An Approach to the Automated Determination of Host Information Value

Justin M. Beaver
Applied Software Engineering Research Group
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
beaverjm@ornl.gov

Robert M. Patton
Applied Software Engineering Research Group
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
pattonrm@ornl.gov

***Abstract*—Enterprise networks are comprised of thousands of interconnected computer hosts, each of which is capable of creating, removing, and exchanging data according to the needs of their users. Thus, the distribution of high-value, sensitive, and proprietary information across enterprise networks is poorly managed and understood. A significant technology gap in information security is the inability to automatically quantify the value of the information contained on each host in a network. Such insight would allow an enterprise to scale its defenses, react intelligently to an intrusion, manage its configuration audits, and understand the leak potential in the event that a host is compromised. This paper outlines a novel approach to the automated determination of the value of the information contained on a host computer. It involves the classification of each text document on the host machine using the frequency of the document's terms and phrases. A host information value is computed using an enterprise-defined weighting schema and applying it to a host's document distribution. The method is adaptable to specific organizational information needs, requires manual intervention only during schema creation, and is repeatable and consistent regardless of changes in information on the host machines.**

Keywords-host information value; information asset profiling; security informatics; cyber security risk management; document classification

I. INTRODUCTION

Modern organizations rely on electronically stored data for most aspects of their operations. The accessibility of the data is typically controlled through a system of credentials, authorizations, and file permissions. However, once the files are accessible and/or available on a local user account, ownership and control becomes less certain. Users have the ability to change permissions, redistribute files, and allow access to the downloaded data in a manner that may be inconsistent with the original intent, or any established security policies. Furthermore, users may edit and borrow text from sensitive documents to create new documents that are still inherently sensitive, yet may no longer be subject to access restrictions. While such user actions are more often for convenience than explicit malicious acts, the result is that most organizations have very little visibility into 1) the inventory of text data on their networked computers, 2) the criticality of the text data, or 3) the distribution and accessibility of the most critical data across their network. At best, the current state of the practice is the creation of broad data protection zones on the organization's network such that a specific user's computer would be placed into a "sensitive data" zone if that user is expected to work with sensitive data. However, there is no additional visibility, mapping, or validation of the actual data into these protection zones.

Developing a scalable computer network defense requires knowledge of the perceived high-value targets in the system, which are typically identified as those computational assets whose role it is to manage large and/or sensitive volumes of data (e.g., file shares and mail servers). However, apart from a computer's role, determining high-value targets becomes more difficult. Does the computer of a staff engineer contain valuable information? Typically, the answer varies with the person, their projects, and their position on those projects. Unfortunately, the fluid nature of staff members, projects, and roles in an organization makes it challenging to determine the value of information on an employee's computer based on these criteria. Current approaches to determining host information value use models and toolsets that are based on interviews of data owners to evaluate the impact of the data asset [10]. While these methods are valuable, they can be expensive and time consuming to collect and maintain the data necessary to compute a reliable value. In addition, the data owners themselves may not have an accurate assessment of the data for which they are responsible or for its value to the organization. Given the flux of information on a given host, manual approaches seem impractical in an operational setting. Thus, we focus on an automated means to determine the value of a computer asset, based on its contained data.

In addition to data management, understanding the value of the information on each computer host can also potentially provide guidance on a course of action when an attack is detected. Knowledge of whether a host holds critical or sensitive data can drive an organization to respond faster, more appropriately, and more accurately. A host machine containing documents detailing peripheral company projects warrants a different course of action when targeted in an attack than a host machine containing strategic information for the organization. Without a means to value the information that a host contains, the appropriate level of response is difficult to ascertain. In addition, in the event that a host is infiltrated, the information

This work was sponsored by the Lockheed Martin Corporation, through a U.S. Department of Energy Work For Others (WFO) contract.

value provides insight into the data that has potentially been compromised and can better quantify the impact.

The contributions of this paper are an approach to automated text data discovery on a network as the basis for scoring the value of the information contained on a host computer. Our approach leverages methods of raw text analysis to classify individual documents and then applies various scoring algorithms to each host's document distribution arrive at an information value score. An evaluation of the defined approach is explored by applying the approach to a set of faux host document corpuses.

II. Related Work

Information Asset Profiling (IAP) is a process that focuses on determining the value that a computational asset provides to a computer system [4]. It is an element of the risk management process that enables an organization to assess, mitigate, and evaluate the risk in a system [12]. Understanding an asset's value allows an organization to design and implement appropriate information security protections, and to develop a plan that proactively addresses impact and recovery should the asset be compromised [4].

