On the Usability of Clustering for Topic-oriented Multi-level Security Models

Paal E. Engelstad
Norwegian Defense Research Establishement (FFI), Kjeller, and
Oslo and Akershus University College (HiOA), Oslo, Norway
e-mail: paal.engelstad@{ffi,hioa}.no

***Abstract*—Security levels used in organizations today are typically course-grained, broad and distinct, using security levels such as "Confidential" and Secret". However, current research is advocating a move towards more fine-grained security models, e.g. such as Attribute-Based Access Control, where information objects and end-users are characterized in terms of complex meta-data. One idea promoted is a topic-oriented approach where information objects are characterized in terms of fine-grained descriptions of the topics of its content. It will lead to higher flexibility, but will also rely on a policy-database to assign a specific security policy to topics and subtopics. Due to increased complexity, it will also require automatic or semi-automatic tools for determining the topics and sub-topics of information objects, and the tools should extract topics that are easily understood by humans, since humans need to control the policy. This paper studies the feasibility of using clustering techniques to help humans in extracting the topics from information objects. A number of clustering methods are discussed, including k-means, Wards hierarchical agglomerative clustering, Correlated Topic Models (CTM) and Latent Dirichlet Allocation (LDA). To the best of our knowledge, an in-depth analysis on the feasibility of using clustering for this problem has not been presented in previous work. Our analysis points out challenges with clustering in particular, which must be addressed before realizing the general vision of topic-oriented policy-driven security models.**

Keywords–Multi-level security, topical clustering, policy-driven, cross-domain information exchange, machine learning.

I. Introduction

Security labels are used by the military, government agencies, international organizations and private corporations to associate security attributes to a specific information object [1]. In a military setting, examples of such security labels include categories such as "Unclassified", "Restricted", "Confidential", "Secret" and "Top Secret", while some private enterprises may use labels such as "public" and "business internal". In modern environments the information objects are digital information, such as word documents, text messages and e-mails. To attach a digital label to the object, various types of digital labeling technologies can be applied (e.g. the XML Confidentiality Label [2]). The security label can be digitally attached and bound to the data object, e.g., by using a cryptographic mechanism such as a digital signature.

The security classification and labeling is useful to determine how information objects shall be handled, as there is a policy associated with each classification level. The policy determines: who is allowed to access the information objects (according to the persons security clearance); in which systems the information object can be stored; to which systems the information object can be sent; and so forth.

At the same time, current research is advocating a move towards more fine-grained topic-oriented security models, where the label might contain extensive information describing the topics covered by the content of the object. It will lead to higher flexibility, but will also rely on a policy-database for easy management of all the specific policies of different topics and subtopics. Furthermore, with the ever-growing amount of digital information, undertaking the actual labeling of information objects is getting more and more a tremendous challenge. Going from course-grained classification levels to fine-grained descriptions of topics of documents will probably increase the workload with an order of magnitude. Thus, the vision of a topic-oriented, policy-driven solution will rely on the existence of good automatic or semi-automatic tools for determining the topics and sub-topics of information objects.

This paper studies the feasibility of using clustering-techniques to address this. Not only are the tools required to extract topics effectively without many errors. The tools should also extract topics that are easily interpreted by humans, since humans will need to control the policy associated with the topics. For instance, if a human wants to implement a policy rule that down-grades all documents belonging on specific topic into unlimited public use, the clustering mechanisms need to be able to form clusters that clearly and consistently separate this topic from the other topics in the data set.

II. Background

A. Related Work

So far, related work has mostly focused on frameworks and architectures for realizing the vision of fine-grained topic-oriented and policy-driven security models. For instance, Kongsgård et. al. [3] provide a framework that allows

This work was supported by the University Graduate Center (UNIK).

978-1-5090-0206-1/15 $31.00 © 2015 IEEE
DOI 10.1109/EMS.2015.13

for security labels that contain complex information (metadata) about an information object. An overview of some important aspects of the framework is outlined in Fig. 1. The architecture is designed for determining which security label attributes are to, and should not, be included within a given data objects security label according to policy. They present a solution for the use of Attribute Based Access Control (ABAC) principles to the process of information labeling [4]. In particular, the framework provides support for pluggable attribute modules (e.g., content checkers) whose output serve as input to the policy decision.

