Query Length Impact on Misuse Detection in Information Retrieval Systems

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

ABSTRACT

Misuse is the abuse of privileges by an authorized user and is the second most common form of computer crime after viruses. Earlier we proposed a misuse detection approach for information retrieval systems that relied on relevance feedback. The central idea focused on the building of a user profile containing both query and feedback terms from prior queries. Our algorithm matched new activities to existing profiles and assigned a likelihood of misuse to an activity. Only initial evaluation was provided.

We now expand and evaluate our system using both short and long queries noting the effect of query length in the accuracy of the detection. The results indicate an overall precision of 83.9% when short queries are used, and 82.2% for long queries. The rate of the undetected misuse for short queries is less than 2% and for long queries less than 6%. Although higher precision score configurations result in a lower false alarm rate, unfortunately, they increase the rate of undetected misuse both for short and long queries. Given this tradeoff, for any particular application constraint, system behavior can be tuned to minimize either false alarms or undetected misuse.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – User Profiles and Alerts.

General Terms

Algorithms, Experimentation, Security

Keywords

Algorithms, Experimentation, Security, User Profile

1 INTRODUCTION

A recent Computer Security Institute/Federal Bureau of Investigation study noted that after viruses, i.e., malicious code, insider abuse, called misuse, is the second most prevailing form of computer crime [1]. We focus the problem on misuse detection in search systems, in general, and information retrieval systems specifically. We evaluate the accuracy of our misuse detection system both for short queries, like that issued in web search engines, as well as long queries like those issued by information analysts.

SAC'05, March 13-17, 2005, Santa Fe, New Mexico, USA.

Recently, for example, a former Chase Financial Corporation employee pleaded guilty to unlawfully accessing bank records [2]. The employee was authorized to access the computer where the records were maintained; however, she abused her privileges and accessed records not within her scope of authorized interest. Similarly, a former Cisco employee was sentenced for accessing a stock-processing computer. Although he did have authority to access the given computer, what he was accessing was the database associated with the granting of employees Cisco stocks [3]. Furthermore, a former FBI investigative analyst pleaded guilty for exceeding his authorized access to protected computers [4]. He searched information from many government database systems for his own benefit and also disclosed classified information to friends and family. All of these cases are examples of information system misuse.

Typically, misuse detection systems rely on techniques that are "systems oriented". Namely, they detect an abuse of access rights based on record identity, permission status, or location. However, in information retrieval systems, misuse can occur by accessing content that is inappropriate - independent of who owns the data record, where it is stored, or what the document's permission status is. Hence, our focus is on the detection of the misuse of content. In this effort we focus on detecting the misuse as the user searches for off-allowed-topics.

2 PREVIOUS WORKS

Misuse detection has generally been employed to complement the shortcomings of other prevention techniques. Prior work on misuse detection mainly focused on usage logs and user profiles. Profile-based detection systems audit the deviation of user activities from normal user profiles. One approach, reviews a user's command history over a specific period of time to detect potential command usage pattern differences [5]. Another scans and then mines the usage logs [6].

In information system, specifically for database applications, Chung et al. in [7] describe their misuse detection system, DEMIDS. DEMIDS logs user access patterns to the database tables, columns, and other structures to build a user profile to track the behavior of the user.

In [8] a misuse detection system was proposed and developed by comparing user behavior in terms of content rather than in terms of commands issued to a developed user profile, learned through clustering, relevance feedback, and fusion methods. Thus a new dimension was created to profile-based misuse detection for search systems. Later in [9] this relevance feedback approach was improved and evaluated on 300 cases.

The work herein investigates the effects of query length on the quality of misuse detection using Relevance Feedback. We evaluated our approach on 1300 cases in comparison to a team of

four human user auditors and demonstrated comparable misuse detection capability.

The remainder of this paper is organized as follows: We give an overview of our improved relevance feedback approach in section 3. In Sections 4 and 5, we present our experimentation setup, evaluation metrics, and results. In Section 6, we elaborate a framework for our misuse detection system to adapt to legitimate user interest change. Finally, we conclude and outline our proposed future work.

3 APPROACH

A potential misuse of an information retrieval (IR) system can be indicated by comparing a user's actions against his/her profile. As initially proposed in [8], detecting misuse in an information retrieval system is a process that has two phases described in detail below, and presented in figure 3.1.

