

Final Report: Machine Learning for Facilitating Analysis at Regional Fusion Centers

SERRI Project: Regional Data Analysis

**Christopher Symons, Brian Klump,
Rowan Chakoumakos, and Derek Rose**

This material is based upon work supported by the U.S. Department of Homeland Security under U.S. Department of Energy Interagency Agreement 43WT10301. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

SERRI Project: Regional Data Analysis

**Final Report: Machine Learning for Facilitating Analysis at
Regional Fusion Centers**

Christopher Symons, Brian Klump, Rowan Chakoumakos, and Derek Rose
Oak Ridge National Laboratory

Date Published:

September 2009

Prepared for
U.S. Department of Homeland Security
under U.S. Department of Energy Interagency Agreement 43WT10301

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

ACKNOWLEDGEMENTS

The authors would like to express their great appreciation for the assistance and support of the Tennessee Fusion Center. In particular, we would like to thank David Hawtin, Steve Hewitt, and Oxana Munson. Without their crucial input, this project could not have been completed. In addition, we would like to thank Frank DeNap of ORNL for his invaluable support throughout the project.

CONTENTS

ACRONYMS vii

SOUTHEAST REGION RESEARCH INITIATIVE ix

1. Introduction 1

 1.1 Background 1

 1.2 Objective 2

2. Regional Analytics 2

 2.1 Identifying and Accessing Relevant Information..... 2

 2.2 Data Overload 2

 2.3 Scarcity of Trained Analysts..... 3

 2.4 A Changing Stream of Analytical Questions 3

 2.5 Integration of New Analysis Tools..... 3

3. Machine Learning for Regional Analysis 4

 3.1 Problem Formulation..... 4

 3.2 Unsupervised Learning..... 5

 3.3 Supervised Learning..... 6

 3.4 Adaptive Learning Methodologies..... 6

 3.4.1 Semi-Supervised Learning 6

 3.4.2 Active Learning 8

 3.4.3 Task-Transfer Learning 8

 3.4.4 Reinforcement Learning 9

 3.5 Privacy Preserving Data Access..... 9

4. Realistic Application Scenarios and Recommendations 9

 4.1 Moving Beyond a Clustering Engine 9

 4.2 Develop a Common Interface into Known Sources 10

 4.3 A Browser Plug-In Designed to Learn Relevance 10

 4.4 Algorithms that Learn Quickly 10

5. Conclusion 11

6. References 11

ACRONYMS

ASO	Alternating Structure Optimization
DHS	Department of Homeland Security
ORNL	Oak Ridge National Laboratory
SERRI	Southeast Region Research Initiative
SVM	Support Vector Machine
TBI	Tennessee Bureau of Investigation
TFC	Tennessee Fusion Center

SOUTHEAST REGION RESEARCH INITIATIVE

In 2006, the U.S. Department of Homeland Security commissioned UT-Battelle at the Oak Ridge National Laboratory (ORNL) to establish and manage a program to develop regional systems and solutions to address homeland security issues that can have national implications. The project, called the Southeast Region Research Initiative (SERRI), is intended to combine science and technology with validated operational approaches to address regionally unique requirements and suggest regional solutions with potential national implications. As a principal activity, SERRI will sponsor university research directed toward important homeland security problems of regional and national interest.

SERRI's regional approach capitalizes on the inherent power resident in the southeastern United States. The project partners, ORNL, the Y-12 National Security Complex, the Savannah River National Laboratory, and a host of regional research universities and industrial partners, are all tightly linked to the full spectrum of regional and national research universities and organizations, thus providing a gateway to cutting-edge science and technology unmatched by any other homeland security organization.

As part of its mission, SERRI supports technology transfer and implementation of innovations based upon SERRI-sponsored research to ensure research results are transitioned to useful products and services available to homeland security responders and practitioners.

For more information on SERRI, go to the SERRI Web site: www.serri.org.

