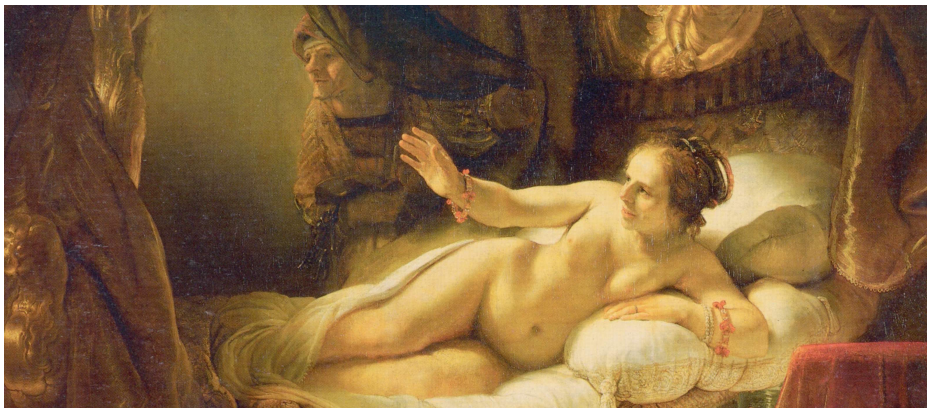

Search and Exploration of X-rated Information

WSDM'16 Workshop Proceedings

Vanessa Murdock, Charles L.A. Clarke, Jaap Kamps, Jussi Karlgren (editors)

February 22, 2016
San Francisco, USA
http://sexi2016.org/

Cover image: Rembrandt Harmenszoon van Rijn. Danaë. 1636. Oil on canvas. Hermitage, St. Petersburg, Russian Federation.

Preface

These proceedings contain the research and position papers of the WSDM'16 Workshop on *Search and Exploration of X-rated Information*, held in San Francisco, USA, on February 22, 2016.

Adult content is pervasive on the web, has been a driving factor in the adoption of the Internet medium, and is responsible for a significant fraction of traffic and revenues, yet rarely attracts the attention of researchers. This half-day workshop on Search and Exploration of X-Rated Information at the 2016 WSDM conference will discuss questions for information access tasks and search behavior related to adult content. While the scope of the workshop remains broad, we will devote special attention to the privacy and security issues with respect to adult content. The recent release of the personal data belonging to customers of the adult dating site Ashley Madison provides a timely context for the focus on privacy and security. The data collected by adult sites, derived from both visitors to the site and providers of content, is arguably more sensitive than other commercial data, because of the controversial nature of the sites themselves.

The workshop consisted of three main parts:

- First, introduction by the organizers to frame the problems, and outline potential solutions.
- Second, paper sessions with two papers selected by the program committee. Each paper was reviewed by at least two members of the program committee.
- Third, break out groups on different aspects of information access tasks related to adult content, and a panel discussing the results of the workshop in the final slot.

When reading this volume it is necessary to keep in mind that these papers represent the ideas and opinions of the authors who are trying to stimulate debate. It is the combination of these papers and the debate that made the workshop a success.

We like to thank the ACM and the WSDM for hosting the workshop, and local chairs for their outstanding support in the organization. Thanks also go to the program committee, the authors of the papers, and all the participants, for without these people there would be no workshop.

January 2016

Vanessa Murdock
Charlie Clarke
Jaap Kamps
Jussi Karlgren

Table of Contents

Front Matter.

Preface iii

Table of Contents v

Overview.

Current Research on Search and Exploration of X-rated Information 7
Vanessa Murdock, Charles L. A. Clarke, Jaap Kamps, Jussi Karlgren

Submitted Papers.

Multimedia Metadata-based Forensics in Human Trafficking Web Data . . 10
Chris Mattmann, Grace Yang, Harshavardhan Manjunatha, Thamme Gowda N, Andrew Jie Zhou, Jiyun Luo and Lewis John McGibbney

Efficient filtering of adult content using textual information 14
Thomas Largillier, Guillaume Peyronnet and Sylvain Peyronnet

Back Matter.

Author Index 18

Current Research on Search and Exploration of X-Rated Information

Vanessa Murdock
Microsoft
vanmur@microsoft.com

Charles L.A. Clarke
Waterloo University
claclark@plg.uwaterloo.ca

Jaap Kamps
University of Amsterdam
kamps@uva.nl

Jussi Karlgren
Gavagai & KTH Stockholm
jussi@kth.se

ABSTRACT

Adult content is pervasive on the web, has been a driving factor in the adoption of the Internet medium, and is responsible for a significant fraction of traffic and revenues, yet rarely attracts attention in research. The research questions surrounding adult content access behaviors are unique, and interesting and valuable research in this area can be done ethically. WSDM 2016 features a half day workshop on *Search and Exploration of X-Rated Information* (SEXI) for information access tasks related to adult content. While the scope of the workshop remains broad, special attention is devoted to the privacy and security issues surrounding adult content by inviting keynote speakers with extensive experience on these topics. The recent release of the personal data belonging to customers of the adult dating site Ashley Madison provides a timely context for the focus on privacy and security.