Phase 1: Building Profile

User profiles are initially built either by prior knowledge such as a job description or are built from user queries that are monitored and approved by a systems administrator.

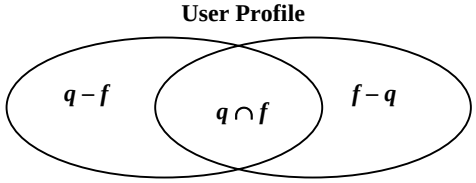
Phase 2: Detecting Misuse

In the detection phase, any user query is tested against the user's profile to determine the level of potential misuse by generating a degree of warning. A misuse warning is computed by comparing the difference between the new queries and the user profile.

1. Build User Profile
   profile := null
   For each query
   profile := query ∪ Feedback(query)
2. Detection Phase
   For each query
   Warning w :=0
   terms := query ∪ Feedback(query)
   Generate Warning w (terms, profile)

Figure 3.1: Algorithm to Detect Misuse in Search Systems using Relevance Feedback

When a user submits a query, the relevance feedback mechanism automatically selects potentially *M* "good" terms retrieved out of top *N* documents and adds both query terms and relevance feedback terms to the user profile. In the detection phase, any user query is tested against the user's profile.



$q \cap f$: Profile terms in both query and feedback subsets
$q - f$: Query terms not in profile feedback subset
$f - q$: Feedback terms not in profile query subset

Figure 3.2: User Term Profile

We illustrate the components of a user profile in Figure 3.2. The user profile is the union of query term subset *q* and feedback term subset *f*, namely, $q \cup f$.

Previously in [8], the misuse warning *w* was generated by definition RF1. In RF1 the query terms for detection phase are not the union of the query and its relevance feedback terms, but the query alone. RF1 is defined as:

RF1: $$w = \frac{|Q_A|}{|Q|}$$

where $|Q_A|$ is the number of query terms absent from profile and $|Q|$ is the size of the query *Q*.

Definition RF1 does not consider the effect of presence or absence of user query relevance feedback terms (*F*) in the user profile, namely, the importance of the fact that the user query feedback terms in the profile can also indicate whether a user query matches the content of a user profile. Thus, we modified RF1 to RF2 to generate lower warnings when:

- Query terms are part of the profile. **(C1)**
- Feedback terms from the query are part of the profile. **(C2)**

In RF2, warning *w* is high only if neither conditions C1 nor C2 are true. Let Q_q be the number of query terms present in profile's *q* set, and Q_{f-q} be the number of query terms present in profile's $f - q$ set. To evaluate the relative importance of Q_q versus Q_{f-q}, we add a weighting factor β for the warning generated from Q_{f-q}. In RF2, we treat the terms in $q \cap f$ and $q - f$ sets identically.

RF2:

$$w = w_Q \; w_F$$

$$w_Q = \Phi\left(|Q_A| - |Q_q| - \beta |Q_{f-q}| \right)$$

$$w_F = \Phi\left(|F_A| - |F_q| - |F_{f-q}| \right)$$

$$\Phi(w_Q) = \frac{w_Q + |Q|}{2|Q|} \qquad \Phi(w_F) = \frac{w_F + |F|}{2|F|}$$

The variables in RF2 are defined as follows:

w : misuse warning

w_Q : warning from a user query Q

w_F: warning from feedback terms F of the user query

Q_A : set of user query terms absent from profile

Q_q : query terms present in profile's *q* set

Q_{f-q} : query terms present in profile's *f–q* set

F_A : feedback terms of user query that are absent from profile

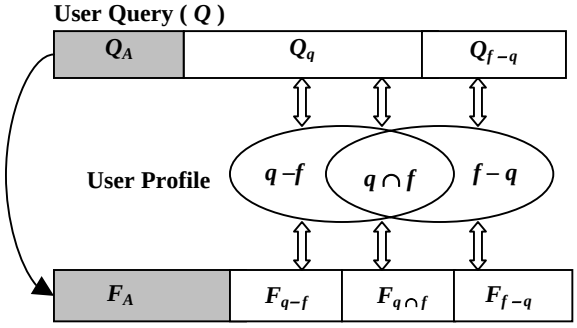
F_q : feedback terms of user query present in profile's *q* set

F_{f-q} : feedback terms of user query present in profile's *f–q* set

β : term weight between 0 and 1 that is associated with Q_{f-q}

$\Phi(w_Q)$, $\Phi(w_F)$: functions normalizing w_Q, w_F between 0 and 1

We further modified RF2 to RF3 and gave the user query feedback terms in $q \cap f$, $q - f$, and $f - q$ different weights. By varying the emphasize on respective feedback term subsets in RF3, we found improvement either in precision and false alarm or in undetected misuse —the named metrics are defined in section 4.2. Figure 3.3 shows RF3 elements in generating the warning.



