Technical report on

Automatic Identification of Paedophile Keywords

Measurement and Analysis of P2P Activity Against Paedophile Content project
`http://antipaedo.lip6.fr`

Christian Belbèze, David Chavalarias, Ludovic Denoyer, Raphaël Fournier,
Jean-Loup Guillaume, Matthieu Latapy [1], Clémence Magnien,
Guillaume Valadon, Vasja Vehovar, and Aleš Žiberna

Abstract

Accurate and up-to-date knowledge of keywords entered by users who search or provide paedophile content is a key resource for filtering purposes and for monitoring by law enforcement institutions. However, such keywords are often hidden and may change frequently, and our current knowledge about them relies on manual inspection and field expertise. We explore here the possibility to help in improving this situation by applying various keyword analysis methods. Using a large-scale real-world collection of paedophile and non-paedophile file names, we construct lists of keywords suspected to be used as paedophile keywords. We evaluate the relevance and interest of these lists by submitting them to experts, thus showing that automatic approaches are indeed of great interest for this task.

1 Introduction.

People interested in paedophile content use common-sense (for instance *child porn*) and specific (for instance *qqaazz*) keywords to search for files and to name them. These keywords may be unknown to other users (including law enforcement personnel monitoring paedophile activity) and kept secret in this community. In addition, new keywords may appear over time, as previous ones become well known (by law enforcement personnel, filters, and users who distribute *fakes*, *i.e.* files with paedophile names but non-paedophile content). One may also guess that paedophile users try to create keywords which are difficult to detect, and so avoid to disclose them and change them frequently.

Accurate and up-to-date knowledge of these keywords is however a key resource for law enforcement (machine inspection, online investigation, monitoring), filtering (in P2P systems or web search engines for instance), and in general for studying paedocriminal activity. This knowledge nowadays relies on domain expertise of personnel involved in fighting against paedocriminality. As a consequence, and despite their effort in monitoring this activity, people involved in fighting it still have difficulties in maintaining an accurate and up-to-date list of paedophile keywords.

[1] Contact author: Matthieu.Latapy@lip6.fr

Our aim is to apply various high-level computer science techniques in order to evaluate their ability to help in this task. We design realistic scenarii for applying these methods to automatic paedophile keyword detection. We then apply them independently and submit the obtained results to experts for assessment. The selected methods span the variety of currently available techniques for automatic keyword detection. We therefore perform, for the first time, a detailed comparison of these methods when they are applied to automatic detection of paedophile keywords. We also obtain relevant lists of paedophile keywords, which constitute a significant contribution in themselves.

2 Methodology.

This section presents the method followed to assess automatic keyword detection methods. It relies on a real-world dataset, two scenarii describing existing knowledge usable to detect paedophile keywords, and a method for assessment by experts.

2.1 Dataset.

The data used for this study is described in detail in [5]. It was obtained by a modified *eDonkey* client which sent, during approximately 150 continuous days (5 months), a set of keyword-based queries to all reachable *eDonkey* servers approximately every 12 hours. It recorded the answers to these queries, mostly lists of filenames matching the keywords it sent. In order to preserve user privacy, file names were normalised (only alphanumeric characters were kept and translated to lower case) and the words which appeared in less than 100 distinct file names in the measurement were anonymised.

In order to obtain large amounts of data related to paedophile activity, the queries sent by the client were known typical paedophile keywords: *qqaazz*, *aabbccddee*, *babyshivid*, *hussyfan*, *pthc*, *ptsc*, *r@ygold*, and *kingpass*. In addition, the client also sent queries with non-paedophile keywords: *porn*, *madonna*, *linux*, *batman*, *cnrs*, *mickael jackson*, and *sex*. In this way, data on both paedophile and non-paedophile file names has been collected.

The obtained data, used in the following, finally consists in a list of 1 250 537 distinct file names, among which 103 110 contain at least one explicit paedophile keyword (belonging to the above list).

2.2 Scenarii.

All methods rely on a file names data set and an input paedophile keywords list. Therefore, in theory, to completely assess a method, the results obtained with all possible combinations of known paedophile keywords as input lists should be studied.

In order to evaluate all automatic keywords detection methods and explore their strengths and weaknesses, while keeping the number of situations to compare reasonable, two situations were considered.

First, we suppose no specific domain knowledge, and so we use *common sense* keywords as indicators of paedophile content. We chose the following: *child* and *sex*; *child* and *porn*; and age indications: *1yo*, *2yo*, ..., *12yo*.

The second situation we consider is the one in which we have some expertise, which will help in identifying keywords which we do not know. We suppose here that we knew the following keywords: *qqaazz, aabbccddee, babyshivid, hussyfan, pthc, ptsc, r@ygold,* and *kingpass* (these are the ones used in the measurement).

In both situations, each method must provide a ranked list of keywords, in decreasing order of estimated relevance as paedophile keyword. We thus expect to observe strongly paedophile keywords and keywords with paedophile connotations at the beginning of these lists.

