

A Text Analysis Approach to Motivate Knowledge Sharing via Microsoft SharePoint

Robert M. Patton, Wade McNair, Christopher T. Symons, Jim N. Treadwell, Thomas E. Potok
Oak Ridge National Laboratory
{pattonrm,mcnairaw,symonst,treadwelljn,potokte}@ornl.gov

Abstract

Creating incentives for knowledge workers to share their knowledge within an organization continues to be a challenging task. Strong, innate behaviors of the knowledge worker, such as self-preservation and self-advancement, are difficult to overcome, regardless of the level of knowledge. Many incentive policies simply focus on providing external pressure to promote knowledge sharing. This work describes a technical approach to motivate sharing. Utilizing text analysis and machine learning techniques to create an enhanced knowledge sharing experience, a prototype system was developed and tested at Oak Ridge National Laboratory that reduces the overhead cost of sharing while providing a quick, positive payoff for the knowledge worker. This work describes the implementation and experiences of using the prototype in a corporate production environment.

1. Introduction

Knowledge management practitioners wrestle with how to extract, retain and exchange information about experiences within an organization. People struggle to do for their employers what seems to happen naturally at a personal level outside of their workplace.

There are two consistent pillars of knowledge management that need to be stated. First, knowledge management is used by and inside organizations. Knowledge management is inherently a process by which organizations seek to capture and distribute insights gained from stakeholder experiences. Second, it desires an organic engagement with the knowledge worker. Value to the organization relies entirely upon adoption and use by the workers. Finding incentives for people to engage in knowledge sharing is a reasonable and likely outcome of these two pillars. Deriving clever incentive schemes, implementing the latest information systems, or crafting “participatory” policies are all touted as solutions among the chatter. The problem is that people -- the knowledge workers - get lost in the conversation.

People develop a sense about who in the firm has expertise, and that expertise creates both positional and relational power. “Knowledge is power,” is an American euphemism for “Joe is more valuable than Jim.” This paper describes a simple approach to provide value to all stakeholders who share information. The simple act of sharing information can provide value to both the knowledge worker and the organization in a manner that reinforces organizational culture, provides incentives to increase sharing, and facilitates firm-wide learning and analytics.

Unfortunately, there are three primary barriers to knowledge sharing. First, knowledge workers must receive something of equal or greater value than what is given. Second, there is often a high overhead cost associated with sharing, or the act of sharing is outside the workflow, thus creating extra work. An overhead costs is generally in terms of additional time required on the part of the knowledge worker to engage in the act of sharing. Third, the user interface must be simple, yet powerful, and should not require an advanced technical degree in order to use it. Consequently, in order to encourage adoption of knowledge sharing, the overhead cost should be minimized; the value received should be maximized, and the interface should intuitively facilitate the exchange of knowledge.

To achieve this goal, the work described here utilizes text analysis and machine learning techniques to create an enhanced knowledge sharing experience. A prototype system was developed and tested at Oak Ridge National Laboratory (ORNL). By reducing the cost of sharing while providing a quick, positive payoff for the knowledge worker, this pilot provides incentive for knowledge workers to share. This work describes the implementation and the experiences of using the prototype in a corporate production environment.

2. Background

One of the primary problems that many organizations have is that there is a multitude of data in documents, but search doesn't work well and people

don't know where to look. Consequently, they keep duplicating efforts or have only part of the information. At first glance, it is easy to launch into technology solutions focused on improving search or KM-oriented, process solutions aimed at improving the location and awareness of information repositories. Both of those approaches fail to address the "people" problem. The first task has to be centered on how people within the organization naturally share and access information, and the factors that either motivate or discourage them from sharing. Several works has been performed on knowledge sharing and incentives, and they highlight the necessity of incorporating intrinsic incentives [1][8][19].