Content-based Protection and Release (CPR) [5] represents a variation of ABAC that has been proposed for future use in NATO. In CPR, the attributes within a content label are used to convey the properties of an information object. Access decisions are then based on protection and release policies effectively expressing requirements (in terms of attributes) on the user and her terminal and/or environment in order to be granted access to information objects with such properties. CPR depends on the ability to assign content properties to information objects, and the work proposed in this paper is therefore relevant.

Other works, such as [6], [7], [8] and [9], focus more on the specific firewalls (usually referred to as *guards*) that are used to enforce the policy in cross-domain information exchange.

To the best of our knowledge, the current paper is the first work to consider clustering as a tool for topic selection and topic classification in the context of multi-level security. The contribution of the paper is not only to clustering for on the research agenda in general as a potential solution to this problem We also go one step further and deeper, and explore to which extent it is credible that clustering can be applied to a fine-grained topic-oriented security model of the future.

B. Supervised vs unsupervised methods

Clustering represents a family of unsupervised machine learning methods, designed to extract various features of a data set and group the different observations together into different clusters, e.g. for further analysis and processing. Examples of popular clustering technologies that are relevant for topical clustering includes k-means [10], Wards hierarchical agglomerative clustering [11], Correlated Topic Models (CTM) [12] and Latent Dirichlet Allocation (LDA) [13]. These will be discussed below. (Due to space limitations, we will only go into the details of a clustering method whenever necessary for our analysis. The reader is referred to the references for further details.)

Typical for the majority of clustering techniques is that the selection of features that contributes the most to the assignment of different observations into different clusters is not controlled in a supervised way. After the clusters have been formed, there are a number of supplementary methods

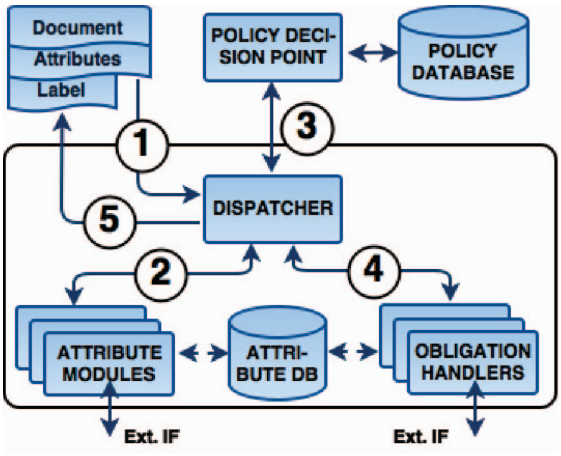

Figure 1. A flexible framework for trusted policy-based data labeling. Due to space limitations, the reader is referred to [3] for a detailed explanation of the framework.

that can be used for labeling the clusters, such as associating meaningful keywords to each cluster (e.g., see [14] or

In this paper we analyse if clustering methods are able to group together documents into clusters that correspond to topics that are easily conceived by humans as specific topics. The reason for this requirement is that the security policy will set rules associated with a specific topic. If the clustering technique defines topics that are not easily understood by humans, it is difficult to associate meaningful security policies to the machine-generated topics. The focus is not on the subsequent labeling. Here we assume that an ideal tool with perfect performance (which of course does not exist in reality) to do this exists. This assumption allows us to focus separately on the potential quality of the clustering part.

In contrast to clustering, there is also a family of supervised machine learning techniques, typically used for classification. Examples includes Naïve Bayes, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Lasso, etc. (Cf. [8] for examples of using these for automatic classification of course-grained classification levels.) An alternative approach to clustering can be to manually assign topics and subtopics to various parts of each documents in a training set, and use supervised classification techniques to train a machine learner. Then, the machine learner can subsequently be used to assist a human in the topical classification of new documents. After the correctness of the topical classification of a new document has been controlled by a human, the new document is added to the training set, and the machine learner is trained again with the expanded training set, and so on. First, we expect that the work load of this approach will be much higher. Second, even with a method of gradually increasing the training set with supervised classification, clustering techniques might be a

necessary complimentary tool to detect new topics that emerges in new documents over time and to make sure that the new topics are included in subsequent classification.

