Detecting Misuse of Information Retrieval Systems Using Data Mining Techniques

Nazli Goharian, Ling Ma, Chris Meyers

Information Retrieval Laboratory, Illinois Institute of Technology
goharian@iit.edu

Misuse detection is often based on file permissions. That is, each authorized user can only access certain files. Predetermining the mapping of documents to allowable users, however, is highly difficult in large document collections. Initially [1], we utilized information retrieval techniques to warn of potential misuse. Here, we describe some data mining extensions used in our detection approach.

Initially, for each user, we obtain a profile. A system administrator assigns profiles in cases where allowable task vocabularies are known a priori. Otherwise, profiles are generated via relevance feedback recording schemes during an initial proper use period. Any potential misuse is then detected by comparing the new user queries against the user profile. The existing system requires a manual adjustment of the weights emphasizing various components of the user profile and the user query in this detection process. The manual human adjustment to the parameters is a cumbersome process. Our hypothesis is: **Data mining techniques can eliminate the need for the manual adjustment of weights without affecting the ability of the system to detect misuse.** The classifier learns the weights to be placed on the various components using the training data. Experimental results demonstrate that using classifiers to detect misuse of an information retrieval system achieves a high recall and acceptable precision without the manual tuning.

Our test data contained 1300 instances, each assessed by four Computer Science graduate students. We ran a 10-fold cross validation using the commonly available freeware tool, WEKA on classifiers such as support vector machine (SMO), neural network (MLP), Naïve Bayes Multinomial (NB), and decision tree (C4.5). The misuse detection systems used throughout our experimentation are based on the nature of the user query length. That is, in different applications the user queries may be short (Title) or longer (Descriptive). Thus, we considered the following systems: 1) short queries are used for building profile and detection (T/T); 2) long queries are used for building profile and detection (D/D); and 3) long queries are used for building profile and short queries are used for detection (D/T). For each system setup, we chose top M=10, 20, 30 feedback terms from top N=5, 10, 20 documents, based on BM25 term weighting. The distribution of the a priori known class labels are 40.9% "Misuse", 49.3% "Normal Use", and 8.7% "Undecided". "Undecided" cases are the cases that the human evaluators were unable to determine otherwise. The pool of queries creating the instances contains 100 TREC 6-7 Title and Descriptive ad hoc topics.

[1] Ling Ma, Nazli Goharian, *Query Length Impact on Misuse Detection in Information Retrieval Systems*, ACM Symposium on Applied Computing, Santa Fe, NM, March 2005.

IEEE International Conference on Intelligence and Security Informatics, Atlanta, May 2005,

The disks 4-5 2GB collection was used. Unfortunately, there is no standard benchmark to use in evaluating misuse detection systems. Thus, we had to build our own benchmark. We evaluate the accuracy of misuse detection using both *Precision* (correctly detected misuse/detected as misuse); and *Recall* (correctly detected misuse/total misuse).

Precision (%)		**Title Build & Title Detect (T/T)**					**Desc. Build & Desc. Detect (D/D)**					**Desc. Build & Title Detect (D/T)**				
N	***M***	**MA**	**SMO**	**MLP**	**NB**	**C4.5**	**MA**	**SMO**	**MLP**	**NB**	**C4.5**	**MA**	**SMO**	**MLP**	**NB**	**C4.5**
5	***10***	68	70	71	67	73*	69	69	70	68	73*	69	64	71	62~	72*
	20	69	70	71	67	72+	69	69	71	67	71	69	70	71	62~	73
	30	69	69	70	69	72	68	69	71	67	71	70	70	70	62~	73
10	***10***	70	71	72+	68	73+	69	69	71	67-	70	70	70	70	63~	71
	20	70	69	72	69	73	70	67	70	67	70	71	71	72	62~	72
	30	70	69	73	69	72	71	67~	71	67~	72	71	71	71	62~	71
20	***10***	70	71	72	69	73	70	67~	71	67-	72	71	70	73	63~	74
	20	71	69	72	69	72	71	67~	71	69~	71	72	71	72	63~	73
	30	71	70	73	69	72	72	67~	71	68~	70	72	71	72	64~	72

Recall (%)		**Title Build & Title Detect (T/T)**					**Desc. Build & Desc. Detect (D/D)**					**Desc. Build & Title Detect (D/T)**				
N	***M***	**MA**	**SMO**	**MLP**	**NB**	**C4.5**	**MA**	**SMO**	**MLP**	**NB**	**C4.5**	**MA**	**SMO**	**MLP**	**NB**	**C4.5**
5	***10***	98	97	96	99	95~	97	98	96	98	94	95	97	96	99*	95
	20	98	97	95~	99	94~	97	98	96	98	94	95	96	95	99*	93
	30	97	98	95	98	95	97	98	96	98	95	94	97+	95	99*	94
10	***10***	98	95~	95-	98	94~	95	98	96	96	94	95	97	95	99*	94
	20	98	97	95~	97	96	96	99+	95	97	95	93	95+	95	99*	95
	30	98	98	95-	98	95-	95	99+	95	98	93	92	98*	95	99*	94
20	***10***	98	96	95~	97	95-	94	99*	95	97+	94	94	98	94	99*	91
	20	97	98	95	97	94	93	99*	94	98*	93	92	97*	94	99*	94
	30	95	98	94	97	93	91	99*	93	97*	93	90	97*	94	99*	93+

We illustrate the precision and recall of four classifier based (SMO, MLP, NB and C4.5) and our baseline, manually adjusted (MA) detection system. To systematically compare, 10 trials of 10-fold cross-validated paired T-test of classifiers versus our MA baseline were conducted over the precision and recall of the "Misuse" class. In the tables shown, statistically significant entries at the 0.05 and 0.01 significance level are designated +/- and */~, respectively. Markers - and ~ indicate that the manual adjustment performed better than the classifiers. All entries without a marker are statistically equivalent. As the results demonstrate for each of the systems T/T, D/D, D/T, there is always a classifier that performs statistically equivalent to or better than the manual adjustment approach, eliminating the need for manual intervention. Examples of such are SMO and NB for T/T, MLP and C4.5 for D/D and D/T in regards to both Precision and Recall. Furthermore, some classifiers such as NB for D/T favor recall over precision and vice versa in the case of C4.5 in T/T. Hence, depending on the application and organization, a classifier can be chosen that optimizes either recall or precision over the other.