1. Introduction

This report is a summary of findings in a study of data access- and overload-related problems facing regional analysts. The project's focus on regional analysis revealed areas where a one-size-fits-all solution is not capable of addressing data analysis needs that are either too region or mission specific. As it turns out, many of these same requirements are likely not being met for non-regional analysts due to the fact that the specific information of interest at any given time is a moving target that varies with respect to each analytical question. The report concludes with general recommendations and a potential path forward for addressing some of the most pressing needs in the near term. It presents several solvable issues of importance to regional analysts and presents methods that already exist within the machine-learning research community that have the potential to greatly enhance analysts' capabilities.

1.1 Background

The Southeast Region Research Initiative (SERRI) Program has worked to support various types of data analysis and fusion at the regional level. Much of this work has been designed to support Department of Homeland Security (DHS) interests in facilitating the anticipation and prevention of terrorist activities by supporting information sharing.

One of the objectives of the SERRI program has been to support the functions of many of the information fusion centers in the region. Through its involvement with several such regional centers, ORNL has become familiar with many of the obstacles preventing them from fulfilling all aspects of their mission, particularly with regard to the analysis of terrorism and other threats that require broad bases of information, including national and international, to recognize.

The Tennessee Fusion Center is one such regional analysis center that was formed jointly by the Tennessee Bureau of Investigation and the Tennessee Office of Homeland Security. It is meant to serve as a clearinghouse for all information that might be relevant to law enforcement in the region. A key challenge for analysts at such centers is the inability to acquire all relevant data despite its potential availability. However, their task of regional information support can also be overwhelmed if they must sift through too much national, international, or cross-region information. Of particular concern is the inability to recognize obvious threats, connections, trends, etc. due to either a lack of information or due to information overload.

In particular, due to the fact that regional analysis is often more targeted, adding in additional relevant data without introducing data overload is a serious issue. It typically falls to analysts to identify specific subject areas or target pieces of information that are important. At that point, even when more global information is available, finding the informational components of interest is often overly difficult.

Automated forms of data analysis have improved in recent years, but they remain at a level where fine grain approaches are too error-prone to be of significant help to analysts. However, intelligent, but coarse-grain data analysis (such as clustering, categorization, coarse natural language processing, etc.) can significantly aid analysts in dealing with large amounts of information, without losing credibility of analysis (due to

the heavy supervision of the analyst). Despite the potential, it is clear that most analysts do not often take advantage of these capabilities. There are often many reasons for this, but in the case of regional analysis one such reason is the inability of analysis tools to be taught or otherwise easily modified in ways that would aid the investigation of very specific analytical targets, subjects, areas, etc.

1.2 Objective

In order to assist in alleviating the aforementioned concerns, this project's main objective was to discover methods that could support the regional analysis of the full range of threats and vulnerabilities that are the mandate of local analysis centers. The focus of our study was on the latest advances in machine-learning, and the goal was to match needs identified through interactions with regional analysts with methods that are within reach. The biggest potential payoffs can be found whenever the potential capability is adequate for filling the need but simply lacks the right problem formulation. The work was performed in the manner of a pilot project in conjunction with the Tennessee Fusion Center with the purpose of working directly with regional analysts in order to allow them to focus the study and keep it relevant.

2. Regional Analytics

Through this project, the ORNL researchers were able to interface directly with regional analysts in order to understand the daily issues that make it difficult for them to find relevant information. The analysts at the Tennessee Fusion Center were extremely accommodating, and their input was used to drive the direction of the study.

2.1 Identifying and Accessing Relevant Information

Regional data analysts are tasked with protecting region-specific infrastructure, tracking regional hate groups, etc. Therefore, they have very specific missions for which their own knowledge and expertise is often the only good resource. At the same time, they are expected to uncover virtually all relevant data from the news, intra- and inter-agency reports, open-source web sites, and other data sources relevant to their latest investigation. In order to anticipate and head-off a threat, the analyst has to have access to the information that specifies or hints at the threat. This data access issue is being addressed by the Fusion Centers, but there are still major issues to overcome, particularly in regard to access to potentially classified data. It is particularly difficult to establish a "need-to-know" when the custodian of the information is not focused on specific regional threats. Establishing a connection with data one never sees is a very difficult problem. A path toward a solution is suggested in this report, but that solution is still a long way off.