Thus, the main focus of this paper is on clustering.

C. The k-means clustering method

We select the *k-means* clustering method as a staring point for our analysis in this paper, since it is intuitive and simple to analyze. A compelling features of k-means in terms of interpretation is that how strongly an observation relates to a cluster centroid is given by a squared distance metric.

Assume k-means works on a training set of i observations (documents), $\{x^{(1)}, \ldots, x^{(n)}\}$, where $x^{(i)} \in R^n$. K-means starts by defining k different centroids, $c \in \{1, \ldots, k\}$, each initialized with random "positions" $\mu_1, \ldots, \mu_k \in R^n$. During each repeated iteration of k-means, the algorithm - as a first step - minimizes the distortion by assigning each observation $x^{(i)}$ to a fixed centroid. For every training example document , assign it to the closest centroid μ_j. For convenience, we define an assignment vector, $c = c^{(1)}, \ldots, c^{(n)}, where\ \mathrm{c}^{(i)} \in \{1, \ldots, k\}$. Then, the assignment of document i, $c^{(i)}$, in is found by:

$$c^{(i)} := \arg\min_j \parallel x^{(i)} - \mu_j \parallel^2 \tag{1}$$

Then, in the second step of the iteration, k-means minimizes the distortion further by keeping the assignment fixed, and adjusting the position of each centroid to a new value. For each j, the new position , μ_j, is calculated as a mean value of the positions of the observations that were assigned to the centroid in the first step:

$$\mu_j := \frac{1}{\|\{i : c^{(i)} == j\}\|} \sum_{\{i:c^{(i)}==j\}} x^{(i)} \tag{2}$$

The objective function of k-means is the distortion, J, defined as:

$$J = \sum_{j=1}^{k} \sum_{\{i:c^{(i)}==j\}} \parallel x^{(i)} - \mu_{c^{(i)}} \parallel^2 \tag{3}$$

Since it is minimized in both the first and second step of each iteration, it must monotonically decrease for each iteration, and both the document assignments and the centroid positions will converge to fixed values.

III. Clustering with three main topics ($k = 3$)

A. Analysis of a corpus with three distinct topics

In order to test the ability of clustering to extract topics that are easily conceived by humans, we select a set of documents within three different topics. The documents need to be pertinent to the security domain. Therefore, we have used documents from the Digital National Security Archive (DNSA). This archive contains the most comprehensive collection of declassified US government documents available to the public [15].

The collection includes three unrelated and quite specific topics that we find useful to use for our analyses:

- "AF": Afghanistan: The Making of U.S. Policy, 1973-1990
- "CH": China and the United States: From Hostility to Engagement, 1960-1998
- "PH": The Philippines: U.S. Policy during the Marcos Years, 1965-1986

These topics are easily conceived as very different by humans, relating to different countries, different time periods, different political concerns, different actors involved, and so forth.

For our experiment, we have 2793 documents available, where 1955 of the documents (approximately 70%) are used for the clustering, i.e. these are the documents that were used for most of our analyses. The remaining 838 documents we spared as an independent set, whenever we needed that to investigate our results in further depth (like a test set in supervised learning). All three topics contain an arbitrary mix of both classified and unclassified documents. For picking out the actual documents, we used the same selection criteria that were used in [8].

An advantage of grouping together documents from three different topics in this way is that we have a labeled corpus with respect to the three topics, i.e. we have control with the topic that each document belongs to. We can use this knowledge to study the efficiency of the clustering in detail, and we might apply supervised methods and techniques to learn more from our analyses. However, each of the three topics contains most probably a number of "subtopics", and subtopics might contain a number of "sub-sub-topics" and so forth. The labeling of such subtopics is beyond the reach and scope of our analyses below.

For the machine learning, the document corpus was preprocessed, following a standard bag-of-words approach: We extracted the content of the documents, and formed a Document Term Matrix (DTM), where the rows correspond to the different documents (observations) and the columns correspond to all possible words (or word stems) in the corpus. The DTM is a sparse matrix where an entry in the DTM shows how many times a specific word has been used in a specific document.