An informal survey of KM users at Oak Ridge National Laboratory and the Office of Naval Research was conducted. Both organizations rely on Microsoft SharePoint [12] as the chosen KM platform, yet adoption by users was extremely low in spite of management pressure and training classes. Questions were open-ended such as "What problems have you had in using Microsoft SharePoint?" and users were encouraged to demonstrate their experiences. Numerous insights became clear. The common theme revolved around "high overhead cost, low return value." Most voiced concerns about the inadequacy of finding relevant and valuable information as well as accessing information remotely through virtual private networks (VPN's) and the reduced performance associated with moving data over the network. Access to and use of existing systems required a great deal of "overhead" such as multiple logins, group access requests, menu navigation, document upload tagging. While such overhead costs originated for some purpose, ultimately it serves only to reduce the value of the KM system to the end user. Consequently, users are driven away from KM tools and into the black markets of "shadow IT" -- the rogue systems, insecure processes, and non-sanctioned means of doing work. Users develop and engage these tools as a path of least resistance and greatest return value.

Furthermore, when people share information, the organization learns about who is sharing, what information is growing, where trends are, and what significant terms are surfacing over time. Without having good data in the system, the organization is reliant on self-described subject matter experts and no means of evaluating high versus low performing contributors.

As a result, what is needed is the creation of an environment that pushes information to users that they find valuable and can access easily with minimal overhead costs. For the information to be valuable, it must be both current and relevant. Users' incentives to contribute good information are based on the value

they receive from discovering other data sources and people who are sharing similar items. The more they contribute, the more they discover.

Such a knowledge marketplace is an ideal environment whereby knowledge becomes the currency and each stakeholder can increase their wealth by contributing their particular knowledge to the KM system. For the price of sharing his or her knowledge, each stakeholder receives either additional knowledge they did not already have, or the organizational recognition of being "the" expert of his or her subject domain. In other words, stakeholders "get what they pay for." If stakeholders do not share, then they do not receive. If stakeholders do not share current or relevant knowledge, then they do not receive current or relevant knowledge. In addition, stakeholders have the ability to become wealthier in the amount of knowledge if they continue to share knowledge over time. This encourages stakeholders to evolve and grow in their knowledge in order to maintain their competitive advantage. Such a knowledge marketplace contains the necessary mechanisms to leverage the innate behaviors such as self-preservation and self-advancement of the stakeholder.

In order to create this environment, the challenge becomes the creation of a system that provides the appropriate value (i.e., relevant and personal value to the individual knowledge worker) for the knowledge that is shared. The marketplace could then be regulated and monitored by the organization. Our approach to creating this knowledge marketplace is described in the next section.

3. Prototype system

The prototype system developed at ORNL utilizes Microsoft SharePoint as its basis [12]. Many organizations have adopted this platform as their knowledge management system, and have used it for a variety of use cases. Unfortunately, one of the main uses cases is simply as a file sharing system. One of the main drawbacks of the system is the lack of dynamic content that can be pushed to the user that is also relevant to the user. To address this, an automated analytic engine was developed that interfaces with SharePoint via the WebDAV protocol, and utilizes text analysis and machine learning techniques to provide discovery of SharePoint content.

There were three primary technical implications taken from the requirements and surveys of the users. First, search the enterprise for recently uploaded and relevant documents in the organization defined by the users. Rather than requiring users to follow a process

of moving documents to a specific repository, users are allowed to place documents where it makes sense for them, but must provide awareness of the contribution to other staff. Second, render results to users in a way that requires the least amount of data transfer over the network. For this issue, we used extensible markup language (XML) to transfer results from the data ingest engine to the browser. Using the XML web part as the target space solved two concerns. First, it allows an easy configuration that only consumes a few kilobytes of data for the exchange. This allows a quick transfer of updates so new information can be pushed to users just like really simple syndication (RSS) feeds. Second, since Microsoft provided extensions for SharePoint sites to service mobile devices, users can access the results of this prototype through their existing, corporate smart phones. No additional authentication is required. The final technical implication is that the prototype system must provide both structured and unstructured ways of viewing the results. Since users are not searching for specific information, results need to provide both familiar and provoking ways of viewing.