CCS Concepts

•**Information systems** → Information retrieval;

Keywords

Adult content; Privacy and security; Research ethics; Research practice

1. INTRODUCTION

The second workshop on *Search and Exploration of X-Rated Information* (SEXI) for information access tasks related specifically to adult content was held at WSDM'16 in San Francisco, building on the success of the first workshop at WSDM'13 [5, 6]. The WSDM'13 workshop was the first of this kind that has been presented in the web mining and information retrieval communities, and for WSDM 2016 we proposed a second workshop that brings these issues to the fore. We intend an open and respectful discussion of issues about adult information access, which will illuminate the areas in which adult information access, and user-generated adult content differ from standard information access, and standard user-generated content. We seek to define a set of research areas which represent ethical and legal opportunities to explore the research questions surrounding adult content. Special care is given to ensure that no adult content is actually presented at the workshop, and that the discussion remains respectful and professional, and at the same time fun. While the scope of the workshop remains broad, the workshop has special theme: privacy and security issues surrounding adult content.

The recent release of the personal data belonging to customers of the adult dating site Ashley Madison provides a timely context for the focus on privacy and security [1] The data collected by adult sites, derived from both visitors to the site and providers of content, is arguably more sensitive than other commercial data, because of the controversial nature of the sites themselves.

Adult content is pervasive on the web, has been a driving factor in the adoption of the Internet medium, and is responsible for a significant fraction of traffic and revenues, yet rarely attracts attention in research [1, 2]. The scientific community has spent considerable energy studying user-generated content and information access on the web, to the exclusion of adult content. This is understandable, as the topic is distasteful to some, and requires special legal and ethical considerations when asking employees, contractors and students to analyze and process the data. Furthermore, methods that work for other types of information access behavior are assumed to work for all types of content, including adult content.

We argue that this is an incorrect assumption. In fact, even core concepts such as relevance and diversity, which are fundamental to any application involving information seeking and access, are defined differently for adult content. Adult queries frequently fall outside of the taxonomy of queries (informational, transactional, navigational) that applies to standard web queries. Users searching for adult content frequently have an entertainment need, rather than an information need. Thus, because of the nature of the content, the user may be more satisfied with multiple similar images, than with a set of search results that capture different meanings of the query terms. Furthermore, understanding that a user is searching for a term in an adult context often disambiguates the term.

For example, consider the query "bikini." In a non-adult

[1] http://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/ visited December 2015

context, the user might be querying about a ham and cheese sandwich, or may be interested in viewing catalog photos of bikinis to purchase, or perhaps would like images of the Bikini Islands. Since the search engine cannot be sure, one strategy is to represent multiple senses of the query in the results presented to the user. In an adult context, images of catalog photos, sandwiches and islands will be more than an annoyance. Images of people wearing bikinis, although repetitive and not representing a diverse sense of the query term, is what the user is expecting.

Equal in importance to serving adult content in the best possible way, is the issue of avoiding serving adult content to those who are not looking for it. Many innocuous terms (such as "snake," "cougar," "swimsuit") have adult connotations. Understanding when a person uploading content is uploading adult content is important. Often the only information available to determine the adultness of an image or video are the vague tags the user applies. Complicating the interpretation of the tags is that adult content may be described euphemistically with ordinary nouns that reflect a particular visual imagery. Similarly, when a person issues a query, it is not always clear whether they are searching for adult content, and it is extremely important for the search engine to understand this before serving adult content to a person who is not expecting it.

2. OPEN QUESTIONS

The workshop seeks a greater understanding of the particular issues in accessing adult content, especially user-generated adult content on the web. The discussion is limited to adult content that is legal, although topics such as identifying online predators, child pornography, or human trafficking are within of the scope.

The focus of the workshop on privacy and security issues surrounding adult content, was chosen in order to put this area of research on the agenda, and explore the basic research questions that should be addressed in the field, the types of data needed for research, and and the barriers to doing research this area. Due to the lack of attention to this area of research there are many open questions. These questions include:

Classification.

Even researchers and search applications not interested in adult content will have to deal with it in order to avoid it—presenting adult content to innocuous searchers is clearly a massive failure both for the individual searcher as well as for the reputation of the service. What are automatic methods for identifying adult content, in particular adult user-generated content? How can we identify adult content in video, images, and text? What is the best way to identify adult query intent, and deal with ambiguous requests? What are the appropriate ad placement strategies in adult content?

Access.

Access to adult content seems to require a different approach than the ubiquitous navigation search—with searchers exhibiting an exploratory information seeking behavior, characterized by a diverse set of relevance criteria. How should adult content be ranked? How should search, exploration, and recommendation be balanced? How does searching adult content relate to search on adult chat sites and social networks? Is there a benefit to personalizing adult content?

Evaluation.

Given the distinct nature of adult content and the diverse relevance criteria, appropriate evaluation is crucial. What is a relevant result, and what are suitable metrics for relevance? Is adult content a recall-oriented, or precision-oriented task? What is the right level of evaluation—individual requests or whole search sessions? What is similarity and diversity in adult content? How important is the avoidance of failure, relative to success? Are searchers for adult content more tolerant of non-relevant results?

Ethics.

What are the ethical issues in working with adult content in an academic environment? What are the ethical implications for the search industry, given that it partly facilitates the online adult industry? How can adult material be made available so as to promote responsible behavior through the whole chain from production to consumption? Is adult user-generated content more ethical than professionally produced media?

Security and Privacy.

Adult content is one of the primary vehicles for malware on the internet. In addition, as many adult content sites collect personally identifying data (such as user names and credit card numbers), particular care must be taken to protect the data collected from visitors to the sites, as well as content providers. While users of these sites are not typically doing anything illegal, they provide an enticing target to thieves because of the controversial nature of the industry. In addition, as the adult entertainment industry moves toward content generated by private citizens, and away from large commercial producers, the potential for harm to private citizens producing the content increases.