Standards and processes exist [12][15][10][14] that detail best practices in information security risk management, including IAP. These documents provide guidance from a process perspective, focusing on what data and actions are recommended for sound information risk management, and leaving how to gather that information and perform those actions at the discretion of the implementing organization. The common theme across these standards/processes is that a host's value to an organization must be reliably characterized, as it is the basis for 1) understanding vulnerabilities and threat likelihoods, 2) establishing appropriate access controls, and 3) determining the impact in the event of a loss.

Stevens' IAP process, described originally in [4] and incorporated into the OCTAVE Allegro RM process in [14], gives more in-depth guidance for the specific practices of asset profiling with the goal of establishing a standardized and repeatable approach. Stevens' process is comprised of six steps that provide a consistent framework for documenting, evaluating, and maintaining the value of information assets. It addresses issues such as defining the computational asset, understanding ownership and security requirements, and deriving an appropriate information value. Stevens' process involves creating an Information Asset Profile for each host in the network, which is a collection of metrics that characterize the computational asset [14]. Both creating the IAP, and performing the subsequent steps towards assigning a value to the host, is a manual process of data collection and qualitative analysis resulting in an expert assignment of information value, risk, and impacts.

Manual and qualitative approaches to IAP are prevalent in other current information security risk management work. For example, Fortson [2] describes a process framework for damage assessment and mission impact in cyber defense that includes a step where critical information assets are identified and quantified according to their utility with respect to the organizational mission. In Fortson's approach, the author proposes a worksheet that can serve as an aid in establishing the value of each information asset. Fortson's work was improved by Hellesen [3] who proposed another manual methodology but provided a more standardized approach by computing the value of assets through a weighted sum of factors such as availability, confidentiality, and contextual. Soohoo [5] describes IAP through both the manual process and a decision model for risk management in computer security systems. The model evaluates the cost of different security measures, and attempts to quantitatively identify a baseline sufficiency in the level of security employed.

Addressing the lack of IAP automation, Grimaila, et al., [1] proposed a system for information asset tagging that begins to automate some of the manual processes from earlier works. Specifically, they put forward a system of intelligent agents that maps mission processes to information assets, provides frameworks for applying valuation contexts to information assets, and tracks the change in information value over time based on the mission plans. The application of this technology in the command and control domain is further documented in [13]. This related technology addresses the mission aspect of host value, and is progressive in terms of automating the damage/impact assessment to mission assets. However, it does not address the specific problem of automatically determining the value of information on a host system as the described processes for determining information value are assigned classifications.

Despite this body of valuable work in formulating feasible IAP processes, we view the lack of automation in the implementation those processes as a major barrier to both their efficacy and consistency in an operational setting. While existing human-intensive IAP processes are viable as an approach to initial information security design, they are impractical as a maintenance process that must adapt to changing data, users, roles, and projects. Furthermore, the reliance of expert opinion in the assignment of asset value seems to undermine the goal of a standardized and repeatable approach to asset valuation – a more quantitative and repeatable approach is required. Without a reliable and automated approach, the investment to establish and maintain IAP for an enterprise organization is too great.

We address the technology gaps in IAP by proposing an automated means of assessing the value of the information contained on a host machine. Our approach is to leverage a form of supervised machine learning to classify documents into one of the predetermined information categories that are unique to the organization. The resultant distribution of document classes is the basis for quantifying an individual host's information value. Our intent is for the host information value to be adopted as an element of each host's IAP that is updated regularly so that the score can adapt as the data, users, and projects in an organization change.

III. Methodology

Our approach to automated host information valuation centers on the document corpus stored on a host machine, and the classification of documents as the basis for quantifying the

value of the textual information contained on that host. The process for automated IAP is shown in Figure 1.

In Step 1, an enterprise-specific schema comprised of information categories is developed. This schema represents the different topics that are relevant and appropriate for the organization, and the terms/phrases that describe each topic. These information topics/categories are the basis for the classification of all documents on a host, as is performed in Step 2 (described Section III.B). In Steps 3 and 4, a scoring algorithm is applied to each host's distribution of categorized documents to compute the host information value (described in Section III.C). Thus, the resultant information value for each host machine is derived from its contained textual information.