2.2 Data Overload

Addressing the data identification and access issues leads to the nearly opposite problem of data overload. Even in the case of region-specific analysis, the amount of

data that might contain relevant information is typically on a scale that makes human sorting of the data impossible. A standard search using terms of relevance can return millions of potential documents. If the information is not found among the first several results, it will almost inevitably be overlooked. The regional analysts have hundreds of potential data sources to which they regularly turn, and each of them can return large amounts of data. The result is a recurring “needle in a haystack” problem that greatly impedes successful regional analysis.

2.3 Scarcity of Trained Analysts

Compounding the issue of data overload is the scarcity of trained analysts. The regional analyst’s mission is very focused and often unique to a single analyst, so while they benefit from information produced by and for other analysts, there is typically a large information gap that must be covered. This specialization is likely necessary, but it becomes particularly problematic when an experienced analyst leaves their assignment.

2.4 A Changing Stream of Analytical Questions

Although there are many tools that can make the analyst’s life easier, some of their most basic needs are not currently being met, despite the clear potential of machine learning to aid the analysts. According to the analysts, sorting through irrelevant information still consumes the large majority of their time. Existing tools that are available or might be made available to them would not solve the problems that they point to as being most critical. We discuss the issue further elsewhere in this report, but the essence of the problem is that a tool trained or designed to find a specific type of information is not flexible enough for the analyst when addressing a constant stream of different analytical questions. Furthermore, while coarse, unsupervised methods of data organization can clearly help, the inability to target specific issues still leaves the analyst with more information than they can possibly pursue, most of which is still likely to be completely irrelevant.

2.5 Integration of New Analysis Tools

One of the major challenges involved in the deployment of any tool designed to facilitate analysis is the practical issue of adoption. Any solution that would be readily adopted by the analyst would be one that works seamlessly with tools they already use and that is simple and intuitive to apply. Since the analyst is already over-burdened, if they have to go to a separate tool, or learn to use a complex one, then the chance of adoption diminishes dramatically, no matter how much potential a tool might have to help.

Particularly substantial questions that remain are the following: How can the amount of data be pruned and pared back in a way that is relevant to the precise problem that the analyst is currently investigating? Is there a way to let the analyst tell the computer what is relevant and let the computer present the data to them? And, if this can be done, can it be done unobtrusively, such that the burden on the analyst is not increased (even in the early adoption phase)?

3. Machine Learning for Regional Analysis

Machine Learning (Mitchell 1997) is a well-developed area of artificial intelligence that focuses on finding computational solutions that can be learned through observations. It is most applicable in situations where the computer scientist does not understand the procedures they would like to emulate in the detail that would be required to define a stable set of steps as in traditional algorithm design. This description is very apt for the analytical process, and therefore, it is not surprising that machine learning is very popular in the design of tools intended to support analysts.

3.1 Problem Formulation

The analytical process is one that contains many steps. While many of these steps can and are being aided by computational tools, other steps are either beyond current computational abilities or else the problem has not been formulated in a way that makes it obvious that computational tools can be of assistance. Most of the work that is being performed by computers is organizational. Some tools are available that focus on the “search” aspect of information gathering in an attempt to find relevant information. As an example, the Alabama Fusion Center uses a tool called *Riverglass*. It is designed to help an organization keep track of relevant material by continually scanning information sources and pushing the most relevant to the user. It even has a learning component. It is certainly a very valuable tool, and it can solve some data issues that the regional analyst has, but it does not address the moving target issue that each analyst faces. In other words, if every case is different, then learning over time means little. The analyst will likely still be forced to scan tremendous amounts of data that is irrelevant to the particular problem of the day. They might not miss as much, and their focus may initially be better, but the data overload problem is so significant that they really need a tool that learns what they are looking for right now and assumes it is different than what they were looking for last time.

In order to apply computational algorithms to the data analysis issues facing regional analysts, it is first important to cast the problems in terms that allow the link to be made. In the case of the regional analysts at the Tennessee Fusion Center, there are really two major areas where computational methods based on machine learning could be applied in the near term in ways that directly address the biggest time drains on the analysts. Because the regional analyst has his or her own area of concern that may not be completely shared by any other analyst at any level, there are not tools available that are organizing the data that pertain precisely to their problem. More than that, each question that arises is specific, and it is not possible to have relevant information gathered and filtered a priori in such cases. Currently, standard keyword-based search tools are the main tool that the analysts rely on to meet these ever-changing needs.