The workshop aims to define a set of research areas, to elucidate the special issues surrounding the access of user-generated adult content. A set of best-practices for working with this data in an academic environment was discussed, and a research agenda for the near future was proposed. Special care is taken to select and moderate the program to present the information in a respectful and scientific manner.

3. FORMAT AND PROGRAM

The workshop is planned to feature two keynote speakers, the first keynote addressing the technical engineering challenges of given access to, or avoiding, adult content on the Web, and the second keynote addressing the social aspects of online adult content.

Accepted papers include Largillier et al. [3], who investigate the efficient filtering of adult web content, aiming to prevent surfacing this content in the wrong context. Mattmann et al. [4] study detecting human trafficking based on crawling ads related to them, revealing invaluable cues for law enforcement and governmental organizations to identify victims and intervene to aid them.

Short presentations by participants are also planned, along with a closing panel. A full report of the workshop will appear in the June 2016 issue of SIGIR Forum.

REFERENCES

[1] L. Azzopardi. Searching for unlawful carnal knowledge. In N. J. Belkin, C. L. A. Clarke, N. Gao, J. Kamps, and J. Karlgren, editors, *Proceedings of the SIGIR'11 Workshop on "entertain me" : Supporting Complex Search Tasks*, pages 17–18. ACM Press, 2011.

[2] A. C. Halavais. Small pornographies. *SIGGROUP Bulletin*, 25(2):19–22, 2005. URL http://doi.acm.org/10.1145/1067721.1067725.

[3] T. Largillier, G. Peyronnet, and S. Peyronnet. Efficient filtering of adult content using textual information. In Murdock et al. [7], pages 14–17.

[4] C. Mattmann, G. Yang, H. Manjunatha, T. G. N, A. J. Zhou, J. Luo, and L. J. McGibbney. Multimedia metadata-based forensics in human trafficking web data. In Murdock et al. [7], pages 10–13.

[5] V. Murdock, C. L. A. Clarke, J. Kamps, and J. Karlgren. Report on the workshop on search and exploration of x-rated information (SEXI 2013). *SIGIR Forum*, 47(1): 31–37, June 2013. http://dx.doi.org/10.1145/2433396.2433507.

[6] V. Murdock, C. L. A. Clarke, J. Kamps, and J. Karlgren, editors. *SEXI'13: Proceedings of the WSDM'13 Workshop on Search and Exploration of X-rated Information*, 2013. ACM Press.

[7] V. Murdock, C. L. A. Clarke, J. Kamps, and J. Karlgren, editors. *SEXI'16: Proceedings of the WSDM'16 Workshop on Search and Exploration of X-rated Information*, 2016.

Multimedia Metadata-based Forensics in Human Trafficking Web Data

Chris A. Mattmann[1,2], Grace Hui Yang[3], Harshavardhan Manjunatha[2], Thamme Gowda N[2], Andrew Jie Zhou[3], Jiyun Luo[3], Lewis John McGibbney[1]

[1]Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA

[2]Computer Science Department
University of Southern California
Los Angeles, CA 90089 USA

[3] Department of Computer Science
Georgetown University
Washington, DC, 20057 USA

mattmann@jpl.nasa.gov, huiyang@cs.georgetown.edu,

ABSTRACT

Crawling the web for ads related to human trafficking yields a wealth of traditional static and dynamic web page content. Ads in which victims are trafficked by their predators are often littered with textual signals such as physical characteristics (hair, body type, race, ethnicity, etc.), and location information (city, state) that can aid law enforcement and non governmental organizations in identifying victims and intervening to aid them. The ads also carry a strong multimedia footprint, as there are increasingly images and videos to accompany the ad text. Many groups have researched computer-vision (CV) based approaches to analyze these images and videos to extract meaningful features that when combined with the physical characteristics and location information can greatly aid in thwarting these activities. However CV approaches are computationally expensive and further they may not be able to discriminate between images and videos taken in similar lighting and color situations as they rely greatly on image analysis which is limited to scene properties. Our team has investigated and created a preliminary system that takes advantage of image and video multimedia metadata – or information about the actual images and videos typically recorded when they are created. This metadata can be leveraged to provide valuable discriminatory signal to aid in search and relation of multimedia data while saving computational cost and also providing a novel complimentary feature set when CV techniques are unable to differentiate between images and videos.

CCS Concepts

• **Information systems~Information retrieval** • **Information systems~Content analysis and feature selection** • Information systems~Similarity measures • Information systems~Information extraction.

Keywords

Apache Tika, Jaccard Similarity, Metadata, Multimedia, Forensics

1. INTRODUCTION

Human trafficking (HT) is a global modern phenomenon in which modern day slavery is enabled through the public Internet. Women, men, and children are sold for sexual slavery, labor-based slavery and for other potentially criminal activities [1]. The ease of use of the Internet as a large bulletin board complete with the ability to upload images and videos to describe the victims has contributed greatly to the growth of HT throughout the world. The size of publicly available HT data is estimated to be around 60 million ads, and 40 million images and 10s of thousands of videos already eclipsing the ability for single computers to process and analyze the information.

Law enforcement, non governmental organizations (NGOs) and other groups are interested in collecting publicly available ad postings, multimedia information (images, video, etc.) and other data to assist in identifying human trafficking victims, and in finding and prosecuting predators responsible for the trafficking. This data can be collected by using web crawler software to download public ad and bulletin board data from HT websites. Crawling the web for ads related to human trafficking presents an interesting challenge both in scale in terms of the number of web sites and number of ads and multimedia, but additionally in terms of search and information retrieval. Ads for trafficking victims regularly have textual based signals that can aid law enforcement and NGOs including physical information about the victim such as race, ethnicity, gender, body and hair type, etc., along with physical and forensic information about the location in order to connect buyers or "johns" through the Internet to the trafficking victim. Textual signals are an important element that can be used to identify victims and relate them together. Manual approaches for analyzing textual signals in HT ads remain in use by NGOs and law enforcement today.

