BY DURSUN DELEN AND SULIMAN AL-HAWAMDEH

A Holistic Framework for Knowledge Discovery and Management

THE INCREASED INTEREST IN KNOWLEDGE DISCOVERY, knowledge management, and knowledge transfer can be attributed to many factors including the advances in information and communication technologies; data explosion and information overload; the expected significant loss in the workforce as the baby boomers retire; and, the need for organizations to better utilize their intellectual capital to stay ahead of the competition. As with the massive amounts of information being added to corporate databases and the Internet everyday, effective and efficient knowledge discovery and its management has become an imminent problem. In spite of addressing a special part of the problem, which has been the case in a vast amount of the recently published research articles on knowledge management, in this paper we propose a holistic framework for knowledge management. This highly integrated framework is composed of a number of interdependent modules designed to perform the activities of the knowledge management cycle

including creating, extracting, storing and using/reusing knowledge. By exploiting various existing technologies such as data warehousing, data mining and text mining, along with Web crawling and federated search engines, this integrated system provides the knowledge worker with the most relevant information to make the best possible decision in a timely manner. This article emphasizes the importance of knowledge management, makes the case for an integrated knowledge management framework, displays and discusses a holistic knowledge management framework, elaborates on the required capabilities and functionalities of such a framework, and concludes with the future directions and final recommendations.

The information age that we are living in is characterized by rapid growth in the amount of data collected and made available in electronic media. This swift pace of growth in digitized information makes it imperative for us to seek alternative methods to effectively and efficiently utilize these invaluable assets. In order to assist the user with the problem of information overload and not exacerbate it, these new methods should go beyond what is already available to the knowledge worker today. The most commonly used method for managing unstructured information has been to employ keywords indexing. While keywords are good at identifying information, they often fall short on telling how relevant such information is to the queried situation and might lead to the assumption that the more often a keyword is mentioned the more relevant the information is to the query.

Most of the current information retrieval systems available on the Web today are based on keyword indexing with very little emphasis on context or textual information. There is a need for a more intelligent information retrieval system that can take into account the semantic information and go beyond the simple keywords searching. But given the limitation of artificial intelligence based systems and the fact that most knowledge exists in the minds of

people in the form of tacit knowledge, it is imperative to build collaborative information systems capable of integrating this valuable knowledge into the information retrieval environment. To integrate tacit knowledge transfer processes into information systems there is a need to understand the complexity associated with tacit knowledge capture and transfer.[5, 7, 10]

Many believe that tacit knowledge residing within an individual is only observable via an individual's actions.[6, 7] But there have also been suggestions in tapping/harvesting tacit knowledge. Some of these methods include interview sessions, narrations or story telling, knowledge exchange protocols or action protocols using the repertory grid and analogies and metaphors.[8] The discussion on whether tacit knowledge can be captured (codified) or transferred from one person to another will remain arguable as far as we continue to view knowledge in two dimensional space as either tacit or explicit. Al-Hawamdeh[1] introduced a third dimension to the knowledge definition by distinguishing between actionable knowledge (tacit knowledge) in the form of behavior, skills, competencies and experiences and articulated knowledge (implicit knowledge) in the form of individual thoughts and language use. Viewing tacit knowledge as actionable knowledge and articulated knowledge help us in designing information systems that support the capture and transfer of these two types of knowledge. While articulated knowledge in the form of "Know How" and "Know Who" can be transferred easily using information systems technologies such as email, collaboration tools, discussion board and chat room technology, actionable knowledge on the other hand requires more advanced technologies such as video and other synchronous multimedia technologies.

The challenge that most information and knowledge management systems face today is the lack of ability to integrate the capture and transfer of actionable knowledge, articulated knowledge as well explicit knowledge. In this paper, we report on the conceptual design of a highly interactive holistic framework for knowledge management. The main idea behind the system can be explained in a three-step process. First is to try and satisfy the user request by lo-

cating and extracting information from diverse information resources. At this end, technologies such as text mining, web crawlers and sophisticated search engines are used to identify and locate relevant information and make it available to the user.[3, 4] In this step the system is dealing with explicit knowledge or information stored in databases, data warehousing and on the Web. If the user is still not satisfied with the outcome from the first step, then the next step is to engage an intelligent broker that can match the user request with pre-identified expertise and post the user request to an expert.

