

In Search of A New Generation of Knowledge Management Applications

Edy S. Liongosari, Kelly L. Dempski, Kishore S. Swaminathan

Center for Strategic Technology Research

Andersen Consulting

3773 Willow Drive, Northbrook, IL 60062, USA

E-mail: edy@cstar.ac.com

ABSTRACT

Today's typical Knowledge Management systems are not much different from document management systems. In both cases, the retrieval process involves entering a set of keywords and then browsing through a list of documents related to those keywords found by the systems. If Knowledge Management is to live up to its promises, a new generation of Knowledge Management-enabled applications has to be developed. The information has to be presented beyond just a list of documents. Applying data mining techniques to these systems is one of the few promising avenues that may yield a new set of applications. This paper describes our on-going research effort to extract and mine information from one of the largest private Knowledge Management systems in the world.

INTRODUCTION

With well over 3,000 databases containing millions of documents accessed by over 50,000 users across 78 countries, Andersen Consulting has the largest Lotus Notes installation in the world. This system, called the *Knowledge Xchange*[™] or simply KX, is one of the strategic assets of Andersen Consulting. It is designed to support communities of practice that share and reuse knowledge. It contains discussion databases, library databases, and various directories. The library databases contain a wide variety of documents such as project proposals, project deliverables, case studies, credentials, résumés, newsletters and prototypes.

The KX is used primarily for two purposes: finding documents and finding subject matter experts. Imagine a typical scenario where we need to quickly write a proposal for a client. We need to gather our credentials on the subject matter, previous proposals related to that subject, case studies, estimating guidelines, and the names of the subject matter experts. With the KX, you can find all of that information without leaving your desk.

THE EXPLORATION

The KX has been in place for over five years and it stores over 200 GB of information accumulated by tens of thousands of Andersen Consulting's employees over that period of time. If you view the KX as a sea of raw data, performing data mining over the KX may indeed provide much insightful information such as how Andersen

Consulting operates as an organization, how it manages its relationships with its clients, and how it responds to market demands. These insights in turn can be used for various planning and decision making processes, as well as tracking the progress of certain strategic objectives.

With this in mind, we started an exploration to find out what type of new information we can discover from the KX as it is today. Obviously this is very much a bottom-up approach.

In this paper, we will describe the scope of our inquiry and the initial challenges we encountered. We will then cover some of the quick wins, followed by a more in-depth examination of one facet of the exploration: the construction of *The Old Boys Network*. We conclude this paper with a brief discussion on other parts of this investigation and the business benefits the exploration has delivered so far.

EXPLORATION SCOPE: DATA SELECTION

In order to scope this exploration effort, we have selected the twenty largest and most widely used databases in the KX. Ten of them are document libraries, five discussion databases, and five directories. Together they comprise about 1.5GB of data excluding the attachments. With these twenty databases, we have a manageable size of data to mine and at the same time it is large enough to represent the rest of the KX.

The document libraries contain information such as the authors of the documents, creation dates, abstracts, and related keywords. The discussion databases contains the topics of the discussions, the participants, and the dates of discussions. The directories include information about Andersen Consulting's employees such as their phone numbers, e-mail addresses, locations, and the groups to which they belong.

INITIAL CHALLENGES

Mining directly from Lotus Notes databases is arduous. Unlike DBMSs, the underlying storage of Lotus Notes is structurally weak. Even though Lotus Notes has things called databases, they are not really databases in the traditional sense. They are tables. Each row represents a document and each column represents an attribute associated with the document.

There is no notion of relationships across tables. This makes consistency maintenance across tables a difficult thing to do. Many of our databases, especially the older ones, do not have any consistency maintenance mechanisms in place.

This creates many problems. For example, the names of document authors are not validated. David S. Smith can be entered as Dave Smith. If David S. Smith has authored two documents, but one of the documents has Dave Smith as its author and the other has David S. Smith, there is no trivial way to reconcile the two documents and conclude that they are in fact, written by the same person. This inconsistency may cause incorrect results in our mining process.

While the solution we came up with does not solve all of the inconsistencies, it is good enough to enable us to move forward to the next step. Our solution can be found in [1]. The process we went through to cleanse and integrate the data is very similar to the KDD process as described in [3,4]. Suffice to say that the central idea of our solution is based on a data model similar to the one shown in Figure 1.

This model serves as an index to the underlying information. The information of a person, for example, is derived from the list of document authors in the library databases, the project member listings, the employee telephone listing, and so on. Similarly the relationships between the entities are also derived. The *Has skills in* relationship between the entities *Person* and *Subject Area*, is derived from the number of documents the person has authored in that subject area.

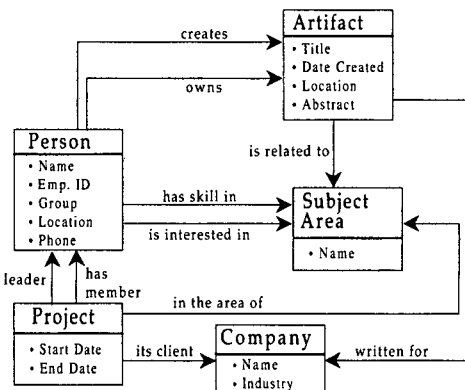


Figure 1. The data model used in our experiment.

