International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:5, No:5, 2011

Discovery and Capture of Organizational Knowledge from Unstructured Information

J. Gu, W.B. Lee, C.F. Cheung, E. Tsui, W.M. Wang

***Abstract*—** Knowledge of an organization does not merely reside in structured form of information and data; it is also embedded in unstructured form. The discovery of such knowledge is particularly difficult as the characteristic is dynamic, scattered, massive and multiplying at high speed. Conventional methods of managing unstructured information are considered too resource demanding and time consuming to cope with the rapid information growth.

In this paper, a Multi-faceted and Automatic Knowledge Elicitation System (MAKES) is introduced for the purpose of discovery and capture of organizational knowledge. A trial implementation has been conducted in a public organization to achieve the objective of decision capture and navigation from a number of meeting minutes which are autonomously organized, classified and presented in a multi-faceted taxonomy map in both document and content level. Key concepts such as critical decision made, key knowledge workers, knowledge flow and the relationship among them are elicited and displayed in predefined knowledge model and maps. Hence, the structured knowledge can be retained, shared and reused.

Conducting Knowledge Management with MAKES reduces work in searching and retrieving the target decision, saves a great deal of time and manpower, and also enables an organization to keep pace with the knowledge life cycle. This is particularly important when the amount of unstructured information and data grows extremely quickly. This system approach of knowledge management can accelerate value extraction and creation cycles of organizations.

***Keywords*—**Knowledge-Based System, Knowledge Elicitation, Knowledge Management, Taxonomy, Unstructured Information Management

I. Introduction

Due to advancement of ICT, the amount of data and information in an organization is massive and continue growing at high speed. How to elicit the needed knowledge from data and information for decision making in organization is a prime concern in knowledge management (KM). Knowledge assets do not merely reside in structured form of information and data; it also embedded in scattered unstructured form such as meeting minutes, reports, forums, emails or Short Message System (SMS) messages, etc. Merrill Lynch estimates that more than 85 percent of all business information exists as such unstructured form (Shilakes and Tylman, 1998). None of this unstructured information can be found in formal organizational chart or process diagram; rather, it is usually changing with people and environment, hidden in daily knowledge supply and demand in high speed transactions. This is especially true in knowledge intensive industry since knowledge workers highly rely on such unstructured knowledge exchanges in both internal and external environment. The unstructured information contains organizational knowledge which is regarded as critical intellectual capital for value creation. Unfortunately, many professionals, decision makers, experts and managers have to spend much time in non-value adding knowledge retrieval. Additionally, if staff resigned or retired, it is highly like the organizational knowledge will be buried deeply as a myth. This indicates a need for proper and systematic handling of such form of knowledge assets. The knowledge elicitation and auditing process regarding to such unstructured nature is more difficult as the source is dynamic, continuously changing, and multiplying at high speed.

Traditional information management deals with formal, order and structured information in various databases and repositories, whereas knowledge management deals with informal and unstructured information. The knowledge elicitation process from unstructured information is mostly realized by human mind. Conventional methods and tools of managing unstructured information are usually concentrated on “file and document management” level rather than content level for knowledge elicitation. A new approach is desired which is dynamic and flexible enough to support the knowledge discovery, capture, navigation, storage, update and monitoring.

The Case Study conducted in a public organization aims at introducing a knowledge-based system: Multi-faceted Automatic Knowledge Elicitation System (MAKES) in discovery and capturing the organizational knowledge from unstructured information such as meeting minutes and related documents; linking issues, decisions, status and progress with knowledge workers; offering multi-facet taxonomy of structured knowledge representation and flexible knowledge retrieval, search and navigation without causing any interruption of the daily operation of staff.

II. Unstructured Information

Unstructured information represents the largest and fastest

J. Gu is with the The Hong Kong Polytechnic University, Hung Hom, KLN, Hong Kong (phone: 852-98396159; e-mail: jessica.gu@polyu.edu.hk).
W.B. Lee is with The Hong Kong Polytechnic University, Hung Hom, KLN, Hong Kong. (e-mail: wb.lee@inet.polyu.edu.hk).
C.F. Cheung is with The Hong Kong Polytechnic University, Hung Hom, KLN, Hong Kong. (e-mail: benny.cheung@inet.polyu.edu.hk).
E. Tsui is with The Hong Kong Polytechnic University, Hung Hom, KLN, Hong Kong. (e-mail: eric.tsui@inet.polyu.edu.hk).
W.M. Wang is with The Hong Kong Polytechnic University, Hung Hom, KLN, Hong Kong. (e-mail: wm.wang@inet.polyu.edu.hk).