The analytic process that is performed in addressing a target problem can roughly be boiled down to the following: A keyword search is performed across the databases and other information sources that are available; The returned data is manually searched (and in the case of large amounts of data, little of the data is actually covered); The analyst follows any leads and repeats the process using new information. The key assets are the analyst’s background knowledge and domain expertise. The key

obstacles are the inability to uncover the “right” data, either because they don’t have access to it or, more commonly, because they are overwhelmed by redundant or irrelevant information.

In essence, the analyst is performing manual text categorization (classifying documents as being relevant or irrelevant) as a first step in uncovering information of interest. While this may be a moving target, they are essentially performing a repetitive task and doing so in a way that is observable. This problem formulation allows a computer to learn from the analyst and continue on to emulate their behavior after observing it for a period of time. The following issues arise, but all are manageable. First, if we consider the problem to be text categorization, then the two classes are potentially different for every problem the analyst addresses (since relevancy changes depending on the current question). Thus, if we need to learn a classifier for every problem, then a second issue is how many examples need to be “labeled” by the analyst before the machine can learn enough to truly be of assistance. Finally, is there a way to observe the analyst’s categorizing without being obtrusive; i.e. without the analyst having to change or augment their work process in any significant way? If these issues can be addressed, then it is a fairly straightforward process to use a computer to quickly take over for the analyst in sorting the information. Since sorting through irrelevant information is the most time consuming task for the analyst, and since it is clear that a problem formulation exists to address this task computationally, we attempt to lay out connections in greater detail below.

In the following section we present some of the necessary underlying ideas from the field of machine learning, and attempt to describe them in a way that is roughly understandable to a lay person. In the final section, we take these concepts and use them to lay out a recommended course of action in providing answers to the above-mentioned issues.

3.2 Unsupervised Learning

Unsupervised learning is most often simply a form of clustering of the data. A set of features is devised (such as document terms in a text categorization problem), and similarity between two items (e.g. text documents) is evaluated using some form of comparison based on these features.

There are many forms of clustering (Zamir and Etzioni 1998; Jain, Murty et al. 1999), and many possible similarity scores (Lesot, Rifqi et al. 2008). One of the most common methods for assigning similarity in text processing is begun by preprocessing the information via removal of meaningless terms followed by some form of lemmatization; i.e. removing endings such as “s” and “ing.” This is followed by term weighting, which is a process that determines the significance of the remaining words by scoring them such that those that occur in a smaller subset of the documents are weighted more heavily (since they likely indicate distinguishing content more than common words). Then, the terms are scored and turned into a vector so that a cosine similarity score can be calculated. Note that there are many other forms of document comparison, but the above process is used for illustrative purposes since it is quite common.

By grouping similar documents in an unsupervised fashion (a human does not need to look at any of the documents ahead of time), a clustering program can allow an analyst to quickly find information groups that are similar.

3.3 Supervised Learning

Supervised learning (Mitchell 1997; Duda, Hart et al. 2001) is the most common form associated with the field of machine learning. It consists of having a human label information, and then having the computer learn to do the same thing. This simply requires one to define a set of classes (or a scoring mechanism in the case of regression) and a set of observables (typically referred to as *features*). Based on examples classified by a human, the machine then learns the significance of the features in determining the class to which an example belongs. There are too many forms of supervised learning to mention here, but some of the more popular forms are Support Vector Machines (SVMs) (Vapnik 1998), a nonparametric form of learning that seeks to maximize the margin between classes, and Graphical Models (Bishop 2006), a form of learning that seeks to constrain the search space for a solution by using links between random variables to assert independence assumptions among them.

If an analyst knows the categories (or groups) into which they want information sorted, then providing labels and using supervised learning will target the grouping of interest, unlike in the case of unsupervised learning.