At this stage we are dealing with a type of knowledge in the form of "Know How" that can be articulated and captured using tools such as an email system, discussion board or any other collaborative tools. If all that failed and the user request still is not addressed, then we might be dealing with an actionable knowledge that other means of communication is necessary. Actionable knowledge is a type of knowledge that requires face-to-face interaction, apprenticeship, and socialization. In this case, the intelligent broker could suggest ways of addressing the user request by facilitating that interaction to take place.

Integrating Diverse Information Sources

One of the biggest problems that organizations are facing today is the lack of ability to capture and integrate information residing in diverse sources, some of which are internal to the organization (for example, transaction databases, data warehouses, knowledge portals, and document management systems) while others are external (for example, commercial databases, credit reports, market research excerpts, news agency announcements, Web pages, etc.). Tightly integrating multiple diverse sources into a single cohesive system for the sake of centralization of the information sources often creates rather large extremely rigid systems that are not practically manageable.

The alternative method to the tightly integrated systems is the use of federated systems, which are based on a loose interfacing of applications and data resources.[2] However, since most federated systems employ a central data management dictionary with transla-

tors that convert search queries into something that can be understood by the diverse information sources, without frequent updates to the dictionary and the translators, the entire data sources can quickly become obsolete and unusable for searching. Additionally, as the number of source systems becomes larger, the maintenance of the dictionary and translators becomes harder, if not insurmountable.

The other issue with integrating information from multiple sources is the high cost associated with such a process. Information only becomes valuable when people need it and can make use of it. Therefore, whether centralized or federated the construction of any large information repository should be driven by the obvious needs of the users. That is, creating a knowledge repository to store information that is often used or sought after is normally a good idea, as opposed to creating knowledge repositories because we can technically and financially afford to do so. Thanks to the increasingly lower cost of storage devices and the advances in data warehousing technology, large amounts of information collected over time can now be retained for a longer period of time. Some of the information retained today might be for legal purposes, but the majority of it is retained with the objective of mining it in the hope of discovering valuable knowledge that can be used to assist in the decision making processes to enhance an organization's competitiveness.

According to recent trends, extracting useful information from large amounts of diverse information sources requires the use of data and text mining tools. Data Mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases, where the data are organized in records structured by categorical, ordinal and continuous variables. However, a vast majority of business data are stored in documents that are virtually unstructured.

According to a recent study by Merrill Lynch and Gartner, up to 90% of all organizational data is stored in some sort of unstructured text.[9] Therefore there is a need for tools to process unstructured data sources. Text mining (also known as text data mining and knowledge discovery in textual data-

bases) can be described as the process of deriving novel information from a collection of texts (also known as a corpus) which aims to address this need. By novel information, we mean associations, hypotheses or trends that are not explicitly present in the text sources being analyzed. Even though text mining is considered a part of the general field of data mining, it differs from regular data mining, such that in text mining, the patterns are extracted from natural language text rather than from structured databases of facts.

Databases are designed for programs to process automatically; text is written for people to read and understand. We do not have programs that can "read" and "understand" text (at least not in the manner that human beings do). Furthermore, despite the phenomenal advances achieved in the field of natural language processing, we will not have such "intelligent" programs for the foreseeable future. Many think that it will require a full simulation of how the mind works before we can write programs that read and understand the way people do. If so, then what does text mining do? On the most basic level, it numericizes and characterizes the unstructured text documents into structured table representations and then, using data mining tools and techniques, extracts patterns/knowledge from them.

## A Holistic Framework for Knowledge Discovery Management

The limitations of the current keyword based information retrieval systems and the complexity associated with integrating information from diverse information sources intensify the search for a better inquiry system, which is capable of taking advantage of the latest technologies in knowledge discovery and knowledge re-use. Since technology is merely an enabler in knowledge management environment, there is a need to achieve a certain degree of harmonization between people, technology, and information. Figure 1 shows the main components of the integrated knowledge management platform. The platform is divided into two distinct subsystems (knowledge creation sub-system and knowledge utilization subsystem) that are integrated into each other through the knowledge depository. The level of integration between these subsystems (and the modules within them) determines the performance of the whole system and the quality of services provided to the users. In the knowledge creation subsystem, the main components are the text and data mining tools and the WebCrawler. The text and data mining tools module is used to extract useful information from diverse information systems such as large data warehousing systems and textual information systems that includes the Web. Extracted information in the form of knowledge nuggets is then stored in the knowledge depository to be used by the knowledge utilization subsystem.

