DOTS: Detection of Off-Topic Search via Result Clustering

Nazli Goharian and Alana Platt
Information Retrieval Laboratory
Illinois Institute of Technology
{goharian, platt} @ ir.iit.edu

Abstract— **Often document dissemination is limited to a "need to know" basis so as to better maintain organizational trade secrets. Retrieving documents that are off-topic to a user's pre-defined area of information need (task) via a search engine is potentially a violation of access rights and is a concern to every private, commercial, and governmental organization. Such misuse, defined as "off-topic access to sensitive data by an authorized user", is the second most prevalent form of computer crime after viruses per a recent Computer Security Institute/Federal Bureau of Investigation study.**

We present a content-based off-topic detection approach that uses *query result clustering* to detect off-topic searches. This approach supports higher detection precision than the state of the art. Multiple methods for picking the "good" clusters are proposed, and their effect on the detection rate and precision is evaluated. A high detection precision is critical as a false access violation accusation unfairly and inappropriately subjects the user to scrutiny. Our empirical results show that using *clustering query results* can significantly reduce such false positives.

Index Terms—**Clustering, Information Retrieval, Off Topic Search, misuse detection**

I. Introduction

Illegitimate access to document collections in an organization by an insider, i.e., an authorized user, is a risk for every organization. We focus on the detection of misuse of information retrieval systems. That is, we focus on detecting an authorized user accessing data via a search engine that is considered "off-topic" to their predefined area of interest. This has become an increasingly prevalent problem in today's society. Organizations such as Chase Financial Corporation [15], Cisco [16] and the FBI [17] have all had employees face legal action due to their accessing documents within their organizations that were outside their scopes of information need (task). Likewise, the ability to inappropriately access internal documents without being detected can result in undetected insider trading, a clear violation of SEC regulations. Thus, mechanisms that detect the inappropriate access of content are needed.

Off topic (misuse) detection is a two-phase process, namely profile creation followed by on-line continuous detection. The detection phase is based on dissimilarity of user's information needs (tasks) to one's profile.

User profile (user model) creation as an active research topic is studied vastly in past years in the fields of information filtering, personalized web, collaborative web, and information retrieval. User profiles have also been used in the intrusion detection efforts. There have been very few and only very recent studies on the topic of off-topic search detection. We are interested in investigating how by picking "good" cluster representative(s) for user query results and comparing that with the user profile, the task of off-topic search detection can produce less false positives, i.e., reducing unfair accusation and scrutinization of the users.

We assume that, for each authorized user, a profile exists that defines his/her legitimate scope of information need (task), and like in [21], remains relatively constant for lengthy periods of time. This profile is either assigned, for example, as in the case of an information analyst, or learned over a period of use, as, for example, described in [3, 9]. Regardless of how the profile is determined, misuse (off-topic search) is defined as a user querying the information retrieval database for material that is not relevant to his/her profile.

Misuse detection techniques can generally be categorized into *system based* and *content based* approaches. *System-based* approaches rely on system characteristics to detect a deviation from normal behavior. For example, a user that seldom accesses a calendar file might cause an alarm, if he/she accesses this same calendar file 25 times during a session, since this is a clear deviation from the user's standard behavior. *Content-based* approaches verify that the content being accessed matches a perceived valid scope of information need. For example, a law enforcement officer whose valid scope of information need is "SEC insider trading" may cause an alarm, if the documents retrieved focus on "gun trafficking". Although ideally the officer may need to access information on gun trafficking, if such access is not part of his/her daily activities, such behavior might indicate potentially inappropriate actions. Our efforts are *content-based* detection schemes as they evaluate the document content rather than its system characteristics, e.g., name, size, storage location, usage, etc. We note that the described detection scheme issues ***suggestions*** of inappropriate access in the form of ***warnings*** and not ***accusations***. Ultimately, the final decision of whether a violation of access rights occurred is that of a human in the loop, namely the human who monitors and evaluates the validity and severity of each of the generated warnings.