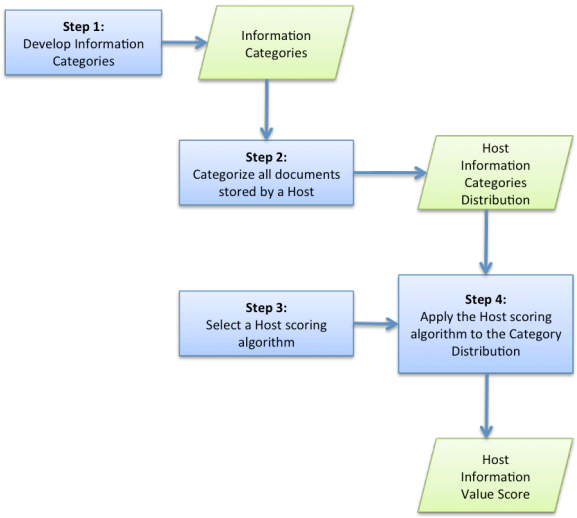

Figure 1. Automated Host Information Value Process

We envision an operational model where the information value for each host is determined automatically and updated regularly. Once the enterprise schema, or set of information categories, is defined by the organization, this method requires no manual intervention and periodic updates can be automated in order to maintain a current information value for each host in the organization's networked environment. Thus, the organization's computer network defense can adapt its protection schemes according the to changing distribution of critical information. Similarly, attack response processes (courses of action) can consider a current host information value should the host be targeted in an attack. The subsections below describe in detail each step in the process shown in Figure 1 in order to more completely describe the underlying methods.

A. *Information Category Development*

The first step in the automated host information value process is the development of the set of information categories that are significant to an organization. An *information category* is simply a collection of terms and/or phrases that characterize a specific topic of information. Information categories may be as generic or specific as necessary to meet the organization's needs. For example, an information category called "Anatomy" might be characterized by specific terms such as "body", "structure", or "morphology". Additionally, the "Anatomy" category might include terms for all of the human body parts and organs. The creation of information categories can be accomplished by either manually constructing the term/phrase lists, or by automatically building term/phrase lists from exemplar documents. So, continuing the "Anatomy" example, an organization may choose to build the category by using an anatomy textbook as an exemplar document.

The size and scope of information categories is completely configurable by an organization. For example, one embodiment may be to create categories based on the sensitivity of information. In this model, categories might include topics such as "Public Domain", "Business-Sensitive", "Sensitive But Unclassified", and "Classified". Exemplars of these types of documents would be used to build the category lists. Other embodiments of categories could focus on business areas, organizational units, product lines, or capabilities. The

B. *Host Document Classification*

Once the information categories are defined, they serve as the basis for the classification of documents in a host corpus. We use a supervised classification technique in which each document is allocated to one of the information categories based on the similarity between the terms/phrases in the document's text and the terms/phrases defined for each information category. The number of documents on a host allocated to each information category is then used to quantify the information value.

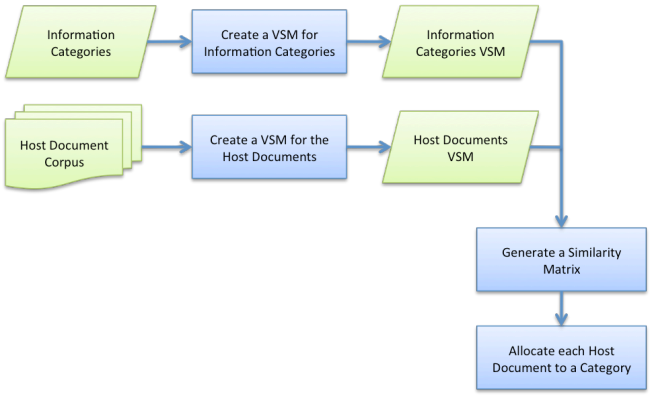

Figure 2. Host Document Classification Process

Figure 2 describes the classification process. The process takes as inputs the set of information categories and the set of host documents that contain machine-readable text. At the core of the classification process is the creation of a Vector Space Model (VSM), a mathematical representation of document contents, for each host document. A VSM is created for each information category and also for each document in the host corpus. A similarity matrix is built which enables the comparison of host documents to the information categories. Each host document is allocated to the information category with which it has the highest calculated similarity. The details for each of these steps are described below.