The WebCrawler is an intelligent agent used to monitor relevant information on the Web and captures and stores information into the knowledge depository in the form of knowledge nuggets which are updated for accuracy at all the times. The knowledge depository is a key component of the knowledge management system. It stores and organizes information coming from the modules within the knowledge creation subsystem as well as capturing the transactional logs of questions and answers handled by ask the expert module. Organizing information in a way that makes it easily and quickly retrievable is the key for any successful information system, since the choices made for the system organization in the knowledge depository will have a major impact on the performance of the whole knowledge management system.

The Intelligent Broker component of knowledge utilization subsystem is designed to facilitate collaboration and interaction between the user community and knowledge depository as well as the human experts (which are identified by the system via the "know-who" knowledge nuggets within the depository). With the integrated components such

Figure 1. integrated knowledge management platform
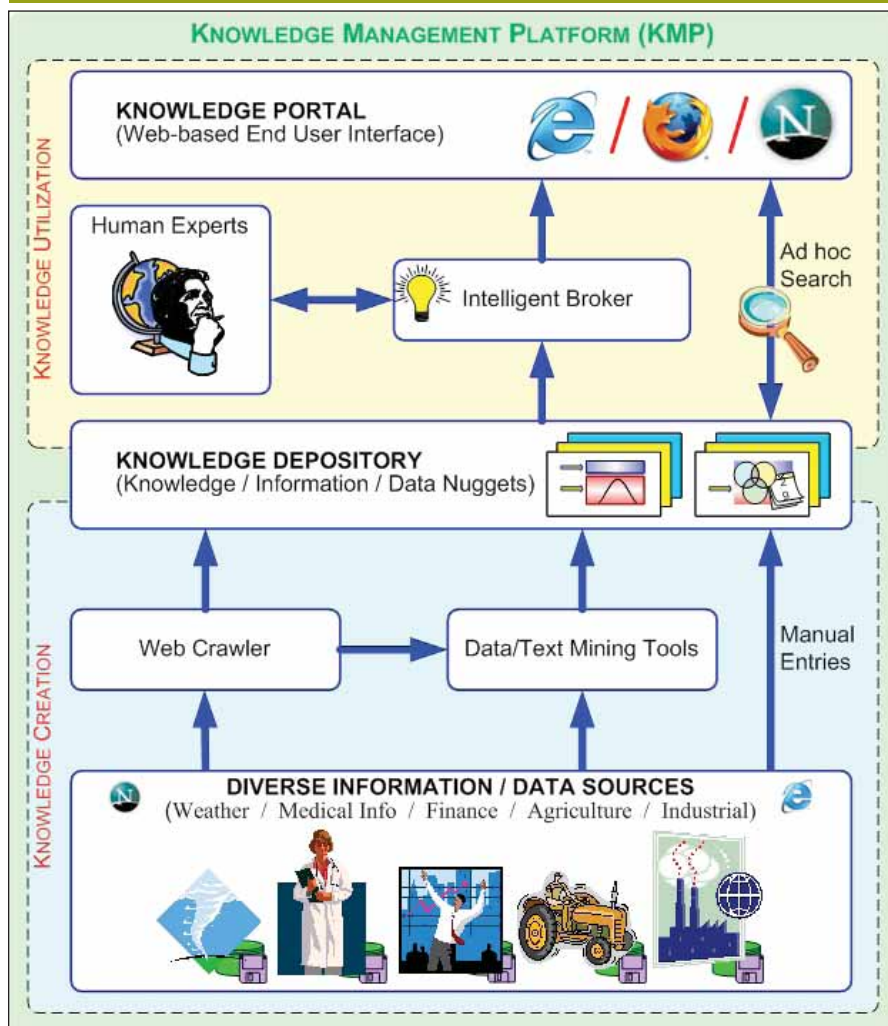
as ask the expert, knowledge portal, problem profiling, and personalization, the knowledge utilization subsystem provides users with access to the most relevant information via the knowledge depository and/or direct contact to human experts. It will also provide direct access to the knowledge depository via the ad-hoc keyword searches.

Knowledge Discovery and Re-Use
One of the ways to minimize knowledge loss is to enhance knowledge capture and re-use within the organization. Knowledge loss within the organization could happen as a result of many internal and external factors. Some of these factors include retirements, resignations, and restructuring or even outsourcing activities.[11] The challenge in minimizing knowledge loss is the ability to identify the knowledge sources and then take the necessary measures to ensure knowledge retention and utilization. Unlike information, where information systems can be used to collect process and store information, most knowledge resides in the minds of people and therefore it is necessary to involve these people in the process. This means that we acknowledge the

fact that certain types of knowledge are not codifiable and the best way to tap into that type of knowledge is to create a knowledge management environment by which these sources of valuable knowledge are identified and made available as and when the knowledge is needed.