For further details of this standard method for text analysis, the reader may refer to e.g. [8]. The only specific note about the analyses presented below, is that we do not make a transformation of the term frequency entries in the DTM (i.e. no tf-idf transformation), and we do not normalize the term frequencies with the document length, either [8]. The reason is that some of the clustering methods we use below, such as LDA, require integer-value entries in the DTM.

B. Naïve approach: Creating three clusters with k-means

We want to explore to which extent clustering is able to extract the three topics that are obviously distinct from a

TABLE I
SUPERVISED LABELING OF k-MEANS CLUSTERS

	#	Unlabeled - k=3			Labeled - k=3			Labeled - k=25			Labeled - k=50			Labeled - k=100		
	docs	C1	C2	C3	AF	CH	PH	AF	CH	PH	AF	CH	PH	AF	CH	PH
AF	576	2	535	39	0	39	537	123	15	438	200	266	110	235	115	226
CH	664	0	526	138	0	138	526	4	129	531	1	557	106	2	478	184
PH	715	26	637	52	0	52	663	1	25	689	0	284	431	3	133	579
Sum	1955	28	1698	229	0	229	1726	128	169	1658	201	1107	647	240	726	989
Assgnment accuracy.:		0.358			0.410			0.481			0.559			0.664		
Assgn.acc. (new docs):					0.40 (0.37, 0.44)			0.46 (0.43, 0.50)			0.57 (0.54, 0.61)			0.64 (0.61, 0.67)		

human perspective.

As a naïve approach, we apply the k-means clustering method to our DTM in order to create exactly three clusters, i.e., by setting $k = 3$ as an input parameter to k-means and cluster the documents (DTM rows) with respect to the word frequencies (DTM entries) of each document.

K-means is guaranteed to converge, since the distortion (cf. eq. 3) decreases monotonically for each iteration (Section II-C). However, it might end in a *local* minimum solution, since the distortion function is a non-convex function. Usually this is not a problem. Nevertheless, to increase the confidence that a *global* minimum is found, we run k-means a number of times, each time using different random initial values for the cluster centroid positions. (Actually, we ran the algorithm with as much as 500 different sets of initial values, since the algorithm is quick to run). Then, looking at all the different solutions, we pick the one with the lowest distortion.

The results of creating three clusters with k-means are shown in the left pane of Table I, i.e, the pane with the heading "Unlabelled - k=3". There are three clusters, "C1", "C2" and "C3", and each column corresponds to one cluster. The column shows how many documents of each topic the cluster contains, i.e. one row corresponds to one topic. Since we have a corpus of documents labeled with the topic ("AF" vs "CH" vs "PH"), it is possible to arrange the results in this way for further analysis. Let us assume that a human getting the topic clusters from a clustering tool, evaluates that "C1", "C2" and "C3" correspond to the topics "AF", "CH" and "PH", respectively. Then, the number of correct assignments are the sum of the numbers along the diagonal (i.e. $2 + 526 + 52 = 580$), and the "Assignment Accuracy", which is defined as the share of correct cluster assignments with respect to the total number of assigned documents, will then be: "Assign.acc"=$580/1955 = 0.297$. Given as a baseline that random assignment would have an expected average assignment accuracy of 0.33, it is not particularly convincing to group the clusters in this way.

The assignment accuracy depends on how humans interpret the results of the clustering, and defines how the different clusters correspond to different topics. For instance, say that the human instead label the clusters differently, so that "C1", "C2" and "C3" correspond to the topics "PH", "AF" and "CH", respectively, the number of correct assignments would increase to a value of 699, and the assignment accuracy would raise to 0.358.

C. Labeling of the topic clusters

To explore what the best assignment accuracy that the clustering method can ideally perform, let us assume that the human that will interpret the clustering results will investigate the documents in each cluster (manually or assisted by a tool), and assign a topic label to each cluster. The human counts the number of documents of each topic in the cluster, and assign the cluster to the topic with the most documents. Looking at the same example as above (left pane of Table I), ideally both "C1" and "C2" should be assigned to the topic "PH", while "C3" should be assigned to "CH". This assignment is shown in the second pane from the left in Table I, where the values of "C1" and "C2" has been aggregated into the "PH" column, while the "CH" column holds the values of "C3".

