Technical report on the

Automatic Detection of Paedophile Queries

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

Matthieu Latapy [1], Clémence Magnien and Raphaël Fournier

Abstract

Filtering or identifying paedophile queries is a key issue for law enforcement and search engines. However, these queries are in general mixed with a huge amount of other queries. Moreover, little is known on their characteristics. We address here these two issues in order to design the first tool for automatic detection of paedophile queries. Using domain expertise, we select some paedophile queries in a set of hundreds of millions of queries entered in a general public P2P system. We extend this set by manually inspecting the queries it contains and the words composing them. We then design a tool which tags any query as paedophile or not. We run it on our dataset and evaluate its performances by submitting appropriate samples to external experts. This assessment shows that the tool performs very well. Going further, the assessment makes it possible to estimate precisely and rigorously its error rates, which we compute and provide.

1 Introduction.

Most available search engines rely on keyword-based queries submitted by users. These queries are descriptions of the content (web pages, files, database entries, etc.) searched by users. Usually, these queries aim at matching words which occur in the description. Automatic classification of these queries into thematic categories is difficult because of the ambiguity of the terms used, errors in keyword spelling, the use of multiple languages, etc.

However, this is an important issue in several situations. In the context of paedocriminality fighting, automatically detecting queries is a key tool for filtering (the system may refuse to answer paedophile queries, and/or record and/or signal them), for a better knowledge of paedophile activities (examining such queries is crucial for this), and other important tasks. For instance, web search engines like *Google* try to filter paedophile queries; administrators of *eDonkey* servers and other file distribution services try to reject paedophile files by automatically inspecting filenames; etc.

The classical approach to this problem would be to start with a set of queries known to be paedophile, a set of queries known not to be, and then use machine learning techniques

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

to obtain an automatic filter based on statistical properties inferred from these two training sets.

In the case of paedophile activity, though, such data does not exist. In addition, as paedophile queries are relatively rare (typically less than a percent of the whole), constructing a set of paedophile queries would be very difficult: it would mean that experts would have to manually inspect a huge set of queries, which is extremely time consuming, difficult to assess, and prone to errors.

We propose here a solution to this problem. To do this, we use a large corpus of queries entered in a P2P system which we collected beforehand [1]. We then use our domain expertise (knowledge of typical paedophile keywords, in particular) to manually inspect this dataset. We gradually construct a filter based on this expertise and observations of the data (co-occurrences in particular). The goal of this filter is to be able to automatically decide if a given query is paedophile or not. We finally assess the performances of this filter by confronting its results to estimations of experts of the field.

2 Data.

The dataset used for this study contains more than 127 millions of queries entered in the *eDonkey* system in 2007, and captured during a measurement described in [1]. This is the largest set of such queries ever collected, and therefore it suits our purposes. Other datasets may be relevant too (in particular web search engine queries), but as our primary interest is in P2P activity we leave this for future work.

In order to protect user privacy, this dataset has been normalised and anonymised.

More precisely, we kept only one occurrence of each query (we removed all duplicate queries). We then broke down these unique strings into words (words are sequences of letters and/or numbers only, containing no space or punctuation characters). We then normalised the obtained words by converting all letters to lowercase. After this normalisation, some queries which were originally different become indistinguishable. For instance, queries *madonna.mp3* and *Madonna MP3* both become *madonna mp3*.

Then, we anonymised our dataset. Words appearing in a very small number of strings have a significant chance of representing personal information. We set a threshold of 100 occurrences to distinguish between rare and common words: all words appearing in less than 100 different queries are replaced by an integer, while others are kept in clear in the dataset. Note that during the normalisation process we only considered distinct queries, therefore if a client enters 100 times the same query, this is considered as a single query by our anonymisation process.

Notice that such normalisation and anonymisation, though necessary, constitute a significant obstacle in our task. For instance, one may guess that some specific paedophile keywords are rather rare; we will be unable to observe them. Even more worrying, some keywords like *k!ds* or *r@ygold*, often used in paedophile queries, will be splitted by the normalisation procedure (into *k ds* and *r ygold* respectively, which appear as two-word queries). This kind of difficulties however is inherent to work on real-world data, which

must be properly anonymised in order to preserve user privacy and be in accordance with law. We will deal with the implications of this preprocessing of the data below.

Notice finally that we removed from the initial set all empty queries and queries which contained only anonymised words. This led to a final set of 100 968 115 queries, which is the dataset considered in the rest of the paper.

3 Filter construction.

According to experts of paedocriminality fighting, some keywords are characteristic of paedophile activity in P2P systems. These keywords have no other meaning and are dedicated to the search of such content. Typical examples include *qqaazz*, *r@ygold*, or *hussyfan*. One may therefore use them as a first basis for searching for paedophile queries in a set of queries.

