Using Relevance Feedback to Detect Misuse for Information Retrieval Systems

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

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

General Terms
Algorithms, Experimentation, Security

Keywords
Misuse, Detection, Security, Relevance Feedback

1. INTRODUCTION
Misuse is the abuse of privileges by an authorized user and is the second most common form of computer crime after viruses [1]. In [2], we developed a misuse detection system for search systems that compared user behavior to user profile learned through clustering, relevance feedback, and finally the fusion of results of these methods. Here we improve the relevance feedback method used to detect misuse. As compared to the approach described in [2], the presented approach yields higher detection accuracy with a lower rate of undetected misuse.

2. APPROACH
The overall detection algorithm, details found in [2], is:

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

User profile terms are obtained either from prior knowledge or are built from user queries monitored and approved by a systems administrator. When a user submits a query, pseudo-relevance feedback 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.

Previously, in [2], we measured the misuse warning w generated by the algorithm as defined in the definition RF1:

CIKM'04, NOV 9-11, 2004, Arlington, VA, USA

RF1:

$$w = \frac{|A|}{|Q|}$$

where A is the set of query terms absent from profile and Q is the set of query terms.

This definition, however, does not consider the effect of presence or absence of user query relevance feedback terms in the user profile. Thus, we modified RF1 to generate lower warnings when:

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

The user profile is the union of query term subset P and feedback term subset R. $P \cap R$ is the set of terms that appear in both P and R subsets of the profile. $P–R$ is profile query terms that are not in the profile relevance feedback term subset. $R–P$ is profile relevance feedback terms that are not in query term subset of the profile.

In the revised warning definition, RF2, we treat terms in the $P \cap R$ and the $P–R$ sets identically. That is, there is no distinction among terms in the query terms subset of the user profile.

RF2:

$$w = w_p \bullet w_r$$
$$w_p = \Phi(|A_p| - |P_P| - \beta \bullet |P_{R-P}|)$$
$$w_r = \Phi(|A_r| - |R_P| - |R_{R-P}|)$$
$$\Phi(x) = (x + |Z|) / (2 \bullet |Z|)$$

w : misuse warning

w_p : warning from a user query

w_r : warning from relevance feedback terms of the user query

A_p : set of user query terms absent from profile $P \cup R$

P_P, P_{R-P}: query terms present in profile's P, R-P set, respectively

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

R_P : feedback terms of user query present in profile's P set

R_{R-P} : feedback terms of user query present in profile's R–P set

β : term weights between 0 and 1 that is associated with P_{R-P}

$\Phi(x)$: function normalizes warning level x between 0 and 1

Z : Query size for w_p normalization, relevance feedback size for w_r normalization

In Definition RF2, warning w is high only if neither C1 nor C2 are true. To evaluate the relative importance of P_P versus P_{R-P}, we add a weighting factor β for the warning generated from P_{R-P}.

We further modified RF2 to Definition RF3 and gave the terms in $P \cap R$, $P-R$, and $R-P$ different weights. Our experimental results for RF3 show that different weights on respective profile term subsets can improve either precision or the rate of undetected misuse.

RF3:

$$w = w_p \cdot w_r$$

$$w_p = \Phi(|A_p| - |P_P| - \beta \cdot |P_{R-P}|)$$

$$w_r = \Phi(|A_r| - \alpha \cdot |R_{P \cap R}| - \delta \cdot |R_{P-R}| - \gamma \cdot |R_{R-P}|)$$

$$\Phi(x) = MAX(0, (x + |Z|) / (2 \cdot |Z|))$$

P_P, P_{R-P}, A_p, A_r, Z: the same as in RF2.

$R_{P \cap R}$, R_{P-R}, R_{R-P}: feedback terms of user query present in profile subset $P \cap R$, $P-R$, $R-P$, respectively.

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

Φ(x): normalization function similar to Φ(x) in RF2, except for a chance that the warning is slightly smaller than 0, which is resolved by a MAX function.

3. EXPERIMENTATION & RESULTS

We used the TREC 2GB collection and title only TREC topics (301 to 400). These topics were manually separated into 22 categories according to their content coverage. Each profile was built with 60 queries from which at least 20 queries were distinct and randomly sampled from 6 random categories. We used top 20 terms from top 5, 10, and 20 documents for relevance feedback terms. In our misuse detection system, a misuse warning is rated as one of the five levels according to its severity, "strong misuse", "misuse", "undetermined", "almost normal use" and "normal use".