3.4 Adaptive Learning Methodologies

While supervised and unsupervised learning represent the foundations of machine learning, the present and the future belongs to adaptive learning methods, such as semi-supervised learning, active learning, multi-task learning, transfer learning, and reinforcement learning. These and similar subfields have arisen due to a need to adapt machine learning to real problem areas. It is these areas that potentially offer the right fit to the real problems addressing the analysts mentioned above.

3.4.1 Semi-Supervised Learning

Semi-supervised learning (Chapelle, Scholkopf et al. 2006) is a more real-world-oriented learning platform that attempts to benefit from the advantages of both supervised and unsupervised learning. In essence, it typically attempts to find structure (e.g. in the feature space) in the unlabeled data that can be used to make learning using the labels more effective.

Since the training (labeled) data in supervised learning is used to judge the relevance of the observations (features), if there are too many features compared to the number of labeled examples, learning a good classifier is very difficult. The principle of dimensionality reduction is one in which the goal is to find a smaller set of features on which to learn to make judgments. The objective is to find and remove redundant and irrelevant features while perhaps finding more meaningful combinations of other features. Some of the more successful forms of semi-supervised learning are designed to use the unlabeled data to perform dimensionality reduction.

One such approach is manifold learning. A manifold is a reduced-dimension space in which classification is much easier. It often represents a space that allows you to easily separate the classes among which you are trying to make a distinction. Some of the more theoretically sound versions of manifold discovery are based on the Graph Laplacian from spectral graph theory (Chung 1997). For example, assume you construct a graph G (a set of vertices (examples) and the links between them) in the original feature space that joins examples together. If the graph varies smoothly with respect to your problem (examples from different classes are rarely linked, similar examples from the same class are linked, etc.), then you can use it to represent a manifold. The Laplacian of the graph can then be used to find a space that roughly represents that 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, if you construct the graph so that it truly represents a manifold, then the eigenvectors associated with the smallest eigenvalues can be used as a map into a space that is smooth with respect to the connections in your graph (Belkin and Niyogi 2003). This approach has been successfully applied to text categorization (among other applications) (Belkin and Niyogi 2004) in order to increase the effectiveness of a small number of labeled examples.

This shows promise in terms of making use of a small number of labeled examples from the analyst. We considered a two-category text-classification problem designed to emulate the classification of analyst data into relevant and irrelevant in order to test the hypothesis that we could learn quickly enough to effectively help the analyst in sorting data even very early in the data-sorting process. We used two somewhat similar categories from the 20 Newsgroup dataset, which is a publicly available resource routinely used to test text-categorization algorithms, and tested to see how well we could learn after only ten examples had been labeled. The binary problem we chose was to distinguish between the documents from the Hockey and Baseball newsgroups. These document sets should be similar enough that the categorization is nontrivial, while being different enough to allow a clear, meaningful division. As indicated in Table 1, it does seem to be possible to learn such tasks quickly enough to benefit the analyst greatly in any similar data bifurcation they attempt. The manifold learning algorithm was one based on the work described in (Belkin and Niyogi 2004), but the algorithm was modified to allow the few labels to guide the construction of the graph in order to make it smoother with respect to the target problem. The results shown are the average of experiments run on 3 random partitions of the data into a labeled set (of size 10) and an unlabeled set (of size 1983). The data was preprocessed

using the text processing mentioned in the section on unsupervised learning to construct the original feature space.

Table 1: Binary Text Categorization Error Rates After 10 Labels

	SVM (supervised)	Improved Laplacian Manifold (semi-supervised)
Average Error Rate (Using 10 Labeled Examples)	46.97 %	5.26 %

Manifold learning is only one form of semi-supervised learning, but other forms have also been previously shown to be very effective in improving text categorization and other tasks. One of the most significant is an algorithm described as Alternating Structure Optimization (ASO) (Ando and Zhang 2005). ASO is a process by which auxiliary tasks are used to learn useful predictive structures over the unlabeled data in order to allow learning from the labeled examples to be more effective. Auxiliary tasks are ones that are potentially related to the target task, but they are designed such that labels are always available. For example, if the target task is text categorization, then an auxiliary task might be to predict the most common term in half of the document, given the set of features from other half of the document. In short, there are multiple ways to construct successful semi-supervised learning algorithms, and there is great promise in using them to learn quickly (with little supervision).

