ORNL/TM- 2012

Knowledge Discovery, Knowledge Management and Enterprise-Wide Information Technology Tools Final Report

March 2012

Prepared by
Robert Patton, Chris Symons, Bryan L. Gorman, and Jim Treadwell
Computational Data Analytics Group
Computational Sciences and Engineering Division
Oak Ridge National Laboratory

ORNL/TM-2012

Computational Sciences and Engineering Division

Knowledge Discovery, Knowledge Management and Enterprise-Wide Information Technology Tools
Final Report

Robert Patton
Chris Symons
Bryan L. Gorman
Jim Treadwell

March 2012

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831-6283
managed by
UT-BATTELLE, LLC
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

Knowledge Discovery, Knowledge Management and Enterprise-Wide Information Technology Tools Final Report

Background

The U.S. Office of Naval Research Global (ONR Global) is a strategic command organization that provides science and technology (S&T) solutions for the Navy and Marines Corps. With almost 100 scientists, technologists and engineers worldwide, the command is the interface between the global S&T community and the operational fleet and forces of the Navy and Marine Corps.

The Computational Data Analytics (CDA) research group at the Oak Ridge National Laboratory (ORNL) conducts innovative basic and applied computer science research on challenges of national interest. The research focus is in the areas of intelligent agents, emergent behavior, pervasive computing, machine learning, information retrieval, and knowledge discovery.

Objectives

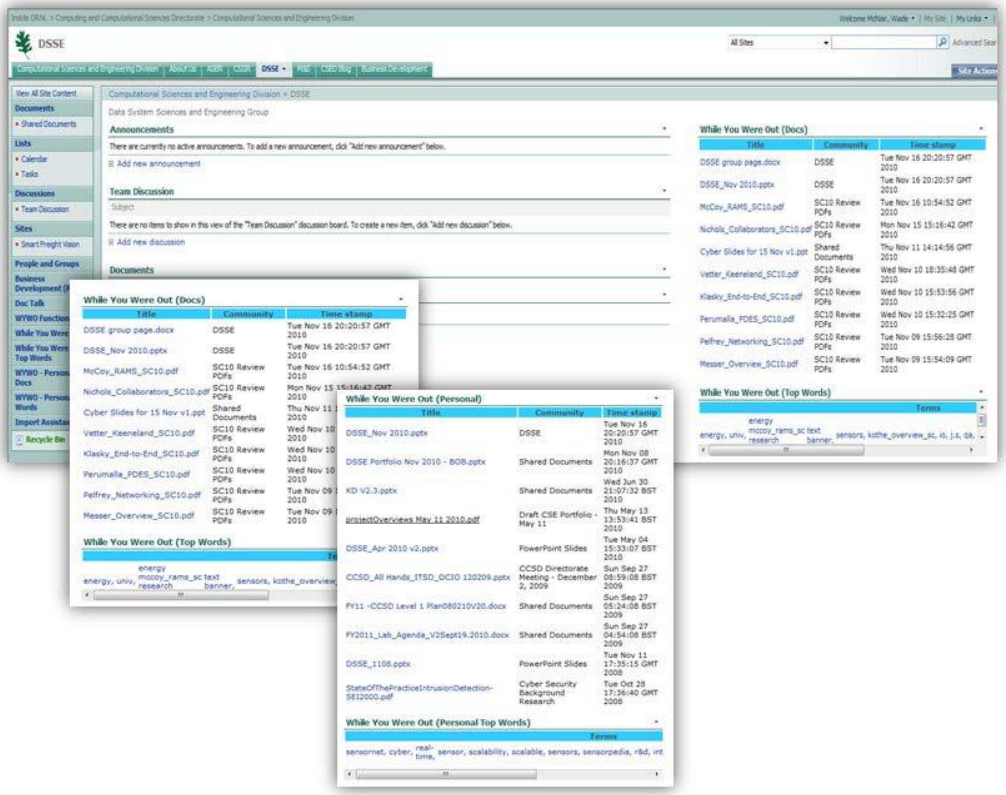
In 2010, ONR Global requested assistance from ORNL's CDA group to develop a knowledge discovery and management utility for its S&T documentation. Although ONR Global has access to and use of ONR's enterprise-wide SharePoint file system, there is no mechanism in the enterprise that allows ONR Global researchers to "connect the dots" between their research interests and the publications, notes, documentation, and interests of their colleagues in different regions of the global enterprise. The search mechanisms in SharePoint and commercial enterprise search engines do not provide the "push" function that ONR Global requires to keep its knowledge workers up-to-date and informed of developments and findings within the enterprise knowledge base.