growing source of knowledge available to businesses and governments world-wide. The amount of unstructured and semi-structured information in enterprises is growing rapidly, doubling every year, by some estimates (Waters, 2005; Moore, 2002). Management of information and knowledge is of two kinds: the management of structured information, and the management of unstructured information. Unstructured information traditionally is stored as documents in local hard disks or in file servers, or in email systems. The documents include research reports, memos, letters, white papers, presentations, etc. Unstructured information is generally represented in various forms. The lack of structure in unstructured information makes it difficult for it to be collected, accessed, categorized, and searched because such information has no effective association with meta-data. Unstructured information is unmanaged information. In most organizations, there is a large amount of unstructured content which often represents the key intellectual assets of organizations. Because of the inherent difficulties of understanding unstructured content, many organizations are beginning to tackle this problem. In an effort to control and manage such information with visualization patterns, some technologies have been developed in KM practices in organizations. Hatch (2007) showed that organizations are now starting to prioritize the use of unstructured data. Morris (2008) thought that the principal challenge with unstructured information is that it needs to be analyzed in order to identify, locate and relate the entities and relationships of interest, and to discover the vital knowledge contained therein. Decomposing the whole process of unstructured information into various phases is a right approach to the management of unstructured information. These phases consist of text mining, categorization, information retrieval, portals, taxonomy generation, and so on. Using a search engine is an effective approach to discovering and indexing documents which contain specific terms. The content management system can manage effectively many kinds of content, provide access and version control, both of which are important aspects of knowledge management. Knowledge portal offers an ideal platform for knowledge workers to explore into the unstructured information sea and gain useful knowledge.

III. The Multi-faceted and Automatic Knowledge Elicitation System (MAKES)

A. The Architecture of MAKES

The architecture of MAKES is shown in Fig. 1. The input tier is the data input of the system. It includes the unstructured information of the company (e.g. meeting minutes). MAKES is composed of four components which are: concept elicitation and maintenance, concept association, multi-faceted navigation, and thesaurus model, respectively. The output tier includes the knowledge inventory, social network analysis (SNA) which illustrates the interaction among the knowledge suppliers and knowledge customers of the organization, and the relationship analysis among the concepts. It provides a convenient view of knowledge among different people and different concepts which supports the concept of a multi-faceted taxonomy and dynamic navigation.

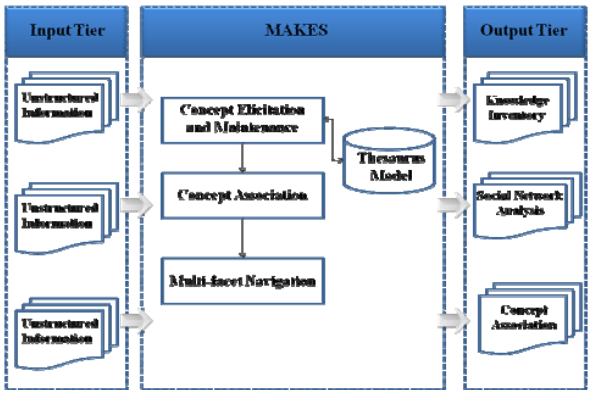

Fig. 1 The Architecture of MAKES

B. Concept elicitation and maintenance

As shown in Fig. 2, the concept elicitation and maintenance module includes the preprocessing of input unstructured data, concept extraction procedure, and the thesaurus model. The thesaurus model contains the controlled vocabularies, synonyms of concepts, and the hierarchies and relationships among the words. The thesaurus model is continuously updating with the self-learning mechanism: through the analysis of new incoming minutes or documents of the company. The unstructured text is firstly pre-processed to filter out irrelevant data and information such as stop words and HTML tags, etc. A concept elicitation algorithm is then developed to extract the key concepts embedded in the texts. Based on the algorithm, the key concepts reside in the text are extracted and consolidated. The concept list is checked against the existing thesaurus model. Concepts which do not exist in the current thesaurus model are regarded as new concepts. The new concepts are evaluated using a rule-based analysis. In the present study, several rules are embedded for making recommendations regarding the new concepts. These rules take into consideration the popularity and density of the new concept. The popularity of the concept is measured by the number of people in the group who share the same concept while the density of the concept is measured by the frequency with which the concept appears in the unstructured information over a certain period of time. If the popularity and density of the new concept achieve a certain threshold, it is suggested as being ready for revision and retention in the thesaurus model. Words that already exist in the thesaurus model are considered to be old concepts. They are normalized based on their relationship with synonyms in the thesaurus model, and then the normalized terms are applied in the indexing of minutes.