2.3 Assessment.

Assessing the lists obtained by each method and in both scenarii is a crucial but subtle task. We asked law enforcement experts for this for their help in this evaluation.

However, it was impossible to submit directly the lists to the experts: there are 14 such lists, each containing dozens of keywords. Instead, we selected the top 30 keywords in each list, then merged all lists and constructed the lists of keywords which appeared in at least one list. This led to 189 keywords.

We then presented this list of 189 keywords sorted in alphabetical order to 8 law enforcement experts (through a web page) and asked them to classify each keyword as: *specific paedophile keyword*, *paedophile keyword*, *I don't know* or *general keyword*. Recommendations for this classification were as follows:

> Please tag each keyword below according to its paedophile nature:
>
> - *specific paedophile keyword* if it is used to search for paedophile content specifically (like 'pthc' for instance),
> - *paedophile keyword* if it may be used to search for paedophile content but may be used in other contexts as well (like 'child' for instance),
> - *I don't know* if you don't know this keyword,
> - *general keyword* if the keyword has no paedophile nature (like 'jpg' for instance).

In the context of this work, interesting keywords are the ones in the two first categories, in particular the first one.

This induces ratings for each keyword (the ratio of experts who tagged it with the different tags), from which various ratings may be inferred for lists themselves.

3 Methods for automatic keyword detection.

This section briefly presents each method tested for the automatic detection of paedophile keywords. Full details are available in the cited references.

3.1 Relative frequency method (FREQ).

Given a list of keywords indicating paedophile content, the goal of this method is to detect similar, but unknown, words. The idea is to select words appearing frequently in filenames containing words from the list.

The first step consists in selecting all filenames containing at least one word from the studied list, to obtain a list of paedophile file names. Then it is possible to extract all words appearing in at least one paedophile file name, and compute their frequency, i.e. the number of names in which a given word appears. The words with the highest frequencies in the list are the ones which appear the most frequently in the same file names as the words from the given list.

This method does succeed in extracting paedophile keywords which do not belong to the list. However, it has one drawback: it also tends to select words which appear frequently in *all* file names, regardless of the context, such as 'jpg' or 'mpg'.

To solve this problem, we compute the *relative frequency* of each word appearing in a paedophile file name. It is equal to the number of paedophile file names this word belongs to, divided by the total number of files it belongs to. A word with a relative frequency equal to 1 or close to 1 therefore appears almost exclusively in the same file names as the words from the list, and is thus probably specific to the paedophile context.

Finally, we select all words appearing in the same file name as at least one word from the initial list, compute their relative frequency, and sort them by decreasing relative frequency.

3.2 Statistical co-occurrence method (COOC).

This method consists in constructing a weighted co-occurrence network and in using its statistical properties.

First, we constructed the weighted co-occurrence network in which nodes are words appearing in any filename in the dataset; a link exists between two nodes if the corresponding words appear together in a same filename (multiple occurrences of a same word in a same filename being ignored). Moreover, the link is weighted with the number of filenames in which the two words co-occur.

This direct weighting is not satisfying, though: the links between frequent words naturally tend to have high weights, which does not mean that the two corresponding words are strongly related. To improve this, we considered two possible normalisations of the weight.

First, the Jaccard normalisation consists in dividing the weight of the link as defined above by the sum of the number of occurrences of the two corresponding words. This

means that the Jaccard weight captures the fraction of occurrences of the words which led to co-occurrences. In this way, rare words may have significant weights.

Notice that the Jaccard network is symmetric (the Jaccard weight is defined on unordered pairs of words). However, a link between two words may be important for one of them and much less important for the other. This may occur for instance when a rare word always occur with a more frequent word. In order to capture this, we considered the probability variant of the weights: a link between w and w' is weighted both by the probability to observe w' in a filename if it contains w, and by the converse probability. The obtained network is no longer symmetric.

Based on this weighted networks, we computed word rankings as follows. First we considered the normalised sum of the weights of this word's links to words in the initial list of paedophile keywords. Second, we considered the alpha-centrality [3] using endogenous factors which indicate the initial set of paedophile keywords. We then ranked words according to the results of these computations.

We finally obtained four lists, for Jaccard and probability networks, and for both ranking methods. Manual inspection showed that all are relevant, but we delivered the list based on Jaccard coefficient and alpha-centrality, which seemed to be the most interesting.

3.3 Community-based method (COMM).

A *community* in a network is defined as a subset of nodes which are strongly connected with each other but only poorly connected to nodes outside the community. Partitioning a network into relevant communities gives a lot of information on its structure and may help in identifying nodes which play similar roles. The underlying computation is however complex and time-consuming. There currently exists only one method able to produce good quality results in networks of millions of nodes or more [2]. This is the method used for this study.

Like in the COOC method (Section 3.2), we construct a co-occurrence network between words in filenames of the dataset as follows: two nodes are linked if they appear in a same filename. In addition, we associate to each link a weight known as the Jaccard coefficient: it is the number of filenames in which the two words co-occur divided by the total number of filenames in which at least one of them occur. We explored other weightings of the links (no weight, number of filenames in which the words co-occur, and a vote weight in which each filename distributes equal fractions of its unit vote to all the links it induces) but the community detection performed best with the Jaccard weights.