Another, complementary approach consists in inspecting queries which fit common sense about paedophilia. For instance, one may observe queries containing keywords *child* and *sex*, or similar associations. One may also inspect queries containing only one of these keywords, in particular *child*, in order to gain intuition on what kinds of queries are entered with this keyword (probably paedophile queries, but also queries for games or movies for children, song titles containing this word, etc.).

Manual inspection of the queries containing typical paedophile keywords like the ones cited above clearly shows that these queries are of paedophile nature. Indeed, such keywords appear in very explicit queries, but are also used in single-keyword queries. This is true also for queries containing pairs of keywords like *child* or *kid*, and *sex* or *porn*.

In addition, manual inspection of these queries raised an unexpected observation: many queries contain age indication under the form *n yo* where *n* is a number lower than 16, meaning that the user is seeking content involving *n years old* children. Other suffixes are also used in place of *yo*: *yr*, *years old*, etc. Age indications are strong indicators of paedophile queries, but they are not sufficient in themselves: they also occur in many non-paedophile queries.

Finally, we added queries which contain words related to family, denoting parents *and* children, and a word such as *sex*.

In all situations above, local language variations also occur, in particular French, German, Spanish, and Italian versions. Some queries, which are however much rarer, are formulated in other languages, such as Russian. We included the most frequent translations in our sets of keywords.

We finally decided to construct our filter as follows. See appendix A for the source code of our filter, including lists of keywords cited below (with the same name as the one appearing in the source code).

First, we considered that any query containing a keyword in a specific keyword list, called *explicit* in the source code, is a paedophile query. This reflects the fact that some keywords are used strictly by paedophiles in this context. The list of keywords is very small, as they are very specific, but our confidence in them is very high.

Second, we constructed a set of keywords related to childhood, called *child* in the source code, and a set of keywords related to sexuality, called *sex* in the source code. Any query which contains a keyword in both sets is considered as paedophile. Notice that this may be false in some cases, as for instance in queries like *destinys child sexy daddy*. We will discuss this in the next section.

Third, we identified suffixes related to age indications and stored them in the list named *agesuffix* in the source code. As age indication is not sufficient in itself, we decided that a query is paedophile if it contains age indication *and* a word in the *sex* or *child* list.

As any set of rules, these rules lead to some false positives and negatives, *i.e.* respectively non-paedophile queries identified as paedophile by the rules, and paedophile queries identified as non paedophile by the rules. We manually inspected a first set of results and identified a few such situations, which we corrected by adding minor variants to these general rules. This slightly improved the results. These variants are visible in the source code in appendix.

We finally reached a situation where the changes made to the filter had negligible effect, or induced an increase in its false positive and negative rates. Despite this, we clearly still miss some paedophile queries, and tag some non-paedophile query as paedophile. Inspection of these errors however indicate that going further would need much more intricate techniques, without ensuring better results. Before this, it therefore seemed natural to assess the results obtained with this filter.

4 Assessment methodology.

Let us consider a set Q of queries, and let us denote by P^+ (resp. P^-) the set of paedophiles (resp. non-paedophile) queries in Q. Let us denote by F^+ (resp. F^-) the subset of Q which is tagged as paedophile (resp. non-paedophile) by our filter.

Ideally, we would have $F^+ = P^+$, which means that our filter makes no mistakes. In practice, though, there are in general paedophile queries which our filter mis-identifies, *i.e.* queries in $P^+ \cap F^-$. Such queries are called *false negatives* (the filter produces an erroneous negative answer for them). *False positives* are defined dually.

The numbers of false positives and negatives describe the performance of our filter on Q. Notice however that they strongly depend on the size of P^+ and P^-. In our situation, we expect P^+ to be much smaller than P^- (most queries are not paedophile), which automatically leads to small number of false negatives, even in the extreme case where the filter would give only negative answers. Moreover, the fractions of false positives and false negatives depend on the fraction of paedophile queries in Q, which is an unknown quantity.

In such situations, two natural definitions of false positive and negative rates coexist. Both will prove to be useful here.

First, one may consider the rate of false negatives (resp. positives) when all inspected queries are paedophile (resp. non-paedophile). If we run the filter on Q, this leads to:

$$f^+ = \frac{|F^+ \cap P^-|}{|P^-|} \quad \text{and} \quad f^- = \frac{|F^- \cap P^+|}{|P^+|}$$

An estimate of f^+ may therefore be obtained by sampling a random subset X of P^- (*i.e.* random non-paedophile queries) and manually inspect the results of the filter on X. Constructing X is easy: as most queries are non-paedophile, one may sample random queries and then manually discard the ones which are paedophile. As long as X is small, this has a reasonable cost. However, the fraction of queries in X which will be tagged as paedophile by our filter will be extremely small, probably even null. As a consequence, the estimate of f^+ obtained this way will be of poor quality.

Conversely, an estimate of f^- may be obtained by sampling a random subset X of P^+ (*i.e.* random paedophile queries) and manually inspect the results of the filter on X. As P^+ is very small and unknown, sampling X is a difficult task. We may however approximate it using the notion of *neighbour* queries as follows.