MINING

Having a data model like the one depicted in Figure 1 makes our mining effort a lot easier. The model transformed the isolated, disparate data in Lotus Notes into one cohesive, integrated and well-structured body of knowledge. Even with just five entities, we are able to perform many analyses.

Quick Wins

With well-structured data, we can now deliver new functionalities beyond the ones provided by the traditional

document management systems. Below are two examples of such functionalities that do not require data analysis.

Biography Generator

Remember the time when you wished you could quickly find out the background of the person you were talking to over the phone. With the traditional Knowledge Management system, you have to enter the name of the person as a keyword; the output is a list of documents that you have scanned through. This is a very time-consuming process and it requires time you do not have – especially when you are in the middle of a telephone conversation.

With the Biography Generator, one can generate a chart that outlines the subject areas that a given person has been involved with over time. Figure 2 shows a bar chart with a list of subject areas in the Y-axis and time in X-axis. This summarized information will provide you with a quick glance of the recent activities of that person. The chart is generated based on the keywords and the creation date of the documents that person has authored and the projects he or she has previously been involved in.

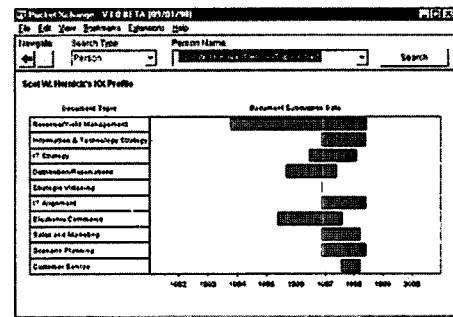


Figure 2. Summary of one's work over time

Rate of Absorption

Given a new subject area such as Electronic Commerce, Rate of Absorption will be able to demonstrate how quickly each group within an organization become aware of and begin incorporating this new area into their work. This information is important to many executives to understand the dynamics of their organizations.

The rate of absorption is derived by identifying the readers and authors of documents in this subject area and tracking them over time based on their groups (see Figure 3).

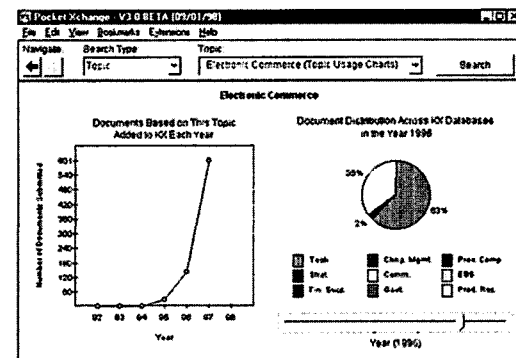


Figure 3. Tracking the distribution of Electronic Commerce.

The Old Boys Network: a more in-depth coverage

While the two features above can be implemented rather easily on the top of the data model shown in Figure 1, other features we are exploring require further analysis on the data and one of them is what we call the *Old Boys Network*.

The Old Boys Network is a network of who-knows-whom and many executives consider a good working knowledge of this network complemented by the information about who-knows-what as an extremely important asset. Many of these executives spend a large sum of money every year to join organizations like Young President Organization (YPO), to maintain and enlarge their social networks.

While the KX does not have much information about the people outside our firm, it does contain a lot of information about our internal people. Our task is to figure out who-knows-whom internally based on the information we have in the KX. The input is the name of two people: A (i.e., you) and B (i.e., the person you want to reach), and the output is a list of A's and B's mutual acquaintances.

This problem is pretty much a graph analysis problem. Each node of the graph represents a person and the link between two nodes represents the relationship between two people.

The links are created based on the following information:

- People who have co-authored a document;
- People who have significant discussions in the discussion database, and
- People who belong to the same group.

The links are weighted to show the degree of "closeness". For example, people who have co-authored many documents are considered a lot closer than those who merely belong to the same group. We use the degree of closeness in the traversal and matching process. The link with the highest degree of closeness is traversed first.

In order to make this tool simple and fast, we utilize one simplifying assumption. Our algorithm limits the search to two degrees of separation for the following reasons:

- We believe that a communication path with more than two intermediaries is ineffective in most cases.
- Given that we have about 80,000 nodes, the output can become unmanageably large beyond two degrees of separation.
- It improves the performance and the complexity of our algorithm since we have a large graph and it consists of many disconnected sub-graphs. Instead of trying to find out the shortest path between any two nodes, we can now use a simple graph traversal algorithm to find the path to the target node.

Unfortunately due to privacy and security reasons, we do not have access to e-mail or voice mail information among the people in our firm. Otherwise, we would be able to significantly increase the completeness of our data. To compensate for this handicap, whenever the algorithm cannot find a path between A and B, it generates a list of people within two separations from B. This way, A can manually scan through the output and may find somebody he or she knows.