From this weighted network, our methodology for automatic paedophile keywords detection consists in the following steps:

1. compute communities in the network;
2. find communities which contain at least one well know paedophile keyword;
3. if a community contains more than 100 unanonymised words, then build a subnetwork out of this community and go back to step 1. Otherwise, stop.

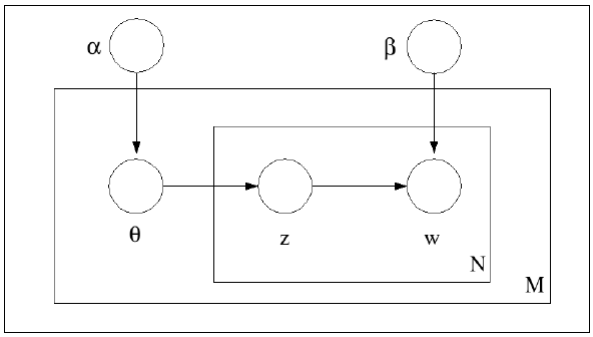

Figure 1: Graphical representation of the LDA model.

By the end of this process, we obtain a list of *candidate* words, *i.e.* unanonymised words which belong to a community containing at least one known paedophile keyword.

In order to rank these candidates, we need to rate them. Given a paedophile keyword list (obtained from the considered scenario), we first extract from the data set all filenames which contain at least one keyword of the list. We then compute the frequency of each candidate word in these filenames (*i.e.* the number of such filenames they belong to). This gives a rating for each candidate word [2]. We then sort the candidate keyword by decreasing frequency values.

3.4 Machine learning method (MALE).

We implemented a machine learning method based on Latent Dirichlet Allocation (LDA) [1]. LDA is a model aiming both at detecting latent topics in a set of textual documents and at associating words and documents to the detected categories.

LDA is a probabilistic model with latent variables which models the documents generation process through conditional dependencies and dirichlet distributions. The generative process, which is illustrated by random variables in Figure 1 considers that, when writing a textual document:

1. we first choose to what extent a document will be related to each topic. This corresponds to the value of the θ random variable which is a distribution of probabilities over the topics;

2. we then, *(a)* choose a particular topic (variable z), and then *(b)* choose a word corresponding to this topic (variable w).

Steps *2.(a)* and *2.(b)* are then repeated until the end of the document. Parameters α and β are hyper parameters of the model that have to be manually tuned. The number of topics is also chosen by hand.

[2]Note that some candidate words obtained a rating of 0 because they do not co-occur with any paedophile keyword from the list.

Considering this process, machine learning methods allow us, knowing a set of documents, to compute the parameters of the model that best fit the collection. These parameters correspond to the distribution of probabilities of the documents over the categories $P(z|\theta)$, and to the distribution of probabilities of the words among the categories $P(w|z)$. The learning step is made by using a *Gibbs-EM* algorithm that is not described here.

While LDA is usually used for detecting a set of categories into a collection, we consider here a different use of the model. Instead of allowing LDA to detect relevant categories within the documents, we force the model to find at least one topic with a high probability of generating paedophile keywords. In order to achieve this, during the learning step, each time we find a known paedophile keyword, we force the model to associate this word to the topic number 1. By doing that, we expect the model to group into this category both known keywords, and *new* paedophile keywords. As our primary interest is in category number 1, we arbitrarily ran the learning with 10 different topics, in both scenarii.

3.5 Aggregate-based method (AGGR).

We construct a co-occurrence network, as in the COOC and COMM methods (Sections 3.2 and 3.2), but focus on the initial set of keywords (depending on the scenario). More precisely, we consider this set of keywords and all keywords which co-occur with them; a link exist between two nodes if they co-occur in a same filename. In addition, each node and link has a weight, equal to its number of occurrences. See Figure 2 (left).

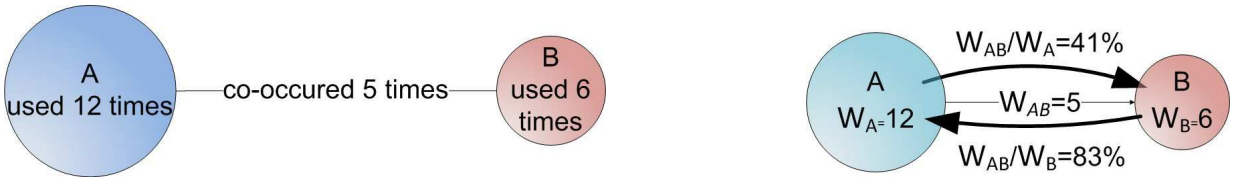

Figure 2: Left: weighted graph constructed from the initial data. Right: its directed weighted version.

We also want to take into account the fact that the importance of word A for word B is not necessarily the same as the importance of word B for word A, *i.e.* associate *directed* weights to links. We define the weight of the link from word A to word B as the ratio between the undirected weight of the link and the weight of A. This reflects the importance of this link from the viewpoint of A. The directed weight from B to A is defined similarly. See Figure 2 (right).

