On Off-Topic Access Detection In Information Systems

Nazli Goharian and Ling Ma
Information Retrieval Lab, Illinois Institute of Technology
{goharian@iit.edu, ma@iit.edu}

Abstract
We focus on detecting insider access violations to off-topic documents. Previously, we utilized information retrieval techniques, e.g., clustering and relevance feedback, to warn of potential misuse. For the relevance feedback approach, we minimize the indicative features needed for detection using data mining techniques. We show that the derived reduced feature subset achieves equivalent performance to that of the previously derived full set of features.

Categories and Subject Descriptors
K.6.5 [**Management of Computing and Information Systems**]: Security and Protection - Unauthorized Access

Keywords
Algorithms, Experimentation, Security

1 Introduction
For each authorized user as in [5], a profile that defines their legitimate scope of interest is defined. This profile, which is either assigned or learned over a period of use, remains relatively constant for lengthy periods of time. Regardless of how the profile is determined, misuse is defined as a user querying the information retrieval database for material that is not relevant to their profile.

Misuse detection, particularly as it relates to information retrieval systems, is only of recent interest. In [2, 3, 4], we presented approaches based on information retrieval and classification processing techniques to detect the misuse of information retrieval systems. Our efforts are *content-oriented* detection schemes as they evaluate the document content rather than its system characteristics, e.g., name, size, storage location, usage, etc. Note, we do not advocate using only content-oriented detection approaches; rather, we believe that such content-oriented approaches are complimentary to systems oriented approaches and should be used in conjunction with one another. Our evaluation results, however, strictly evaluate our content-oriented approaches in isolation. Clearly, combining our detection schemes with other schemes is likely to yield a lower rate of false alarm and a higher rate of detection.

CIKM'05, October 31-November 5, 2005, Bremem, Germany.
ACM 1-59593-140-6/05/0010.

Other content-oriented misuse detections exist. Aleman-Meza, et al [1] describe an ontology-based solution to detect an insider threat, namely misuse. In [5], the authors propose the fusing of role based monitoring methods, social network analysis, and semantic content analysis to detect inappropriate information exchange. In both efforts, no evaluation results are presented. Finally, in [6], a natural language processing approach involving entity tagging for detecting misuse is presented. The authors favorably compare their approach against a described "bag of words" solution. The reported accuracy for their approach is similar to ours as reported in [4]; they use a different collection to evaluate their approach.

2 Data Set Generation and Evaluation
As a misuse detection benchmark collection does not exist, using TREC data and queries, thirteen profiles were constructed. Readers are referred to [4] for a detailed description of the data, profile formation, and evaluation approach. In short, we consider five levels of ranking. We used four human evaluators to evaluate all 1300 cases. The distribution of levels is 40.9% for *strong misuse* L_5 or *misuse* L_4, 49.3% for *almost normal use* L_2 or *normal use* L_1, and 9.8% for *undecided* L_3.

Table 1: Contingency Matrix

		Prediction				
		L5	**L4**	**L3**	**L2**	**L1**
Evaluation	**L5**	TA	TA	UM	UM	UM
	L4	TA	TA	TA	UM	UM
	L3	FA	TA	-	-	-
	L2	FA	FA	-	-	-
	L1	FA	FA	-	-	-

We calculated the recall and precision based on the above contingency matrix. Any query that actually is evaluated as misuse (L4 & L5) and is predicted as such is considered a *true alarm* (TA). Similarly, an evaluated *undecided* (L3) query that is predicted as L4 and vice versa is likewise determined as TA. All other predictions of evaluated L4 & L5 queries are defined as *undetected misuse* (UM). Likewise all queries predicted as L4 & L5 that are not covered by the above are considered as *false alarms* (FA). All other cases are not considered, as they are legitimate use and are considered as such; these are marked as "-".

The accuracy of the system is measured based on *recall* and *precision* and is defined as Recall = TA / (TA + UM), and Precision = TA / (TA + FA).

3 RF2-Based Classifier

RF2, as formally defined in [4], considers the absence of query terms in the profile, i.e., Q_A; the presence of the query terms in the query subset of the profile, i.e., Q_q; the presence of query terms in the feedback subset of profile, i.e., $Q_{f\text{-}q}$; the absence of user query feedback terms from profile, i.e., F_A; the presence of user feedback terms in query subset of profile, i.e., F_q; and finally, the presence of user feedback terms in the feedback subset of the profile, i.e., $F_{f\text{-}q}$. We used a tuning parameter to adjust the emphasis of the query terms versus the feedback terms. Represented as a weighted vector, $<Q_A, Q_q, Q_{f\text{-}q}, F_A, F_q, F_{f\text{-}q}>$, the RF2 elements are defined as in Table 2.