II. RELATED WORK

The topic of off-topic search detection, particularly as it relates to information retrieval systems, is only of recent interest. This is a two-phase process, namely profile creation followed by on-line continuous detection. The first phase, namely building the user profile (or "user model") is an active research topic in past recent years in different research communities, such as information filtering [18], collaborative Web search [25], Personalization [19, 20, 22], Peer-to-peer information retrieval [12], and misuse detection [4, 6]. User profile represents the valid scope of information need (tasks) of a user.

For every user or user group (or role, as for role based access control structures), a profile must be established. A profile can be created manually, i.e., simply obtained by definition. That is, a system administrator or a job supervisor based on the job can simply assign each user or user group a "ready built" profile; or it can be generated automatically from user queries and feedback terms extracted from documents (pseudo feedback, implicit feedback, explicit feedback). Thus, commonly, the profile is bag-of-words, mostly weighted, or it consists of pre-defined ontologies. The profiles may be modified over time by either a system administrator, automatically, or a hybrid-approach that combines the two. We, like in [6], continue to build profile with query terms and terms from the retrieved documents. We also incorporate the terms defining the task description (interest) of a user. The summary of this process is given in section 4.

The second phase, namely the detection phase is based on dissimilarity of the information accessed by a user to his/her profile. An approach based on information retrieval relevance feedback technique that targeted towards a high recall (detection rate) was presented in [6, 7]. The query terms and query feedback terms are used to measure the similarity (dissimilarity) of a user query with one's profile. A detection rate of 96-98% with a corresponding detection precision of only around 69-72% depending on the query length and system characteristics was reported. Note that such a low level of detection precision is potentially dangerous as it might introduce many false positives subjecting the users to unwarranted scrutiny. Our contribution in this paper is enhancing the detection precision by using *clustering query results* and comparing the "good" clusters, as user search representative, to the user profile.

Elovici et al. [5] and Last et al. [10] use also clustering. However, they use clustering to group users' Web search results to form user profiles to perform anomaly detection. Similarly, the authors in [12, 24] cluster the queries so as to group past queries thus modeling user interest. In contrast, we cluster query results in the detection phase to detect the users' off-topic search. Thus, clustering is not used to build the profiles but to detect such off-topic search (misuse).

Among other efforts in off-topic search detection is an ontology-based approach. The access to a document is considered illegitimate if a user's profile does not have a semantic association with the documents retrieved by the search [1]. Symonenko et al., in [21], propose a fusion-based (hybrid) approach to detect the off-topic search of information content. By fusing role based monitoring methods, social network analysis, and semantic content analysis, an approach for detecting inappropriate information exchange is developed. A continuation of the effort described in [21], a natural language processing approach involving entity tagging for detecting off-topic search is presented in [26], and the authors favorably compare their approach against a described "bag of words" solution. The reported accuracy for their approach is similar to an earlier approach, using relevance feedback (called RF2), reported in [6]; however, due to the use of different collections to evaluate their approaches, the comparison across systems is not easily accomplished.

III. OUR APPROACH

The second phase, as mentioned in the previous section, detects misuse based on dissimilarity of user's retrieved information to one's profile. We are interested to obtain a representative for such information to facilitate the detection. Thus, the retrieved documents to the user query are of our interest. However, a user's query that is genuinely on-topic still may retrieve documents that are off-topic. Earlier Lu [11] showed that clustering similar documents together and then choosing the largest two clusters, as the final retrieval results of the user query, improve retrieval results for the *search* task. The reason for picking the largest two clusters lies in the topic similarity of larger clusters to the query.

Our motivation in clustering query results to detect off-topic search lies in the well known fact that even highly accurate search engines, as part of the retrieval process, return documents that are off topic. By clustering the documents and then only considering documents from the ***best*** clusters, the rate of false positives is reduced (e.g., higher detection precision). We propose to use terms chosen from "good" clusters (we will define "good clusters" later in the section) consisting of top retrieved result sets to the user query.