The first implication is addressed through the use of unsupervised text analysis and machine learning techniques, and will be discussed in more detail in the following sections. One of the major drawbacks of most KM systems is the lack or poorly implemented search capabilities. Furthermore, requiring users to rely on a multitude of keyword searches only inhibits the knowledge sharing and usage. What is needed is the ability to search for “interesting” information without explicitly defining what “interesting” means to each user. In addition, each user is allowed to organize their data in a way that is meaningful to them. Therefore, the search capability must transcend the structure imposed by each user, and simply identify relevant information based on content alone. To accomplish this challenge, the prototype system employs both unsupervised and semi-supervised approaches to text analysis. In using an unsupervised approach, the prototype system can transcend any structure applied by the user as well as avoid the lack of or inaccurately applied metadata. It allows relevant, valuable results to be delivered rapidly, which speeds adoption by the users. In addition, this approach is also particular useful when the KM system is not well populated with documents. Over time, however, a much more refined and possibly more powerful approach is through the use of semi-supervised machine learning where the system can learn key characteristics about the knowledge being shared and leverage that information to enhance the searching of relevant, valuable knowledge. For the initial system, the semi-supervised machine learning is used to assist

users in auto-tagging documents according to knowledge already stored in the KM system. This reduces the overhead cost of providing metadata by the users, and also speeds user adoption of the KM system.

The second and third implications are addressed by the user interface. As will be described, a very lightweight, easy to use interface was developed that conveys the results of the text analysis in a meaningful way.

3.1. Unsupervised text analysis

Our approach is to categorize documents according to a profile that is defined by the knowledge worker. First, the raw text is converted into a collection of terms and associated weights using the vector space model method. The vector space model (VSM) is a recognized approach to document content representation [16] in which the text in a document is characterized as a collection (vector) of unique terms/phrases and their corresponding normalized significance.

Developing a VSM is a multi-step process. The first step in the VSM process is to create a list of unique terms and phrases. This involves parsing the text and analyzing each term/phrase individually for uniqueness. The weight associated with each unique term/phrase is the degree of significance that the term or phrase has, relative to the other terms/phrases. For example, if the term “plan” is common across all or most documents, it will have a low significance, or weight value. Conversely, if “strategic” is a fairly unique term across the set of documents, it will have a higher weight value. The VSM for any document is the combination of the unique term/phrase and its associated weight as defined by a term weighting scheme.

In our approach, the term frequency-inverse corpus frequency (TF-ICF) is used as the term weighting scheme [15]. Over the last three decades, numerous term weighting schemes have been proposed and compared [3][6][9][10][17][18]. The primary advantage of using TF-ICF is the ability to process documents in $O(N)$ time rather than $O(N^2)$ like many term weighting schemes, while also maintaining a high level of accuracy. For convenience, the TF-ICF equation is provided here:

$$w_{ij} = \log(1 + f_{ij}) \times \log\left(\frac{N+1}{n_j+1}\right) \quad (1)$$

In this equation, f_{ij} represents the frequency of occurrence of a term j in document i . The variable N

represents the total number of documents in the static corpus of documents, and n_j represents the number of documents in which term j occurs in that static corpus. For a given frequency f_{ij} , the weight, w_{ij} , increases as the value of n decreases, and vice versa. Terms with a very high weight will have a high frequency f_{ij} , and a low value of n .

For the prototype system described here, a corpus of 258,231 documents from the LATIMES and FBIS data collections was used for the ICF table. In the ICF table, we store N , which is the total number of documents in the corpus. Also, for each unique term j , after removing the stop words and applying Porter's Stemming Algorithm [14], we store n_j , which is the number documents in the corpus where term j occurred one or more times. As a result, the task of generating a weighted document vector for a document in a dynamic data stream is as simple as one table lookup. The computational complexity of processing N documents is therefore, $O(N)$.