Given a query q in Q, its *backward neighbour* is the query q_- in Q such that q_- was the last query in Q which was received from the same IP address as q less than two hours before q, if it exists. In this way, we expect that q_- was entered by the same user[2] as q, seeking the same kind of content. Likewise, we define the *forward neighbour* q_+ of q as the first query in Q which was received from the same IP address as q less than two hours after q, which we similarly expect to be entered by the same user as q.

We denote by $N(S) = \{q^+,\ q_-,\ q \in S\}$ the set of neighbour queries of all queries in any set S. We assume that queries in $N(P^+)$, *i.e.* the neighbours of paedophile queries, are also paedophile with high probability (much higher than for random queries in Q). We expect this to be also true for queries in $N(F^+)$.

Obviously, $N(P^+) \cap P^+ \subseteq P^+$ and $N(F^+) \cap P^+ \subseteq P^+$. But $N(F^+) \cap P^+ \not\subseteq F^+$ in general. In other words, $N(F^+)$ probably contains queries in P^+ (*i.e.* paedophile queries) which are not detected by our filter. If we consider the queries in $N(F^+) \cap P^+$ as random paedophile queries, then they may be sampled to construct a set X of random paedophile queries suitable for estimating f^-. As X contains only paedophile queries, this estimate is equal to the number of queries in X not detected as paedophile by our filter divided by the size of X.

Notice that the queries in X may actually be biased by the fact that they are derived from F^+: the probability that a user enters a paedophile query which the filter is able to detect probably is higher if this user already entered one such query (he/she may enter in both cases keywords detected by our filter). As a consequence, our estimate of f^- may be an under-estimate.

Finally, one cannot, in our context, evaluate f^+ properly; on the contrary, we are able to give a reasonable (under)estimate for f^-. But both f^+ and f^- are needed to evaluate the performance of our filter.

In order to bypass this issue, we will consider the following variants of false positive and negative ratios, which capture the probability that the filter gives an erroneous answer

[2]It must be clear that IP addresses are not sufficient to distinguish between users; in this context, however, this approximation will prove to be satisfactory.

when it gives a positive (resp. negative) one:

$$f'^{+} = \frac{|F^+ \cap P^-|}{|F^+|} \quad \text{and} \quad f'^{-} = \frac{|F^- \cap P^+|}{|F^-|}$$

An estimate of f'^{+} may therefore be obtained by sampling a random subset X of F^+ (*i.e.* a random set of queries for which our filter gives a positive answer) and by manually inspecting this subset in order to obtain the number of false positives. We expect all sets involved in these computations to be of significant size, so there is no obstacle in computing a reasonable estimate for f'^{+}.

Conversely, an estimate of f'^{-} may be obtained by sampling a random subset X of F^- and inspect it to determine the number of false negatives, *i.e.* the number of queries in X which actually are paedophile (*i.e.* in P^+). However, as already noticed, paedophile queries are expected to be very rare, so the number of observed false negatives will be extremely small as long as X is of reasonable size.

Therefore, one may easily compute a significant estimate of f'^{+}, but computing a reasonable estimate for f'^{-} is not practically doable in our case.

Finally, the quantities we will use for evaluating the performances of our filter are f'^{+} (the rate of errors when our filter decides that a query is paedophile) and f^- (the rate of paedophile queries that our filter misclassifies as non-paedophile), which we are able to properly estimate. We describe the practical protocol we used to compute these estimates in the following section.

5 Assessment protocol.

We denote by Q the set of queries described in Section 2, and use the formalism of Section 4. We divide Q into three sets (with overlap): F^- (the set of queries which our filter tags as not paedophile), F^+, and $N(F^+)$ (the set of neighbours of queries tagged as paedophile by our filter). These three sets are easy to compute from Q using our filter.

Notice that some queries in F^+, *i.e.* some queries which are tagged as paedophile by the filter, are composed of only one word. Then, this word is necessarily a word in the *explicit* paedophile keywords list in the filter source code: *babyj, babyshivid, childlover, childporn, childsex, childfugga, ddoggprn, hussyfan, kdquality, kidzilla, kingpass, mafiasex, pedo, pedofilia, pedofilo, pedoland, pedophile, pedophilia, pedophilie, pthc, ptsc, qqaazz, raygold, reelkiddymov, yamad, youngvideomodels*. These keywords are known to have a very strong paedophile nature. Therefore, if such a keyword appears alone in a query, then there is little chance, if any, that this query is not paedophile. One may then increase the efficiency of our assessment by assuming that these one-keyword queries are indeed paedophile queries and not submitting them to experts. We denote by F_1^+ the set of queries in this set, and by $F_{>1}^+$ the queries in F^+ composed of more than one word. Our assumption then consists in saying that $F_1^+ \subseteq P^+$, and therefore we will assess $F_{>1}^+$ only.