Finally, we select only links between a word in the initial paedophile list and a word which does not belong to it. We do not take into account links with weights lower than 0.1; this threshold was observed to be relevant in practice, and removal of links is important for the complexity of computations. The importance of each keyword is computed as the sum of the weight of its links, multiplied by its number of links. This increases the importance of words co-occurring with many different known paedophile keywords. We then sort the words according to this importance.

3.6 Language analysis method (LAAN).

Language analysis methods define notions like the *context* of a keyword, its *specificity*, its *proximity* to other keywords, etc. [4]. They make it possible to examine situations where the keywords we seek are specialisations of a given set of keywords (Figure 3, left) or evolutions from a known set of keywords (Figure 3, right). In the context of paedophile keywords situations, both cases may occur: paedophile users may enter keywords which give more precise information on the content they describe; they may also introduce new, confidential keywords, which will co-occur with known paedophile keywords in a more subtle way over time.

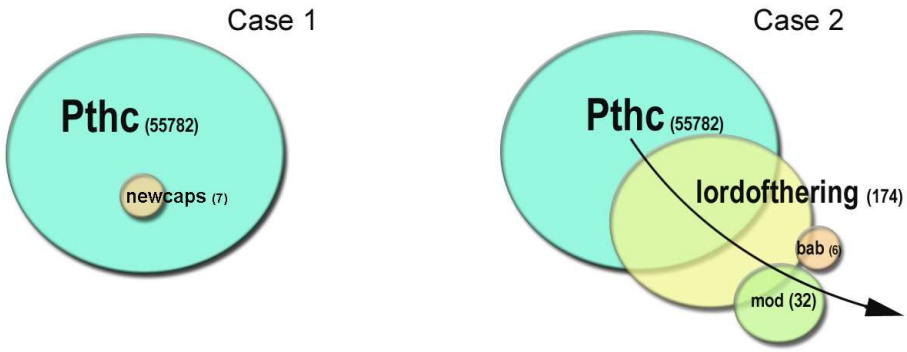

Figure 3: Graphical representation of the two cases we consider.

To find keywords corresponding to the first situation, we have to find terms that are more specific in their uses than some well-known paedophile keywords, in the sense that the files in which they appear are most often also flagged with these well known keywords. To find keywords corresponding to the second situation, we have to examine how many paedophile keywords appear in their closest contexts.

Finally, this method is divided into three steps:

1. extend the initial set of paedophile keywords (depending on the scenario) by selecting terms with about the same level of paedophile content, *i.e.* have a proximity to the initial set higher than a given threshold; this is done by using the *pseudo-inclusion measure* [4];

2. compute the terms which are better contextualised by this extended set;

3. finally assign a rank to each word which is the lowest rank of keywords from the extended set among the closest neighbours of the target word.

4 Results.

Each method described in the previous section produced a keyword list, displayed in Tables 5 and 6 for *common-sense* and *expert* scenarii respectively. As explained in Section 2.3,

we submitted all keywords appearing in these lists to ten experts (law enforcement personnel involved in fighting online paedocriminality). This section presents the results of this assessment.

First notice that not all experts evaluated all keywords: 3 keywords received answers from 7 experts; 25 from 8; 138 from 9; and 23 from 10. All these numbers of answers are significant for the evaluation. We however need to normalise the number of answers of each type for a given keyword with respect to the total number of answers it received. We thus obtain a rating for all keywords, provided in Table 7 at the end of this report.

These results clearly show that the produced lists of keywords contain relevant information (many keywords are tagged as specific paedophile keywords by many experts), although a significant portion of uninteresting keywords are also present (other words are tagged as general keywords). As the vast majority of keywords in the original dataset has no paedophile nature (and actually there may be very few specifically paedophile keywords), these results may be considered as excellent.

We now compute for each list the four average ratings of the top 10, top 20, and top 30 keywords. This gives a first assessment of their respective quality and interest, see Tables 1 and 2.

	common-sense previous knowledge scenario					
	COMM	AGGR	FREQ	MALE	LAAN	COOC
	top 10 keywords					
general	11.1	6.8	18.6	29.9	34.9	5.9
unknown	7.7	8.5	34.2	4.0	46.8	21.3
paedo	41.1	28.2	32.8	23.6	8.2	22.3
specific	40.1	56.6	14.4	42.6	10.0	50.5
	top 20 keywords					
general	16.0	17.8	24.1	31.0	31.8	10.2
unknown	14.4	21.1	33.4	4.8	43.3	14.9
paedo	37.8	24.0	30.1	38.7	17.7	29.9
specific	31.8	37.1	11.5	25.4	7.2	45.0
	top 30 keywords					
general	23.6	19.0	25.7	29.3	32.6	15.0
unknown	17.0	25.7	33.5	4.8	40.4	17.1
paedo	31.2	21.5	30.1	44.7	17.2	29.9
specific	28.2	33.8	9.8	21.2	9.8	38.0

Table 1: Global ratings for top 10, 20 and 30 keywords in each list, in the *common-sense* scenario.