Beyond textual signals are the rich multimedia images and videos present that can be leveraged along with the textual ad data for search and retrieval in the HT domain. Current research is focused on computer vision (CV) techniques [2] and in advanced approaches such as machine learning and deep learning on CV [3] to relate together images and videos and to relate objects in both to scenes, and to places and things. Some limitations of CV based approaches to multimedia search are their computational cost. Even if resources are available to leverage CV, CV techniques are typically based on image analysis and thus relations made between images, videos, and further they may not be able to discriminate between dissimilar images and videos taken in similar lighting and color situations. This is most often seen in cropped images and videos that CV techniques have difficulty discerning.

Our team has investigated and created a preliminary system that takes advantage of image and video multimedia *metadata* – or information about the actual images and videos typically recorded when they are created. The system exploits content creation metadata as it propagates throughout the image and video lifecycle: from creation to editing and manipulation and to dissemination. Metadata includes useful properties related to both the physical properties of the content (RGB color space; whether or not the flash fired;) to geo-location, to information about the instrument that captured the multimedia (camera/phone Make/Model; Serial number;) to other information such as creator, date/time the multimedia was generated. Our system, *Image Space*, leverages metadata to overcome CV-only oriented approaches, and to compliment image and video analysis techniques, with metadata-based forensics to better combine multimedia with ad-based data in the HT domain. In particular, our techniques have shown to identify image and video relationships to ads, and to identify relationships between victims and predators in ways otherwise not possible without our approach.

The rest of this paper is organized as follows. Section 2 describes Image Space, our metadata forensics toolkit for multimedia. Sections 3 and 4 describes our algorithmic approach for relating images and videos together based on multimedia metadata and on domain dynamics. Section 5 concludes the paper by identifying next steps and future work.

2. A METADATA FORENSICS TOOLKIT FOR MULTIMEDIA

We have constructed the ImageCat and ImageSpace architecture depicted in Figure 1. ImageCat – short for "Image Catalog" and shown in the bottom middle of Figure 1 – is an Extract, Transform and Load (ETL) system to automatically create a search index of Image and Video similarity metadata descriptors by extracting that information from a collected set of multimedia data (10s of millions of images/videos). Though we have fielded ImageCat in the HT domain, it is also applicable to any multimedia data and images collected on the Internet. ImageCat uses Apache Tika [7] to automatically identify multimedia files and their type, and in turn to invoke open source, third party parsing libraries on the multimedia data. For example, EXIFTool is a useful Perl-based program that extracts EXIF image and video metadata – Camera and scene properties, make, model, color space information, etc. – and is integrated into Tika. FFMPEG extracts video information such as bitrate, tracks, scene properties, etc. – and is integrated into Tika. Besides EXIF metadata and other descriptors, Apache Tika integrates the Tesseract [4] library to perform Optical Character Recognition (OCR) and to extract text descriptors from multimedia data as well – as shown in the middle right portion of Figure 1. We will detail the use of OCR later in this section.

Apache Tika is integrated into the Apache Solr search indexing system via a plugin called "SolrCell" that automatically runs Tika on the server side during the indexing process. To perform indexing of the information from crawled web data, a list of image file paths, possibly as large as tens of millions of images, is presented to ImageCat (left side of Figure 1), and those images are run through Apache OODT [5] a data-flow oriented ETL system. OODT's File Manager (FM) tracks and records file

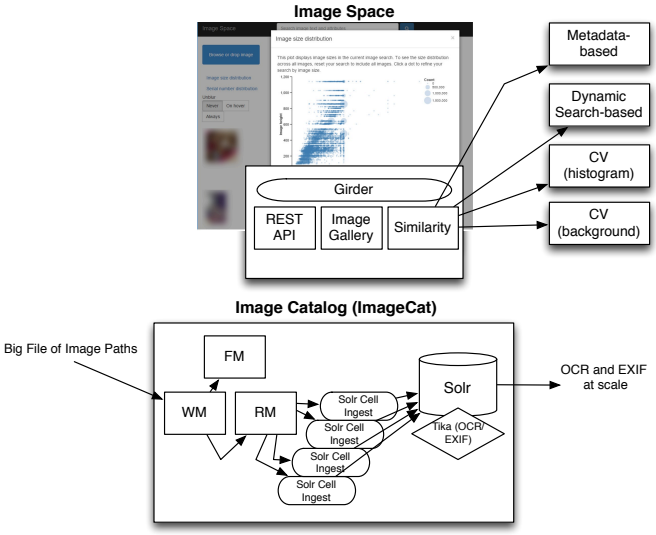

Figure 1. The Image Space and Image Cat architecture.

locations and metadata; its Workflow Manager (WM) manages provenance of the ETL pipelines, and the Resource Manager (RM) schedules the ETL jobs on a large cluster or the cloud. Each job in ImageCat is an ingest job that sends the image or video file to SolrCell for server-side Tika-based extraction and for building the inverted index and catalog. OODT records all the provenance information from file to ETL to resource scheduling. This is useful information especially since ImageCat can quickly be constructed, and/or torn down depending on adaptations to Tika and when changes or fine-tuning are necessary.