Other Mining Efforts

Here are the highlights of some of our on-going efforts:

- *Community of Practice*: Discover communities of practice by grouping people based on their recent involvement in a particular subject area and their level of expertise. This will help people within a large organization become aware of and ultimately communicate with those people that are working in a similar area across organizational and geographical boundaries. Furthermore, if the same information is analyzed over time, we can see how a community of practice evolved. We will be able to identify when a particular practice is slowing down and which new practice the people are moving into.
- *Related subject areas*: The ability to group subject areas into clusters may yield several interesting insights. One trivial way is to find out how closely each subject area is related to each other. This information can be derived from the number of co-occurrences of each subject area in the documents. However, if we analyze the clusters based on time, we will be able to find out how one subject area is evolving to another. This information is rather important for the knowledge managers to be able to organize categories and hierarchies of subject areas properly.
- *Suggested reading*: By analyzing what documents are ordered by other people with a similar background (i.e., skills and project experience), the system is able to suggest new reading and other training materials.

CURRENT STATUS AND FUTURE WORK

Many of the features mentioned in this paper are still being implemented. The ones listed in the *Quick Wins* section have been implemented and distributed as part of our knowledge management tool suite [2]. The first version of these features was released around June of 1998 and is currently being used by about 4,000 Andersen Consulting's employees worldwide.

The benefits delivered by the analysis based on the simple 5-entity data model are tremendous. For example, we have developed a tool called *Lead Finder* for our internal marketing group. *Lead Finder* looks for people with a set of interests in certain subject areas. Prior to using *Lead Finder*, the marketing group spent more than 80% of its time looking for the right people within our organization to contact. With *Lead Finder*, the group spends its time

working with the target audiences instead of looking for them.

We believe we are at the beginning of a long journey. There is a lot more information we can extract, mine, and discover. The data model needs to be enlarged to cover a more diverse set of information. The potential payoffs of this effort are high. However, there are very few products out there that provide this type of functionality. The great challenge here is that the information extracted from documents in the knowledge stores is subjective and open to interpretation. This can yield inaccurate and unexpected results from the mining process if not done carefully.

REFERENCES

[1] Brody, A.B., et.al. Integrating Disparate Knowledge Sources. To appear in Proceedings of the Second

International Conference on Practical Application of Knowledge Management (Apr. 1999).

[2] Davenport, T.H., Hansen, M.T. Knowledge Management at Andersen Consulting. In Harvard Business School Case Study, N9-499-032, Aug. 1998.

[3] Fayyad, U., Piatetsky-Shapiro G., Smyth P. The KDD Process for Extracting useful Knowledge from Volumes of Data. In the Communications of the ACM, 39, 11 (Nov. 1996), pp. 27-34.

[4] Gardner, S. (1998): Building the Data Warehouse, Communications of the ACM, 41, 9 (Sep. 1998), pp. 53-60.

The Social Affordances⁸ of E-Mail

Barry Wellman⁹

Professor of Sociology, Centre for Urban and Community Studies, University of Toronto,
Toronto Canada M5S 2G8. wellman@chass.utoronto.ca

There has been a tradition in the CSCW literature that e-mail is always an inferior substitute for face-to-face communication. The question has been about the extent to which e-mail can substitute for face-to-face.

My students and I were meditating on this the other day. We realized that there were a number of ways in which e-mail was preferable to face-to-face communication. Here they are, in no particular order. E-mail gives you:

Focus

- Time to think in the preparation of statements.
- Allows you to say what you want to say rather than be subjected to interruptions and re-interpretations.
- Greater focus on content rather than on externalities.
- Simpler emotional communication.
- More control of your work flow.

Privacy and Control

- No intrusions by others into a private conversation.
- Unless you are under covert surveillance, third parties cannot observe your communications with someone.

- A relatively enduring record of your communication, time and date stamped.

Speed and Ease

- Some of us type faster than we talk.
- Almost all of us read faster than we listen.
- If you are already sitting at a terminal, e-mail is quicker and less disruptive of work flow than walking over to someone or dialing a telephone number and waiting for it to be answered (by humans or voice-mail.)
- Email is more likely to reach the other party -- wherever s/he is -- than telephone calls or doorbell ringing.
- Less need to get bogged down in the social amenities when you quickly want to convey information.
- No need to remember people's addresses or directions to their homes.
- Easy to include attachments and hypertext links.

In short, e-mail is not just a lame version of face-to-face communication. *E-mail is e-mail.*

⁸ Thanks to Erin Bradner (Univ. of California, Irvine) for suggesting the term „social affordances“.

⁹ This column was prepared when I was an Visiting Professor (Spring, 1999) at the School of Information Management and Systems, University of California, Berkeley. The students in the „Information in Society“ course who contributed ideas to this article are: Carol Butler, Richard Chen, Ame Elliott, Danyel Fisher, Malo Hutson, Petter Johnstad, Nalini Kotamraju, Megan Thomas, and Cathrine Torgersen.