

# Web Data Mining Techniques for Expertise-Locator Knowledge Management Systems

Irma Becerra-Fernandez, Ph.D.\*

Juan Rodriguez\*\*

\* College of Business Administration, Decision Sciences and Information Systems

\*\* Knowledge Management Lab

Florida International University

Miami, FL 33199

{becferi, jrodri36}@fiu.edu

## Abstract

Knowledge management systems are making inroads into organizations that want to get a handle on their intellectual capital. To this end, we have designed and implemented Expert Seeker, an Expertise-Locator knowledge management system that will be used in several NASA Centers to identify experts within the organization. This paper presents the use of web data mining techniques to construct and maintain an employee's profile.

## Introduction

Knowledge management systems (KMS) have been defined as "an emerging line of systems [which] target professional and managerial activities by focusing on creating, gathering, organizing, and disseminating an organization's 'knowledge' as opposed to 'information' or 'data'" (Alavi and Leidner, 1999). Based on the KM Life Cycle models (Nissen, 2000) and on a study of the KM systems underway at many organizations (Becerra-Fernandez, 1998a) a framework emerges for classification of KMS (Becerra-Fernandez and Stevenson, 2000). The framework includes the following:

1. *Knowledge Preservation*: Refers to systems that preserve and formalize the knowledge of experts so it can be shared with others. As such, these systems aim to elicit and catalog the tacit knowledge of experts, and serve to transfer their knowledge. Knowledge preservation systems formalize knowledge in models such as concept maps, which allow others to learn the domain (Cañas et. al., 1999).
2. *Knowledge Application*: Refers to systems that assist in solving problems. Organizations with significant intellectual capital require eliciting and capturing knowledge for reuse in solving new problems as well as recurring old problems.

New problems could be similar to old problems or even consist of a combination of old problems (Becerra-Fernandez & Aha, 1999).

3. *Knowledge Discovery*: Refers to systems that create new knowledge through the implementation of intelligent algorithms such as data mining, and through the inference of data relationships.
4. *Knowledge Repositories*: Refers to systems that organize and distribute knowledge. Knowledge repositories comprise the majority of the KMS currently in place. Under the auspices of KM, tools historically used for singular unrelated purposes are integrated to address the corporate memory problem. Expertise-Locator Systems (also called knowledge yellow pages or People-Finder systems) is a special type of knowledge repository that point to experts, those that have the knowledge within the organization.

Expertise-Locator Systems are knowledge repositories that attempt to organize knowledge by identifying experts who possess specific knowledge within an organization. Several organizations in different business categories have identified the need to develop systems to help locate intellectual capital, or Expertise-Locator KMS. The intent in developing these systems is to catalog knowledge competencies, in a way that could later be queried across the organization. A literary review and a table comparing the characteristics of hallmark Expertise-Locator KMS in use in organizations today appears in (Becerra-Fernandez, 2000a). The results presented in this study confirm that most of the Expertise-Locator KMS in place today rely on each employee completing a self-assessment of competencies, which is later used when searching for specific knowledge areas. The issue of a self-assessment is one that offers both advantages and disadvantages. The advantage of self-assessment is that it allows

building a repository of organization-wide competencies quickly. The disadvantage is that the results of self-assessment are subjective, based on each person's self-perception, the results could be hard to normalize, and employees' speculation about its possible use could 'skew' the results. For example, one particular organization conducted a skills self-assessment study during a period of downsizing. This resulted in employees' exaggeration of their competencies; for fear that they might be laid-off. On the other hand, another organization made it clear the self-assessment would be used to contact people with specific competencies to answer related questions, which may result in employees downplaying their abilities, in order to avoid serving as consultants for the organization.

Data mining refers to the extraction of information or the identification of patterns, usually within a large collection of data (Fayyad et. al., 1996; Zaiane et. al., 1999; Ahonen et.al.1997). Web data mining makes use of data mining techniques to extract information from web-related data. There are three types of web data mining. These are web structure mining, web usage mining, and web content mining. Web structure mining examines how the web documents themselves are structured. Web usage mining involves the identification of patterns in user navigation through web pages in a domain. Web content mining, is used to discover what a web page is about and how to uncover new knowledge from it. Web content mining is based on text mining and information retrieval (IR) techniques; which consist of the organization of large amounts of textual data for most efficient retrieval, an important consideration in handling text documents. IR techniques have become increasingly important, as the amount of semi-structured as well as unstructured textual data present in organizations has increased dramatically. IR techniques provide a method to efficiently access these large amounts of information. One application of these methods is in the construction of Expertise-Locator KMS.