OCR and EXIF metadata are particularly important to images and videos crawled from the HT domain. Victims usually hold placards containing vital information like phone number, email address, which can be parsed by Tika and Tesseract. Although the OCR from Tesseract contains noise, partially parsed text can be leveraged as keywords in search against Solr and has been found to be effective.

ImageSpace is a front-end web application for ImageCat. It connects directly to an ImageCat and can be used for searching and querying browsing large collections of multimedia information in an efficient manner using Apache Solr. ImageSpace is built on top of the Girder web framework, and it presents a REST API (shown in the upper left of Figure 1) for interactively extracting multimedia metadata with Tika; for performing similarity comparisons based on metadata and CV techniques, and by temporally associating user's queries as we will explain in Section 4. ImageSpace supports both text-based search and multimedia based searches against any of the extracted metadata properties. ImageSpace provides an interactive density plot feature that visualizes the distribution of images (based on Image Size). Similarly ImageSpace also provides an interactive histogram feature (binned on camera serial numbers) to refine your current image search session or across all images in the index. The Text based approach searches based on the parsed textual content (e.g., OCR) of images and the value of other image metadata attributes. Common text-based multimedia search queries include email address, etc. The image-based approach allows users to search for similar images by uploading a query image onto the ImageSpace application, or by selecting one of the images from the text based search results. Upon uploading or selecting an image, the metadata attributes of that image are displayed along with the various similarity metric options to

search upon using the chosen image as a query. The similarity metrics are indicated in the upper right portion of Figure 1, and correspond to image content (histogram), background of the image, size/resolution of the image, and other metadata attributes. An experimental temporal image search is described in Section 4. Our multimedia similarity metric is described in the next section.

3. METADATA BASED MULTIMEDIA SIMILARITY

Our team has also derived a metadata-based similarity metric allowing for clustering and grouping of multimedia data forensically, without relying on computationally burdensome computer vision (CV) techniques. The approach is derived from an observation that metadata creators and content creators leave a forensic footprint that propagates through the content dissemination process – from authorship and tool (e.g., Camera; Video Recorder; Phone) to editing (Photoshop, GIMP, etc.) to delivery (web-server Headers etc.) to browser or image/video reader.

Algorithm 1. Tika Jaccard Similarity

```
1 input: directory of files d
2 output: scores s for all files in d
3
4 goldSet:= {}
5 allMetadata:= {}
6 scores:= {}
7
8 for file in d:
9   text, metadata:= tika.parse(file)
10     goldSet:= goldSet ∪ metadata.keys
11     allMetadata[file]:= metadata
12
13 goldenSetSize:= |goldSet|
14
15 for file in allMetadata.keys:
16     overlap:=|allMetadata[file]∩ goldSet|
17     score:= overlap / goldenSetSize
18     scores[file]:= score
19
20 return scores
```

The approach, shown in Algorithm 1, builds upon the *Jaccard* similarity algorithm [1]. Leveraging an Image or Video's metadata *footprint*, the algorithm works as follows. We leverage the extracted Tika metadata and text descriptors from ImageCat. A golden feature set is computed by iterating all the files in a given directory and extracting out metadata key names using Tika (e.g., *EXIF Flash*, *RGB Color Space*, *Camera Make*, *Camera Model Serial Number*, etc.) and/or their discretized value space (line 10 in Algorithm 1). The features are collected across all multimedia files, as is the per file metadata. A second pass through the set of files in line 15 shown in Algorithm 1 is performed so that for each set of per file extracted metadata, the intersection of each file's space can be computed and compared with the entire set of features in the golden feature set (read: *all metadata property names and/or values* across the entire set of collected data). This allows a distance metric to be derived that shows each files computed distance from the golden set, allowing for metadata-based multimedia characterization.

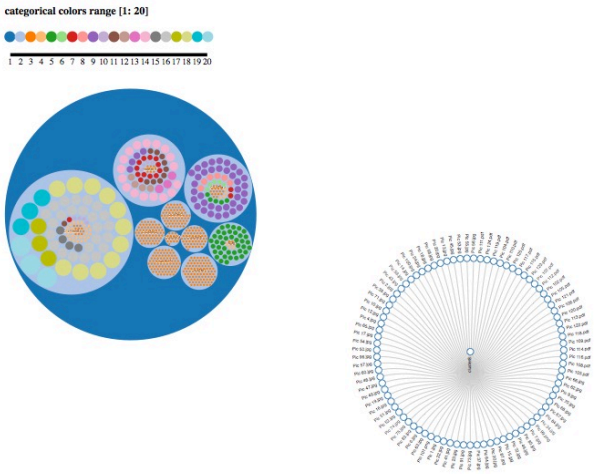

Figure 2. Image Similarity Computed from H/T data.

Once the scores are computed from Algorithm 1, a simple clustering threshold technique can be applied to visually create image, video and multimedia clusters based on metadata names and/or value spaces. These clusters can then be visually represented using our technique built upon the Data-Driven Documents (D^3) framework [8] as shown in Figure 2. In the upper left are the cluster groups derived from Jaccard's coefficient differencing. Each group represents a set of items that share the same metadata features. In the bottom right are the actual images and/or videos present in the cluster. Color in the upper left clusters are per feature and indicate the number of times that metadata field is present in the cluster, and the density of the circles are used to represent the number of files in the cluster.