Once a vector representation is created for each document, similarity comparisons can be made. In our approach, a cosine similarity is used to compare two vectors A and B, as shown in (2).

$$\text{Similarity} = (\mathbf{A} \cdot \mathbf{B}) / (\|\mathbf{A}\| \|\mathbf{B}\|) \quad .2$$

Similarity values ranges between 0 and 1, inclusive. A value of 1 means that vectors A and B are identical; while a value of 0 means that they are not alike at all.

Using the cosine similarity measure, categorization is then performed. As described previously, the prototype system interfaces with Microsoft SharePoint via the WebDAV protocol, which allows data stored in SharePoint to be viewed as if it were a network drive to the client machine. Based on a configuration file, the system first processes a user's "Shared Documents" folder on their SharePoint site. Each document in this folder is transformed into a vector using TF-ICF. Next, these vectors are then added together using vector addition, and the resultant vector is then normalized. This vector is then used as a profile vector that represents this user's knowledge and interests. Next, the system then creates document vectors for each document that is stored in SharePoint that the user has the permissions to see. For each one of these documents, its corresponding vector is compared to the profile vector according to the cosine similarity measure. A record of the M most similar documents is stored. Once the M most similar documents are found, the terms with the highest weights from each document are then extracted. After processing, the prototype system then provides the result (most similar documents and their corresponding top weighted

terms) to the user via web parts configured on their personal SharePoint site.

3.2. Semi-supervised machine learning

Often, many organizations insist that knowledge workers tag or provide metadata about the knowledge or documents being shared via the KM system. Some organizations may even have a pre-defined set of required tags to choose in order for the organization to have consistency of tags across the organization. This is usually done at the time when the document is uploaded to the KM system. While beneficial to the organization, this only adds to the overhead costs associated with knowledge sharing, which generally reduces adoption and usage by users. This is particularly true when there is a complex structure into which the information is required to fit. A lack of organization or metadata can be dealt with in different ways, including the previously described unsupervised approach, but requiring such structure can facilitate much more refined knowledge discovery techniques across a knowledge base. However, given the impediment this can cause in knowledge sharing and the high error rates of automated methods for fine-grained information extraction, a usable solution requires careful planning.

Even simple structures can be beneficial to human navigation of the data. Categorized folders may be set up for sharing of information related to specific subjects. A few simple tags may be required at upload time to improve future searches or automated knowledge sharing. In these cases, it is possible to take advantage of this structure to greatly facilitate future additions and avoid the impediments that tagging requirements can present. In the current system, it is possible to learn a model for deciding relevance to a specific category of data defined by the user or users of a system. This may be something as simple as a set of documents aggregated into a shared folder or a more specific set defined by metadata for the purpose of structured learning. Once a dichotomy can be made between information that belongs in the same group and information that does not, it is possible to provide highly accurate auto-tagging for coarse categorizations, such as those that typically result from human categorization. In other words, since there are multiple ways to divide the same content, we use examples of the desired division whenever available. It is sometimes possible to infer the correct divisions based on existing user folders, etc., depending on the specific deployment and use cases. In order to support tagging, the prototype system employs a semi-supervised machine learning [13] technique in an attempt to generalize based on the encapsulated

knowledge to make decisions on new data. This captured knowledge can then be used to provide human-correctable auto-tagging suggestions.

Even when we have user-defined tags in a system, it is still common that a large majority of the potentially relevant information will not have assigned tags. In other words, we often have labeled data and unlabeled data, with the latter typically being the much larger set. We therefore employ a semi-supervised [4] approach to learning classifications of this information. The large number of examples (documents) of unknown classification can be used to find important structure that makes learning easier. For example, many problems have an innate dimensionality that is much lower than that of the original representation. In simple document classification, the original representation might include a feature for every word in the corpus, while the similarities most relevant to the classification can often be better encapsulated by a non-linear combination of a small subset of these terms. There are many graph-based approaches to semi-supervised learning that attempt to capture this innate dimensionality based mostly on unlabeled data [11]. For example, a graph that connects documents can be used to represent a manifold; an alternate representation that encapsulates the same knowledge but has much lower dimensionality.