Q_A : Query terms absent from profile

F_A : Feedback terms absent from profile

Q_q , Q_{f-q} : Query terms in profile's $q, f-q$ subset, respectively

$F_{q \cap f}, F_{q-f}, F_{f-q}$: Feedback terms from user query in profile $q \cap f$, $q-f$, $f-q$ subsets, respectively

Figure 3.3: RF3 Warning Generation

RF3:

$$w = w_Q \; w_F$$

$$w_Q = \Phi\left(|Q_A| - |Q_q| - \beta |Q_{f-q}| \right)$$

$$w_F = \Phi\left(|F_A| - \alpha |F_{q-f}| - \delta |F_{q \cap f}| - \gamma |F_{f-q}| \right)$$

$$\Phi^*(w_Q) = \max\left(0, \Phi(w_Q)\right) \qquad \Phi^*(w_F) = \max\left(0, \Phi(w_F)\right)$$

The additional variables in RF3 are defined as:

α, δ, γ : term weights with value range between 1 and 2; α, δ, γ are associated with $F_{q \cap f}$, F_{q-f}, F_{f-q}, respectively.

$\Phi^*(w_Q)$: function normalizing w_Q, between 0 and 1.

$\Phi^*(w_F)$: function normalizing w_F between 0 and 1.

4 EXPERIMENTATION

4.1 Experimentation Setup

The pool of queries for profile building/testing contains 100 TREC-6 and TREC-7 [11] ad hoc queries (*topic*s 301- 400). These queries are known to be in a categorical nature. We used TREC-6 and TREC-7 2GB document collection. The queries were manually separated into 22 categories according to their content coverage. Categories of queries cover different area of interest such as crime, medicine, economy, etc. We built user profiles with the procedure described in Figure 3.1. Each profile was built with 60 queries from which at least 20 queries were distinct and randomly sampled from 6 random categories. We ran all experiments for both short queries (*Title*) and long queries (*Descriptive*). For each *Title* query, the corresponding *Descriptive* query provided more descriptive information. The *Title* queries have each 1 to 4 terms per query with a median size of 3. Note that this length is roughly comparable to web queries. The *Descriptive* queries have each 2 to 19 terms per query with a median size of 7.

Throughout our experimentations we used top 10, 20, and 30 terms from top 5, 10, and 20 documents for relevance feedback terms. Rocchio relevance feedback algorithm and BM25 term weighting function are used.

Five-level evaluation is commonly employed in user recommendation systems [12]. In our misuse detection system, a misuse warning is rated as one of the five levels according to its severity, "strong misuse" (L_5), "misuse" (L_4), "undetermined" (L_3), "almost normal use" (L_2) and "normal use" (L_1).

We are aware of the subjectivity of individual human evaluators to make the judgment on a potential misuse. To minimize the error in judgment, four human evaluators each evaluated 1300 test cases. All four evaluators are Computer Science graduate students with prior knowledge in Information Retrieval Systems and Relevance Evaluation. Each of the four evaluators manually read the TREC ad hoc queries used to build the user profiles, as well as all the 1300 test cases that were used to generate the misuse warnings, and then assigned a warning level to each test case. The evaluators assigned a high warning level when a user searched for completely different categories of knowledge not in the user profile. If the evaluators had a different warning rating based on their perception of the test case, we used their rounded average warning rating.

We assessed the judgment of the four evaluators by calculating the mean standard deviation and a pair-wise correlation analysis on their judgments. We found that all four evaluators judged all cases very similarly, with correlation coefficient of judgments in the range of 95-99%.

4.2 Evaluation Metrics

We evaluated our system by evaluating its closeness to the actual ratings, percentage of cases evaluated correctly, the percentage of false alarm, and finally, the percentage of misuse not detected. Thus, we used four measures to evaluate our system, which are explained in this section.

We evaluate the accuracy of our misuse detection system by comparing the system's numerical scores, i.e., the warning levels generated by the detection system, against the actual ratings assigned by the human evaluators. Mean Absolute Error (MAE) is widely used in the recommendation systems to measure error between actual ratings and system predictions [12][13].