What we find in practice using this technique is that it naturally groups multimedia files that were either edited, created, captured, or modified in the same vein and/or fashion. This can derive and provide new meaning unable to be discerned visually or using CV techniques. For example, scenes and pictures, and videos in which the objects in the scene and/or background are not related by color or by histogram or by general look and feel can be related in the sense that the pictures were captured using the same camera; or same camera type; or same camera settings which provide information as to the content creator. In the HT domain, this is typically the predator (and also depending on the stage of editing, the victim) and so metadata multimedia forensics can be a useful tool in augmenting existing CV-oriented techniques with new ways to associate multimedia information with textual ad-based features – the domain of trafficked weapons yields similar results. In the next section we will describe another approach that our group is pioneering to relate images together based on the way that users query ImageSpace and ImageCat for multimedia information.

4. DYNAMIC SEARCH OF HUMAN TRAFFICKING DATA

Dynamic search is an emerging topic in Information Retrieval (IR) research [11]. In dynamic search, we model dynamic systems that change over time or a sequence of events using artificial intelligence and reinforcement learning.

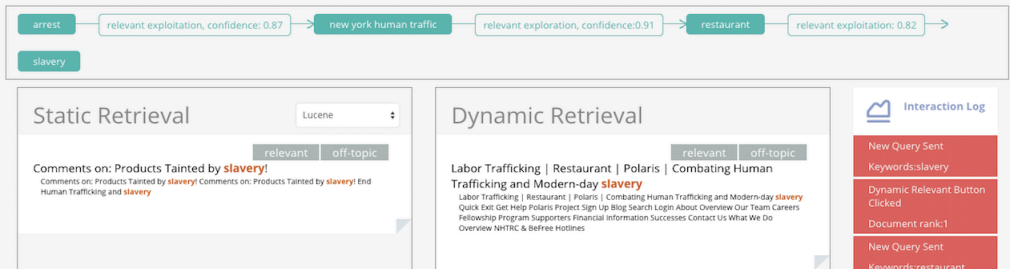

Figure 3. Dynamic Search Engine Built for Human Trafficking Dataset

In a dynamic search setting, a user issues multiple queries during a session to accomplish a search task. The common process is that the user sends a query, gets the top ranked documents, changes her query and sends it again. As a result, a series of queries, a series of retrieved documents, and rich user interactions will be generated, until the session stops when the user finishes her search task. During a session, the temporal dependencies between queries are presented in that way that the previous queries and the previously obtained search results will influence how the user issues the current query and how the search result rankings could be optimized for the current query.

We have preliminary integrated dynamic search into our ImageSpace application. Figure 3 shows the interface of a dynamic search engine developed for the HT dataset and integrated as a similarity metric into ImageSpace. Our dynamic search metric supports the digital forensic process that requires detailed, iterative, and complex queries and is based on the metadata we extract using SolrCell and Tika. We have demonstrated that the dynamic search process is more effective than using a static off-the-shell search engine especially when users are able to use ImageSpace and query it over time.

For instance, in a case to discover all the massage parlor multimedia data related to arrests in the United States, the user entered a series of queries into ImageSpace. The first query is "arrests". The user scans the images and videos and finds that the second image is relevant. After clicking the second image, the user found the phrase "New York" in the extracted second iamge and used it to form the second query "New York human trafficking". In the third returned result, a video, the user found a name "Christopher Robert" – a predator whose name was present in the metadata authorship from a picture taken with his camera - and used it as the next query. Because the dynamic search engine always keeps the context in mind, the search for a person named Christopher is not out of the scope of human trafficking, while a static search engine would return persons outside of this criminal domain on the top of the document list without saving this context. We are excited by the preliminary results from applying this approach to multimedia data and look forward to next steps.

5. CONCLUSION AND FUTURE WORK

We have described our approach to multimedia forensics in the Human Trafficking domain and more generally for relating images and videos together with text when querying information collected from the web. While the early results from our approach appear promising, a number of pertinent areas remain unexplored.

The cosine distance algorithm is under evaluation for more precise metadata clustering. In addition, we are exploring the effectiveness of dynamic search similarity relating to traditional CV approaches and open datasets from the ACM Multimedia conference. We are also adding additional video similarity metrics including the Pooled Time Series approach. Finally we are addressing some limitations of the approach including discerning the optimal metadata for clustering; finding more scalable ways (e.g., REST-ful services) to invoke additional extractors and choosing an optimal set of extractors for the same content types.

6. ACKNOWLEDGMENTS

This work was supported by the DARPA XDATA/Memex program. In addition, the NSF Polar Cyberinfrastructure award numbers PLR-1348450 and PLR-144562 funded a portion of the work. Effort supported in part by JPL, managed by the California Institute of Technology on behalf of NASA.

7. REFERENCES

[1] New Search Engine Exposes the Dark Web, http://www.cbsnews.com/news/new-search-engine-exposes-the-dark-web/, Accessed: November 2015.

[2] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision* 8.2 (1992): 99-111.

[3] M. Abadi, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. tensorflow.org, 2015.

[4] R. Smith. An overview of the Tesseract OCR engine. *IEEE ICDAR* 2007.

[5] C. Mattmann, et al. A reusable process control system framework for the orbiting carbon observatory and NPP Sounder PEATE missions. *IEEE Space Mission Challenges for Information Technology*, 2009.

[6] Jaccard Index, https://en.wikipedia.org/wiki/Jaccard_index, Accessed: November 2015.

[7] C. Mattmann and J. Zitting. *Tika in Action*. Manning Publications, 2011, 256 pages.

[8] M. Bostock, O. Vadim and J.effrey Heer. D³ data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17.12 (2011): 2301-2309.

[9] Foto Forensics, http://fotoforensics.com/tutorial-meta.php, Accessed: November 2015.