C. Multi-facet navigation

Concerning the variety of knowledge needs in organization, the multi-faceted navigation allows knowledge workers to connect with target knowledge directly. People from different

fields may request different knowledge representation and navigation. For example, the research and development (R&D) department may focus on the knowledge about new technologies and products. The human resources management (HRM) may prefer to focus on the interaction (knowledge demand and supply) among the staff of the company. The multi-faceted navigation allows users with such need and function. After the analysis of the concept association, unstructured information can be navigated in different dimensions. Users can select the interested field of data to browse. Not only unstructured information is displayed in a ranked list, but also the concept relationships and the social network of the retrieved information are presented in a graphical view as shown in Fig. 3. When a search query is entered into the system, it retrieves the corresponding knowledge from the knowledge inventory. The knowledge assets are indexed in the concept elicitation module and stored in the inventory. There are two types of enquiries: "Concept Relationships" and "Social Network". The major difference between the two types is in the meanings of the nodes and edges. For concept relationships, the nodes represent the concepts among the retrieved knowledge assets, and the edges represent the strengths of the relationship among the concepts. In the Social Network, the nodes represent the personnel among the retrieved knowledge assets, and the edges represent the volume of the interaction among the concepts. Both the concept relationships and social network are constructed in a self associated concept mapping (SACM) (Wang et al., 2008) format based on the concept associations.

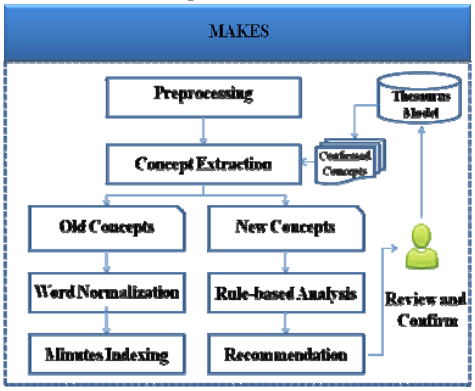

Fig. 2 Concept elicitation and maintenance module

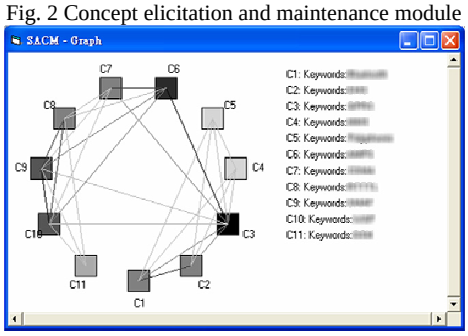

Fig. 3 An Example of SACM

D. Analysis of the results of knowledge elicitation

After social network and concept relationships are discovered, they are evaluated by a rule-based inference engine. The criteria for the classifications of the concepts elicited from the concept relationships and social network is shown in Fig.4. Basically, the elicited concepts are classified into four categories according to two attributes "Importance" and "Permeability". The categories are: critical concept, focus concept, uncommon concept and general concept. Importance refers to the number of items of unstructured information which contain the elicited concept. The greater the number of items of unstructured information the greater the importance of the elicited concept is. Permeability refers to the number of knowledge workers who have used the elicited concept in the unstructured information. The greater the number of knowledge workers, the greater the permeability. When a concept has a high importance and a high permeability, it is classified as a critical concept, such as market information. Focus concept is identified when a concept has a high importance but low permeability. This kind of concepts is importance but they are only shared among small group of people, such as expert knowledge, trade secrets, etc. When a concept has a low to moderate importance but a high permeability, it is classified as a common concept, such as routine practice, etc. When a concept has a low to moderate importance and a low to moderate permeability, it is classified as a general concept such as correspondence. The classifications are automatically determined by the rule-based inference engine and reported in a knowledge inventory report which shows a list of classified concepts together with their classifications. The average number of items of unstructured information and the average number of knowledge workers who have used the unstructured information are used as indicative criteria for the classification of the elicited concepts. On other hand, the results of the social network analysis are used at the level of individuals, departments or organizations to identify teams and individuals who are playing central roles. As shown in Fig. 5, the knowledge workers are classified into four categories according to two attributes "Knowledgeable" and "Impact". The four categories are: critical user, focus user, general user, and brokering user, respectively. Knowledgeable refers to the number of identified concepts provided by the knowledge worker. The greater the number of concepts that are identified as being linked to a knowledge worker means the more knowledgeable the worker is. Impact refers to the number of users who cite the identified concept provided by the knowledge worker. The higher the number of citations, the greater the impact is. When a worker has a high impact and is more knowledgeable, he/she is classified as a critical user. For example, sales people in a trading company must know a lot of different concepts and people; they play a very important role to the company. For focus user, it is identified when a staff is more knowledgeable and has fewer interactions with other people. For example, experts are identified as focus users. When a staff possesses moderate level of knowledgeable but always interacts with others, he/she is classified as a brokering user, such as secretaries and coordinators who serves as a knowledge agent