To measure system accuracy with the MAE metric, we first convert the system generated numerical misuse warnings to five levels, with each level covering the misuse-warning interval of 0.2 between 0 and 1. L_1 indicates the lowest and L_5 the highest misuse potential. MAE is formally defined as follows:

$$MAE = \sum_{i=1}^{n} \frac{|L_i - I_i|}{n}$$

where i is the test case identifier, L_i is the predicted warning level by the misuse detection system on a test case; I_i is the warning level derived by taking the rounded average of the four human evaluators' warning ratings. n is the total number of test cases. The lower the MAE, the more accurately the misuse detection system predicts misuse warnings. MAE demonstrates how much

on average the predicted misuse level deviates from the actual misuse level.

Another measure is the *Precision*. In as much there is some gray level of difference between ratings of two consecutive levels, we accept one level of difference. We define precision *P* as:

$$P = \frac{m}{n}$$

where *m* is the number of test cases that their predicted misuse level is at most within 1 level of difference with the actual human assessor misuse level. We consider these as valid detections. *n* is the number of cases that the system assigned a misuse level to them. We measure the overall precision and the precision in levels L_4 and L_5 corresponding to *n.*

Furthermore, we measure the percentage of *Undetected Misuse*, UM, defined as the ratio of undetected misuse to the number of misuse cases assessed by the evaluators. We are not concerned about undetected misuse warning at level L_3, namely "undetermined", since warning at this level is not indicative and covers an unclear area between "almost normal use" and "misuse". Undetected misuse is the notion of *Recall* with the difference that *UM=1-Recall*. Recall is the ratio of the correctly detected misuse to the number of misuse cases assessed by the evaluators.

Finally, we measure the percentage of *False Alarm*, FA, defined as the number of false alarms in all levels, divided by total number of test cases.

5 RESULTS

In sections 5.1 to 5.2, we present our results using Relevance Feedback warning definitions RF1, RF2, and RF3 for both short (*Title*) and long (*Descriptive*) queries. We described the Title and Descriptive queries in section 4.1.

Our results are based on 1300 test cases with top ten, twenty, and thirty (M=10, M=20, M=30) relevance feedback terms from top five, ten, and twenty documents (N=5, N=10, N=20). The rate of Undetected Misuse (UM), False Alarm (FA), Precision (P), the precision of detection in levels L_4 and L_5 of misuse, and finally Mean Absolute Error (MAE) are presented.

5.1 Title (short) Queries

The comparison of results on Title (short) queries for RF1, RF2, and RF3 shows that throughout all experimentations, the results are consistent for most configurations and parameters, with the exception of a few anomalies. The comparison and analysis follows:

The lowest rate of *undetected misuse* for all three misuse warning definitions, occurs when relevance feedback terms from top five documents (N=5) are used for building user profile and detection, independent of the number of top feedback terms M. Generally, the lower the value of M, (M=10), the better is the rate of the undetected misuse, regardless of N. However, there are some anomalies such as the case with N=5 and M=20 that shows a lower UM rate compared to M=10. Furthermore, in RF2 and RF3, reducing the weight on feedback terms for generating warning (β=0.1) was shown to have the least amount of undetected misuse. Moreover, in RF3, putting more weight on feedback terms that appear in the query but not feedback subset of profile (α = 2); or appear in both query and feedback subsets of profile (δ = 2), achieves the lowest undetected misuse. We present the lowest rate of undetected misuse for RF1, RF2, and RF3 for various values of M, in Tables 5.1 (a), (c), and (e).

The overall *precision* and precision in levels four and five improves by increasing N (N=20) and M (M=30). In RF2 and RF3, the higher β (β=0.9) also increases the precision in levels four and five. However, in many configurations, the best overall precision is achieved by β=0.5. Moreover, putting more emphasis on feedback terms that appear in the feedback subset of the profile (γ=2) achieves higher overall precision in majority of configurations, and higher precision in levels four and five in almost all the configurations. Tables 5.1 (b), (d), and (f) illustrate the highest overall precision with the best setup for RF1, RF2, and RF3 for various values of M. Both RF2 and RF3 have higher precision than RF1. Although the "best" setups illustrated in tables 5.1 are not isolated, a few setups do not follow the general trend in parameter values as discussed.