In this system, the semi-supervised approach [4] that is employed relies on the graph Laplacian from spectral graph theory [5]. In this case, a graph is constructed that joins together documents that are similar to each other. The graph can be constructed using a nearest neighbor or similar approach based on proximity in the original feature space. We use essentially the same preprocessing as we performed in the unsupervised approach described above in order to obtain the original high dimensional vector representation. This graph represents a manifold. The Graph Laplacian is a matrix defined as follows:

$$L(u,v) = \begin{cases} 1, & \text{if } u=v \text{ and } d_v \neq 0 \\ -\frac{1}{\sqrt{d_u d_v}}, & \text{if } u \text{ and } v \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases}$$

where d is the degree (number of incident edges) of a vertex, and adjacency refers to a neighboring connection in the graph.

This matrix has some very interesting properties. The one of interest here is the fact that eigenvectors associated with the smallest non-zero eigenvalues of the matrix are smoothest with respect to the original graph. So, these eigenvectors can be used as a map into a low-dimensional space that is smooth with

respect to the connections in your graph. This approach has been successfully applied to text categorization (among other applications) [2] in order to increase the effectiveness of a small number of labeled examples. The approach is particularly useful when the feature space can be made less noisy because we understand the problem and the original features are processed based on this knowledge. Thus, text preprocessing that includes the use of stop word removal, stemming, etc. allows the technique to be more effective.

The biggest motivation for the dimensionality reduction is the hope of representing the same information in fewer variables, which in turn makes it easier to account for the role of each of these dimensions in making the categorization we care about. Ease of learning in this sense equates to high-accuracy classification models built with a relatively small number of labeled examples. In other words, the presumed presence of a low innate dimensionality is used to combat the “curse of dimensionality” [7].

3.3. User interface & performance

In order to leverage the power of the unsupervised and semi-supervised techniques, the system needed a simple to use, intuitive user interface. Thus, several XML-based web parts were developed. Utilizing the WebDAV interface, the prototype system writes out the results of the text analysis to files stored on SharePoint. Each user’s personal SharePoint site is then configured with web parts that are capable of reading each user’s corresponding output files.

There are 2 types of web parts. The first web part is styled to facilitate structured information; that is, it provides the column structures of document title, author, location (directory), or other data that describes how the author catalogued the information. Each file name is a hyperlink directly to the document, so that the user may open a document of interest. A second web part describes the unstructured data. Based on the tag cloud concept, significant terms extracted via the text mining of each document are given a hyperlink that leads the user to the list of files where that word or phrase was most significant. Stated simply, within two steps, users can see the most current information shared with the organization and a subset of documents based on their specific terms of interest. This happens without the user entering a single search term.

Furthermore, there are 2 sections of web parts with each section consisting of 1 web part that shows a list of files, and 1 web part that shows the top words or phrases from those files, as described previously. In the first section, only the files most recently added to SharePoint are shown, along with their corresponding

most significant words and phrases. This section enables users to see the new material being added, and provide an encouragement to those users who are not yet sharing their knowledge on SharePoint. Examples of the web parts from this section are shown in Figures 1 and 2.

While You Were Out (Docs)		
Title	Community	Time stamp
FY12 Budget Highlights.pdf		Mon Feb 14 17:59:13 GMT 2011
CSIIR_Feb 2011.ppt	CSIIR	Mon Feb 14 17:34:25 GMT 2011
CSE Division Portfolio 1st page.ppbx	CSE	Mon Feb 14 17:32:52 GMT 2011
DSSE website - updated.pptx	Shared Documents	Mon Feb 14 16:29:22 GMT 2011
DSSE Group Meeting 2-9-2011.pptx	Shared Documents	Wed Feb 09 13:07:49 GMT 2011

Figure 1. Results of prototype showing most recent documents

While You Were Out (Top Words)				
Terms				
rd,	cyber,	real-time,	scalable,	poc:

Figure 2. Results of prototype showing most significant words from most recent documents

In the second section, only those files that are most similar to the user's profile of "Shared Documents" are shown, along with their corresponding most significant words and phrases. This is where the user receives the value of having shared their knowledge via documents uploaded to SharePoint. Examples of the web parts from this section are shown in Figures 3 and 4.

