazine.
 - [6] Dejan & Viljan, Intelligent agent aided use of unstructured information in decision support.
 - [7] David A Maluf & Peter B Taran, Managing Unstructured Data with Structured Legacy Systems. NASA Aims Research Center Intelligent System Divisions.
 - [8] Eliassi-Rad, T.: 2001, 'Building Intelligent Agents that Learn to Retrieve and Extract Information' Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison, WI. (Also appears as UW Technical Report CS-TR-01-1431)..
 - [9] Maes, P. Agents that reduce work and Information overload. Communications of the ACM, 37(7), 1994, pp. 31-40.
 - [10] Sebastiani, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR). Volume 34 Issue 1, March 2002
 - [11] Soderland, S.: 1999, 'Learning information extraction rules for semi-structured and free text' Machine Learning: Special Issue on Natural Language Learning 34 (1/3), 233-272.
 - [12] Seymore, K., McCallum, A. and Rosenfeld, R.: 1999, 'Learning Hidden Markov Model Structure for Information Extraction' In: Proceedings of the Sixteenth National Conference on Artificial Intelligence Workshop on Machine Learning for Information Extraction. Orlando, FL, pp. 37-42.
 - [13] Shavlik, J. and Eliassi-Rad, T.: 1998a, 'Building intelligent agents for web-based tasks: A theory-refinement approach' In: Proceedings of the Conference on Automated Learning and Discovery Workshop on Learning from Text and the Web. Pittsburgh, PA.
 - [14] Tony Pfitzner & Tyson Lloyd Thwaites, Unstructured data: A management overview. www.allianza.com.au
 - [15] Vinita Gupta, Managing unstructured data, www.expresscomputeronline.com dated 21/01/2008