The rate of *false alarm* is shown to decrease with larger N (N=20) and larger M (M=30). The reason is that with the larger N and M the user profile has a wider scope of content. Furthermore, as β increases in RF2 and RF3, (β=0.9), false alarm rate decreases. In addition to that, in RF3, putting more emphasis on feedback terms that appear in the feedback subset of profile (γ = 2) achieves often a lower false alarm.

Clearly, a tradeoff between the rate of undetected misuse and false alarm; and between the rate of undetected misuse and precision in levels four and five is evident throughout our experimentations.

Table 5.1(a): *Title*- **Lowest Undetected Misuse for RF1, RF2** (β=0.1)**, and RF3** (β=0.1, α=2, δ=1, γ = 1), **M=10**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	1.6%	21.8%	77.3%	72.6%	57.8%	0.79
RF2	1.6%	22.2%	77.0%	71.8%	57.4%	0.79
RF3	1.6%	22.2%	77.0%	71.8%	56.8%	0.79

Table 5.1(b): *Title*- **Highest Overall Precision for RF1, RF2** (β=0.5)**, and RF3** (β=0.5, α = 1, δ = 2, γ = 1), **M=10**

(N=20)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	2.3%	20.6%	78.3%	73.7%	63.3%	0.76
RF2	2.4%	17.7%	81.1%	76.2%	72.2%	0.70
RF3	3.0%	17.5%	81.1%	76.3%	72.8%	0.70

Table 5.1(c): *Title*- **Lowest Undetected Misuse for RF1, RF2 (**β=0.1)**, and RF3** (β=0.1, α=2, δ =1, γ=1), **M=20**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	3.3%	21.3%	77.1%	73.8%	55.7%	0.78
RF2	1.0%	21.8%	77.7%	72.8%	60.7%	0.78
RF3	1.0%	21.5%	78.1%	73.4%	59.6%	0.77

Table 5.1(d): *Title*- **Highest Overall Precision for RF1**(N=10)**, RF2** (N=20, β=0.5)**, and RF3** (N=20, β=0.1, α = 1, δ = 2, γ = 1), **M=20**

	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	3.9%	20.1%	77.8%	74.3%	63.5%	0.77
RF2	6.3%	15.8%	81.2%	77.3%	74.3%	0.70
RF3	5.7%	15.1%	82.2%	77.5%	75.5%	0.69

Table 5.1(e): *Title*- **Lowest Undetected Misuse for RF1, RF2 (**β=0.1**), and RF3** (β=0.1, α=2, δ=1, γ=1), **M=30**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	4.1%	20.7%	77.3%	73.6%	60.8%	0.78
RF2	1.2%	20.5%	78.8%	73.2%	66.0%	0.76
RF3	1.2%	20.5%	78.8%	73.2%	66.0%	0.76

Table 5.1(f): *Title*- **Highest Overall Precision for RF1** (N=10)**, RF2 (**N=20, β=0.1**), and RF3** (N=20,β=0.1, α = 1, δ = 1, γ = 2)**, M=30**

	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	5.9%	18.8%	78.2%	74.4%	70.1%	0.77
RF2	5.7%	15.2%	82.0%	78.8%	75.6%	0.70
RF3	7.7%	12.3%	83.9%	80.3%	81.3%	0.67

5.2 Descriptive (long) Queries

We present the comparison among all three misuse-warnings RF1, RF2, and RF3 using Descriptive (long) queries. Similar to earlier experimentations, we evaluated the accuracy of our system when a different number of top documents (N=5, 10, and 20) and top terms (M=10, 20, and 30) are used for relevance feedback terms. Moreover, various parameters explained earlier (β, α, δ, γ) are evaluated.

Similar to Title queries, the rate of undetected misuse is dependant on the value of N. Lower the N (N=5), lower is the rate of undetected misuse. Similarly, there is a tradeoff between the rate of undetected misuse and false alarm as in Title queries.

Across all three definitions RF1, RF2, and RF3, the *Descriptive* queries, similar to the *Title* queries, achieve a lower rate of undetected misuse when less number of terms (M=10) is used for the relevance feedback. An anomaly is observed when a lower UM rate is achieved for RF2 and RF3 by M=20 compare to M=10. As a tradeoff between undetected misuse and false alarm, the larger value of M, i.e., M=30 reduces the amount of false alarm. Similar to Title queries, the larger the β (β=0.9), the less is the false alarm. The smaller the β, i.e., (β=0.1), the less is the rate of undetected misuse

Similar to Title queries, For RF3, with any given N, M, β configurations, putting more weight, (γ=2), on the query feedback terms that exist in feedback subset of the user profile, $f - q$, achieves lower amount of false alarm. In addition to that, in RF3, putting more weight on feedback terms that appear in query subset of profile, $q - f$, (α=2), achieves often the lowest undetected misuse.