While You Were Out (Personal)		
Title	Community	Time stamp
KD V2.3.pptx	Shared Documents	Wed Jun 30 21:07:32 BST 2010
KD V2.3.pptx	Shared Documents	Wed Jun 30 21:07:32 BST 2010
AdvMethodsForCyberEventCorrAndAssetVal-DASHNetORNL-January2009.doc	Documents	Wed Jan 14 15:19:30 GMT 2009
AdvMethodsForCyberEventCorrAndAssetVal-DASHNetORNL-January2009.doc	Documents	Wed Jan 14 15:19:30 GMT 2009
SentAniMultLangWebForums.pdf	Shared Documents	Tue Nov 11 15:07:22 GMT 2008

Figure 3. Results of prototype showing most similar documents to user

While You Were Out (Personal Top Words)				
Terms				
cyber,	pattonrm@ornl.gov,	agent-based,	exfiltrating,	exfiltration

Figure 4. Results of prototype showing most significant words of the most similar documents to user

In Figure 1, the system is showing the most recently uploaded documents to the system that are

accessible by the user. While not necessarily relevant to every user, it is, at a minimum, making the user aware of the dynamic nature of the KM system and activity level. For example, each year, ORNL participates in the Supercomputing conference. Consequently, in early November of each year, the KM system shows that supercomputing related documents are the most actively shared in the system. A long-term goal of this prototype system is to provide categorical trending information that would highlight activities and changes in the knowledge of the organization.

In Figure 3, the system is showing the most recently upload documents that are also the most similar to the user's profile as defined by the documents that the user has shared. In this case, the user's profile was predominantly defined by knowledge discovery and cyber security related documents. Unfortunately, a particular weakness of our current approach is the need for a similarity measure that has a temporal decay component. As shown in Figure 3, the most recent document is more than six months old. This is potentially not useful or relevant. The challenge to be addressed by future work is to identify whether the knowledge contained within that document is expired or is still current. Such information could be incorporated into the similarity calculation to the user's profile.

Next, Figures 2 and 4 show the corresponding significant terms of the documents shown in Figures 1 and 3, respectively. This selection is based on the TF-IDF term weighting scheme. Each term is a hyperlink to a list of documents in which that significant term occurred. While Figures 1 and 3 show a document level view of the KM system, Figures 2 and 4 begin to show content level view of the KM system. Ultimately, this is really what users find valuable, as they can begin to learn and discover additional terms that are relevant to the documents that they have shared. Future work will explore the possibility of creating RSS feeds for these terms such that user's can subscribe to content where these terms are significant, regardless of where they are stored in the KM system.

Finally, the prototype system has been tested on a machine with dual quad-core processors running at 3.0 GHz with a Raid-0 file system for the data. A data corpus of 4.1 million text documents was stored on the Raid-0 file system. The prototype system was able to categorize the entire corpus in 2 hours and 40 minutes while using 550 MB of RAM. Obviously, this is an optimal case, but demonstrates the system's ability to quickly categorize a large number of documents. In practice, the network latency will affect performance.

4. Example use cases

In a large organization, there are numerous use cases as a result of the variety of knowledge workers involved. This section will highlight some of our own experience with using the prototype system in a few different use cases.