We finally construct the set of queries to assess by select 1000 random queries in each of the sets F^-, $F^+_{>1}$ and $N(F^+)$ (thus 3000 queries in total[3]). This leads to three subsets which we denote by $\overline{F^-}$, $\overline{F^+_{>1}}$, and $\overline{N(F^+)}$ respectively.

We then asked to 71 individuals to contribute to our assessment procedure by tagging each of these 3000 queries as *paedophile*, *probably paedophile*, *probably not paedophile*, or *not paedophile*. They may also choose the *I don't know* answer. In order to help their choice, we provided for each query its backward and forward neighbours, when they existed. See Figure 1.

These individuals were chosen among law enforcement personnel, NGO personnel, and consultants, all involved in fighting online paedocriminality. Among them, 37 provided some feedback through our web interface, among which 20 provided more than 300 answers (thus assessing more than 10% of the whole). In order to ensure the relevance of statistics below, we considered only these contributors. Their answers are summarised in Table 1.

id	*paedo*	*probably*	*don't know*	*probably not*	*not paedo*	total	relevance
1	1530	149	25	66	1230	3000	99.5
2	1381	247	125	580	667	3000	98.5
3	1679	89	2	113	1117	3000	99.1
4	769	83	50	95	420	1417	99.3
9	1598	5	15	1	1381	3000	98.8
10	128	81	1	26	124	360	100.0
11	179	94	0	106	94	473	98.9
12	1624	126	16	165	581	2512	99.8
14	351	16	2	16	27	412	100.0
15	647	119	71	40	439	1316	98.4
16	1174	111	20	64	789	2158	99.1
19	335	17	1	70	166	589	97.5
20	641	383	4	112	753	1893	97.8
28	1071	546	2	453	928	3000	88.4
39	1554	197	28	327	894	3000	97.6
43	1506	120	6	25	393	2050	98.3
57	371	1017	496	570	546	3000	95.7
59	976	936	405	594	89	3000	96.6
64	344	12	10	70	156	592	98.3
68	550	91	208	122	99	1070	98.1

Table 1: Assessement results for each relevant expert.

Some contributors may have an inadequate knowledge of our particular context (identification of paedophile queries), and invalidate our assessment by entering erroneous answers.

[3]Since there is an overlap between $N(F^+)$ and the other sets, we could have sampled some queries more than once, leading to less than 3000 queries in total. Since 1000 is small compared to the total sizes of the three sets, this did not happen here.