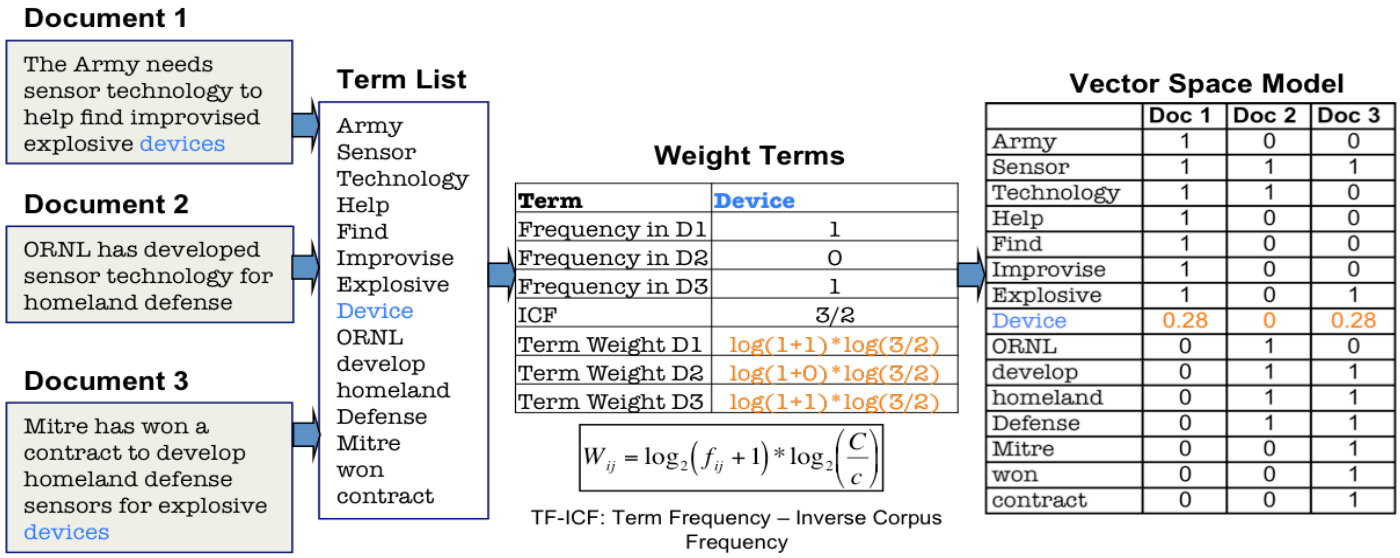

Figure 3. Vector Space Model Creation Process

1) Vector Space Models

The Vector Space Model is a recognized approach to document content representation [7] in which the text in a document is characterized as a collection (vector) of unique terms/phrases and their corresponding normalized significance weight. Developing a VSM is a multi-step process, a simple example of which is shown in Figure 3.

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.

2) Selecting a Weighting Algorithm

The weight associated with each term in a document is an indicator of significance and can be computed using several available term-weighting algorithms. In this application, the primary concern is selecting a weighting algorithm that can be parallelized, due to practical constraints in its operational environment. Several popular term weighting algorithms, including TF-IDF [16] and Okapi [18], are dependent on a static corpus for their weight calculations. That is, these algorithms consider term/phrase frequencies across the entire corpus of documents as the basis for determining the significance of those terms/phrases in individual documents. However, in the application of term-weighting for the automated determination of information value, analysis of the static corpus is not practical. One reason is that the document corpus across an entire enterprise network is not static – it is always changing. A more compelling reason is that it is intractable to analyze the entire corpus of enterprise documents in order to calculate the significance of terms/phrases in individual documents. An approximation of the corpus term weights must be made.

The Term Frequency-Inverse Corpus Frequency (TF-ICF) term-weighting method [6] is very similar to TF-IDF in its mathematical formula, yet uses an independent static corpus as a means to determine term/phrase significance in individual documents. The TF-ICF approach uses a large, independent text corpus and leverages Zipf's Law [17] to produce an accurate approximation of term/phrase significance. TF-ICF was selected as the weighting algorithm for automated host information valuation because it provides a parallelized variant of TF-IDF that does not require analysis of the entire enterprise data set. We do not claim that TF-ICF is the only term-weighting method that can be applied to this problem, but justify our selection of TF-ICF in order to explicitly describe the need to be independent of a static corpus analysis.