As addressed in Section 2, for every user or user group (or role, as for role based access control structures), a profile must be established. Our profile is a set of words and phrases, collectively referred to as *terms*, which accurately, to the extent possible, portrays the valid scope of interest (task) of the user. In the case that the profile is hand-crafted (a system administrator or a job supervisor based on the job simply assigns each user or user group a "ready built" profile), such a profile is task oriented and not system developed; we forgo further discussion of this approach. If a profile is to be created by the system rather than assigned, there ultimately must be a "human in the loop" during the creation phase. Users issue queries that are monitored by the system's administrator. Words and phrases from retrieved documents that are deemed as valid by the monitor along with the query terms determine the profile. The detailed description of this process (building user profile/user model) is given in [6], and we simply adopt it as our approach. A summarized description of this process is provided in Section 4 of this paper.

It should be noted that a profile represents strictly the defined valid scope of interest. Only positive examples are included. A strictly positive "training" example is possible since, by definition, all non-matching documents are considered as *possible* off-topic. We again note that non-matching retrieved documents are not necessarily off-topic since there might be many valid reasons why they were retrieved. Thus, the described approach only suggests *possible* off-topic search by issuing weighted ratings (level of deviation from on-topic search). As stated earlier, the ultimate decision is left to the system monitor (administrator).

Once a profile is obtained, the detection phase commences. That is, query results are continuously monitored to guarantee strict adherence to valid access. In operational mode (detection phase), for each query issued, the top ranked *d* returned documents are clustered. From the top resulting clusters (the selection of top is discussed below), the top *t* terms from each document belonging to those clusters are picked. Top terms are selected based on tf-nidf (normalized *tf-idf)* weighting. *tf-nidf* considers both the frequency of the term T_j in a given document D_i and the uniqueness of that term in the whole collection to select a term. The values are normalized between zero and one [8]. The selected terms represent the retrieved content. These terms are compared against the user's profile to determine the level of potential off-topic search.

As for the "top clusters", the premise we use is that the most populated clusters among the generated clusters represent the topic of the search. That is, the smaller clusters are considered as "noise" and non-representative of the results obtained. This assumption is based on prior search related work [11] that demonstrated that only retaining the top few clusters improves search accuracy.

The clustering algorithm we chose is an agglomerative hierarchal clustering algorithm [8]. Hierarchical clustering algorithm is an order-independent algorithm, meaning, it generates the same clusters when repeated, and also generally provides internally tighter and more disjoint clusters, namely is viewed as more accurate than many of the more efficient clustering approaches. The negative associated with hierarchical clustering is its runtime complexity. Since we cluster only the result set, only a small number of documents are involved. Thus, the higher computational complexity is not operationally prohibitive for our task.

Having selected a result clustering approach, we needed to determine which clusters to retain as the representative of the search topic of the user. There are several parameters that are taken into effect. First, clusters with relatively larger number of documents are of interest as representative of search topic. To guarantee larger cluster sizes, we defined a threshold $\sqrt{d}$ (*d* is the number of top retrieved documents to be clustered), as the threshold of the hierarchical algorithm for the number of clusters to be generated. Second, we eliminate "noise documents" by only retaining selected "better" clusters. The question that remained was, "Which and how many are the better clusters?" Towards this end, we experimented with various methods to determine which document clusters to keep. Our goal was to choose the clusters that are most representative of the intent of the user's query. The two properties we investigated are 1) the similarity of the cluster centroid, as the topic representative of the cluster, to the query, and 2) the number of documents in the cluster. By comparing the *similarity* of the cluster centroid to the query, we attempted to choose the cluster(s) whose "topic" is most similar to that of the query. We also chose cluster *size* as a property to determine the quality of a cluster. Since the clusters contain top ranked documents, it can be said that the more documents that are clustered as similar together, the stronger is the indication that the cluster represents the topic of the search. Third, the number of clusters to keep is also a parameter to be tested. We experimented to identify the number of clusters that leads to the best accuracy. Thus, starting with one cluster, we continued adding clusters to define a threshold that adding more clusters did not help the accuracy improvement. As mentioned earlier, both size and the similarity of cluster to the query, as two factors, independent or in combination, are considered. Based on these criteria we define methods for our "good" cluster selection (listed in Table 1). We elaborate on each method when we present the corresponding accuracy results in Section 5. The similarity between a cluster and the query is measured based on a distance measure using the *tf-nidf* (term frequency and normalized inverse document frequency) weight.