3.4.2 Active Learning

Active learning (sometimes referred to as selective sampling) (Cohn, Atlas et al. 1994; Freund, Seung et al. 1997) describes any methodology by which the machine *actively* chooses the examples that are to be labeled. By controlling the order in which examples are presented to the human, it is possible to speed up the learning process. A typical active learning algorithm will look at an unlabeled example and judge the usefulness of seeing it labeled based on things like *representativeness* (i.e. does this example represent a large portion of the data or is it an outlier?), *informativeness* (e.g. am I uncertain about its label, and will seeing it potentially change my classifier dramatically?), and *diversity* (i.e. is this example different from ones for which I have already seen the labels).

Carefully choosing the order in which information is presented to the analyst can therefore have a dramatic effect on the learning rate of any classifier. It is not obvious how best to integrate active learning with a semi-supervised learning tool designed to continue a sorting procedure for an analyst, but at the very least, it is simple to ensure that an initial batch of information is not completely redundant, and doing so should improve the initial learning rate.

3.4.3 Task-Transfer Learning

Task-transfer learning (Taylor, Fern et al. 2008) is also particularly relevant to the support of regional analytics. This type of learning is intended to make use of common factors learned in previous tasks that could make learning easier on a new task. There

are certainly common factors that an analyst looks for in assessing potential threats, and transfer learning is one way of formulating the goal of capturing and applying these commonalities. In this project, we looked at this type of learning but concluded that it was a longer and more difficult road to apply it usefully to analyst support. In addition, the research in this area is not as mature as in some of the other areas mentioned, and there is significant research that would need to be performed to solve some of the issues preventing its useful application.

3.4.4 Reinforcement Learning

Reinforcement learning (Sutton and Barto 1998) is the type of learning most commonly associated with robotics. It is based on the concept of using positive or negative feedback to learn an optimal set of actions in a given scenario. This type of learning is certainly relevant to learning to perform actions that an analyst might take, but the design of a reward system is difficult without introducing an obtrusive element into the learning process. Furthermore, the most successful reinforcement algorithms rely on the set of possible states being reasonably small, and this presents another issue in formulating problems for reinforcement learning from the analytical process.

3.5 Privacy Preserving Data Access

One last research area that we wish to note is that of privacy preserving data sharing (Shaneck, Kim et al. 2006). Although our exploration of this area was limited, it does appear that there are algorithms (not yet scalable enough) that could potentially be used to develop automatic methods for establishing a need-to-know without compromising private or classified data. This would be a major effort, but it is one that might be considered in the future as a way to support regional analysts without comprising secure data.

4. Realistic Application Scenarios and Recommendations

As a result of this study, our major recommendation is to put into place a system that will learn to quickly judge relevance to the current question under investigation by the regional analyst. In the preceding, we have mentioned several issues that could make this difficult, but we have hinted at solutions as well. In this section, we will lay out potential solutions that are clear and fairly straightforward to implement. By doing so, we hope to outline a path by which the analysts can realistically be supported in a way that does not disturb their current workflow, but that significantly increases their potential effectiveness. The general idea and method was developed in response to needs expressed by the analysts at the Tennessee Fusion Center, and a great desire for exactly this capability was expressed.

4.1 Moving Beyond a Clustering Engine

At the very least, a clustering engine could be put into place to help the analyst organize and sort data. However, we present another suggestion that is just as

straightforward, but which has a much greater potential for dramatically assisting the analysts. The main recommendation we outline is designed to take advantage of previous research in adaptive learning and use it to help an analyst to sift through large amounts of data to find information that has a high probability of being relevant to the specific problem that is currently being investigated. We suggest solutions to all of the issues related to this that were raised in preceding sections.

4.2 Develop a Common Interface into Known Data Sources

The first step would be to put in place a common interface into the data sources of interest to the analyst. This would allow them to run one search across all (or at least most) sources. At that point the computer can run their search across multiple sources and gather the results. If the number of results is minimal, they can simply be presented to the analyst. However, when the data is overwhelming, the system can begin a learning and sorting process specific to the current investigation.