3) Document Classification Using Similarity

In this classification process, a VSM is created for each document in the host corpus and also for each enterprise-specific information category. Once created, the process of comparing VSMs to determine similarity requires the use of matrix algebra to develop a similarity matrix as shown in Figure 4.

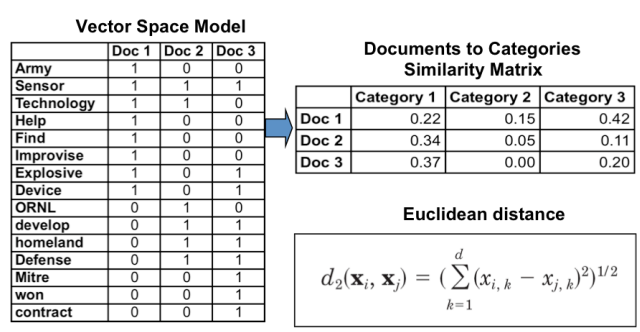

Figure 4. Similarity Matrix Calculation Process

The similarity matrix includes calculated similarity values for all combinations of documents and categories. Similarity values are calculated using the recognized Euclidean distance approach detailed in [8]. Each document from the host corpus is allocated to the information category based on the highest similarity value, and those with no similarity to any categories are binned to a default category called "Unknown".

C. Host Information Value

The *host information value* is a quantification of the significance of the information on a host computer, relative to the information categories defined by the organization. The classification of the host document corpus described in Section III.B results in a distribution of the raw number of documents allocated to each information category. Considering this distribution as the data input, this section explores three different potential scoring methods and discusses how each one meets different organizational needs. These scoring methods are intended to be representative, but by no means exhaustive.

1) Weighted Normalized Scoring

The Weighted Normalized Scoring method takes a user-defined numerical weighting for each category, and applies it to the proportion of the host document corpus that was allocated to each category. The formula for Weighted Normalized Scoring is shown in Equation 1.

$$IV = \sum_{i=1}^{C} \frac{n_i}{N} * w_i \tag{1}$$

where,

IV = the calculated Information Value for the Host,
C = the number of information categories,
n_i = the number of documents from the host corpus allocated to category i,
N = the number of documents on the Host, and
w_i = the organization-defined weight assigned to category i.

The Weighted Normalized Scoring method is an effective means of valuation where the proportion of documents allocated to categories is augmented by the organization-defined weighting scheme. This approach best fits information value use cases where the quantity of classified documents is important. If an organization bases their cyber defense on computational assets dealing with specific business areas, those assets with a larger proportion of documents classified to the highest value business area would be scored higher.

For example, a research organization may have several business areas including computational sciences, biological sciences, chemical sciences, etc. However, supposing that the organization is known for its work in nuclear technologies, and the bulk of the business-sensitive information is contained in this business area, the weighting for this business area would be assigned higher values in the enterprise schema. The operational result is that those host machines with higher proportions of documents that have been classified as "nuclear technology" documents would have higher information values.

2) Weighted Relative Scoring

The Weighted Relative Scoring method takes an organization-defined numerical weighting for each category, and applies it to the proportion of the total documents associated with a category that a specific host contains. The formula for Weighted Relative Scoring is shown in Equation 2.

$$IV = \sum_{i=1}^{C} \frac{n_i}{N_i} * w_i \tag{2}$$

where,

IV = the calculated Information Value for the Host,
C = the number of information categories,
n_i = the number of documents from the host corpus allocated to category i,
N_i = the number of documents across all hosts allocated to category i, and
w_i = the organization-defined weight assigned to category i.

The Weighted Relative Scoring method places higher value on those computational assets that have higher relative proportions of categorical documents, enhanced by the organization-defined weighting scheme. This approach also fits information value use cases where the quantity of classified documents is important, however the quantity is relative to the information category instead of the host. If an organization bases their cyber defense on the distribution of critical data across their network, and scales that defense based on targets with larger percentages of critical information, then the Weighted Relative Scoring is a practical option.

3) Binary Representation Scoring

Where the Weighted Normalized Scoring method focuses on the quantity of categorized documents, the Binary Representation Scoring method focuses on the presence of categories of documents. This approach addresses the use case where the presence of at least one document in a particular category is sufficient to affect a host's information value. An example is the determination of different levels of sensitive documents - the presence of just one "sensitive" document is sufficient to score the information value of the host at the "sensitive" level.