First, the *specific* ratings for the *expert* scenario are in general lower than for the *common-sense* scenario. This might seem counter-intuitive, but is caused by the fact that specific keywords are very rare. Since previously known words from the lists are removed from the results, in the case of the *expert* scenario there are fewer specific words

	expert previous knowledge scenario					
	COMM	AGGR	FREQ	MALE	LAAN	COOC
	top 10 keywords					
general	12.2	11.2	17.9	44.2	20.1	6.5
unknown	6.0	7.6	51.9	1.2	45.3	11.4
paedo	58.8	30.6	17.8	41.0	18.9	48.1
specific	23.0	50.6	12.4	13.6	15.7	34.1
	top 20 keywords					
general	19.8	19.0	15.6	31.6	25.3	13.0
unknown	12.8	24.0	55.9	4.7	49.0	17.7
paedo	43.9	23.0	13.3	47.5	16.2	39.1
specific	23.4	34.1	15.1	16.2	9.5	30.2
	top 30 keywords					
general	25.3	21.6	16.7	38.8	19.5	18.5
unknown	17.7	25.8	53.8	4.6	51.2	15.1
paedo	37.9	20.8	13.5	45.0	16.6	37.9
specific	19.1	31.8	16.0	11.6	12.8	26.5

Table 2: Global ratings for top 10, 20 and 30 keywords in each list, in the *expert* scenario.

that the methods are susceptible to detect than in the case of the *common-sense* scenario. Notice that these words are indeed detected in a significant way in the *common-sense* scenario. Conversely, the *paedo* ratings are in general higher for the *expert* scenario than for the *common-sense* scenario, which confirms this intuition: the top words detected by the methods belong to the paedophile context, but since there are less specific words to detect, the words that remain have a higher *paedo* rating.

The fact that there are few interesting words to detect also causes the lists to become less and less selective when they grow: the ratings for the top 30 (resp. top 20) lists are lower than the ones for top 20 (resp. top 10) in almost all cases. Indeed, the paedophile keywords are near the top of the list, and the ratings decrease when the list becomes too long.

Finally, some methods perform very well: for the *common-sense* scenario, AGGR and COOC reach 50 as *specific* rating, which means that the fraction of experts which considered any keyword as specifically paedophile in the corresponding lists is in average 50%. This is excellent, and surpasses significantly other methods. These ratings become lower when we consider more keywords (top 20 and top 30), but they remain significantly larger than the others.

Note also that the methods which have the lowest *specific* and *paedo* ratings combined, FREQ and LAAN, are the ones which have the highest *unknown* rating. Though these methods seem to perform less well than the others at first glance, this may be worthy of further study, to determine if some of the unknown keywords are paedophile.

Notice that only 25 keywords have less than 50% of identical answers; conversely, 72

have more than 75% of identical results. Given the fact that four answers were possible, this shows that experts are in very good accordance with each other, which is an important fact. This makes it possible to define the *category* of a keyword as the answer on which the largest number of experts agree (in most cases, this is more than 50% of experts, and when several categories reach the maximum we decided to choose the most paedophile category). The number of keywords in each category is given for each list in Tables 3 and 4.

	common-sense previous knowledge scenario					
	COMM	AGGR	FREQ	MALE	LAAN	COOC
	top 10 keywords					
general	1	0	0	3	4	0
unknown	0	1	4	0	5	3
paedo	4	1	5	1	0	1
specific	5	8	1	6	1	6
	top 20 keywords					
general	3	2	3	6	6	1
unknown	2	5	6	0	10	3
paedo	6	2	9	7	3	3
specific	9	11	2	7	1	13
	top 30 keywords					
general	8	4	7	7	11	4
unknown	4	8	9	0	12	4
paedo	7	3	12	15	4	5
specific	11	15	2	8	3	17

Table 3: For each list produced in the *common-sense* scenario, the number of keywords classified in each category by experts. We selected top 10, top 20 and top 30 keywords in each list.

These results are in accordance with global ratings of lists. They show that, even at the word level, AGGR and COOC significantly surpass other methods. They are able to construct lists of 30 keywords, half of which are classified as specific paedophile keywords by more than half our experts. Notice that they also produce a significant rate of keywords which our experts do not recognise, indicating that the meaning of these keywords should be explored further.

The ratings displayed in this report also make it possible to enter into more subtle considerations. For instance, a list with many very specific keywords but others which are generic may be as interesting as a list having all its keywords reasonably relevant. We provide a complete set of indicators which make it possible to assess precisely the quality and interest of each list, depending on the context of use.

	common-sense previous knowledge scenario					
	COMM	AGGR	FREQ	MALE	LAAN	COOC
	top 10 keywords					
general	1	0	2	4	1	0
unknown	0	1	5	0	6	1
paedo	6	2	2	4	1	4
specific	3	7	1	2	2	5
	top 20 keywords					
general	3	2	2	5	3	1
unknown	3	6	13	0	12	4
paedo	8	2	2	11	3	6
specific	6	10	3	4	2	9
	top 30 keywords					
general	6	5	2	11	3	4
unknown	5	8	19	0	18	4
paedo	11	3	4	15	4	10
specific	8	14	5	4	5	12

Table 4: For each list produced in the *expert* scenario, the number of keywords classified in each category by experts. We selected top 10, top 20 and top 30 keywords in each list.