With this method, we have derived with a confusion matrix for the assignments, where the rows correspond to the actual topics and the columns correspond to the assigned topics. We observe that the "Assignment accuracy" is now increased to 41.0%, which seems to be the maximum that we can ideally achieve with the given clustering algorithm. Now, we can use the same centroids that was derived with the training set, to study how a new documents (e.g taken from the separate document sets) is expected to be assigned into different clusters after a re-clustering. Then we will derive with a similar confusion matrix, and the assignment accuracy of the new document is 0.40. (See the bottom row

TABLE II
ASSIGNMENT ACCURACY WITH VARIOUS CLUSTERING METHODS

Clustering Method	Assignment accuracy $k = 3$	$k = 100$
k-means	0.372	0.664
Ward (D)	0.381	0.474
Ward (D2)	0.379	0.571
CTM	0.367	(0.718)
LDA w/Gibbs	0.525	0.792
LDA w/VEMfixed	0.507	0.836
LDA w/VEM	0.517	0.856

of Table I, where the 95% confidence interval is shown in brackets).

D. Testing various clustering methods ($k = 3$)

It is clear that the naïve approach of expecting k-means to extract three clusters that matches well to the three selected topics is not very successful. Now, we perform the same analysis using a number of other clustering methods to create exactly three clusters. The methods selected are Wards hierarchical agglomerative clustering both with Ward's original criteria ("D2") and without ("D"), Correlated Topic Models ("CTM") and Latent Dirichlet Allocation ("LDA"). For LDA, we test it both with Gibbs sampling ("Gibbs") and with variational expectation-maximization, both with an estimated α value ("VEM") and with a fixed, default α value of 1.666 ("VEMfixed"). Results are shown in the first column ($k = 3$) of Table II. The results show that a topic-oriented model like LDA performs considerably better than generic models like k-means and Ward. However, LDA is still not capable of extracting effectively the three topics directly out of the corpus.

In Table III, we see that the most nominant terms for the clusters C1 and C3 formed by LDA-VEM are closely related to "PH" and "CH" respectively, while C2 seems to dominated by a mix of documents from both "PH" and "AF".

IV. Investigating subtopics ($k > 3$)

A. Determining the number of relevant sub-topics, k

Looking at the example in the previous section, we observed that the clustering detected "C1" as a small subtopic of "PH". The subtopic was so distinct from the rest of the documents in "PH" that this small subtopic was allocated one of the three main clusters of k-means with $k = 3$. The result is that the majority of remaining "PH" documents will be allocated to "C2" and "C3". There are two remaining clusters to hold the majority of documents from all three topics. Due to this argument, it is clear that when doing the clustering we should account for a large number of subtopics within each of the three topics. A better approach might be to cluster with a large $k > 3$, and then use automatic labeling of the clusters to see how well the clustering performs.

There are many methods to select from when first trying to estimate the appropriate number of clusters to extract [16]. Amongst the simplest methods are two rule-of-thumb rules. The first "Rule of thumb 1" is a general notion that the number of clusters should relate to the square root of the number of observations n as follows: $k = \sqrt{(n/2)}$. The second "Rule of thumb 2" is more adapted to the sparse DTM matrix of text analysis, and claims that $k = (m*n)/|\{(i,j) : DTM_i, j \neq 0\}|$, where m is the number of terms (columns) in the DTM, while $|\{(i,j) : DTM_i, j \neq 0\}|$ refers to the number of non-zero entries in the DTM. Investigating the details of the DTM, the resulting estimations of the optimal number of clusters k are shown in Table IV. Values are shown both for the entire corpus, as well as a corpus only consisting of either of the topics "AF", "CH" or "PH". A sum of the values found for the three topics are shown in the rightmost column.

Another popular approach is to study the explained variance. Increasing the number clusters, increases the explained variance (or decreases the "un-explained variance"). An "elbow" in the curve indicates that the benefit of adding more clusters starts to decrease. The relevant curves are shown in Fig. 2, and the corresponding estimated elbow values are shown in IV.