in the company to streamline the knowledge flow. When a staff possesses moderate levels of impact and knowledgeable, he/she is classified as a general user. The classification are automatically determined and reported in a critical user report. The average number of concepts and the average number of citations are used as indicative criteria for the classification of the knowledge workers.

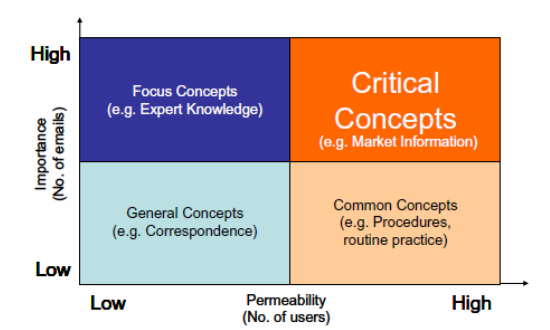

Fig. 4 Classification of elicited concepts

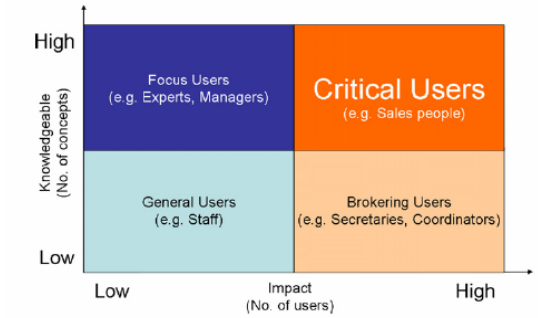

Fig. 5 Classification of knowledge workers

E. Conventional Approach vs. MAKES

One of the major differences between the conventional approach and the MAKES approach is summarized in Table 1. In comparison, MAKES provides a larger coverage of concept and is supported by artificial intelligence (AI) technologies so as to achieve automatic classification, categorization, intelligent searching and navigation, personalization and self-maintenance. A traditional taxonomy is static after the development process, and requires human intervention for making any later changes. The construction of the maps is difficult, time consuming and expensive. In contrast, MAKES, integrated natural language processing techniques, SACM, and multi-faceted taxonomy, has the advantage of higher learning capability, smaller data size and faster speed in knowledge elicitation process. It is more useful for simulating human learning activities and in work with more difficult and unstructured information areas. One of the advantages includes dramatic reduction in time and human effort when classify and retrieve large quantities of knowledge assets.

TABLE I SUMMARY OF CONVENTIONAL AND MAKES APPROACHES COMPARISON

Characteristics	Conventional	MAKES
Dimension of categorization of knowledge	Single dimension	Multi-dimension
Knowledge Representation	Only one concept	Several different concepts at many levels of abstraction
Taxonomy Structure	Static (Unchangeable)	Dynamic (Changeable)
AI Support	No	Yes
Automatic classification	No	Yes
Intelligent searching and navigation	No	Yes
Personalization of taxonomy	No	Yes
Automatic knowledge elicitation	No	Yes
Self-maintenance of taxonomy	No	Yes

III. THE CASE STUDY

The capability of MAKES was evaluated through a trial Case Study in a public organization of Hong Kong. The organization consists of a chairman and members representing various sectors of the industry including employers, professionals, academics, contractors, workers, independent persons and Government officials. The main functions of the organization are to forge consensus on long-term strategic issues, convey the industry's needs and aspirations to Government, as well as provide a communication channel for Government to solicit advice on all industry-related matters. The organization has set up Committees to pursue initiatives that will be conducive to the long-term development of the industry. Meetings as a communication platform are frequently held for opinion collection, discussion and decision making. According to the KM needs, the Case Study mainly focuses on the communications among major committees, especially on the discovery and capture decisions made in series of meetings through unstructured meeting minutes.