TABLE 1
DEFINITION OF METHODS

Methods	Definition
CR1a	Keeping the largest cluster.
CR1b	Keeping the cluster with the highest similarity to the query.
CR2a	Keeping the two largest clusters.
CR2b	Keeping the largest cluster, and the second or third largest cluster, based on which is more similar to the query.
CR3a	Keeping the three largest clusters.
CR3b	Like CR2b, keeping the largest cluster, but keeping two of the three next largest clusters based on their similarity to the query.
CR3c	Keeping the three largest clusters. Ranking them based on their similarity to the query, and using that rank as a weight function. The cluster with the highest rank has all of its terms count as usual (term*1), the second highest ranked cluster is (term*2/3), and the third highest cluster is (term*1/3).
CR4	Like CR3c, but with the 4 largest clusters.

IV. EVALUATION

To accurately evaluate the performance of our approach, we used a previously published testbed, evaluation framework and measure for off-topic search detection [6]. The detailed description of building user profiles /user model is given in [6]. We provide the readers a summarized overview of this dataset, profile creation and evaluation:

A. Data Set & Profiles

The testbed given in [6] includes the profiles, documents, and queries. The data that were used to create this testbed are the NIST TREC 2GB dataset [23] containing roughly half a million documents (disks 4 & 5), and a hundred queries, (query numbers 301-400) with both short (Title) and long

(Descriptive) descriptions. These TREC queries were manually separated into subject categories and cover different areas of interest such as crime, security, disaster, medicine, biology, economy, business, politics, environment, etc. To create each of the 13 profiles for experimentation and evaluation, several of the aforementioned categories were chosen "on topic" for a user. As mentioned, each category is mapped to a subset of the hundred queries. Thus, based on each of the assigned categories of each user, a subset of queries is chosen for the user's search. Those queries were then issued, and the terms from the top ranked documents were retained. Those terms (query terms and feedback terms from top documents) were then stored as a bag of words, which models the user's interest.

B: Evaluating Detection Outcome

As described in [6], five levels of rating the off-topic search (misuse) are considered, namely the detection outcome according to user's search deviation from a valid profile. The five-level is determined based on human evaluators. The levels are "off-topic" (L_5), "probably off-topic" (L_4), "undetermined" (L_3), "probably on-topic" (L_2) and "on-topic" (L_1). The distribution of levels is 40.9% for "off-topic" or "probably off-topic", 49.3% for "probably on-topic" or "on-topic", and 9.8% for "undetermined".

The ranking levels generated by the detection system, i.e., *predicted level*, are compared against the *actual level*. Two modes of detection are defined, namely *tolerant* (the uncertainty about a search being off-topic or not is not a penalty in evaluating the accuracy of the system; and *stringent* (the uncertainty about a search being off-topic or not is considered a penalty in evaluating the accuracy of the system). The formal specification for each mode of operation is defined in the contingency matrices shown in Tables 2 and 3. The rows of tables indicate the human evaluated off-topic search levels (actual) and the columns indicate the system predicted off-topic search levels.

As shown in *Tolerant* mode, Table 2, any query that is labeled by human evaluators (actual level) as off-topic L_4 or L_5, and is classified by the system as such is considered a *true positive* (TP). Similarly, a query that its actual level is L_3, indicating uncertainty, and is classified as L_4, or vise versa (L_4 as actual level and classified as L_3), is likewise a *true positive* (TP*)*. All other cases that the actual level is L_4 or L_5 are considered as *false negative* (FN). These are the cases that the system does not correctly predict the off-topic searches. Likewise all predictions that indicate level L_4 or L_5 that are not covered by the above are considered as *false positive* (FP). These are all actual levels L_1 or L_2 that are predicted falsely as L_4 or L_5; and those that their actual level is L_3 but predicted falsely as L_5. All other cases (actual levels L_1, L_2, and L_3 predicted as any of L_1, L_2, and L_3) are *True negatives* (TN) and are not considered as they are legitimate use and are considered as such.