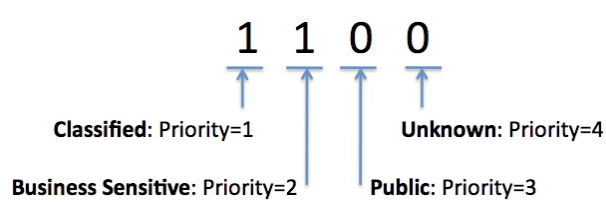

Figure 5. Binary Representation Scoring Method

The Binary Representation Scoring method requires that the operator prioritize categories from highest to lowest. Each category is represented in the information value score as a binary number in a column of the score's value. Consider an example where there are 3 prioritized categories: the "Classified" category is the highest priority, the "Business-Sensitive" category has a medium priority, and the "Public"

category is the lowest priority. As shown in Figure 5, a 4-digit number is used to represent the score, with each category corresponding to a column (the 4th column is typically reserved for the "Unknown" category). The presence of at least one document in a given category will be indicated by a '1' in that column, and the absence of any documents for that category will be indicated by a '0' in that column. So the score shown in Figure 5 indicates that the scanner found at least one document in both the "Classified" and "Business Sensitive" categories, and no documents in the other two categories.

Although the Binary Representation Scoring method provides no insight into the quantity of documents allocated to each category, it provides a very concise mapping of score values to categorical representation. A weakness in quantity-focused scoring methods, such as the Weighted Normalized Scoring method, is that there is some uncertainty in how an information value was derived. That is, a high information value score could be the result of a marginal quantity of highly weighted categories, or it could be the result of a significant proportion of a host's corpus comprised of several mid-range weighted categories. By giving each category a column in the number, the Binary Representation Scoring method clearly communicates the derivation of the information value.

IV. ANALYSIS

In this section, we evaluate and analyze the application of the approach to automated host information valuation presented in Section III. We first evaluate the document classification process to quantify the extent to which host documents are binned as expected. Secondly, we analyze the resultant host information values, and the extent to which they accurately score hosts.

We elected to use the 20 Newsgroups data set [9] as the subject data for our analysis, which is a recognized resource for classification research and is freely available for confirmation of our results. While a collection of news articles is not necessarily representative of typical host data in an enterprise network, it will serve to measure the accuracy of our host scoring process. In our evaluation, we selected each newsgroup to be a unique information category and randomly selected 100 documents (~10%) from each to build the information categories. The evaluation and analysis activities were then performed on the remaining ~90% of the documents from each category to separate the training and test data sets.

A. *Document Classifier Evaluation*

Computing a valid host information value is predicated on the reasonably accurate classification of documents residing on that host. We applied the VSM creation and comparison process described in Section III.B: each document was classified based on the similarity matrix and then the classification was compared to the actual newsgroup to which the document belonged. The accuracy results, measured in terms of the proportion of documents in each news group that were correctly classified, are presented in Table I. We are satisfied with the average accuracy of the 0.69 when using 100 (~10%) exemplar documents. In the automation of host information value, this level of accuracy is sufficient to characterize the distribution of the textual data on a host machine.

TABLE I. DOCUMENT CLASSIFICATION ACCURACY

Newsgroup name	No. Documents	Accuracy
talk.politics.mideast	840	0.66
talk.politics.misc	675	0.55
talk.politics.guns	810	0.80
sci.crypt	891	0.72
sci.electronics	884	0.70
sci.med	890	0.63
sci.space	887	0.76
talk.religion.misc	528	0.48
soc.religion.christian	897	0.88
alt.atheism	699	0.80
comp.graphics	873	0.49
comp.windows.x	888	0.69
comp.sys.mac.hardware	863	0.73
comp.os.ms-windows.misc	885	0.11
comp.sys.ibm.pc.hardware	882	0.65
rec.autos	890	0.81
rec.motorcycles	896	0.83
rec.sport.hockey	899	0.88
rec.sport.baseball	894	0.79
misc.forsale	875	0.76
Average	**842.3**	**0.69**

B. *Host Information Value Analysis*

This section analyzes the three proposed methods for quantifying the information value of hosts. Specifically, we are interested in determining how well the calculated host information value represents the actual underlying text data on a given computer. Theoretically, the host information value should always be representative. However, Section IV.A revealed that our document classification approach is imperfect and therefore it remains to be seen how misclassified documents affect the reliability of the proposed scoring methods.