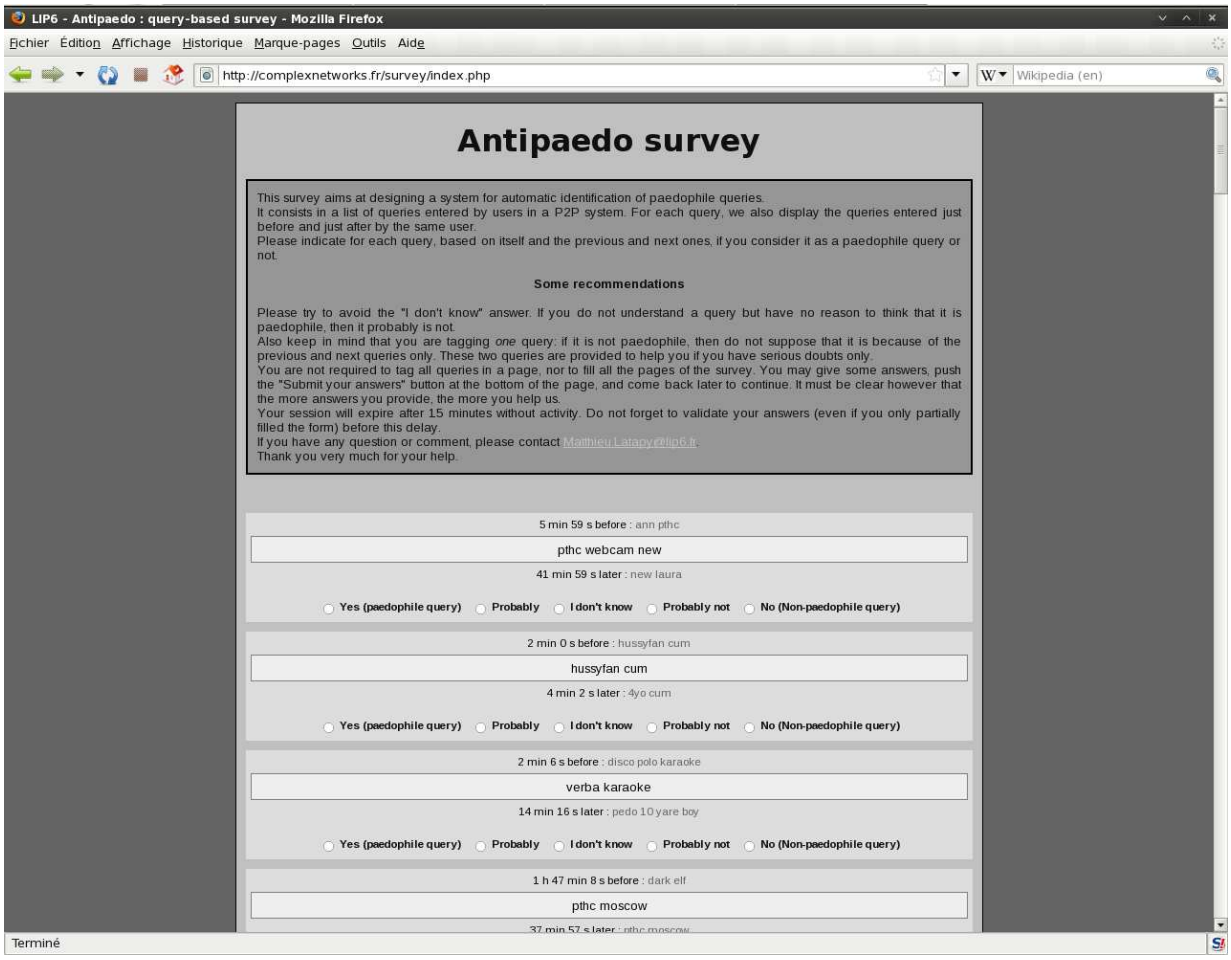

Figure 1: Snapshot of the web interface proposed to experts for the assessment of our filter. A list of queries is presented, each with its backward and forward neighbours, when they exist. Five choices are possible: *paedophile*, *probably paedophile*, *I don't know*, *probably not paedophile*, or *not paedophile*. Queries are ordered randomly and independently for each contributor. A total of 3000 queries are presented, but contributors may stop after providing any number of answers.

In order to identify such contributors, we examined the answers of each contributor to the queries which contain an *explicit* paedophile keyword, *i.e.* a word in our *explicit* list in the filter code. As already said, these keywords are well acknowledged paedophile keywords, which experts of the field consider as strong indicators of paedophile queries. The set $\overline{F^+_{>1}}$ of queries submitted to contributors contains 588 queries which contain at least one keyword in this list. We provide in Table 1 the percentage of these queries (rightmost column) which the corresponding contributor tagged as *paedophile* or *probably paedophile*. For all contributors except contributor number 28, this percentage is above 95%, thus showing that our contributors recognise these keywords. Contributor 28 only slightly disagrees as his/her ratio is 88.4%.

The ratios discussed above may be misleading if a contributor tags all or almost all queries as paedophile. Table 1 gives precise insight on this. The answers of most contributors are well balanced between all possible answers, except for contributors 14, 43, and 68. Manual inspection shows however that these contributors focused preferentially on paedophile queries (they did not assess all queries), therefore we kept them in our expert set.

Finally, we obtained 38 842 answers provided by 20 experts, who contributed at least 300 answers each. These answers are summarised in Table 1. They leads to an average of almost 13 experts assessing each query, which is sufficient for our purpose.

The distribution of these answers among the queries of each considered set is given in Table 2. It is in accordance with what would indeed expect if our filter performs well, and if our assumption that $\overline{N(F^+)}$ should contain many paedophile queries. We analyse this in more details in the next section.

	random subset		
	$\overline{F^-}$	$\overline{F^+_{>1}}$	$\overline{N(F^+)}$
paedophile	52	10670	7686
probably	225	2062	2152
I don't know	865	200	422
probably not	2159	318	1138
not paedophile	8863	231	1799
Total	12164	13481	13197

Table 2: Number of votes of each kind for each considered set.

6 Classification of queries.

For each query q submitted to experts in our assessment procedure, we will denote by q^{++} the fraction of experts (among the ones who provided an answer for q) which tagged it as *paedophile* and by q^+ the fraction of experts which tagged it as *paedophile* or *probably paedophile*. We define q^- and q^{--} dually. Notice that $q^+ + q^- < 1$ in general, as some *I don't know* answers were provided (the fraction of such answers is $1 - q^+ - q^-$). Moreover, $q^+ \geq q^{++}$ and $q^- \geq q^{--}$ for all q.

In order to classify queries according to expert answers, we expect to observe that queries q have either a high q^+ (resp. q^{++}) or a high q^- (resp. q^{--}), but not both or neither, meaning that experts agree on the nature of q. Figure 2 gives much insight on this. They display the correlations between these values. Queries on which many experts disagree are close to the diagonal (the fraction of experts considering them as paedophile is close to the fraction of experts considering them as non-paedophile). We observe that only few queries are in this situation. Notice that dots far from the diagonal (in particular in the upper left corner and in the lower right one) actually represent many queries with the same ratios; in such situations the dots overlap.