The first use case involved a group leader of a research group at Oak Ridge National Laboratory. One of the goals of this group leader is to hire college graduates that are an ideal fit for the group. The Human Resources (HR) group at the lab routinely scouts university campuses for applicants and collects resumes. After collecting over 3,000 resumes, the HR group created a SharePoint site for viewing all the resumes. Resumes were organized according to the university. No additional tagging or organization was provided such as organizing resumes by major. An email was sent to management notifying them of the availability of the new resumes via the SharePoint site. From the HR perspective of this use case, any additional tagging or organization would have been an increased cost of knowledge sharing with very little direct value in return, representing diminishing returns. From the group leader's perspective, the availability of 3,000 resumes organized by university represents a daunting task of finding the right candidate. Consequently, the value of knowledge sharing was non-existent while the cost of improving that value was exceptionally high.

The prototype system provided the value that is missing from this use case. The group leader had already shared a number of documents on his personal SharePoint site. These documents consisted of various presentations, reports, and publications that pertained to the work performed by his group. The system simply used these documents to create a representative profile of the group leader and then used this profile to identify the 10 most similar resumes from the list of 3,000 resumes. Within a few minutes, the group leader was able to review the results of the system, and was pleasantly surprised to see that, based on his opinion and experience, there were at least 3 people that he was willing to contact immediately for an initial interview. In addition, the group leader also realized that the documents he had shared were outdated, and newer material needed to be shared via his personal SharePoint site in order to increase the value of the results he received. He was now willing to share additional knowledge on a more frequent basis. The group leader received the benefit of finding good job candidates simply by sharing the documents he was already producing in his workflow.

The second use case involved two researchers at the ORNL and the Office of Naval Research. For both researchers, their primary responsibility focuses on expanding their expertise in their respective fields. In

order to accomplish this, it is essential to have "research domain awareness" about their particular field and tangential fields, in terms of the people and work that is being performed. Without the prototype system in place, there was little intrinsic motivation to using the SharePoint system. The researchers needed to rely on keyword searches to find information, and the value returned was often very low.

Again, the prototype system provided the value that is missing from this use case. With the prototype system in place, both researchers shared publications that either they had authored, or publications that they felt were relevant and significant to their own research or field of study. The prototype system then used these documents to create a representative profile, and used this profile to find the most similar documents that were also being shared via SharePoint. With little overhead costs or modification to their respective workflows, the researchers now receive the value of discovering other people and research that is similar to their own research within their own organizations. As a result, the researchers were now intrinsically motivated to do several things. First, they were now motivated to use SharePoint and the prototype system on a more frequent basis in order to review the newest results and remain current with the new knowledge being added to the system. Second, they were now motivated to keep the publications that they were sharing more current. Finally, if they are not doing so already, the researchers now have the option of collaborating with other researchers whom they did not previously have visibility or awareness of their work. As in the first use case, the users receive the benefit of discovery additional information simply by sharing the documents they were already producing in their workflow.

5. Future work

The work described here shows that, much like an open market, knowledge sharing can be eagerly and effectively adopted when the value received exceeds the cost of sharing through the use of an interface that facilitates the exchange. Furthermore, it also demonstrates that intrinsic incentives outweigh extrinsic incentives. People are naturally driven by greed, self-preservation, and self-advancement. Knowledge management systems must effectively leverage those qualities to facilitate the exchange of knowledge.

While the current prototype provides significant capability in its current form, several opportunities for future work and to improve the prototype still remain. One such area is delivering that value at scale. Value at scale is the ability to provide meaningful value to a

multiple of people. Since corporations and organizations have a multitude of functional groups, divisions, and staff, providing each person or group the ability to tailor results that suit their needs proves difficult. Having the ability to mine documents across an enterprise, sometimes rendering multiple views of the same documents, and return valuable, fresh information requires constantly refreshing and evaluating the data.

Another area of future research is addressing the ability to evaluate knowledge for “freshness”. As an organization changes over time, so does the corresponding knowledge. In some cases, the knowledge does not change and is always relevant. In other cases, the knowledge is only fresh and relevant for a specific period to a specific group of people. After this period of time, the knowledge is expired and its value diminishes significantly.