[10] M. A. Anoop. Image forgery and its detection: A survey. IEEE International Conference on Innovations in Information, Embedded and Communication Systems, 2015.

[11] Jiyun Luo, Sicong Zhang, Hui Yang. Win-Win Search: Dual-Agent Stochastic Game in Session Search. In Proceedings of the 37th Annual ACM SIGIR Conference (SIGIR 2014).

[12] Hui Yang, Marc Sloan, Jun Wang. Dynamic Information Retrieval Modeling. Tutorial in the 37th Annual ACM SIGIR Conference 2014 (SIGIR 2014). Gold Coast, Australia.

Efficient filtering of adult content using textual information

Thomas Largillier
Normandy University
Caen, France

Guillaume Peyronnet
Nalrem Medias
Paris, France

Sylvain Peyronnet
Qwant & ix-labs
Rouen, France

ABSTRACT

Nowadays adult content represents a non negligible proportion of the Web content. It is of the utmost importance to protect children from this content. Search engines, as an entry point for Web navigation are ideally placed to deal with this issue.

In this paper, we propose a method that builds a safe index *i.e.* adult-content free for search engines. This method is based on a filter that uses only textual information from the web page and the associated URL.

Keywords

Adult content filtering

1. INTRODUCTION

Protecting youth from adult content on the web is an important issue. A study by Sabina *et al.* [10] shows that 93% of boys and 62% of girls are exposed to online pornography during their adolescence. The mean age at first exposure to adult content is around 14.

It is important to filter out adult content for many reasons. We see in [10] that not all exposures to pornography are on purpose. More precisely, almost 7% of boys and 42% of girls answering to the study stated that they never looked for pornography on purpose. We also see in this study that the exposure to deviant sexual activity and child pornography (which the viewing is illegal) is far from being negligible. There is thus a real issue about filtering pornography, and more generally adult content. It is also worth noting that it is legally forbidden in most countries to allows or facilitate the access to pornography to minors.

To ensure a safer Web experience search engines decided to offer a *safe search* option which objective is to remove adult content from search engines results pages.

In this paper, we propose a methodology to construct an adult content free index for a search engine. This index leads to a *safe search engine* rather than an option that can be deactivated.

SEXI 2016, San Francisco, USA

DOI: 10.1145/1235

Our methodology consists in an algorithmic pipeline which main element is a decision forest generated by a supervised machine learning algorithm.

We chose to construct a Web index without any adult websites rather that flagging in a standard Web index websites containing unsafe content for a specific reason. Indeed, when a user enters a query into a search engine, the pages that are output as relevant for this query are those that have high popularity (in term of pagerank for instance) and that have a content relevant to the query. By removing most of the adult websites from the index, we nullify the popularity of the remaining ones (websites tend to receive links mostly from pages in the same topical cluster). Moreover, our filter is based on the textual content of the websites, meaning that adult websites that are missed by the filter does not have an enough adult content to rank on adult queries. However, adult websites consisting of only images and videos, or with very little textual content will not be filtered by our method.

Finally, Our approach is interesting for several reasons: it is efficient, it is fast and the few false negatives are demoted in the search engines results pages, so most of the time nobody will see them.

The structure of the paper is as follows, section 2 presents the related work. In section 3 we give the general architecture of our filter. In section 4 we describe our methodology, the attributes we analyse on Web pages and the experimental results we obtained.

2. RELATED WORK

Identifying adult content on the Web is an active topic since the democratization of the Web. It is a problem of the utmost importance since children have an ever easier access to online resources that was extensively addressed in the literature.

Most techniques focus on adult content detection in media like images. Chan *et al* introduce in [2] the idea of using skin related features in images to detect pornography. However this first step prove efficient to detect images containing a lot of "skin" pixels but was not precise enough to efficiently detect pornographic content.

Rowley *et al* present in [9] the filtering system used in Google at the time. Their objective is to provide a very fast method since the volume of data they have to classify is tremendous. They train a SVM on 27 features based on skin and form detection. Their results are not spectacular but they are working on a real life dataset of over 1 billion images that is really difficult.

Hammami *et al* developed Webguard [4, 3], a tool for fil-

tering adult content on the Web. Webguard uses textual, structural and visual information on a Web page before taking a decision. It was the most complete tool at the time and it outperformed all its competitors. The authors show in [3] that textual and structural information already do a fantastic job at detecting adult content. Visual information only being used to detect false negative and improving the efficiency of the method by a margin. In their papers the authors give no indication on the performances of their tool.

Jansohn *et al* in [5] focus on detecting adult content in videos. Their approach consists on working on images extracted from the video as well as motion features extracted from the video. Using motion histograms together with a bag of visual approach yields very good results for classifying videos.

The following two papers have the same objective as ours, protecting children from pornography. They both focus on adult content accessed using mobile devices.

Amato *et al* introduces in [1] a parental control tool that tests images received on a mobile device before granting access to it. Their method intercepts images notification and transfer the image to a remote server that can classify the image before returning the result. If the image is inoffensive the notification is put back in the mobile device queue, if the image is not suited for a child it is simply deleted and the user wont even know he has received offensive content. Their process rely on a existing image classification system and runs in seconds per picture which is totally untractable for running during indexation.

Park and Kim proposes in [6] an authentication system to access restricted content in Mobile RFID service environment. Their system proposes a better anonymity for users as well as a better protection for minors. The proposed system only takes into account the access to restricted content and requires an already classified collection of content.