Four human evaluators each evaluated 300 test cases on our misuse detection system. Each of the four evaluators manually read the queries used to build the user profiles, as well as all the 300 test queries that were used to generate the misuse warnings, and then assigned a warning level to each of the 300 test cases. We assessed the judgment of the four evaluators by calculating the mean standard deviation and a pair-wise correlation analysis on their judgments, which indicated that all four evaluators judged all cases very similarly.

We evaluated our system by evaluating its closeness to the actual ratings (MAE), percentage of cases evaluated correctly (P), the percentage of false alarm (FA), and finally, the percentage of misuse not detected (UM). Precision *P* allows for at most one level difference between system prediction and human evaluation. We measure the percentage of Undetected Misuse, defined as the number of undetected misuse in level L4 (misuse) and L5 (strong misuse) divided by total number of test cases. We are not concerned about undetected misuse warning at level L3 since level 3 (undetermined) is not a high misuse warning and covers an unclear area between almost normal use and misuse.

In [2], we built profiles and tested the detection accuracy for RF1 using queries containing proper nouns. Thus, our system produced a very high accuracy of almost 92% to detect a potential misuse. We show the new results for 300 cases with RF1, RF2, and RF3 with top 20 relevance feedback terms from top 5, 10, and 20 documents. The Precision (P) and Mean Absolute Error (MAE), the rate of the Undetected Misuse (UM) and False Alarm (FA), and finally, the precision of detection in each of the five levels of misuse (P_{Li}) are presented (see tables 1, 2 and 3).

Table 1: RF1 and RF2 (β = 0.9) **Results**

	RF1			RF2		
	N=5	N=10	N=20	N=5	N=10	N=20
P	62.7%	65.0%	65.0%	70.3%	73.7%	77.3%
MAE	1.17	1.14	1.10	1.07	0.99	0.88
UM	4.7%	5.7%	6.3%	5.0%	7.0%	9.3%
FA	32.7%	29.0%	28.0%	24.7%	19.0%	12.7%
P_{L5}	64.1%	65.7%	69.8%	67.5%	70.9%	84.8%
P_{L4}	54.1%	59.3%	57.5%	57.9%	73.2%	76.4%
P_{L3}	42.0%	44.9%	40.7%	67.7%	60.0%	59.2%
P_{L2}	72.7%	68.8%	76.2%	76.2%	76.2%	78.9%
P_{L1}	100.0%	97.6%	95.2%	100.0%	97.6%	95.7%

The lowest rate of undetected misuse, for all three definitions, RF1, RF2, and RF3, occurs when feedback terms from top five documents retrieved (N=5) are used for building user profile and detection. In addition, in RF2 and RF3, lowering the emphasis on relevance feedback terms for generating warning (β=0.1) was shown to contribute to the least amount of undetected misuse compare to higher value of β. Furthermore, to achieve the lowest rate of undetected misuse in RF3, highest weight is given to profile query terms that are not in profile relevance feedback term subset (α = 1, δ = 2, γ = 1). Putting more weight on *P-R* subset can constrain the scope of search during detection phase, thus reduces the chance of not detecting a misuse.

The highest precision and lowest false alarm, for all three definitions RF1, RF2, and RF3, occur when feedback terms from top twenty documents (N=20), and β =0.9 for RF2 and RF3, are used. In addition to that, in RF3, the emphasis on profile terms that appear in both query terms and relevance feedback terms subsets (α=2, δ = 1, γ = 1) achieves the highest precision of 79.3% and lowest false alarm. Terms in $P \cap R$ subset of the profile are more indicative of user interest shown in profile building phase, since the user must have searched for these terms and these terms must have been ranked high in the retrieved documents.

Clearly, a trade off between the precision and the rate of undetected misuse is evident.

Table 2: Lowest Undetected Misuse Setup of Definitions RF1, RF2 (β=0.1)**, and RF3** (β=0.1, δ = 2, α = 1, γ = 1)

N=5	**FA**	**UM**	**P**	**MAE**	**L_5**	**L_4**
RF1	32.7%	4.7%	62.7%	1.17	64.1%	54.1%
RF2	29.3%	3.3%	67.3%	1.14	66.4%	53.8%
RF3	29.0%	3.3%	67.7%	1.13	66.4%	53.9%

Table 3: Highest Precision Setup of Definitions RF1, RF2 (β=0.9)**, and RF3** (β=0.9, α = 2, δ = 1, γ = 1)

N=20	**FA**	**UM**	**P**	**MAE**	**L_5**	**L_4**
RF1	20.0%	6.3%	65.0%	1.10	69.8%	57.5%
RF2	12.7%	9.3%	77.3%	0.88	84.8%	76.4%
RF3	9.7%	10.3%	79.3%	0.84	86.2%	81.6%

4. REFERENCE

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

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