The overall *precision* in many configurations improves with (N=5); and precision in levels four and five improves with higher N (N=20). Both the overall precision and precision in levels four and five improve with (M=30). The precision at levels four and five improve in RF2 and RF3 by increasing value of β (β=0.9). Furthermore in RF3, (γ=2) improves the precision for levels four and five. We present the highest overall precision with the best setup for RF1, RF2, and RF3 for various values of M in Tables 5.1 (b), (d), and (f). Note that these configurations are not necessarily the best configurations for precision at levels L_4 and L_5.

Table 5.2(a): *Descriptive*- **Lowest Undetected Misuse for RF1, RF2 (**β=0.1**), and RF3** (β=0.1, α=2, δ=1, γ=1), **M=10**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	5.5%	17.5%	79.8%	76.9%	71.7%	0.80
RF2	6.5%	16.8%	80.0%	76.2%	76.7%	0.77
RF3	6.5%	16.8%	80.0%	76.2%	76.7%	0.77

Table 5.2(b): *Descriptive*- **Highest Overall Precision for RF1, RF2** (β=0.9)**, and RF3** (β=0.9, α=1, δ=1, γ=2)**, M=10**

(N=20)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	6.3%	16.7%	80.2%	77.6%	73.2%	0.79
RF2	10.8%	14.2%	80.5%	78.9%	80.4%	0.76
RF3	12.0%	13.1%	81.0%	79.4%	80.9%	0.76

Table 5.2(c): *Descriptive*- **Lowest Undetected Misuse for RF1, RF2** (β=0.1)**, and RF3** (β=0.1, α=2, δ=1, γ=1), **M=20**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	5.7%	16.6%	80.6%	77.9%	73.1%	0.79
RF2	5.7%	16.1%	81.1%	78.3%	77.3%	0.76
RF3	5.7%	16.1%	81.1%	78.2%	77.5%	0.76

Table 5.2(d): *Descriptive*- **Highest Overall Precision for RF1, RF2** (β=0.9)**, and RF3** (β=0.9, α=1, δ=1, γ=2), **M=20**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	5.7%	16.6%	80.6%	77.9%	73.1%	0.79
RF2	6.3%	15.2%	81.8%	79.4%	78.0%	0.75
RF3	6.7%	14.5%	82.2%	80.5%	77.9%	0.74

Table 5.2(e): *Descriptive*- **Lowest Undetected Misuse for RF1, RF2** (β=0.1)**, and RF3** (β=0.1, α=2, δ=1, γ=1), **M=30**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	7.9%	15.9%	80.2%	77.6%	73.6%	0.80
RF2	6.1%	15.8%	81.2%	78.5%	76.7%	0.76
RF3	6.1%	15.8%	81.2%	78.5%	76.9%	0.76

Table 5.2(f): *Descriptive*- **Highest Overall Precision for RF1, RF2** (β=0.9), **and RF3** (β=0.9, α = 1, δ = 1, γ = 2), **M=30**

(N=5)	UM	FA	P	P_{L5}	P_{L4}	MAE
RF1	7.9%	15.9%	80.2%	77.6%	73.6%	0.80
RF2	7.5%	14.5%	81.8%	79.8%	77.3%	0.76
RF3	7.7%	14.2%	81.9%	79.6%	78.8%	0.75

6 A Framework for User Profile Update

We now describe a framework for profile currency maintenance. Misuse detection systems treat sudden change versus gradual drift of user interest differently. Experimentation of such change framework will be addressed in future work.

Misuse detection systems identify sudden change of interest as high-level misuse, unless a systems administrator validates it. Thus, misuse detection systems should not automatically adapt to such sudden changes, to guarantee that potential misuse is properly detected. This is unlike information filtering systems that adapt to both sudden and gradual changes in user interest [14], [15].