As mentioned earlier, hidden knowledge in large amount of unstructured information like meeting minutes is vital for decision making and operation efficiency. There is an urgent need for managing the decisions in series of meetings. Among the committees, frequent and complex meetings and circulation documents are used for communication, decision making and recording. Table 2 summarized the current problems and possible consequences in KM practice of the organization. All the decisions and the progress of issues

being discussed are managed by the Secretariat staff alone which has a hidden risk of knowledge loss if the staff are on leave or resigned. This also inhibits knowledge sharing and transfer which causes ineffective corporate since the staff are depending on each other in performing value adding jobs. Additionally, the handling activities of fast growing minutes are handled manually which indicates a heavy, repeatable and complex workload of the Secretariat staff. It is also difficult for timely knowledge retention which implies huge risks. Usually the Secretariat staffs rely on their memories for knowledge retrieval and search which increases the probability of knowledge duplication, mishandling, and loss. Moreover, staffs often find difficulties in recalling important decisions made long time ago and linking the people involved in the meeting. It takes a long time in connecting every scattered piece of knowledge back in one complete story. If the critical knowledge worker is not in the original position, the decisions will be kept in a black box.

TABLE II
CHALLENGES IN CURRENT DECISION MANAGEMENT SYSTEM

	Detailed Problems Descriptions	Possible Consequences
1	The knowledge discovery and capture from meeting minutes is solely depending on The Secretariat staff.	Inhibit knowledge sharing and transfer. Ineffectiveness in staff performance.
2	The meeting minutes are massive and fast multiplying at high speed. Knowledge hidden is scattered.	Heavy workload. Inefficient knowledge retention and management. Difficult in knowledge discovery by human effort. May cost knowledge loss or other risks.
3	The physical documents of meeting minutes are difficult to retain and retrieve as time goes.	Knowledge hidden in physical documents is hardly retrievable or reusable. Critical decision may lose over time.
4	It is extremely difficult to link up all the progress with the same issue in series of meeting minutes.	May cost knowledge distortion. Violent the accuracy in knowledge exchange.
5	The knowledge discovery process is always a push process. No monitoring or alert is provided to the staff.	May cost work delay or job forgotten.

With the implementation of MAKES, the knowledge discovery and capture from meeting minutes increase the operational efficiency and effectiveness dramatically. As compared in Fig. 6, the knowledge retrieval and navigation of users shorten the lead time, reduce the human effort and increase the accuracy. This is particularly true when the target knowledge in searching was created long time ago and continuing in series of meetings. Now the secretaries are only need to store the meeting minutes to the system and let it capture the critical knowledge automatically.

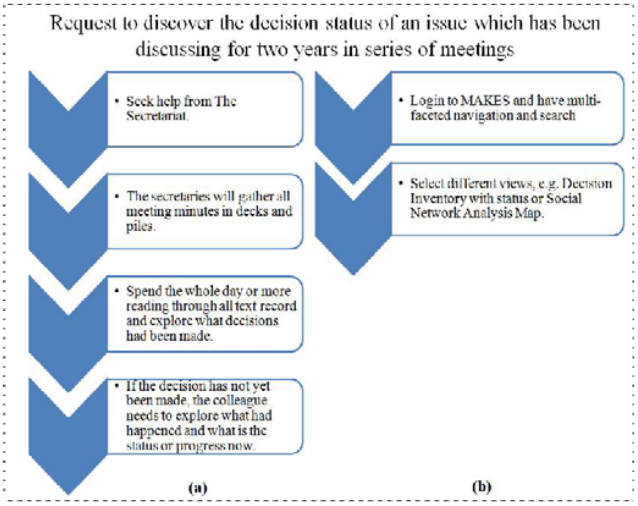

Fig. 6 The traditional approach (a) and MAKES approach (b) in knowledge discovery process

IV. RESULT AND DISCUSSION

The proposed system MAKES through the Case Study demonstrates the automatic decision capturing, concept elicitation, social network presentation, knowledge activities alert and analytical function showing the critical knowledge and concept in meeting minutes and the critical knowledge worker. Examples of the results and deliverables are shown in Fig.6. The system also provides insights to senior management through visualizing the knowledge demand and supply activities among staff and indicates future KM initiatives such as harvesting expert knowledge from critical knowledge workers or introducing collaboration tools to encourage knowledge sharing, etc.

With the trial implementation with MAKES, a number of benefits can be realized which includes:

(i) The time and human effort in unstructured information management is reduced dramatically. The non-value adding and repeated activities and idles in work are minimized.

(ii) The effectiveness of the system in coping with huge amount and fast growing data enhances the competitive advantage of the organization.