We simulated the file systems for 20 unique hosts by drawing documents from the "test data" subset of the 20 newsgroups as described above. For each faux host, we used a random number to generator to determine 1) if a newsgroup would be included in the faux host corpus, 2) the number of newsgroup documents to include, and 3) the specific newsgroup documents to include. The resultant set of faux file systems included hosts with as few as 6 and as many as 15 newsgroup categories, with the number of documents in each category ranging from 17 to 878 files. To enhance the

simulation, we applied a weighting scheme to the newsgroups that attempted to reflect priorities for an intelligence gathering organization. The range of weight values was [1, 100] where newsgroups such as talks.politics.mideast were assigned higher weight values and newsgroups such as misc.forsale were assigned lower weight values.

In the evaluation, we applied the document classification process to each faux host corpus, and calculated the host's information score using each of the three methods proposed in Section III.C. In addition, since the 20 Newsgroups corpus affords us ground truth, we compared the calculated information value to the actual information value. The percentage difference in host information score for each of the three scoring methods and for each host is shown in Table II.

TABLE II. HOST INFORMATION VALUE ANALYSIS RESULTS

Host	**Percentage Difference of Classified vs. Actual Host Information Value by Scoring Method**		
	Weighted Normalized	***Weighted Relative***	***Binary Representation***
1	5.51%	52.22%	40.00%
2	6.53%	47.44%	35.00%
3	11.88%	57.21%	45.00%
4	5.32%	59.71%	50.00%
5	5.26%	50.99%	55.00%
6	1.63%	41.67%	65.00%
7	2.41%	52.20%	55.00%
8	6.04%	51.16%	75.00%
9	6.61%	51.30%	50.00%
10	11.56%	52.47%	50.00%
11	5.06%	53.99%	70.00%
12	8.67%	49.35%	45.00%
13	5.74%	55.79%	60.00%
14	8.00%	51.15%	40.00%
15	8.89%	55.74%	45.00%
16	3.91%	59.09%	45.00%
17	12.37%	52.25%	50.00%
18	4.62%	51.53%	55.00%
19	9.09%	51.93%	45.00%
20	10.56%	52.55%	55.00%
Average	**6.98%**	**52.49%**	**51.50%**

As shown, the Weighted Normalized scoring method for host information value outperformed both the Weighted Relative and Binary Representation scoring methods. The average percentage difference between host information values calculated using the Weighted Normalized scoring method and the actual host information values was 6.98%, which we consider to be an excellent result. It demonstrates an ability to reliably characterize a host document distribution. Furthermore, inspecting the results of all hosts reveals a worst-case percentage difference of 12.37% in this experiment. These results suggest that this technology, using the Weighted Normalized scoring method, is viable for operational use. Unfortunately, both the Weighted Relative and Binary Representation scoring methods were found to be less accurate with both being within 50% on average.

We attribute the performance discrepancy between the scoring methods to the manner by which each method normalizes the document distribution. In the case of the Weighted Normalized method, documents are normalized within each host corpus. Thus, the significance of each individual classification (and, more importantly, misclassification) is tempered by the size of the host's document corpus. The Weighted Relative method also normalizes the document distribution, but in terms of the network-wide prevalence of each information category rather than localized to the host. Thus, information categories for which the classification accuracies are low (See Table I) produce information value scores that are widely variant from the actual values, and mask the significance of those information categories with better classification accuracies. An improvement to this scoring method would be to consider the classification accuracy in the weighting.

In the case of the Binary Representation scoring method, we found this approach to be extremely intolerant of the document classification error. A single misclassification can significantly alter host information value, particularly if it is a high priority category. A finding in our analysis is that the calculated Binary Representation score for a majority of hosts was comprised of all '1' values. The average percentage difference of ~50% reflects the random number generator used to create the host corpus more than a scoring accuracy based on classified documents. One extension to this work will be to explore possible improvements in the Binary Representation method that would allow for more tolerance of document classification error.

V. CONCLUSION

This work addresses the gap in information security technology where there is an absence of automated methods to quantify the information value of a host, as part of IAP in an enterprise architecture. Our approach quantifies the information value of a host machine based on the textual data it contains. An organization's knowledge of the high-value hosts in its network enables the appropriate scaling of its computer network defense systems. Additionally, the information value of a host machine is important for determining an effective course of action when an attack is detected, and useful for understanding the data that was compromised in the event of an infiltration.