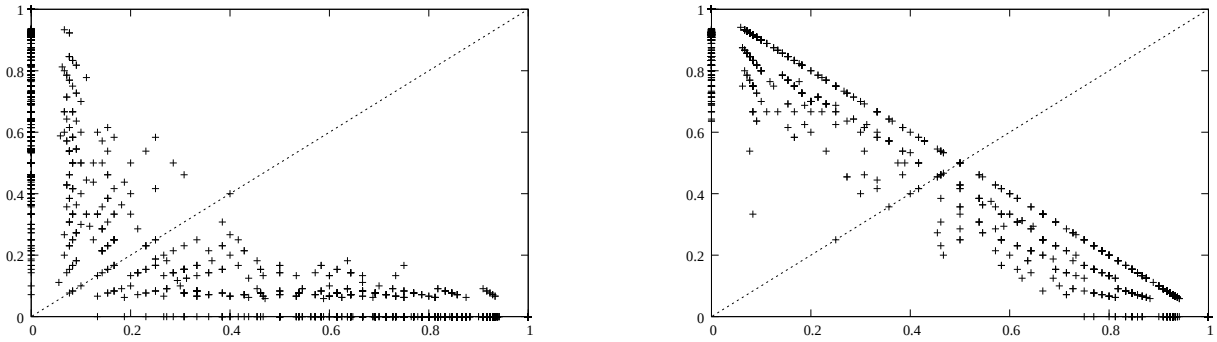

Figure 2: Correlations between different types of answers. Left: for each query, we draw a dot at $(x = q^{++}, y = q^{--})$, *i.e.* if it received a fraction x of *paedophile* and a fraction y of *not paedophile* answers from experts. Right: we draw a dot at $(x = q^{+}, y = q^{-})$, if it received a fraction x of *paedophile* or *probably* answers, and a fraction y of *not paedophile* or *probably not* answers. In both plots, we also display the diagonal $y = x$ for comparison.

In order to deepen this, we display in Figure 3 the difference between q^{+} and q^{-} and between q^{++} and q^{--} for all queries, ordered by increasing order of difference. These plots grow very rapidly for small values on the horizontal axis, meaning that only very few queries have a small difference (this is in accordance with the fact that there are only few points close to the diagonal in Figure 2). On the contrary, for many queries, the difference is very large: it is larger than 80% for approximately 1500 queries (over 3000) in the case of q^{++} and q^{--}, and for 2000 queries in the case of q^{+} and q^{-}.

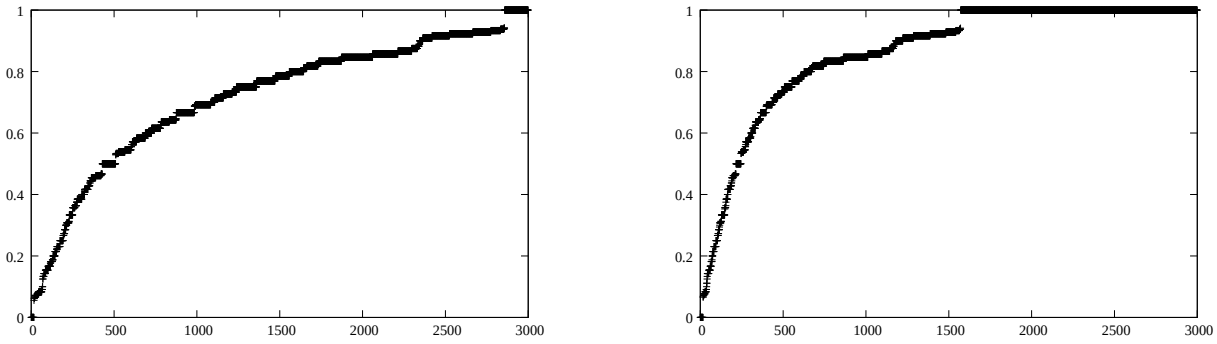

Figure 3: For each query q, the absolute value of the difference between q^{++} and q^{--} (left), and between q^{+} and q^{-} (right). Queries are ordered by increasing order of this difference. A point at coordinate (x, y) therefore means that x queries have a difference lower than y.

Our goal here is to classify as many queries as possible. Notice that only 38 queries have a difference $|q^{+} - q^{-}|$ smaller than or equal to 10%, which already is significant. Moreover, this number increases very slowly when the difference grows. We will therefore classify a query as paedophile if $q^{+} - q^{-} > 0.1$ and as non-paedophile otherwise. We obtain the assessment results presented in Table 3.

	random subset		
	$\overline{F^-}$	$\overline{F^+_{>1}}$	$\overline{N(F^+)}$
paedophile queries	3	985	759
non-paedophile queries	997	15	241

Table 3: Number of queries classified in each category for each considered set.

7 Results.

Thanks to the assessment results in Table 3 and the expressions given in Section 4, we may now compute estimates of the false positive and negative rates which describe the performances of our filter.