This paper presents insights and lessons learned from the development of Expert Seeker, an organizational Expertise-Locator KMS that will be used to locate experts at the National Aeronautics and Space Administration (NASA). The paper also presents the role of data mining techniques in automating the maintenance of the expert's profiles.

## System Overview

The NASA Faculty Awards for Research (FAR) for NASA-Kennedy Space Center (KSC), as well as the Center of Excellence in Space Data and Information Sciences (CESDIS) are funding the development of Expert Seeker at Florida International University Knowledge Management (KM) Lab. Previous Knowledge Management studies at KSC affirm the need for a center wide repository which will provide KSC with Intranet-based access to experts with specific backgrounds (Becerra-Fernandez, 1998a). Expert Seeker aims to help locate intellectual capital within NASA-KSC and GSFC, and its use is expected to expand to other NASA Centers. The Expert Seeker KMS is accessed via KSC's Intranet. The Expert Seeker KMS provides access to competencies available within the organization, including items that are not typically captured by traditional Human Resource applications, such as completed past projects, patents, and other relevant knowledge. This Expertise-Locator KMS will be especially useful when organizing cross-functional teams. The main interfaces on the query engine in Expert Seeker uses text fields to search the proposed data for keywords, fields of expertise, names or other applicable search fields. Expert Seeker will offer NASA experts more visibility, and at the same time allow interested parties to identify available expertise within NASA.

The development of Expert Seeker requires the utilization of existing structured data as well as semi-structured and unstructured web-based information as much as possible. Expert Seeker uses the data in existing Human Resources databases for information such as employee's formal educational background, the X.500 Directory for point-of-contact information, a Skills Database which profiles each employee's competency areas, and the Goal Performance Evaluation System (GPES). Information regarding skills and competencies, as well as proficiency levels for the skills and competencies needs to be collected, to a large extent, through self-assessment. Recognizing that there are significant shortcomings in self-assessment, we propose an increased reliance in technology to update employees' profiles, and thus place less reliance on self-assessed data. Figure 1 presents Expert Seeker's system architecture. A complete description of Expert Seeker, including the technologies used to implement the KMS and the system architecture

appear in (Becerra-Fernandez, 2000b).

### **Challenges in the Development of Expertise-Locator KMS**

One of the principal challenges in the development of Expertise-Locator KMS deals with the development of *knowledge taxonomies*. Taxonomy is the study of the general principles of scientific classification. Knowledge taxonomies allow organizing knowledge or competency areas in the organization. In the case of Expertise-Locator systems, the taxonomy is used to describe and catalog people's knowledge, an important design consideration. Furthermore, knowledge taxonomies could be critical in the successful implementation of the Expertise-Locator systems (Becerra-Fernandez, 2000a).

The use of web data mining can mitigate some of the problems inherent to relying on biased self-reporting required to keep employee profiles up to date, or the need to develop an accurate organizational skill taxonomy a-priori. This technique draws from an existing pool of information that provides a detailed picture of what the employee knows, based on what s/he already publishes as part of their job, including their web pages. A web data mining approach requires minimal user effort to maintain the records accuracy, eliminating the need for "nagging" systems that prompt users to maintain their profiles up-to-date. Through web data mining the collection of expertise data is based on published documents, eliminating the need for possibly biased self-reporting. Using web data mining this information can be collected automatically, and employee skill information can be kept up to date through periodic reprocessing of the document body for documents that are new or have been updated.

### **The Web Text Mining Process**

A KMS system that locates experts based on published documents requires an automatic method for identifying employee names, as well as a method to associate employee names with skill keywords embedded in those documents. For this purpose, Expert Seeker required the development of a name-finding algorithm to identify names of NASA employees. Traditional

IR techniques<sup>1</sup> were then used to identify and match skill keywords with the identified employee names. An IR system typically uses as input a set of inverted files, which is a sequence of words that reference the group of documents the words appear in. These words are chosen according to a selection algorithm that determines which words in the document are good index terms. In a traditional IR system, the user enters a query, and the system retrieves all documents that match that keyword entry. Expert Seeker is based on an IR technique that goes one step further. When a user enters a query, the system initially performs a document search based on user input. However, since the user is looking for experts in a specific subject area, the system returns the names of those employees whose names appear in the matching documents (excluding webmasters and curators). The employee name results are ranked according to the number of matching documents each individual name appears in. The employee information is then displayed to the user.