5 Conclusion.

We presented an experimental study aimed at evaluating the relevance of a large panel of automatic paedophile keyword detection. We designed two scenarii relying on real-world data, and then submitted the obtained results to experts. Using their feedback, we computed ratings for words and lists, which provided strong insight on their relevance.

First, it appears clearly that applying automatic keyword detection techniques makes sense in this context, and would significantly improve the current situation. In particular, running such detection periodically during long periods of time would make it possible to observe the emergence of new paedophile keywords. It must be clear however that obtained results are far from exact; they may be seen as an help for quickening and improving manual inspection.

Among the techniques we tried, which span quite well the variety of available techniques for automatic keyword detection, it appears that methods based on co-occurrence networks with appropriate weights perform best. Among these methods, the ones which rely on direct neighbourhood of known paedophile keywords perform better than more intricate approaches, which argues for the use of methods of moderate complexity. This indicates that most information is actually captured in the weighted co-occurrence network.

Finally, the obtained results are promising. They show that automatic methods may help significantly manual inspection, and identify the most relevant approaches. Tuning further these methods for this particular application would probably improve their results, but one must keep in mind that the amount of specific paedophile keywords is rather low,

and it is embedded in a huge amount of general keywords. In this situation, there is certainly a limit which automatic techniques cannot bypass, and the results presented here may be quite close to this limit.

Notice also that we produced lists of relevant keywords, with expert assessment, which represent a significant contribution in themselves.

Acknowledgements. We warmly thank the experts who helped in assessing the results, in particular Philippe Jarlov who also contributed significantly to data collection. This work is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Journal of Statstical Mechanics: Theory and Experiment*, page P10008, 2008.

[3] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23:19120, 2001.

[4] D. Chavalarias and J. P. Cointet. Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study. *Scientometric*, 75(1):37–50, 2008.

[5] Philippe Jarlov, Matthieu Latapy, Frédéric Aidouni, Clémence Magnien, Christophe Berger, and Firas Bessadok. Monitoring paedophile activity in a P2P network. *Forensic Science International*, 2009.

COMM	AGGR	FREQ	MALE	LAAN	COOC
pthc	pedo	7o	pthc	mara	lsn
pedo	kinderficker	nimbus	jpg	inna	lsm
ptsc	ygold	babyshvid	pedo	mellony	qqaazz
hussyfan	mafiasex	7e	mpg	jenny	lsbar
preteen	childlover	nude01	hussyfan	qqaazz	hussyfan
ygold	lsm	pretten	lolita	spreading	childlover
child	babyj	preteenz	ygold	jackie	ptsc
lsm	lolitaguy	hyman	avi	kleuterkutje	ls
new	ls	mellony	preteen	vdbest	babyshivid
daughter	mylola	nudis	ptsc	november	babyj
underage	moscow	kindergarden	girl	baby	pedo
childlover	valya	doggyfuck	new	maryanne	pthc
little	kleuterkutje	2005new	incest	desidee	mylola
lsn	childfugga	newer	young	novinhas	ygold
sandra	kdquality	bambina	daughter	diaper	child
bd	lsbar	inna	teen	lsbar	vicky
kids	vicky	infant	underage	until	magazine
tori	bd	teal	little	inga	kdquality
mafiasex	sandra	witch	and	change	childfugga
kdquality	cries	lolita2	cum	nimphets	daughter
ls	ver	jenniefer	kids	ptsc	underage
childfugga	little	jr	kiddy	tori	island
kiddy	hussyfa	tori	lsm	ye	rbv
company	kidzilla	novinhas	old	ck	kdv
jenny	yg	peepee	childlover	inces	jenny
kleuterkutje	qsh	boylover	pussy	nancy	preteen
qqaazz	yamad	sofie	boy	weekend	sandra
newer	magazine	lso	girls	babyshvid	boy
magazine	company	olds	model	thor	mafiasex
torture	pedofilia	child	rape	bella	kingpass

Table 5: Lists produced by each method, with common-sense initial knowledge.