To meet its requirement, ONR Global tasked CDA with the following three tasks:

> *Task 1 -- Deliver a knowledge discovery and management tool that retrieves information from existing ONRG systems and documents. This tool serves as the foundation of a KD/M platform for ONRG staff and leaders.*
>
> *Task 2 -- Extend the KD/M platform to receive input from additional ONRG users, devices, and data sources. This includes improving the ability for ONRG staff to find information relevant to their jobs to reduce duplication of effort and costs.*
>
> *Task 3 -- Evaluate the effectiveness of the platform and make improvements in the user interface, visualization, and dissemination of information reporting. The emphasis will be on simplifying the use of the system by ONRG staff.*

Technical Approach

In Phase 1, ORNL's initial approach to the CDA tasks was to integrate the text-analysis algorithms of Piranha with SharePoint, Microsoft's content management and document management platform. The resulting prototype, which was successfully integrated into ORNL's enterprise SharePoint platform, is Raptor (see Figure 1).



ORNL's Raptor (Fig. 1)

In Phase 1, ORNL sought to provide ONR Global the demonstration of a prototype application that would match knowledge workers' domain interests with the corpus of documentation on the ORNL SharePoint site. By seeding Piranha's text mining features with key terms from selected ORNL knowledge workers' profiles, Raptor successfully recommended documents with significantly higher incidences of the text targets. These documents of interest, which were extracted over a scheduled run of the software on ORNL's SharePoint site, were then presented as a "hit list" to the corresponding staff member via an add-on Web Part on the ORNL SharePoint portal. It was demonstrated in Phase 1 that the hit list that Raptor recommended included information that the staff members would not know about across the entire enterprise portal

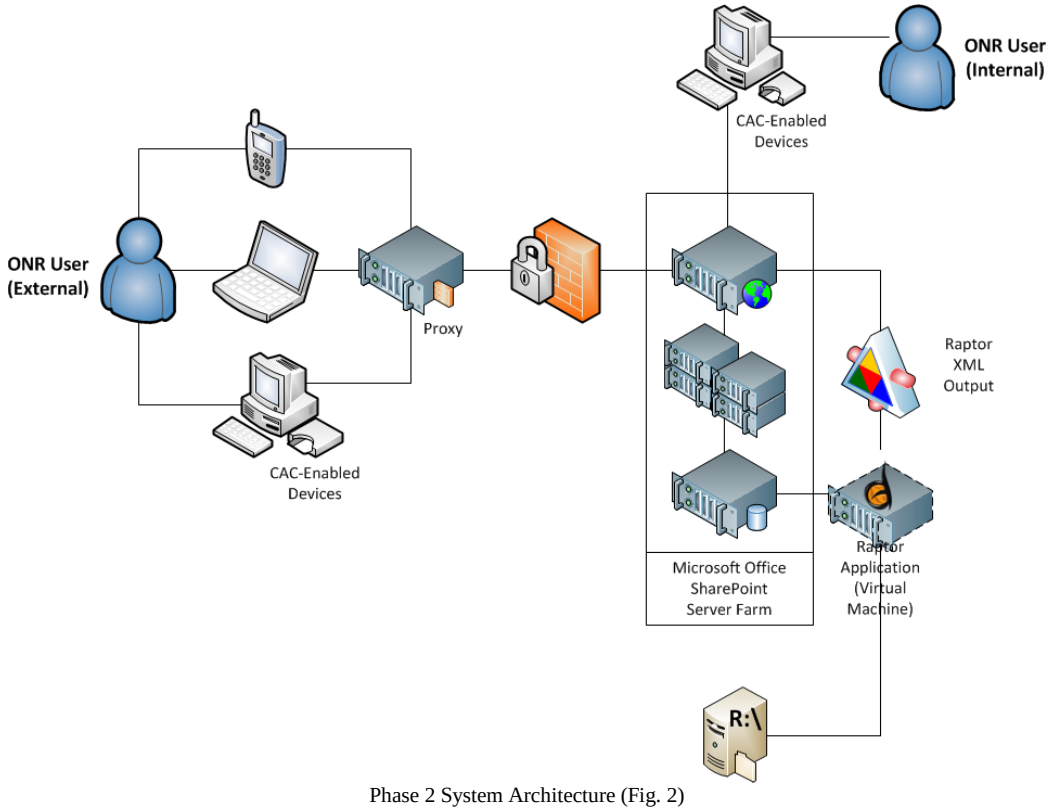
In Phase 1, a successful demonstration of the Raptor prototype was provided using ORNL data.

In Phase 2, an effort was undertaken for a pilot implementation of Raptor at ONR.

Phase 2 requirements included:

1. A demonstration of Raptor using actual ONR data
2. Demonstration of Raptor to the ONR Chief of Naval Research
3. Award of interim authority to test (IATT) in order to conduct user tests and to prepare for eventual full DIACAP certification of the software
4. Review findings

The proposed Phase 2 system architecture for ONR is provided in Figure 2.



Phase 2 System Architecture (Fig. 2)

Results

ORNL successfully acquired and ran Raptor against 749.5GB of ONR data (~300k documents).

The results unexpectedly did not triage the document set to a number or select type that was useful (i.e., the results were often >10k docs).

Although the system worked, it was not found to discover relevant information that was otherwise unknown to the staff members (i.e., documents which were authored by the staff members were often returned on the hit list and unreadable indexes of document and program

titles were also returned.) As well, the time required to process all the documents that were provided for the test case far exceeded expectations (i.e., greater than 10 hours).