(iii) With the learning and continuous updating capability, the system will discover new knowledge in ever evolving environment to support future decision making.

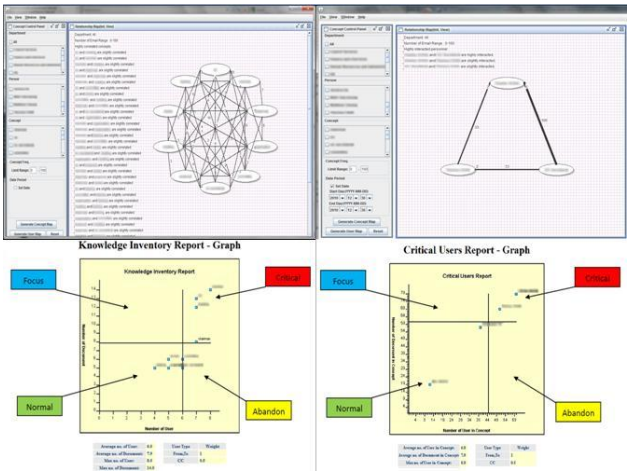

Fig. 6 Examples of knowledge discovery results in the public organization

V. Conclusion

Effective discovery and capture of knowledge assets from unstructured information is essential for an organization. How to locate the right knowledge to the right knowledge worker at the right time is in high priority of daily value creation activities. Many companies realize that valuable knowledge exists in unstructured information such as emails, office documents, PDF-files and many other text based documents in discussion forums, bulletin boards, blogs, etc. This is particularly true for many professional services such as market analysis, investment, research and development, etc. In fact, all of them are trying to get the right knowledge to help them to make the right decisions at the right time. The conventional way of managing unstructured information is inadequate. This paper presents a multi-faceted and automatic knowledge elicitation system (MAKES) which allows for retrieving, automatic classifying, capturing and sharing of appropriate and valuable knowledge from masses of unstructured information which contains multiple concepts at many levels of abstraction. In the present study, the capability and advantages of the MAKES are demonstrated through a successful trial implementation and a verification test conducted in a public organization. According to the results, it is clear that the usefulness and effectiveness of the management of unstructured information can be significantly improved. The time, the cost and the workload on taxonomy development and maintenance are reduced dramatically. It helps an organization to explore new opportunities for value creation. All the mentioned features of MAKES are highly desired in managing unstructured information. With appropriate customization, MAKES can be applicable in a range of professional services such as patent searching, intellectual property management, e-learning, document analysis, customer relationship management and financial analysis.

Acknowledgment

This project was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 5228/09E). The authors would like to express their sincerely thank to Construction Industry Council of Hong Kong for technical support for the work. Many thanks are also due to the support of the Knowledge Management Research Centre of The Hong Kong Polytechnic University.

References

[1] C.F. Cheung, W.B. Lee, W.M. Wang, Y. Wang & W.M. Yeung. “A multi-faceted and automatic knowledge elicitation system (MAKES) for managing unstructured information”, *Expert Systems with Applications*, 2011, vol. 38, pp. 5245 - 5258D.

[2] D. Hatch, “Data Management 2.0: Making Sense of Unstructured Data”, *Aberdeen Group Benchmark Report*, Jul. 2007

[3] C. Moore, “Diving into Data: companies aim to control the rising tide of unstructured data and gain a strategic edge”, *InfoWorld*, http://www.infoworld.com/article/02/10/25/021028feundata_1.html, Oct. 2002 (Accessed 20.3.10)

[4] J. D. Morris, “Unstructured information management – what you don’t know can hurt you!”, *Ezine Articles*, http://ezinearticles.com/?Unstructured-Information- Management---What-You-Dont-Know-Can-Hurt-You!&id=1656140, Nov. 2008 (Accessed 20.3.10)

[5] C.C. Shilakes and J. Tylman, "Enterprise Information Portals", Merrill Lynch, 16 November, 1998.

[6] W. M. Wang; C. F. Cheung; W. B. Lee & S. K. Kwok, “Self-associated concept mapping for representation, elicitation and inference of knowledge”, *Journal of knowledge-based systems*, 2008, vol. 21, pp. 52-61

[7] W.M. Wang & C.F. Cheung, “A narrative-based reasoning with applications in decision support for social service organizations”, *Expert Systems with Applications*, 2011, vol. 38, pp. 3336 - 3345

[8] J. K., Waters, “Managing unstructured information”, *Application Development Trends Articles*, http://www.adtmag.com/article.aspx?id=10542, Jan. 2005 (Accessed 02.10.10)