4.3 A Browser Plug-In Designed to Learn Relevance

The analyst is already looking at pieces of information returned from searches and judging relevance and irrelevance to their current investigation simply by glancing at or reading through the data. At that point an unobtrusive interface is needed through which the analyst can communicate this relevancy to the machine. Through this project, we were able to construct a simple test using a web-browser plug-in to control data presented in the browser and allow an analyst to simply click a button marked relevant or irrelevant to continue to the next piece of data. This plug-in approach would allow an analyst to use their normal process, it would not put any additional burden on the analyst for the training, and it has the potential to be implemented much more cost effectively than the construction of a separate tool. Fortunately, relevancy is somewhat fuzzy in most cases, and any errors in parsing out information from the browser should not overly handicap the learning process. The analysts at the Tennessee Fusion Center expressed support for such an approach as a means of putting a tool in place that would not be difficult to adopt. Thus, learning can take place unobtrusively.

4.4 Algorithms that Learn Quickly

Another issue mentioned was one of being able to learn quickly enough to effect the data-sorting process early enough to be effective. Our previously discussed experiments give us strong reason to believe that a tool can quickly begin reordering information returned from a search to allow the analyst to save tremendous amounts of time, while helping them view more of the potentially relevant information hidden among the search results. In addition, after every click, the learning should improve, and the results can be continuously reordered to encourage optimal use of the analyst's time. One final benefit is that such a system could easily be designed to take care of recording most data provenance information, which is a task that normally falls to the analyst to perform mostly manually.

5. Conclusion

The study conducted under this project revealed that there are several needs of regional analysts that are not being met. Some of these have solutions or potential solutions available in the form of machine learning algorithms. It is our recommendation that at the very least an approach is designed to assist the analyst is dealing with the data overload problem. One such approach is described in detail in this report.

6. References

- Ando, R. K. and T. Zhang (2005). "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data." Journal of Machine Learning Research **6**: 1817-1853.
- Belkin, M. and P. Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation." Neural Computation **15**(6): 1373-1396.
- Belkin, M. and P. Niyogi (2004). "Semi-Supervised Learning on Riemannian Manifolds." Machine Learning **56**: 209-239.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, Springer.
- Chapelle, O., B. Scholkopf, et al., Eds. (2006). Semi-Supervised Learning. Cambridge, MA, MIT Press.
- Chung, F. R. K. (1997). Spectral Graph Theory. Providence, RI, American Mathematical Society.
- Cohn, D., L. Atlas, et al. (1994). "Improving Generalization with Active Learning." Machine Learning **15**(2): 201-221.
- Duda, R. O., P. E. Hart, et al. (2001). Pattern Classification, Wiley-Interscience.
- Freund, Y., H. S. Seung, et al. (1997). "Selective Sampling Using the Query by Committee Algorithm." Machine Learning **28**: 133-168.
- Jain, A. K., M. N. Murty, et al. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**(3).
- Lesot, M.-J., M. Rifqi, et al. (2008). "Similarity measures for binary and numerical data: a survey." International Journal of Knowledge Engineering and Soft Data Paradigms **1**(1): 63-84.
- Mitchell, T. M. (1997). Machine Learning. Singapore, McGraw-Hill.
- Shaneck, M., Y. Kim, et al. (2006). Privacy Preserving Nearest Neighbor Search. ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops.
- Sutton, R. S. and A. G. Barto (1998). Reinforcement Learning: An Introduction. Cambridge, MA, MIT Press.
- Taylor, M. E., A. Fern, et al. (2008). AAI workshop on "Transfer learning for complex tasks".
- Vapnik, V. N. (1998). Statistical Learning Theory, Wiley-Interscience.
- Zamir, O. and O. Etzioni (1998). Web document clustering: a feasibility demonstration. SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.



Southeast Region Research Initiative

National Security Directorate

P.O. Box 6242

Oak Ridge National Laboratory

Oak Ridge, TN 37831-6252

www.serri.org



**Homeland
Security**