TABLE 2
CONTINGENCY MATRIX - TOLERANT

Tolerant		**Prediction**				
		L5	**L4**	**L3**	**L2**	**L1**
Actual	**L5**	TP	TP	FN	FN	FN
	L4	TP	TP	TP	FN	FN
	L3	FP	TP	TN	TN	TN
	L2	FP	FP	TN	TN	TN
	L1	FP	FP	TN	TN	TN

TABLE 3
CONTINGENCY MATRIX - STRINGENT

Stringent		**Prediction**				
		L5	**L4**	**L3**	**L2**	**L1**
Actual	**L5**	TP	TP	FN	FN	FN
	L4	TP	TP	FN	FN	FN
	L3	FP	FP	TN	TN	TN
	L2	FP	FP	TN	TN	TN
	L1	FP	FP	TN	TN	TN

In the *stringent* mode, Table 3, one difference with the tolerant mode is where the actual level is L_4 and the system predicted level is L_3. This case, in the tolerant mode, was considered as a true positive (TP) and in the stringent mode is defined as *false negative* (FN). The second difference is where the actual level is L_3 and the system predicted level is L_4. This case, in the tolerant mode, is considered as a true positive (TP) and in the stringent mode as false negative (FN).

We evaluate the accuracy of our detection system using the standard metrics of recall, precision, and F_1-measure [8]. *Recall* defines the rate of detection, i.e., the ratio of the detected off-topic search to all occurred off-topic searches. As all occurred off-topic searches, i.e., TP+FN equals to 100%, Recall is 1-FN. *Precision* of detection is defined as the ratio of the cases detected correctly as off-topic to the total of the true and false detections. Similarly, *Detection Precision* is 1-FP. The F_1-measure combines both the detection precision and detection rate (recall) measures with equal weighting.

$$\mathrm{Re}call\left(R\right)=\frac{TP}{TP+FN} \tag{1}$$

$$Detection\ \ \mathrm{Pr}ecision\left(P\right)=\frac{TP}{TP+FP} \tag{2}$$

$$F_1-measure=\frac{2PR}{P+R} \tag{3}$$

V. EXPERIMENTAL FRAMEWORK AND RESULTS

For each of our eight methods presented in section 3, we considered thirty configurations, i.e., fifteen based on short (*Title*) queries and fifteen based on long (*Descriptive*) queries. Each set of fifteen configurations is constructed based on the combination of using top d=50, 100, 500 documents from the result set for clustering; and top t=10, 20, 30, 40, 50 ranked terms for each document in the retained top clusters.

The statistics on the number of documents in the selected top clusters is as listed in Table 4 (a, b). The longer queries (b) cover a wider scope of topics; hence their maximum cluster size is smaller than the maximum size of clusters for shorter queries (a).

TABLE 4 (a, b)
POPULATION OF THE TOP CLUSTERS BASED ON TOP *d* RETRIVED DOCUMENTS FOR (a) TITLE, AND (b) DESCRIPTIVE QUERIES

Clustered top *d* retrieved documents	**Top Cluster Size for *Title* (short) queries**		
	Min	*Max*	*Average*
50	8	36	18
100	15	60	26
500	41	120	60

Clustered top *d* retrieved documents	**Top Cluster Size for *Descriptive* (long) queries**		
	Min	*Max*	*Average*
50	12	20	16
100	16	59	37
500	60	76	68

Each configuration represents a stand-alone system, and in practice, a system administrator would select one configuration for deployment.

A. RESULTS AND ANALYSIS

We now describe our results as shown in tables 5-8. In Tables 5 and 6, we depict the best detection precision with its corresponding recall for long (*descriptive*) and short (*title*) queries, respectively, using the *tolerant* mode. In Tables 7 and 8, we depict the same results but for the *stringent* mode.