COMM	AGGR	FREQ	MALE	LAAN	COOC
pedo	pedo	landfill	jpg	bbx	pedo
preteen	kinderficker	kissie	pedo	novinhas	preteen
new	mafiasex	lolifuck	avi	1man	lsm
child	childlover	beyword	lolita	lolifuck	childlover
daughter	lsm	eurololita	preteen	pak	daughter
underage	babyj	nimphets	mpg	jenniefer	v132
little	lolitaguy	jenniefer	new	mellony	child
lsm	ls	laika	girl	doughter	underage
childlover	mylola	suwano	child	luto	lolita
kids	moscow	kurahashi	daughter	childfugga	mafiasex
kiddy	valya	uvs	teen	newstar	new
ls	kleuterkutje	harrier	little	ptff	kinderficker
mafiasex	childfugga	u15	underage	sofie	ls
sandra	kdquality	madnet	childlover	lordofthering	lsn
lsn	lsbar	lolalover	model	playtoy	kids
vicky	vicky	brazuquinha	lsm	lsp	childfugga
childsex	bd	hussyfun	young	maryanne	eine
magazine	sandra	newcaps	porn	nansy	childsex
illegal	cries	ptff	pussy	kacy	vicky
pt	ver	gebusch	kiddy	abt	tochter
mylola	little	rebone	girls	liluplanet	little
st	hussyfa	boylover	rar	jho	petersburg
liluplanet	kidzilla	reallola	ass	nobull	sandra
petersburg	yg	7o	kids	pae	lolitaguy
kleuterkutje	qsh	xlola	cum	childlover	model
ultra	yamad	island03	nude	gebusch	kiddy
moscow	magazine	10of	private	cambodian	old
pre	company	playtoy	sex	nablot	liluplanet
lordofthering	pedofilia	furs	old	kidzilla	babyj
jenny	step	yelitza	incest	nn	moscow

Table 6: Lists produced by each method, with expert initial knowledge.

	expert answers			
word	% general	% unknown	% paedo	% specific
001a	55.6	44.4	–	–
10of	44.4	55.6	–	–
1man	44.4	44.4	11.1	–
2005new	55.6	33.3	11.1	–
7e	50.0	50.0	–	–
7o	22.2	66.7	11.1	–
abt	22.2	66.7	11.1	–
and	88.9	11.1	–	–
ass	75.0	12.5	12.5	–
avi	100.0	–	–	–
baby	11.1	–	88.9	–
babyj	–	–	20.0	80.0
babyshivid	–	–	20.0	80.0
babyshvid	–	–	30.0	70.0
bambina	11.1	–	88.9	–
bbx	11.1	77.8	11.1	–
bd	44.4	55.6	–	–
bella	55.6	22.2	22.2	–
beyword	11.1	88.9	–	–
boy	12.5	–	87.5	–
boylover	–	–	60.0	40.0
brazuquinha	11.1	88.9	–	–
cambodian	33.3	33.3	33.3	–
change	88.9	11.1	–	–
child	–	–	100.0	–
childfugga	–	11.1	22.2	66.7
childlover	–	–	30.0	70.0
childs	–	–	100.0	–
childsex	–	–	30.0	70.0
ck	11.1	77.8	11.1	–
company	77.8	22.2	–	–
cries	55.6	11.1	33.3	–
cum	66.7	–	33.3	–
daughter	11.1	–	88.9	–
desidee	11.1	88.9	–	–
diaper	–	22.2	66.7	11.1
doggyfuck	28.6	28.6	28.6	14.3
doughter	11.1	22.2	55.6	11.1
eine	44.4	55.6	–	–
elli	11.1	88.9	–	–

	expert answers			
word	% general	% unknown	% paedo	% specific
eurololita	–	12.5	50.0	37.5
furs	33.3	55.6	–	11.1
gebusch	11.1	88.9	–	–
girl	44.4	–	55.6	–
girls	44.4	–	55.6	–
harrier	22.2	66.7	11.1	–
hussyfa	–	–	22.2	77.8
hussyfan	–	–	30.0	70.0
hussyfun	–	–	11.1	88.9
hyman	11.1	33.3	44.4	11.1
illegal	33.3	–	66.7	–
inces	11.1	22.2	55.6	11.1
incest	–	–	77.8	22.2
infant	11.1	22.2	66.7	–
inga	10.0	70.0	20.0	–
inna	–	87.5	12.5	–
island	75.0	12.5	12.5	–
island03	44.4	44.4	11.1	–
jackie	66.7	33.3	–	–
jailbait	22.2	44.4	22.2	11.1
jenniefer	66.7	22.2	11.1	–
jenny	71.4	14.3	14.3	–
jho	11.1	66.7	22.2	–
jpg	88.9	–	11.1	–
jr	44.4	55.6	–	–
kacy	44.4	55.6	–	–
kdquality	–	22.2	22.2	55.6
kdv	–	44.4	11.1	44.4
kiddy	11.1	–	55.6	33.3
kids	11.1	–	88.9	–
kidzilla	–	33.3	22.2	44.4
kinderficker	–	–	22.2	77.8
kindergarden	22.2	11.1	44.4	22.2
kingpass	–	22.2	22.2	55.6
kissie	12.5	87.5	–	–
kleuterkutje	–	77.8	11.1	11.1
kurahashi	–	77.8	22.2	–
la2	12.5	87.5	–	–
laika	55.6	33.3	11.1	–
landfill	33.3	66.7	–	–