First notice that, as expected, the number of paedophile queries in the set of queries tagged as non-paedophile by the filter is very low: $|\overline{F^-} \cap P^+| = 3$. As a consequence, approximating $f'^- = \frac{|F^- \cap P^+|}{|F^-|}$ by $\frac{|\overline{F^-} \cap P^+|}{|\overline{F^-}|} = \frac{3}{1000}$ would yield very poor quality result.

The estimation obtained for f'^+ is of much better quality. It relies on the following expression:

$$\begin{aligned} f'^+ &= \frac{|F^+ \cap P^-|}{|F^+|} \\ &= \frac{|F_1^+ \cap P^-| + |F_{>1}^+ \cap P^-|}{|F^+|} \\ &= \frac{|F_{>1}^+ \cap P^-|}{|F^+|} \end{aligned}$$

because of our assumption that all queries in F_1^+ are paedophile, which implies $F_1^+ \cap P^- = \emptyset$.

An estimate of $|F_{>1}^+ \cap P^-|$ is given by $|\overline{F_{>1}^+} \cap P^-| \cdot \frac{|F_{>1}^+|}{|\overline{F_{>1}^+}|}$ which leads to

$$\begin{aligned} f'^+ &\sim \frac{|\overline{F_{>1}^+} \cap P^-|}{|F^+|} \cdot \frac{|F_{>1}^+|}{|\overline{F_{>1}^+}|} \\ &= \frac{15}{204\,846} \cdot \frac{154\,243}{1\,000} \\ &\sim 1.13\,\% \end{aligned}$$

The quality of this estimate is good not only because $|\overline{F_{>1}^+} \cap P^-| = 15$ is significant, but also because we evaluate it using a sample of queries in $F_{>1}^+$, which is much (more than 650 times) smaller than F^-, involved in the estimation of f'^-.

Conversely, the assessment results confirm that estimating $f^+ = \frac{|F^+ \cap P^-|}{|P^-|}$ with our data would yield poor quality approximate, as $|F^+ \cap P^-|$ is small (there are very few paedophile queries), as well as the sample size, compared to the size of P^-.

It is possible to estimate f^- much more accurately. We estimate f^- as follows:

$$\begin{aligned} f^- &= \frac{|F^- \cap P^+|}{|P^+|} \\ &\sim \frac{|F^- \cap (N(F^+) \cap P^+)|}{|N(F^+) \cap P^+|} \\ &= \frac{189}{759} \\ &\sim 24.9\,\% \end{aligned}$$

8 Conclusion and perspectives.

Based on domain expertise and manual inspection of a large dataset, we constructed a filter which tags any query as paedophile or not. In order to assess its performances, we set up an assessment methodology and the associated protocol, which we submitted to experts of the field. Using the answers of 20 such experts, we decided whether each query in various random subsets is paedophile or not, and confronted obtained results with the ones of our filter. Importantly, we were able to rigorously estimate the rates of false positives and negatives, which is crucial for applying our filter to other tasks (like quantification of paedophile activity, for instance [2]).

We finally obtained a rate of false positive $f'^+ \sim 1.13\,\%$, which means that when our filter tags one hundred queries as paedophile one may expect it to be wrong only once. Conversely, the rate of false negative $f^- \sim 24.9$ means that, when one enters paedophile queries into our filter, it will miss one quarter of them. As our goal was to avoid tagging as paedophile queries which are not paedophile, these performances are very good. We succeed in tagging almost no non-paedophile queries as paedophile, while missing only one quarter of paedophile queries, which remains reasonable.

This tool is the first one addressing this important challenge. It succeeds in constructing large sets of paedophile queries, which is of high interest for many applications in fighting paedophile activity. Machine learning methods may be applied to the set of paedophile queries detected by our filter in order to improve it further.

Notice finally that, in some cases, a filter which leads to more false positive but less false negative would have more interest (for instance to explore the variety of paedophile behaviour by pee-filtering the data). Adapting our filter to such contexts is possible, and should lead to variants helpful for other tasks.

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] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an eDonkey server. In *Proceedings of HotP2P'09*, 2009.

[2] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical report on *Quantification of Paedophile Activity in a Large P2P system*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