TABLE 5
DESCRIPTIVE QUERY IN TOLERANT MODE

Algorithm	P	R	F_1
CR1a	87.35	83.24	85.25
CR1b	86.13	81.73	83.87
CR2a	88.63	83.74	86.12
CR2b	88.38	84.38	86.33
CR3a	88.48	84.88	86.64
CR3b	88.23	83.99	86.06
CR3c	89.20	82.97	85.97
CR4	89.50	83.24	86.26

TABLE 6
TITLE QUERY IN TOLERANT MODE

Algorithm	P	R	F_1
CR1a	81.96	85.61	83.75
CR1b	79.81	84.11	81.90
CR2a	83.50	85.50	84.49
CR2b	82.53	86.83	84.63
CR3a	83.08	85.82	84.43
CR3b	83.31	85.01	84.15
CR3c	83.86	84.47	84.16
CR4	83.83	86.13	84.96

TABLE 7
DESCRIPTIVE QUERY IN STRINGENT MODE

Algorithm	P	R	F_1
CR1a	78.43	91.25	84.36
CR1b	76.63	91.96	83.60
CR2a	78.48	92.76	85.02
CR2b	78.69	93.20	85.33
CR3a	78.31	91.00	84.18
CR3b	78.45	92.26	84.80
CR3c	79.19	92.13	85.17
CR4	79.59	92.67	85.63

TABLE 8
TITLE QUERY IN STRINGENT MODE

Algorithm	P	R	F_1
CR1a	72.39	91.04	80.65
CR1b	71.14	90.88	79.81
CR2a	74.50	91.30	82.05
CR2b	73.61	92.64	82.04
CR3a	73.05	92.01	81.44
CR3b	74.25	91.81	82.10
CR3c	74.35	90.69	81.71
CR4	74.52	91.55	82.16

Our initial cluster selection experiment (CR1a) involved only the largest cluster. While our initial results were promising, we were unsure if keeping a cluster based solely on size was the best way to determine the best documents. We next tried a method (CR1b) that computes the centroid of each cluster, and compares it to the query. This proved to be a worse way to detect off-topic search, as precision and recall, and hence consequently the F_1-measure, dropped for both the stringent and tolerant evaluation plans. The only exception to this pattern was that the recall of the descriptive query in the stringent mode improved slightly.

As stated earlier, our focus is to minimize the number of false positives generated by the system particularly since false positives introduce unwarranted scrutiny to the user. Thus, in terms of detection precision and recall, our aim is to improve precision without significantly worsening the recall. We therefore, for the approach in which we retained terms only from documents from within a single cluster, we chose the CR1a approach, namely the processing of documents only from the largest cluster.

We then evaluated the effect of using more clusters to detect off-topic search. Our next method (CR2a) was similar to (CR1a), except that it keeps the two largest clusters rather than only the top cluster. This resulted in a statistically significant improvement over (CR1a) in precision and recall in the tolerant mode (99% confidence level), and statistically equivalent or better than (CR1a) in stringent evaluation plan. Since keeping two clusters was an improvement, we expanded our experimentation using two clusters. Our next method (CR2b) retains the largest cluster and the second or third largest, based on which cluster has a greater similarity to the query. The observed performance of (CR2b) was equivalent to that of (CR2a). That is, retaining terms from documents located within the two largest clusters was equivalent to retaining terms from documents located within the largest cluster or in the second or third largest cluster depending on which of these two clusters is more similar to the query.

Since adding a second cluster improved our results as compared to a single-cluster approach, our next method retained yet an additional third cluster; the first of the three-cluster approaches retains the three largest clusters (CR3a). Overall, this method performed similar to keeping two clusters. We then tried a similar method to (CR2b), except in this method (CR3b) we pick the largest cluster and two of the three next largest clusters based on which ones have better similarity to the query. This method performed equivalent to (CR2a), (CR2b), and (CR3a).

Our next method (CR3c) used both the cluster size as well as cluster similarity in computing off-topic search. We chose the three largest clusters, and then computed their centroids. Based on the centroids' similarity to the query, the clusters were ranked. Their score was computed by weighing the terms as follows:

$$Weighted \quad Score \;= term \;\times \frac{1}{rank} \tag{4}$$

The observed performance of (CR3c) was likewise equivalent to the previous four multi-cluster based methods (CR2a, CR2b, CR3a, and CR3b). We then tried one more variation (CR4) in which we kept the 4 largest clusters, and similar to (CR3c), the clusters were ranked based on their similarity to the query. The performance was equivalent or minimally better than using two or three clusters.