The indexing process was carried out in four stages. First, all the relevant data was transferred to a local directory for further processing. In this case, the data included all the web pages on the NASA domain. This was done with a simple web-mirroring tool called *Wget*<sup>2</sup>.

The second stage identifies all instances of employee names by programmatically examining each HTML file. The name data is taken from the X.500 personnel directory databases. All names in the employee database are organized into a map-like data structure beforehand that is used in the web content mining process. This map consists of all employee names referenced by their last name key. In addition, each full name is stored in every possible form it could appear. For example, the name John A. Smith is stored as John A. Smith, J. A. Smith, J. Smith, Smith, John A., Smith, J.A., and Smith J. An individual document is first searched for all last name keys. Subsequently, the document is again searched using all values of the matching keys. Name data organized in this way can increase the speed of the text search. Using one long sequence containing all names in every possible form as search criteria would slow down processing time.

<sup>1</sup> See for example *Selection by Discriminant Value* in (Frakes and Baeza-Yates, 1992), an algorithm for selecting index terms.

<sup>2</sup> <http://www.gnu.org/software/wget/wget.html>

The third stage involves identifying keywords within the HTML content. This is done using a word frequency calculation. First the text is broken up into individual words, through string pattern matching. Any sequence of alphabetical characters is recognized as a word while punctuation, numbers, and white space characters are ignored. The resulting list of words is processed to determine if a word was included in a *stoplist*<sup>3</sup>. The resulting list of words was then processed with a *stemming* algorithm. A *stemmer* is used to remove the suffix of a word. This is done to group together words that may be spelled differently but have the same semantic meaning. A person who types “astronomical” as a query term would most likely also be interested in documents that match the term “astronomy”.

Once the stemming process is completed, the fourth stage involves calculating the frequency of each term. Word frequency was used during the keyword selection process in the determination of good index terms. However, other indexing algorithms could have been used instead with comparable results. It is important to note that the degree of relation between an employee name and a keyword within an individual document is not considered. Rather, expertise is determined based on the assumption that if an employee recurringly appears in many documents along with a keyword, then that person must have some knowledge of that term. Theoretically, a large document count for a search query should produce more accurate results. The chosen keywords have a twofold purpose. First, they are used to quickly associate employees with recurring skill terms. These keywords can also be used in future work for clustering similar documents into topic areas. Further work includes taxonomy construction from these keywords and the development of a query relevance feedback system that suggests query terms that are related to the query entered by the user.

### **Preliminary Results and Shortcomings of this Research**

Table 1 illustrates preliminary results for a set of skill keyword query terms. Precision was determined by testing whether a keyword entered

<sup>3</sup> A *stoplist* is a group of words that are not considered to have any indexing value. These include common words such as “and”, “the”, and “there”.

as a query term correctly describes the expertise for the corresponding employee. The precision values for the keywords in Table 1 are represented by the percentage of correct matches within the top 15 results for the keyword.

<b>Keyword</b>	<b>Precision (Top 15 results)</b>
Astrophysics	87%
Astronomy	92%
Comet	92%
Climate	92%
Ocean	73%
Atmosphere	87%
Management	64%
Human resources	53%

Table 1 - Precision Results for Sample Skill Keyword Query

Recall was not calculated because it would be hard to determine if the names appearing in the NASA web documents completely reflected all the employees of that organization. The results show a high precision for scientific or research related skill terms, and less precision for the more managerial or administrative related skill terms. This may be due to the nature of the document body as being highly scientific and research oriented. These results show that the system can retrieve experts from the document body with a substantially high degree of precision, in particular for scientific and research related keywords.

One of the shortcomings of this research relates to the fact that the accuracy provided by web data mining depends on the existence of employee web pages and their proper maintenance. Employee web pages must encode some minimal required level of content, including papers or technical documents published, collaborators in the case of multi-author papers, as well as identifying the competencies represented in those papers.

Another of the shortcomings of this research is the possible existence of multiple employees with the same name. Expert Seeker removes all repeat instances of the same name. All information referenced to a particular employee name was indexed without any attempt to distinguish between one of possibly several persons with the same name. Instances of multiple employees with the same name were handled at the time the user queried the database.

The results of the query are a list of hyperlinks referenced by employee names. When a user clicks a name the human resources data of that employee is displayed. In the case of multiple entries in the database for a name, the user is taken to an intermediary web page where a more detailed description of the employee is displayed. The user can then determine which of the employees is most likely to be an expert in the subject area he is making the query for. For example, if the query term is "astrophysics", the name of an employee who works in an astrophysics laboratory is more likely to be an expert in that area than a person who works in an administrative office.