Based on folders associated to specific users (e.g., Bolia, Thorne), ORNL discovered that the ONR document libraries often include broad collections of terms or documents (e.g., "summary documents") that contain lists of keywords. Such files are typically not included on the ORNL enterprise SharePoint portal.

To improve the results, ORNL experimented with options for tuning the system to achieve a more useful hit list. Among these options was employing a supervised machine learning strategy to tag documents to improve the text mining algorithms. ORNL also focused on reducing the amount of time required to run each full index. To improve the overall performance of the system, ORNL implemented an Apache Solr (open source search platform & full text index) pre-processor. These measures were implemented on a test set of data and did indicate performance improvements and a reduction in the size of the resulting hit list; however, the tests were never deployed on the entire ONR test set nor reviewed by ONR subject matter experts to determine usefulness.

Recommendations

Recommendation 1: The first and most basic problem with the pilot Raptor implementation for ONR is that much of the documentation that ends up on a reader's hit list should be excluded from the processing beforehand. While a document such as a directory listing or a file index may have value for listing publications, much like a telephone directory, it should not be considered as appropriate content for the general reading requirements of knowledge workers. A "noise filter" to pre-determine if a document has useful content before it is submitted to text processing is a possible option. Unfortunately, there is not likely an off-the-shelf application that can do such filtering for an enterprise, and if there were, it would certainly require extensive customization because the definition of noise would be different from organization to organization. A supervised learning algorithm to tag a document into recommended classes might be possible, but this will also require development.

Recommendation 2: The second problem with a Raptor solution is that many documents under a single title may actually cover numerous unrelated topics. It is not unusual for some survey or proceedings documents to cover in excess of 20 different subjects. Without some means of isolating specific topics within a single document, in the event that a knowledge worker is looking for a specific topic, returning the entire document will be unhelpful. Discourse analysis or some extension of topic modeling to divide larger documents into coherent pieces for analysis of content/relevance is recommended.

Recommendation 3: Third, a user feedback system based on discriminative learning that allows users to refine how a Raptor capability works for them would be helpful. However, such a capability without the first two recommendations will not be that useful. Also, ONR Global has a unique challenge. Such a feedback system would need to accommodate the complexity of an enterprise with knowledge workers that have domain interests that vary by content and by context across the enterprise. For example, while a specific technology may

not be relevant to some knowledge workers unless it involves a specific country or region, it may have content that is significant to their domain interests. Likewise, a contextual interest specific to a region but outside the domain interest of a knowledge worker may still have significant interest. The nature of the problem is not linear, and a learning tool will have to accommodate multi-dimensional complexity.

Recommendation 4: As observed by Malcolm Gladwell in a recent article on social networks, "there is strength in weak ties." Gladwell goes on to posit that "our acquaintances—not our friends—are our greatest source of new ideas and information. [Social networks] exploit the power of these kinds of distant connections with marvelous efficiency. It's terrific at the diffusion of innovation [and] interdisciplinary collaboration." Ultimately, any push system for information sharing will require knowledge worker engagement to exploit the strength in weak ties. One of the more promising trends in enterprise information technology is the augmentation of enterprise content management systems (e.g., SharePoint) with enterprise social network (ESN) software. ESN software replicates the link sharing, tagging, and signaling features of Facebook and Twitter to allow knowledge workers to easily share information across the enterprise. There are COTS solutions (e.g., Socialcast and SocialText) that are available. While an ESN alone will not meet ONR Global's document sharing requirements, taken together with the other recommendations above, it offers an additional tool for achieving that goal.

Summary of Costs

Under provisions of Interagency Agreement 866-V251-10 between the Department of Energy and ORNL Global, ORNL received $400,000 of incremental funding on May 7, 2010, for Phase 1 of the Raptor project. On April 13, 2011, ONR Global provided ORNL an additional $350,000 of incremental funding. ORNL is presently working with ONR Global on the de-obligation of all unspent funds remaining as of the project end date (March 31, 2012).

Conclusions

Although the overall result of ORNL's research was not completely successful in meeting ONR Global's information and knowledge sharing requirements for the ONR data, the work over the past year and a half did provide valuable insights into the nature and extent of the problem space. Clearly, as evidenced by emerging products like Huddle Synch, there is a growing realization that information and knowledge within enterprises need to be "pushed" as well as searched, and the research collaboration between ONR Global and ORNL was timely and entirely appropriate. A document recommendation platform that can provide the critical and useful information that might not otherwise find its way to knowledge workers will have enormous value in the future. Since ONR Global's interests are not unique, it is increasingly likely that commercial solutions, either on the shelf on their way to the market, will help to address certain aspects of the enterprise's knowledge sharing requirements. Although Raptor alone is not the solution to ONR Global's knowledge sharing requirements, ORNL continues to believe that its technology can be a key component of such a solution.

COTS solutions alone will not provide the complete solution to the more complex and unique problems of document sharing and awareness within ONR Global. A combination of promising technologies and deliberate strategies will certainly enhance knowledge push within the enterprise.