In summary, based on further analysis, using two clusters (CR2a and CR2b) performed better than using one cluster, namely (CR1a and CR1b). The remaining approaches (CR3a, CR3b, CR3c, and CR4) were statistically equivalent to (CR2a and CR2b). Thus, for simplicity and to reduce the number of terms processed, we use (CR2a or CR2b), namely, either of the two- cluster retention approaches, to detect the misuse of information systems.

Having developed a clustering based approach, namely, retaining top terms from documents stored in the two largest or the largest and the most similar cluster from within the next two largest clusters, we compared these clustering approaches (CR2a and CR2b) with the results published using the RF2 relevance feedback approach [6] running on the same data set. *B*ased on their F1 measure, a statistically significant improvement is observed for *clustering query results* in the stringent mode of evaluation plan. In tolerant mode, an equivalent or better F1 measure is observed for CR2 as compared to RF2, based on the query length. Note that in both cases, stringent and tolerant, the detection recall is higher in RF2 and detection precision is higher in CR2.

TABLE 9
COMPARISION OF RELEVANCE FEEDBACK (RF2)
METHOD WITH CLUSTERING QUERY RESULTS
METHOD – TOLERANT MODE

Alg.	Tolerant					
	Descriptive Query			Title Query		
	P	R	F_1	P	R	F_1
CR2a	88.63	83.74	86.12	83.50	85.50	84.49
CR2b	88.38	84.38	86.33	82.53	86.83	84.63
RF2	79.4	94.1	86.13	75.70	93.60	83.70

TABLE 10
COMPARISION OF RELEVANCE FEEDBACK (RF2)
METHOD WITH CLUSTERING QUERY RESULTS
METHOD – STRINGENT MODE

Alg.	Stringent					
	Descriptive Query			Title Query		
	P	R	F_1	P	R	F_1
CR2a	78.48	92.76	85.02	74.50	91.30	82.05
CR2b	78.69	93.20	85.33	73.61	92.64	82.06
RF2	72.40	90.60	80.48	72.30	90.20	80.26

Other observations were likewise noted. For example, it was observed that the results for clustering 500 documents were inferior to those generated by clustering 50 and 100 documents; thus, they are not reported here.

The results obtained using both the *tolerant* and *stringent* matrix demonstrate that for both *descriptive* and *title* queries, as the number of terms is increased, the detection precision increases. This is expected, as a greater number of terms, up to a given threshold, more accurately represents the essence of the document. Surpassing this threshold results in no longer adding "top terms" but rather adding non-discriminatory terms, terms that do not differentiate between documents.

Note that we have not addressed the trend as it relates to number of top documents selected. This is intentionally so. The reason is that since we filter the documents used in the clustering phase by only retaining the documents in the larger clusters, there is no guaranteed subset principle. That is, the documents retained after the clustering phase when 50 documents are used initially are not all necessarily retained when 100 documents are used at the beginning of the clustering phase. Hence, given our current approach, it is not possible to generalize the effects of clustering larger initial document sets.

VI. Conclusion

We presented a content-based off-topic detection approach that uses *query result clustering* to detect off-topic searches. Off-topic search or misuse in information retrieval systems violates the notion of "need to know" that is important to many organizations. We empirically showed that clustering query results approach supports a lower rate of false positives (alarms) while maintaining a good rate of detection. We proposed and evaluated various methods for clustering the query results and showed that using the largest two clusters is a good choice as far as accuracy and simplicity.

Although not addressed in any effort including our own, we believe that fusing high-precision systems with high-recall approaches should yield a detection engine that supports a vastly better F_1-measure than currently exists. As shown in [2], fusing similar systems generally does not improve the overall performance. Thus, as future work, we will fuse one of the previously developed high recall detection systems with the system proposed herein, a higher precision system. We are expanding our effort presented in this paper by incorporating the user's search behavior via a sequence of queries issued in a given window size; initial efforts are described in [13] with the current status described in [14]. We are also investigating the performance of combining both systems-based and content-based detection approaches on the overall detection rate of a misuse of information retrieval search engines.