A Source code of the filter (in *Python*).

```
# input: normalized text queries (one query per line)

# output: supposedly 'paedophile' queries; each line begins
# with an integer which indicates which criteria was used to
# select this query

# Notes:
# . we consider first sub-words, then full words
# . order of tests below is important
# . order influences computation time
# . a query containing _only_ an age is not considered as pedo...

import sys

def contains_word(l1,l2):
for x in l1:
if x in l2:
return(True)
return(False)

def contains_substring(l,s):
for x in l:
if s.find(x) != -1:
return(True)
return(False)

def contains_age_16(suffixes,line):
line = line.replace("you","XXX")
for suffix in suffixes:
l = line.split(suffix)
for a in l[:-1]:
age = a.strip().split()
if len(age) < 1:
continue
age = age[-1]
if age.isdigit():
age = int(age)
if age<16:
return(True)
return(False)


explicit = [
```

```
"babyj", "babyshivid", "childlover", "childporn", "childsex", "childfugga",
"ddoggprn", "hussyfan", "kdquality", "kidzilla", "kingpass", "mafiasex",
"pedo", "pedofilia", "pedofilo", "pedoland", "pedophile", "pedophilia",
"pedophilie", "pthc", "ptsc", "qqaazz",
#"qwerty",
"raygold", "reelkiddymov", "yamad", "youngvideomodels"
]

child = [
"adolescent", "adolescente",
#"baby",
"child", "childern", "children", "childrens", "childs", "enfant", "fillette",
"gamine", "infant", "infantil", "infantile", "infantiles", "kid", "kiddy",
"kids", "kinder", "kindergarten", "menor", "menores", "mineur", "mineure",
"mineures", "mineurs",
#"nina",# caution: nina roberts and others -> porno
"ninas", "nino", "ninos", "nia", "nias", "nio", "nios", "preteen",
"preteens", "underage"
]

sex = [
"abuse", "abused", "abuso", "anal", "animalsex", "ass", "assfuck", "asslick",
"avale", "bath", "bibcam", "bitch", "blowjob", "cum", "cumshot", "defloration",
"dfloration", "dildo", "dogsex", "encule", "enculer", "eurosex", "ficken",
"ficker", "fickt", "fuck", "fucked", "fucks", "fucking", "gay", "groupsex",
"handjob", "hard", "hardcore", "homemade", "incest", "inzest", "kdv",
"lesbian", "lickin", "licking", "loli", "lolita", "lolitaguy", "lolitas",
"lolitasex", "lover", "masterbate", "masterbates", "masterbating",
"masterbation", "masturb", "masturbate", "masturbates", "masturbating",
"masturbation", "masturbe", "masturb", "nackt", "nackte", "nackten", "naked",
"naturist", "nude", "nudist", "nudiste",
#"nu",
#"nue",
"orgasm", "penetration", "pntration", "penis", "pnis", "porn", "porno",
"prostitute", "prostitue", "prostitu", "pussy", "rape", "raped", "salope",
"sado", "sex", "sexe", "sexo", "sexual", "sexually", "shower", "sodom",
"sodomie", "sodomis", "sodomise", "sodomy", "sodomized", "spank", "spanked",
"spanking", "sperm", "suce", "suck", "sucker", "sucks", "swallow", "teensex",
"transexual", "viol", "viole", "viol", "violee", "viole", "webcam", "whore",
"xxx", "zoofilia",
"amamter", "amateur", "amateurs", "amatoriale", "amatrice", "anale", "anus",
"arsch", "baise", "bdsm", "bondage", "chatte", "culo", "cunt", "ejac",
"ejaculation", "erotic", "exhib", "facial", "fellation", "fetish", "fick",
"fisting", "gangbang", "gode", "hentai", "hure", "lesbienne", "lingerie",
"naakt", "orgy", "partouze", "piss", "pornstar", "putas", "pute", "scato",
```

```
"shemale", "sm", "soumise", "sperma", "sperme", "suceuse", "tournante",
"uro", "vicieuse", "voyeur"
]

familychild = [
"baby", "bebe", "boy", "daughter", "fille", "girl", "son", "toddler"
]

familyparents = [
"dad", "daddy", "father", "grandpa", "grandma", "mom", "mommy", "mum",
"mummy"
]

agesuffix = [
"yo", "yr",
#"an",
"ans", "anni", "anos", "aos", "ano", "ao", "jahr", "jahre", "jahres",
"y", "year old", "yearold", "years old", "yearsold"
]

######################
for line in sys.stdin:
if contains_substring(["journal du hard"],line):
continue
line = line.strip()

# first considers the line entirely
if contains_substring(["r ygold"],line):
sys.stdout.write("00 %s\n"%line)
continue;

# then considers the line as a series of words

# very explicit queries
if contains_word(explicit,l):
sys.stdout.write("01 %s\n"%line)
continue

# composed queries
if contains_word(child,l) and contains_word(sex,l):
sys.stdout.write("02 %s\n"%line)
continue
if contains_word(["little","petite"],l) and
           contains_word(["girl","boy","fille","brother","sister"],l)
           and contains_word(sex,l):
```

```
sys.stdout.write("03 %s\n"%line)
continue

# ages
if contains_age_16(agesuffix,line) and
           (contains_word(child+familychild,l) or
            contains_word(sex+["sexy","sexi"],l)):
sys.stdout.write("04 %s\n"%line)
continue

# family
if contains_word(familychild,l) and
           contains_word(familyparents,l) and
           contains_word(sex,l):
sys.stdout.write("05 %s\n"%line)
continue

# too weak
# if contains_age_16(agesuffix,line):
# sys.stdout.write("06 %s\n"%line)
# continue

# # discarded on stderr
# sys.stderr.write("%s\n"%line)
```

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