Table 2: Feature Value Calculations

Q_A	Q_q	$Q_{f\text{-}q}$
$\frac{\sum_{\forall i \in Q_A} tf\ idf_i}{\sum_{\forall i \in O} tf\ idf_i}$	$\frac{\sum_{\forall i \in Q_q} tf\ idf_i}{\sum_{\forall i \in Q} tf\ idf_i}$	$\frac{\sum_{\forall i \in Q_{f-q}} tf\ idf_i}{\sum_{\forall i \in Q} tf\ idf_i}$
F_A	F_q	$F_{f\text{-}q}$
$\frac{\sum_{\forall i \in F_A} tf\ idf_i}{\sum_{\forall i \in F} tf\ idf_i}$	$\frac{\sum_{\forall i \in F_q} tf\ idf_i}{\sum_{\forall i \in F} tf\ idf_i}$	$\frac{\sum_{\forall i \in F_{f-q}} tf\ idf_i}{\sum_{\forall i \in F} tf\ idf_i}$

4 Classification Results and Evaluation

Using a stratified ten-fold cross-validation technique and an SMO Support Vector classification approach, we experimented with query length and with various relevance feedback term selection configurations. Based on the contingency matrix presented in Table 1, the classification based misuse detector SMO achieves statistically equivalent or better accuracy than RF2, as shown in Table 3. D/D and D/T indicate the system accuracy for the long and short queries, respectively. For example, as shown, for D/D both RF2 and SMO have statistically equivalent recall, with 98.1% and 99.3%, respectively. The precision of SMO is statistically significantly better (82.7%) as compared to RF2 (79.4%), marked with "+". Similar observations are made for D/T.

Table 3: Comparison of SMO-6 and RF2, "+" indicates statistically significant gain at 99% significance level

	D/D		D/T	
	RF2	**SMO-6**	**RF2**	**SMO-6**
Best Recall %	98.1	99.3	97.3	96.9
Best Precision%	79.4	82.7+	75.7	83.0+

5 Feature Minimization

To find the minimum feature set that gives equivalent or better misuse detection accuracy, we generated all size k={5, 4, 3, 2, 1} subsets of the entire set of RF2 6 features. The results indicated that all three *feedback-features* F_A, F_q, and $F_{f\text{-}q}$ are important features, i.e., relevance feedback of the query terms is a good indicator for detection; by their elimination from the feature set, the accuracy is degraded. Moreover, it is observed that F_A is almost equivalent to the feature set $\{F_q, F_{f\text{-}q}\}$, i.e., feature set $\{Q_A, Qq, F_A\}$ is generally equivalent to $\{Q_A, Qq, Fq, F_{f\text{-}q}\}$. Feature set $\{Q_A, Qq, F_A\}$ was shown to be the best and generally better than $\{Q_A, Qq, Fq\}$, or $\{Q_A, Qq, F_{f\text{-}q}\}$. This feature set, $\{Q_A, Qq, F_A\}$, was our ultimate minimized feature set with a statistically equivalent accuracy to the full feature set. All smaller subsets produced worse accuracy. The evaluations were done by comparing the results of the SMO detector based on the minimized feature sets (SMO-3) with the SMO detector with the entire RF2 6 features (SMO-6). The results (table 4) indicate that using less features, SMO-3, achieves a statistically equivalent accuracy to the original detector based on six features, SMO-6.

Table 4: Comparison of SMO-6 and SMO-3, "+" indicates statistically significant performance gain at 99% significance level

	D/D		D/T	
	SMO-6	**SMO-3**	**SMO-6**	**SMO-3**
Best Recall %	99.3	99.3	96.9	97.3
Best Precision%	82.7	83.1	83.0	83.4

6 References

[1] B. Aleman-Meza, et al: An Ontological Approach to the Document Access Problem of Insider Threat. *IEEE Intelligence and Security Informatics (ISI)*, 2005.

[2] R. Cathey, et al: Misuse Detection for Information Retrieval Systems. *ACM Conference on Information and Knowledge Management (CIKM),* 2003.

[3] N. Goharian, et al: Detecting Misuse of Information Retrieval Systems Using Data Mining Techniques. *IEEE Intelligence and Security Informatics (ISI),* 2005.

[4] L. Ma and N. Goharian: Query Length Impact on Misuse Detection in Information Retrieval Systems. *ACM Symposium on Applied Computing (SAC),* 2005.

[5] S. Symonenko, et al: Semantic Analysis for Monitoring Insider Threats. *IEEE Intelligence and Security Informatics (ISI),* 2004.

[6] O. Yilmazel, et al: Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content. *IEEE Intelligence and Security Informatics (ISI),* 2005.