References

[1] B. Aleman-Meza, P. Burns, M. Eavenson, D. Palaniswami, A. Sheth: An ontological approach to the document access problem of insider threat, *IEEE International Conference on Intelligence and Security Informatics (ISI),* May 2005.

[2] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, O. Frieder, N. Goharian: On fusion of effective retrieval strategies in the same information retrieval system, *Journal of American Society for Information Science and Technology*, 2004.

[3] E. Bloedorn, I. Mani, T. MacMillan: Machine learning of user profiles, Representational Issues, *American Association for Artificial Intelligence* (AAAI), 1996.

[4] R. Cathey, L. Ma, N. Goharian, D. Grossman, Misuse Detection for Information Retrieval Systems, *ACM 12th Conference on Information and Knowledge Management (CIKM*), November 2003.

[5] Y. Elovici, B. Shapira, M. Last, O. Zaafrany, M. Friedman, M. Schneider, and A. Kandel: Content-based detection of terrorists browsing the web using an advanced terror detection system (ATDS), *IEEE International Conference on Intelligence and Security Informatics (ISI),* May 2005.

[6] N. Goharian and L. Ma: Query length impact on misuse detection in information retrieval systems, *ACM 20th Symposium on Applied Computing (SAC),* March 2005.

[7] N. Goharian, L. Ma, Off-Topic Access Detection In Information Systems, *ACM 14th Conference on Information and Knowledge Management (CIKM*), November 2005

[8] D. Grossman and O. Frieder: Information retrieval algorithms and heuristics, *The Information Retrieval Series*, Springer Publishers, Vol. 15, 2nd ed., 2004.

[9] M. Knepper, K. Fox, and O. Frieder: Query improvement elevation technique (QUIET). *International Conference on Intelligence Analysis*, May 2005.

[10] M. Last, B. Shapira, Y. Elovici, O. Zaafrany, A. Kandel: Content-based methodology for anomaly detection on the Web. *Lecture Notes in Computer Science, International AtlanticWeb Intelligence Conference*, May 2003.

[11] X. A. Lu, M. Ayoub, J. Dong: Ad hoc experiments using EUREKA. Text retrieval Conference (TReC), *National Institute of Standards and Technology*, 1, (http://trec.nist.gov/).

[12] J. Lu, J. Callan: User modeling for full-text federated search in Peer-to-Peer networks, Proceeding of SIGIR 2006.

[13] A. Platt and N. Goharian, Using user query sequence to detect off-topic search, *ACM 22nd Symposium on Applied Computing (SAC),* March 2007.

[14] A. Platt and N. Goharian, Short query sequence in misuse detection, *IEEE Fifth International Conference on Intelligence and Security Informatics (ISI),* May 2007.

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

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

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

[18] S. Robertson and D. Hull: The Track-9 Filtering track report, Proceeding of TREC 2001.

[19] X. Shen, B. Tan, C. Zhai: Context Sensetive informaitn retrieval using implicit feedback, Proc. of SIGIR 2005.

[20] K. Sugiyama, K. Hatano, M. Yoshikawa: Adaptive web search based on user profile constructed without any effort without users, Proc. of WWW 2004.

[21] S. Symonenko, L. Liddy, O. Yilmazel, R. Del Zoppo, E. Brown, M. Downey: Semantic analysis for monitoring insider threaths, *IEEE International Conference on Intelligence and Security Informatics (ISI),* 2004.

[22] J. Teevan, S. Dumais, E. Horvitz: Personalizing search via automated analysis of interests and activities, Proc. of SIGIR 2005.

[23] Text retrieval Conference (TReC), *National Institute of Standards and Technology*, http://trec.nist.gov/

[24] E. Voorhees, N. Gupta, B. Johnson-Laird: Learning collection fusion strategies, Proceeding of SIGIR 1995.

[25] J. Wen, J. Nie, H. Zhang: Query Clustering using user logs, ACM Transactions on Information Systems, 20(1), 2002.

[26] O. Yilmazel, S. Symonenko, N. Balasubramanian, E. Liddy: Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content. *IEEE International Conference on Intelligence and Security Informatics (ISI),* May 2005.