The inquiry platform is an integrated environment by which the users interact with different types of knowledge including the human sources in the form of domain experts. Since the cost associated with eliciting knowledge from the experts is much higher than finding it in structured databases or on the Web,[12] the inquiry system has to be designed in away that users are directed to the human expert as a last resort, and in the event the information retrieved from the knowledge depository is not satisfactory and/or does not answer the user's queries. Figure 2 shows the workflow for posing a query or an unstructured question to the system. When the user submits a new query, the inquiry system first tries to determine if the new query matches any of the previously processed queries and/or the questions answered by the subject experts in the knowledge depository.

If there is a match, the user will be provided with the answer and an additional interaction would take place to determine the level of satisfaction the user had with the given answer. The user can also refine the query and search the knowledge depository for more relevant information. After the interaction with the search facilities, if the user is still not satisfied with the answer or there is no match to the query found in the knowledge depository, the system will then refer the user to a human expert. There are two types of human expert stored in the expert directory: internal experts who are part of the user community of the system and their answers to other users' queries are captured directly and made available in the knowledge depository for future use; or external experts who are not part of the knowledge management platform and in this case other means of communications such as email or chat room is used. As the email or chat gateway can be configured as part of the inquiry system, the communication with the external expert and the user can also be captured and made available in the knowledge depository for future use. In the case where the re-

Figure 2. Workflow for posing a query or an unstructured question to the system
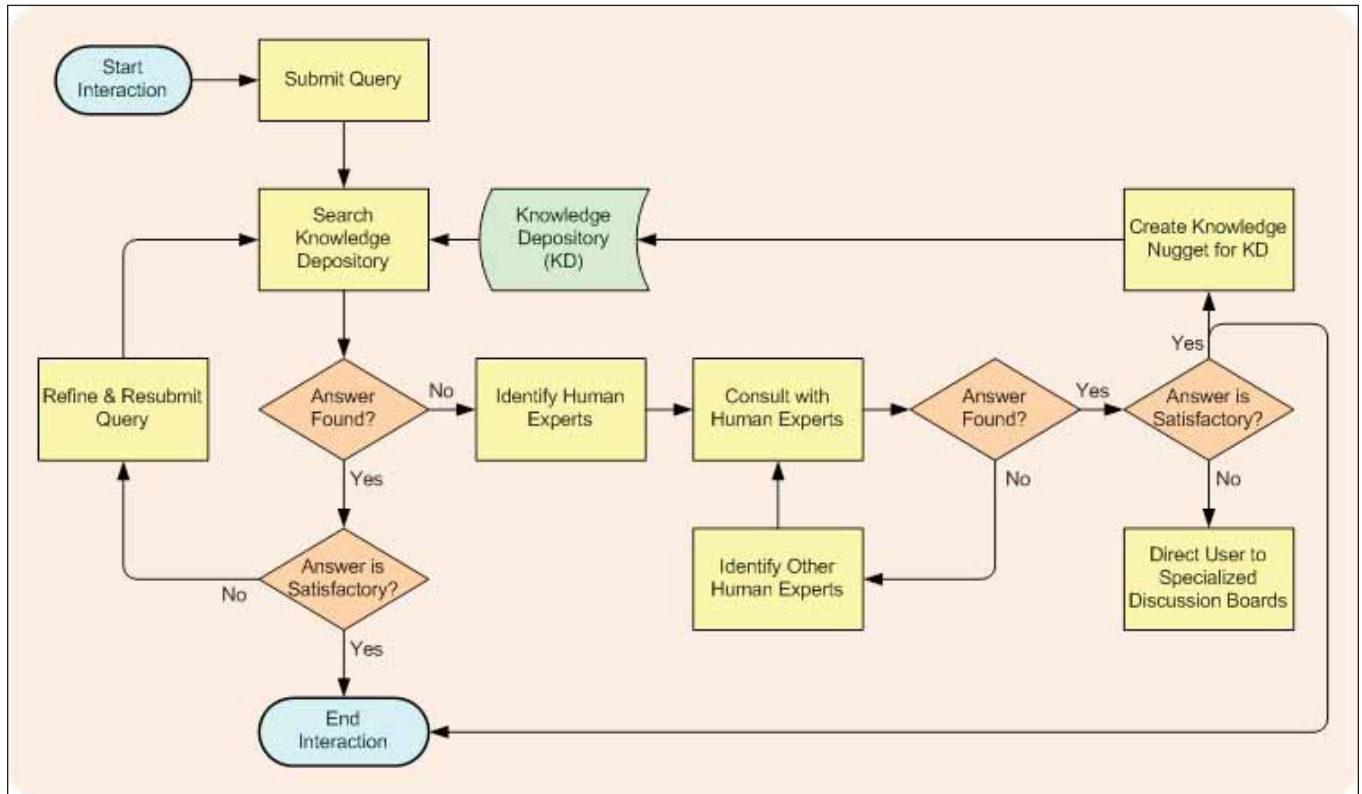
quired knowledge cannot be codified and a human interaction is required, the inquiry system has to facilitate that process by linking the user to the most credible human expert. The communication between the user and human experts and retaining that knowledge will not only minimize knowledge loss but also reduce the cost associated with people asking the same question again and again, especially when that information can easily be captured and stored in the knowledge depository.