Another obstacle is the indexing of keywords that may not be relevant areas of specialization. It is important to note that the query search is performed on the results of an IR keyword indexing process rather than from a predetermined set of skill terms. The system is not designed to discriminate between keywords that describe a skill area and words that are good indexing terms but are not relevant for determining an area of expertise. However, we tend to think this is not a significant problem, because if the system is designed for the determination of expertise in an area of specialization, then it is highly unlikely that a user will enter a query that does not describe a skill area relevant to the areas of specialization within the NASA organization. This point could be argued on the basis of a potential language or terminology barrier between user and data. The flexibility of the indexing process, however, does provide an advantage. This is that a search can be performed using keywords that are not traditionally considered skill terms such as project names and highly specific technical terms. This particular shortcoming also could become a significant issue in the construction of knowledge taxonomies via means of clustering techniques. Keywords that are not skill-related could inadvertently be placed in the taxonomy, creating irregularities within the knowledge hierarchy.

### Conclusions and Future Work

Web content mining can be useful and efficient for the construction of Expertise-Locator KMS in situations where a large text document body exists containing relevant skill information. It provides a solution to some of the challenges faced in the development of these systems,

including how to maintain up-to-date employee skill profiles while minimizing the need for cumbersome and possibly biased self-reporting. A prototype of this system was demonstrated to NASA representatives received a positive assessment. The system is currently under further review for implementation and use within their organization.

**Acknowledgements** The authors wish to acknowledge NASA-Kennedy Space Center and NASA-Headquarters under the Faculty Awards for Research (FAR-99), grant number NAG10-0259, as well as NASA-Goddard Space Flight Center and the CESDIS, contract number NAS5-32337 and subcontract number 5555-97-74, for financial support for the development of Expert Seeker. Special thanks to all NASA employees that collaborated in this effort, including Mr. Chris Carlson, Mr. Steve Chance, Dr. Milt Halem, Dr. Susan Hoban, Mr. James Jennings, Ms. Nancy Laubenthal, Dr. Shannon Roberts, and Mr. Pat Simpkins. The authors also wish to acknowledge the contributions of the students who work in the FIU Knowledge Management Lab and who collaborated in this research, specifically Rigoberto Fernandez, Hector Hartmann, Lusally Mui, Maria Ray, and Hernan Santiesteban.

### References

- Ahonen, H., Heinonen, O., Klemettinen, M., and Verkamo, I. 1997: Applying Data Mining Techniques in Text Analysis, Technical Report, C-1997-23, Dept. of Computer Science, University of Helsinki.
- Alavi, M., Leidner, D. (1999). Knowledge Management Systems: Issues, challenges, and benefits. *Communications of the Association for Information Systems* [online], 1. Available: <http://cais.isworld.org/articles/default.asp?vol=1&art=7> (November 1999).
- Becerra-Fernandez, I. 1998a. Corporate Memory Project, Final Report, NASA grant No. NAG10-0232, 12-25.
- Becerra-Fernandez, I. and Aha, D. Case-Based Problem Solving for Knowledge Management Systems. In Proceedings of the Twelfth Annual International Florida Artificial Intelligence Research symposium (FLAIRS) Orlando, Florida: Knowledge Management Track.
- Becerra-Fernandez, I. 2000a The Role of Artificial Intelligence Technologies in the

Implementation of People-Finder Knowledge Management Systems. *Knowledge Based Systems*, special issue on Artificial Intelligence in Knowledge Management, Vol. 13: No. 5, (October 2000).

Becerra-Fernandez, I. 2000b Facilitating the Online Search of Experts at NASA using Expert Seeker People-Finder. In Proceedings of the Third International Conference on Practical Aspects of Knowledge Management. Basel, Switzerland .

Becerra-Fernandez, I., and Stevenson, J.M. 2000. *Knowledge Management Systems & Solutions for the School Principal as Chief Learning Officer*. Forthcoming.

Cañas, A., Leake, D., Wilson, D. Managing, Mapping, and Manipulating Conceptual Knowledge. In Proceedings of the AAAI-99 Workshop on Exploring Synergies of Knowledge Management and Case Based Reasoning, 10-14. Menlo Park, AAAI Press.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds. 1996. *Advances in Knowledge Discovery and Data Mining*: AAAI Press.

Frakes, W., and Baeza-Yates, R. 1992 *Information Retrieval: Data Structures and Algorithms*. Upper Saddle, NJ: Prentice Hall.

Nissen, M.E. 2000. *Knowledge Based Knowledge management in the Reengineering Domain*. Forthcoming.