The next area of future research is to incorporate the social networking aspect of KM. As described previously, the use case involving the researchers revealed that while discovering new knowledge was significant, what would particular more valuable is associated specific people to specific knowledge. Such a capability would allow people to organized not by a job title or description, but by the type and amount of their knowledge.

Finally, future work will also explore the trending of knowledge within an organization. For business intelligence purposes, a KM system would provide exceptional value given the ability to observe what knowledge has been acquire, is currently being acquired, and provide suggestions as to what knowledge should be acquired in the future. In addition, the KM system should provide this capability at every scale within an organization.

6. Acknowledgements

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285; managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR2225. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

7. References

- [1] Amin, A.; Hassan, M.F.; Ariffin, M.B.; Rehman, M.; , "Theoretical Framework of the Effect of Extrinsic Rewards on Individual's Attitude Towards Knowledge Sharing and the Role of Intrinsic Attributes," Computer Technology and Development, 2009. ICCTD '09. International Conference on , vol.2, no., pp.240-243, 13-15 Nov. 2009
- [2] Belkin, M. and P. Niyogi, *Semi-Supervised Learning on Riemannian Manifolds*. Machine Learning, 2004. 56: p. 209-239.
- [3] Buckley, C., Singhal, A., and Mitra, M. New retrieval approaches using SMART. In *Proc. of the 4th Text Retrieval conference (TREC-4)*, Gaithersburg, 1996.
- [4] Chapelle, O., B. Scholkopf, and A. Zien, eds. *Semi-Supervised Learning*. 2006, MIT Press: Cambridge, MA.
- [5] Chung, F.R.K., *Spectral Graph Theory*. 1997, Providence, RI: American Mathematical Society.
- [6] Hung Chim, and Xiaotie Deng; , "Efficient Phrase-Based Document Similarity for Clustering," *Knowledge and Data Engineering, IEEE Transactions on* , vol.20, no.9, pp.1217-1229, Sept. 2008
- [7] Donoho, D.L., *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century*. 2000.
- [8] An Fengjie; Qiao Fei; Chen Xin; , "Knowledge sharing and Web-based knowledge-sharing platform," *E-Commerce Technology for Dynamic E-Business, 2004. IEEE International Conference on* , vol., no., pp.278-281, 15-15 Sept. 2004
- [9] Jones, K.S. and Willett, P. *Readings in Information Retrieval*, Chap. 3. Morgan Kaufmann Publishers, San Francisco, CA, 305-312, 1997.
- [10] Man Lan; Sam-Yuan Sung; Hwee-Boon Low; Chew-Lim Tan; , "A comparative study on term weighting schemes for text categorization," *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on* , vol.1, no., pp. 546- 551 vol. 1, 31 July-4 Aug. 2005
- [11] Lin, T. and H. Zha, *Riemannian Manifold Learning*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008. **30**(5): p. 796-809.
- [12] Microsoft SharePoint, <http://sharepoint.microsoft.com>, current February 2011.
- [13] Mitchell, T.M., *Machine Learning*. 1997, Singapore: McGraw-Hill.
- [14] Porter, M.F. An algorithm for suffix stripping. *Program*, 14(3), 130-137, 1980.

- [15] J.W. Reed, Y. Jiao, T.E. Potok, B.A. Klump, M.T. Elmore, and A.R. Hurson, "TF-ICF: A new term weighting scheme for clustering dynamic data streams," In Proc. of the 5th International Conference on Machine Learning and Applications (ICMLA'06), pp.258-263, 2006.
- [16] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, nr. 11, pages 613–620.
- [17] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Journal of Information Processing and management*, 24(5): 513-523, 1988.
- [18] Hongzhi Xu, and Chunping Li, "A Novel Term Weighting Scheme for Automated Text Categorization," *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on* , vol., no., pp.759-764, 20-24 Oct. 2007
- [19] Heng-Li Yang; Wu, T.C.T.; , "Knowledge sharing in an organization - Share or not?," *Computing & Informatics, 2006. ICOCI '06. International Conference on* , vol., no., pp.1-7, 6-8 June 2006