Tracking emergent keywords with pedophilic context in peer-to-peer file name

David Chavalarias

▶ To cite this version:

HAL Id: hal-00418582

https://hal.archives-ouvertes.fr/hal-00418582

Submitted on 21 Mar 2010

Tracking emergent keywords with pedophilic context in peer-to-peer file name

David Chavalarias, CREA/CNRS

September 21, 2009

1 Tracking emergent keywords with pedophilic context in peer-to-peer file name

As heuristic for this method, we assumed two different intentions for pedophilic actors creating new pedophilic keywords (fig. 1):

1. **Case 1:** Categorize existing files to share with other actors or for own purpose, with few concern about avoiding being detected by anti-pedophilic organizations. In this case, new keywords should tend to strongly co-occur with existing pedophilic keywords and represent subcategories.

2. **Case 2:** Propose new files on peer-to-peer networks to experts and avoid to be detected by anti-pedophilic organizations. Here, new keywords are more like a code name.
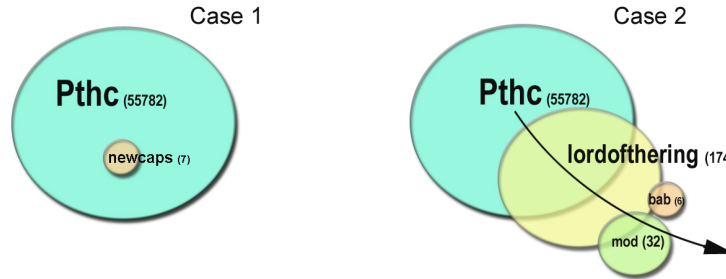
To find keywords corresponding to case 1, we have to find terms that are more specific in their uses than some well known pedophilic keywords, in the sense that the files in which they appear are most often also flagged with these well known keywords. This suggests to take some measure of the inclusion of the extension of a term (*i.e.* the set of files names mentioning the term) in the extension of an other term. This can be done considering measure between terms such that the *pseudo-inclusion measure* ([1]) defined over a period T by[1]: $P_\alpha^T(i,j) = ((\frac{n_{ij}^T}{n_i^T})^\alpha (\frac{n_{ij}^T}{n_j^T})^{1/\alpha})^{max(\alpha, \frac{1}{\alpha})}$.

This measure has the advantage to convey information about the relative position of two terms from the point of view of their use: terms j such that $P_\alpha^T(i,j)$ is close to 1

---

[1] n_i^T (resp. n_j^T and n_{ij}^T) is the number of filenames mentioning the term i (resp. j and both i and j) over the period T.

will contextualize i for $\alpha \gg 1$, and will tend to be more specific in their use relatively to i for $0 < \alpha \ll 1$ [2].

Thus emergent terms corresponding to case 1 will have a strong proximity measure to well know pedophilic keywords for $\alpha \gg 1$.

As for case 2, one of the expected behavior from users is to introduce new keywords that only advanced users will have heard about[3]. To avoid these files being associated to pedophilic content, and thus detected by third parties, these keywords will only weakly co-occur in filenames with well known pedophilic keywords. We don't expect them either to have much occurrences with other commons (non pedophilic) filename terms. Yet, whether intentionally or by mistake, these new keywords should have a larger proportion of co-occurrences with pedophilic keywords than with non-pedophilic ones. The renewal of these pedophilic keyword should lead to a sliding structure of overlapping extensions (fig. 1). To detect these keywords, the important factor is thus not the *strength* of their links to other pedophilic keywords (or the degree of inclusion of their extension in the extension of other pedophilic keywords), but how many pedophilic keywords appear in their closest contexts.

These remarks lead to two methodologies according to wether we are looking for case 1 or case 2 pedophilic terms.

1.1 General methodology

1.1.1 Extending the seed

Given a seed (with or without prior knowledge), we can first improve it by looking for terms with about the same level of pedophilic content. If the seed is made of some pedophilic keywords, this will lead to a bigger seed with other pedophilic keywords most of them having about the same degree of granularity. This can be made taking $\alpha = 1$ in P_α and then retaining all terms at a proximity to the seed higher than a given threshold. In this study, we also limited the minimum number of occurrences (e.g. $P_\alpha > 0.01$ and more than 10 occurrences in our study).

1.1.2 Finding words with pedophilic contexts

We computed the terms that are better contextualized by the extended seed setting $\alpha = 10$. For each word of more than 3 occurrences, we assigned a rank defined as the lowest rank of a keywords from the extended seed in the n closest neighbors of the target word (e.g. $n = 20$). The methodology can be divided into two methods, M1 and M2, which will tend to be more adapted respectively to case 1 and case 2.

- **M 1:** Sort terms in function of their lowest P_α to one word of the seed.

- **M 2:** Sort terms in function of their rank in ascendent order and then in function of their proximity to the seed (in terms of P_α).

In both case, we get for each target keywords diverse indications to guide assessment of pedophilic content:

- **Rank of the target keyword** defined as the lowest rank of a keyword from the extended seed in the n closest neighbors of the target word (e.g. $n = 20$),

---

[2]Note that $P_\alpha^T(i,j) = P_{\frac{1}{\alpha}}^T(j,i)$ so that if j specifies i, i contextualizes j. Moreover, $lim_{\alpha \to \infty}(P_\alpha(i,j))$ is the inclusion measure over the sets of papers mentioning i and j.

[3]An other possibility would be to identify new pedophilic file by a tagging that associates some existing keywords in an unused way.

- **Proximity to the seed**, defined as the proximity to the target keyword of the closest term from the extended seed,

- **Context at** n**:** tells what keywords from the extended seed appear in the first n contexts and what are they rank,

- **Number of occurrences of the target keyword** (Emergent keywords are expected to have few occurrences).

These indicators can be combined after experts feedback to improve the method.

References

[1] 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.