	expert answers			
word	% general	% unknown	% paedo	% specific
liluplanet	–	22.2	22.2	55.6
little	–	22.2	77.8	–
lolalover	–	11.1	22.2	66.7
lolifuck	–	–	33.3	66.7
lolita	10.0	–	40.0	50.0
lolita2	–	10.0	40.0	50.0
lolitaguy	–	–	20.0	80.0
lordofthering	22.2	33.3	44.4	–
ls	10.0	40.0	30.0	20.0
lsbar	25.0	50.0	12.5	12.5
lsm	12.5	25.0	37.5	25.0
lsn	11.1	66.7	22.2	–
lso	11.1	77.8	11.1	–
lsp	11.1	77.8	11.1	–
lsw	11.1	77.8	11.1	–
lucie	33.3	44.4	22.2	–
luto	12.5	75.0	–	12.5
madnet	22.2	77.8	–	–
mafiasex	20.0	–	20.0	60.0
magazine	88.9	11.1	–	–
map	88.9	11.1	–	–
mara	33.3	66.7	–	–
maryanne	44.4	44.4	11.1	–
mellony	11.1	66.7	22.2	–
model	44.4	11.1	44.4	–
moscow	44.4	11.1	44.4	–
mpg	100.0	–	–	–
mylola	25.0	–	37.5	37.5
nablot	–	66.7	11.1	22.2
nancy	55.6	44.4	–	–
nansy	33.3	66.7	–	–
new	87.5	12.5	–	–
newcaps	44.4	44.4	11.1	–
newer	77.8	22.2	–	–
newstar	50.0	37.5	12.5	–
nimbus	33.3	55.6	11.1	–
nimphets	–	30.0	50.0	20.0
nn	–	88.9	11.1	–
nobull	11.1	88.9	–	–
november	88.9	11.1	–	–

	expert answers			
word	% general	% unknown	% paedo	% specific
novinhas	11.1	66.7	22.2	–
nude	44.4	–	55.6	–
nude01	25.0	25.0	50.0	–
nudis	33.3	22.2	44.4	–
old	66.7	–	33.3	–
olds	44.4	22.2	22.2	11.1
pae	11.1	66.7	22.2	–
pak	33.3	66.7	–	–
pedo	–	–	44.4	55.6
pedofilia	–	–	30.0	70.0
peepee	–	37.5	50.0	12.5
petersburg	62.5	12.5	25.0	–
phantom	33.3	55.6	11.1	–
playtoy	22.2	33.3	33.3	11.1
porn	66.7	–	33.3	–
pre	37.5	12.5	50.0	–
preteen	–	–	70.0	30.0
preteenz	–	–	70.0	30.0
pretten	–	22.2	44.4	33.3
private	55.6	22.2	22.2	–
prt	22.2	55.6	11.1	11.1
pt	12.5	37.5	25.0	25.0
ptff	–	77.8	–	22.2
pthc	–	–	10.0	90.0
ptsc	–	20.0	10.0	70.0
pussy	33.3	11.1	55.6	–
qqaazz	–	11.1	11.1	77.8
qsh	–	100.0	–	–
rape	22.2	–	77.8	–
rar	88.9	–	11.1	–
rbv	–	100.0	–	–
reallola	–	50.0	–	50.0
rebone	11.1	88.9	–	–
rizmastar	11.1	77.8	11.1	–
sandra	66.7	22.2	11.1	–
sex	77.8	11.1	11.1	–
sofie	55.6	33.3	11.1	–
spam	77.8	11.1	11.1	–
spreading	77.8	11.1	11.1	–
st	11.1	77.8	11.1	–

	expert answers			
word	% general	% unknown	% paedo	% specific
step	77.8	22.2	–	–
suwano	–	100.0	–	–
teal	22.2	77.8	–	–
teen	11.1	11.1	77.8	–
thor	62.5	37.5	–	–
tochter	11.1	33.3	55.6	–
tori	55.6	22.2	22.2	–
torture	50.0	12.5	37.5	–
u15	11.1	55.6	33.3	–
ultra	88.9	11.1	–	–
underage	–	–	50.0	50.0
until	85.7	14.3	–	–
uvs	11.1	88.9	–	–
v10040	11.1	88.9	–	–
v132	11.1	88.9	–	–
valya	33.3	55.6	11.1	–
vater	22.2	44.4	33.3	–
vdbest	–	88.9	–	11.1
ver	22.2	77.8	–	–
vicky	20.0	20.0	30.0	30.0
weekend	77.8	22.2	–	–
witch	66.7	33.3	–	–
xlola	–	22.2	22.2	55.6
yamad	11.1	33.3	–	55.6
ye	11.1	77.8	11.1	–
yelitza	11.1	77.8	–	11.1
yg	12.5	50.0	12.5	25.0
ygold	–	20.0	20.0	60.0
young	11.1	–	77.8	11.1

Table 7: Assessment results for all keywords.

Project MAPAP SIP-2006-PP-221003.

http://antipaedo.lip6.fr

Supported in part by the European Union
through the *Safer Internet plus Programme*.

http://ec.europa.eu/saferinternet