We proposed an approach by which the information value of a host is calculated from the distribution of classified documents on a host. Document classification was accomplished using a form of supervised machine learning that compares the mathematical representations of both host documents and organization-developed information categories.

The approach to classification was evaluated using a publicly available text document data set, and the Weighted Normalized scoring method was found to produce an accurate host information value despite the misclassification error injected by the document classifier. Our approach to determining host information value is an improvement over existing approaches because it is repeatable, consistent, and is easily automated as part of enterprise operations.

Our intentions for this work going forward include even more focus on the automation of the IAP process. In particular, we are researching methods by which categories may be learned from the documents' contents through applications of topic modeling. We are also expanding our analysis of information value to include more traditional factors, such as the mission of the computational asset and the role of the user, as components in calculating the host information score. Finally, we are exploring domains outside of cyber security that can benefit from this technology, including defense and law enforcement applications.

ACKNOWLEDGMENT

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, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] M.R. Grimaila, R.F. Mills, and L.W. Fortson. "Improving the Cyber Incident Mission Impact Assessment (CIMIA) Process." In *Proceedings of the Cyber Security and Information Intelligence Workshop (CSIIRW 2008),* Oak Ridge, TN, May 2008.

[2] L.W. Fortson. "Towards the Development of a Defensive Cyber Damage and Mission Impact Methodology." Ph.D. thesis for the Graduate School of Engineering and Management. Wright-Patterson Air Force Base, OH, Air Force Institute of Technology.

[3] D.L. Hellesen. "An Analysis of Information Asset Valuation (IAV) Quantification Methodology for Application with Cyber Information Mission Impact Assessment (CIMIA)." Master's thesis for the Graduate School of Engineering and Management. Wright-Patterson Air Force Base, OH, Air Force Institute of Technology.

[4] J.F. Stevens. "Information Asset Profiling." Networked Systems Survivability Program Technical Note CMU/SEI-2005-TN-021. Carnegie Mellon University, June 2005.

[5] K.J. Soohoo. "How much is enough? A risk management approach to computer security." Consortium for Research on Information Security and Policy (CRISP), June 2000.

[6] J.W. Reed, et al. "TF-ICF: A new term weighting scheme for clustering dynamic data streams." In *Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA '06),* pp. 258-263, 2006.

[7] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong. "On Modeling of Information Retrieval Concepts in Vector Spaces." *ACM Transactions on Database Systems,* **12**(2), pp. 299-321, 1987.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data Clustering: a Review." *ACM Computing Surveys,* **31**(3), pp. 264-323, 1999.

[9] K. Lang. 20 newsgroups data set. Available at http://people.csail.mit.edu/jrennie/20Newsgroups/.

[10] Z. Yazar. "A qualitative risk analysis and management tool – CRAMM." Available at http://www.sans.org/reading_room/whitepapers/auditing/a_qualitative_risk_analysis_and_management_tool_cramm_83. SANS Institute, 2002.

[11] D. Wawrzyniak. "Information Security Risk Assessment Model for Risk Management." In *Trust and Privacy in Digital Business,* Springer Berlin/Heidelberg, pp. 21-30, 2006.

[12] G. Stoneburner, A. Goguen, and A. Feringa. "Risk Management Guide for Information Technology Systems: Recommendations of the National Institute of Standards and Technology." NIST Special Publication 800-30, U.S. Department of Commerce, July 2002.

[13] M.R. Grimaila, R.F. Mills, and L.W. Fortson, "An Automated Information Tracking Methodology to Enable Timely Cyber Incident Mission Impact Assessment." In *Proceedings of the 13th International Command and Control Research and Technology Symposium (ICCRTS),* Bellevue, WA, June 2008.

[14] R.A. Caralli, J.F. Stevens, L.R. Young and W.R. Wilson. "Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process." Technical Report CMU/SEI-2007-TR-12, Software Engineering Institute, Carnegie Mellon University. May 2007.

[15] ISO/IEC. "Information technology – security techniques – information security risk management." *ISO/IEC 27005:2008,* 2008.

[16] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval." *Journal of Information Processing and Management,* **24**(5): 513-523, 1988.

[17] G.K. Zipf. *Selective studies and the principle of relative frequency in language.* Harvard University Press. Cambridge, MA, 1932.

[18] K.S. Jones and P. Willett. *Readings in Information Retrieval*. Morgan Kaufman Publishers. San Francisco, CA, 305-312, 1997.