Summarizing the discussion and the results as outlined in

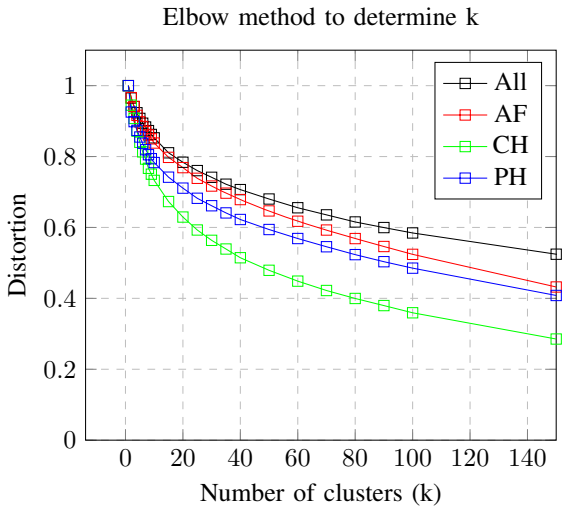

Figure 2. For each curve (corpus), the distortion is shown as a share of its k=1 value, i.e. it reflects the *un*-explained variance.

TABLE III
FIVE MOST DOMINATION TERMS PER TOPIC USING LABELED LDA-VEM CLUSTERS

Word stem	k=3 C1	C2	C3	k=100 AF	CH	PH
1	program	right	china	soviet	visit	manila
2	philippin	secretari	chines	refuge	china	philippin
3	develop	manila	soviet	amin	immedi	them
4	polic	said	militari	afghan	chines	constitut
5	project	afghan	polici	amassador	secretari	martial

TABLE IV
DIFFERENT WAYS TO ESTIMATE k

Method	All	AF	CH	PH	(AF+CH+PH)
Rule of thumb 1	31.3	17.0	18.2	18.9	54.1
Rule of thumb 2	22.8	33.1	20.2	20.1	73.4
Elbow method	∼40	∼40	∼40	∼40	∼120

Table IV, we conclude that estimating k somewhere between 25 and 50 is sensible. However, to be on the safe, we may even select k as high as $k = 100$ to be even more confident that we have included most of the explained variance.

B. Labeling of the sub-topic clusters

Having estimated roughly k as $25 \leq k \leq 100$, we use k-means again, and repeat the same analysis as for $k = 3$ above. Results for varying number of clusters $k = 25$, $k = 50$ and $k = 100$ are shown in the right-most panes of Table I. A problem of the analysis is that with the assumption of a perfect tool for cluster labeling, the assignment accuracy will continue to increase with an increasing number of clusters. (In the extreme case of having one document per cluster, the cluster labeling will force every document into the correct topic category, and the assignment accuracy will reach 100%.) Without an ideal performance, the assignment accuracy will approach the efficiency of the automatic cluster labeling method. However, we observe that when k is limited to a value of 100 or less, the assignment accuracy of a new document is a quite comparable numbers with that of the initial clustering. This gives an indication that the assignment accuracy that we measure within this region will be primarily given by the performance of the actual clustering method. The results in Table I show that it is hard to get a clustering method that will give an assignment accuracy considerably higher than 66%. (Indeed, by running up to $k = 550$, the assignment accuracy with respect to the inclusion of a new document, gave a maximum value of 69%.)

C. Performance of other clustering methods (assuming 100 different sub-topics, $k = 100$)

For the other clustering methods than k-means we perform the same procedure as in the previous subsection on a number of different methods, using $k = 100$. Results are summarized in the right column of Table II. (CTM did not converge for $k = 100$ and $k = 50$, so the value for $k = 25$ is shown instead).

We observe that topic-oriented clustering methods have the potential to achieve above 85%, which occurs for LDA-VEM. This might be sufficient performance in many scenarios. Furthermore, in the right pane of Table III we observe the five most dominant terms derived by LDA-VEM. Looking at the terms, it is not hard for a human being to associate the right groups of clusters to the right topics.

Nevertheless, relying on a high number of clusters, in a realistic scenario where techniques for automatic cluster labeling are not performing ideally (indeed, it is a quite hard problem), the actual performance will be quite lower than levels indicated in the right column of Table II.

D. Supervised machine learning alternatives

As mentioned earlier, it is possible to take an entirely different approach to the problem, reorganize the way this is implemented in an organization, and use supervised classification methods as a starting point instead (e.g., according to the proposal in Section II-B). We have argued that this supervisioned approach will require considerably higher human effort. Nevertheless, the cost must be weighted against potential benefits in terms of increased performance.