In misuse detection systems, the systems administrators are responsible for updating the user profiles as a consequence of sudden change. When new task is assigned to a user, an administrator can either add or remove terms in the profile related to that task, or to rebuild the profile. Furthermore, administrators may themselves update the profile to resolve repeated high-level warnings generated by the system. This case, called manual feedback, results in the updating of the profile that is initiated by a system suggested sudden change. Manual feedback can be applied to lower level of warnings as well.

Unlike information filtering system where the relevance judgments are made available to the system by the user; in misuse detection system the generated warnings are feedback to the system by the administrator.

In comparison to sudden change, gradual drift indicates user legitimate interest change. Therefore, misuse detection system, similar to information filtering system, adapts to gradual change of interest. In the case of low-level misuse, the system updates the profile if there is sufficient number of occurrences of user's given query in enough duration of time. We consider this as a legitimate user interest change and update the user profile by incorporating the terms from the specified query terms. The update to profile is based on pseudo feedback caused by misuse warnings. Obviously, the low level warnings and automated update of profiles are also monitored by administrator.

7 CONCLUSION

The objective of any misuse detection system is to make sure that as many as possible abuses are detected with as few as possible false warnings. We evaluated our misuse detection system for both Title (short) and Descriptive (long) queries and showed promising results in both.

Our experimental results did demonstrate that unfortunately there is a direct correlation between the rate of undetected misuse and false alarm; and between the rate of undetected misuse and precision in levels four and five is evident throughout our experimentations. We are continuing to address new definitions for which the undetected misuse is reduced but precision is still maintained at high level.

We presented a framework to help the misuse detection system learn legitimate user interest drifting with feedback and input. Experiments of such framework will be completed in future work.

Furthermore, we are currently evaluating the accuracy of our detection system on the collection that the documents are tagged based on their sensitivity levels. The results will be presented in our future work.

8 REFERENCE

[1] M. Whitman, *Enemy at the gate: Threats to information security*, CACM, 46(8), 2003.

[2] Press Release, Computer Crime and Intellectual Property section of the Criminal Division of US Dept. of Justice, 2001. http://www.usdoj.gov/criminal/cybercrime/turnerPlea.htm

[3] Press Release, Computer Crime and Intellectual Property section of the Criminal Division of US Dept. of Justice, 2001. http://www.usdoj.gov/criminal/cybercrime/Osowski_TangSent.htm

[4] Press Release, United State Attorney's Office Northern District of Texas, US Department of Justice, November 5, 2003. http://www.usdoj.gov/usao/txn/PressRel03/fudge_ind_pr.html

[5] J. Marin, D. Ragsdale, and J. Surdu, *A hybrid approach to the profile creation and intrusion detection*, DARPA Info. Surv. Conf. and Expo. 2001.

[6] C. Ling, J. Gao, H. Zhang, W. Qian, H. Zhang, *Improving encarta search engine performance by mining user logs*, Int. Journal of Pattern Recognition and Artificial Intelligence, 2002.

[7] C. Y. Chung, M. Gertz, and K. Levitt, *Demids: A misuse detection system for database systems*, In Third Int. IFIP TC-11 WG11.5 Working Conf. on Integrity and Internal Control in Information Systems (1999), 159–178, Kluwer Publishers.

[8] R. Cathey, L. Ma, N. Goharian, D. Grossman, *Misuse* detection for information retrieval systems, ACM CIK*M,* 2003.

[9][*] L. Ma, N. Goharian, *Using Relevance Feedback to Detect Misuse for Information Retrieval Systems*, ACM CIKM, 2004.

[10] David A. Grossman and Ophir Frieder, *Information retrieval algorithms and heuristic*s, 2nd ed., Kluwer Publishers, 2003.

[11] National Institute of Standards and Technology, *Text retrieval conference(trec*), December 2002, http://trec.nist.gov/.

[12] P. Resnick, N. Iacovou et al, Grouplens:*an open architecture for collaborative filtering of net news*, ACM CSCW 1994.

[13] BM Sarwar et al., *Item-Based Collaborative Filtering Recommendation Algorithms*, 10th Int'l W3 Conf., 2001

[14] Dwi H. Widyantoro and John Yen, *Learning User Interest Dynamics with Three-Descriptor Representation*, JASIS 2000.

[15] Dwi H. Widyantoro and John Yen, *Tracking changes in user interest with a few relevance judgments*, CIKM03.

[*] The notation used herein differs slightly from that used initially to improve readership as suggested by multiple reviewers.