In order to have access to all possible knowledge sources, a holistic knowledge management system should not only facilitate the management of traditional knowledge nuggets created from readily available data and information sources using manual or automated means, but also manage the details about the subject experts such as the areas of expertise, contact details, contact preferences and time of availability. This directory can be thought of a sort of "yellow pages" that maintains subject experts and their knowledge profiles from which the tacit knowledge can be obtained on an "as needed" basis and interaction logs can be codified.

Conclusion

Given the increased emphasis on knowledge management practices fuelled largely by the desire of many organizations to eliminate (or to minimize) the knowledge loss, there is a pressing need to go beyond the simple piecemeal information processing activities and focus on a holistic knowledge management process that enhances knowledge capture and re-use in an integrated platform. Designing a collaborative and interactive knowledge management platform requires a good understanding of the complexity associated with the capture and utilization of different types of knowledge and in particular the challenges associated with tacit knowledge. The conceptual architecture presented in this paper recognizes that complexity and provides a practical approach to the tacit knowledge capture and transfer. The ability to record and retain the output of human interaction and provide that as a knowledge source for future use not only helps to improve knowledge capture and re-use but also reduces the cost associated with people seeking answers to the same or similar questions over time.

References
1. Al-Hawamdeh, S. Codifying the "Know How" using CyKnit knowledge integration tools. In *Proceedings of the Intelligent Systems Design and Applications*, Springer Publishing, Tulsa, OK, (2003) 313-319.
2. Avrahami, T.T., Yau, L., Si, L. and Callan, J. The FedLemur project: Federated search in the real world. *Journal of the American Society for Information Science and Technology 57*, 3, (2006) 347-358.
3. Benjamin, P., Delen, D., Ramachandran, S. and Erraguntla, M. Towards a knowledge discovery framework. In *Proceedings of the International Conference on Information and Knowledge Engineering*, CSREA Press, (June 24-27, Las Vegas, NV). (2002) 234-243.
4. Cannataro, M. and Talia, D. The knowledge grid. *Commun. ACM 46*, 1, (Jan. 2003) 89-93.
5. Cordeiro, C.M. and AL-Hawamdeh, S. Social cognition theory and discourse analysis as frameworks in accessing implicit knowledge: the case of Swedish organizations in Singapore. In *Proceedings of the Second International conference on Knowledge management*, (Charlotte, NC, Oct. 2005) 27-28.
6. Davenport, T.H. and Prusak, L. *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, MA., (1997).
7. Desouza, K.C. Facilitating tacit knowledge exchange. *Commun. ACM 46*, 6, (2003) 85-88.
8. Herschel, R.T, Nemati, H. and Steiger, D. Tacit to explicit knowledge conversion: Knowledge exchange protocols. *Journal of Knowledge Management 5*, 1, (2001) 107-116.
9. McKnight, W. Building Business Intelligence: Text Data Mining in Business Intelligence, *DM Review*, (2005) 21-22.
10. Polanyi, M. *The Tacit Dimension*, Doubleday Press, NY, (1967).
11. Wolff, E. The growth of information workers in the U.S. economy. *Commun. ACM 48*, 10, (Oct. 2005) 37-42.
12. Wood, C.A. and Ow, T.T. Corporate data to data derived from the Web. *Commun. ACM 49*, 9, (Sept. 2005) 99-104.

Dursun Delen (dursun.delen@okstate.edu) is an associate professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University.

Suliman Al-Hawamdeh (suliman@ou.edu) is a professor and Program Coordinator of the Master of Science in Knowledge Management at the School of Library and Information Studies at the University of Oklahoma.