TABLE V
PREDICTION ACCURACY OF SUPERVISED MACHINE LEARNING METHODS

Machine learner	Prediction. accuracy	Confidence interval (95%)
kNN (k=1)	0.70	(0.67, 0.73)
kNN (arg max: k=22)	0.78	(0.75, 0.81)
SVM	0.96	(0.95, 0.97)
Lasso	0.96	(0.95, 0.97)

Therefore; to put the clustering performance in perspective, we used supervised multi-class classification to determine the three topics "AF", "CH" and "PH". Table V shows that one can easily achieve a prediction accuracy of 96%, using standard machine learners like SVM or Lasso (cf. [8]).

V. CONCLUDING REMARKS

Topical clustering has been assumed as a building block for realizing a vision of fine-grained policy-oriented multi-level security models. With a coupling to policy, the clustering method must form clusters that are easily interpretable by humans. For instance, if a human wants to implement a policy rule that down-grades all documents belonging the "Afghanistan" topic into unlimited public use, the clustering mechanism needs to form clusters that clearly and consistently are separating this topic from the other topics in the data set.

The most promising method was LDA/VEM, but still, further improvements should be addressed in future work. There is a big research effort on LDA, and including recent advances within this research area seems promising.

For simplicity, we assumed an "ideal"/"perfectly performing" tool for cluster labeling, to focus entirely on the clustering performance alone. However, including a complete analysis that takes both the clustering and the cluster labeling into account at the same time, should also be addressed in future work.

In summary, using clustering for the problem addressed in the paper is promising. However, considerably more work is still needed before the vision of a fine-grained topic-oriented multi-level security model can be realized.

REFERENCES

[1] R. Kissel, *Glossary of Key Information Security Terms*. DIANE Publishing Company, 2011. [Online]. Available: http://books.google.no/books?id=k5H3NsBXIsMC

[2] A. Eggen, R. Haakseth, S. Oudkerk, and A. Thummel, "XML confidentiality label syntax," FFI-rapport 2010/00961, 2010.

[3] K. W. Kongsgård, N. A. Nordbotten, and S. Fauskanger, "Policy-based labelling: A flexible framework for trusted data labelling," International Conference on Military Communications and Information Systems, 2015.

[4] V. C. Hu *et al.*, "Guide to attribute based access control (abac) definitions and considerations," 2014.

[5] K. Wrona and S. Oudkerk, "Content-based protection and release architecture for future nato networks," in *Proc. Military Communications Conference*, 2013, pp. 206–213.

[6] K. Wrona and G. Hallingstad, "Development of high assurance guards for nato," in *Proc. Military Communications and Information Systems Conference (MCC)*, 2012.

[7] R. Haakseth, N. A. Nordbotten, Ø. Jonsson, and B. Kristiansen, "A high assurance guard for use in service-oriented architectures," International Conference on Military Communications and Information Systems, 2015.

[8] P. E. Engelstad *et al.*, "Automatic security classification with lasso," *Proceedings of The 16th International Workshop on Information Security Applications (WISA 2015), Jeju Island, Korea, August 20-22*, 2015.

[9] K. Wrona, S. Oudkerk, and G. Hallingstad, "Development of high assurance guards for nato," in *Proc. Military Communications Conference*, 2010.

[10] J. A. Hartigan and M. A. Wong, "K-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, p. 100108, 1979.

[11] F. Murtagh and P. Legendre, "Wards hierarchical agglomerative clustering method: Which algorithms implement wards criterion?" *Journal of Classification*, vol. 31, pp. 274–295, October 2014.

[12] D. Blei and J. Lafferty, "Correlated topic models," *Advances in Neural Information Processing Systems*, vol. 18, p. 9931022, 2006.

[13] C. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 9931022, January 2003.

[14] A. Popescul and L. H. Ungar, "Automatic labeling of document clusters." [Online]. Available: http://citeseer.nj.nec.com/popescu100automatic.html

[15] Digitial national security archive. "http://nsarchive.chadwyck.com/home.do". Accessed: 2015-03-26.

[16] Wikipedia, "Determining the number of clusters in a data set." [Online]. Available: https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set