3. FAST FILTERING OF ADULT CONTENT

Fig. 1 depicts the principle of our fast filter. The goal of the filter is to construct a search engine index free of adult content. It is impossible to guarantee that there won't be any false negatives, meaning that there will still be some Web pages containing adult content in the final index. However, if the filter remove, for instance, 98% of adult content, we claim that a search engine using this index will almost certainly never show adult content to its users. Indeed, with only very few adult websites in the index, the pagerank of those websites will be low, meaning that to be in top position, a website will need a very relevant content. As we see in section 4.4, the filter fails mainly on websites without content, so it is unlikely that the index contains adult website with relevant content.

We now describe the global architecture of the filter.

Blacklist. To have the most efficient filter, we use a *blacklist* mechanism.The first step is thus to check whether or not the Web page under analysis is in the blacklist.

Adult content disclaimer. For legal or moral reasons, most adult websites declare themselves as such using a disclaimer. Those sites are thus easy to filter.

TLD is ".xxx". Websites whose TLD is ".xxx" contain adult content.

Decision forest. The main part of the filter is a set of decision trees obtained using a statistical classifier. In this

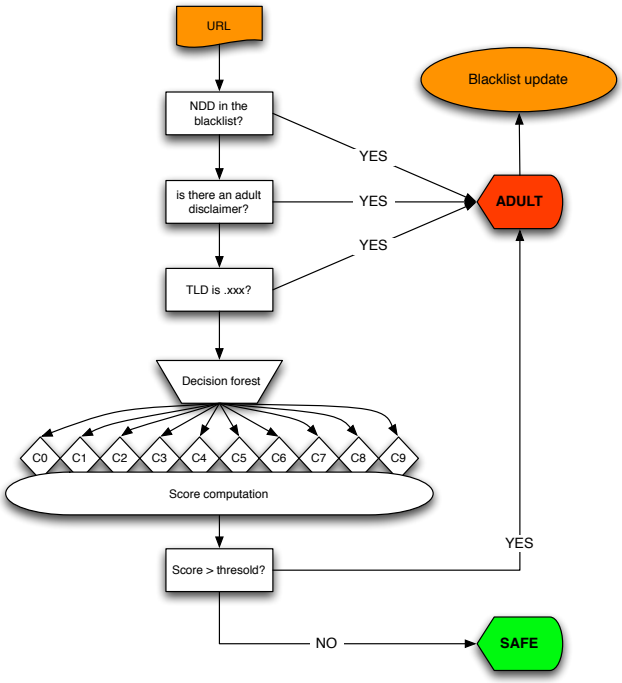

Figure 1: Principle of the filter

paper, we chose to use the C5.0 algorithm, an improved version of the well-known C4.5 (see [7] and [8] for more information). The score of a candidate page (for indexing) is the percentage of decision trees concluding that the page contains adult content. If the score is higher than 50% then the page is considered as unsafe.

Blacklist update. The blacklist is automatically updated as follows: if 3 pages from the same website are considered unsafe, then the domain name of the website is inserted into the blacklist.

4. DECISION FOREST

In this section we first present the methodology we use to train a decision forest to classify Web pages as safe or unsafe (e.g. containing adult content). We then describe the attributes used by the classifiers. Finally, we present experimental evidence that our approach gives satisfying results.

4.1 Methodology

We used the C5.0 algorithm of Quinlan [8] in order to obtain a decision forest that allows for the classification of the content of websites. This means that we obtain several independent decision trees (10 in our case) that will be used concurrently in order to obtain more accurate classification results.

To construct several decision trees, we used the *boosting* option provided by the C5.0. Moreover, our goal is to have a filter with a low percentage of false negatives (adult websites considered as safe), while false positives are less harmful (safe websites considered as containing adult content). Therefore, we penalize the false negatives by making then 20 times more costly than false positives in the C5.0 iterations.

The dataset used to train the decision forest is composed

of 226 Web pages, 120 of them being from adult websites. These websites were chosen manually and are representative of adult websites (youporn-likes sites, discussion forum about pornography/sexuality, swinger listings, erotic fiction and sex stories). Note that creating such a dataset is a matter of trial and error, and this step cannot really be automated.

Once the dataset is created, we gave it as input to the C5.0 implementation of Ross Quinlan, together with features extracted from an analysis of several attributes over pages from the dataset.

4.2 Description of the attributes

We chose to use only attributes that can be analysed quickly. This means that they are internal to pages and belongs to either textual content or quantitative attributes (e.g. number of images). We now describe each attribute and the associated computations.

in-url. We check the URL for the presence of given terms. We decided to use a list of 27 terms, from generic word ("porn") to brands ("cam4", "tube8", etc.).

This attribute is used in two ways. We compute the number of terms from the list that are in the domain name, but also the number of terms that are in the URL. Moreover, we are looking for substring, meaning that, for instance, "sex" is considered as being in sexhungrymoms.com.

The following attributes are also defined by a list of terms. Each being subject to 3 different computations. We first compute the number of occurrences of terms from the file X in the page (denoted nb_X in the following). Then we compute the proportion of words from the file in the page (denoted ratio_X). Last, we make the reciprocal computation, that is the proportion of words from the page that are also in the file X (denoted prop_X).

brand-names. A set of 34 brands related to adult content (content producers, major websites, etc.).
categories-en. 222 english terms that are usual categories of pornographic websites.
categories-fr. 593 terms used by french pornographic websites as categories (terms can be in french or english).
categories-gen. 79 french categories.
en-words. 100 english terms for ultra-sensitive topics (child pornography, rape, etc.).
french-words. A list of 163 french words representative of adult content (the goal of this list is to filter erotic litterature).
pornstars. Names of 8825 adult entertainment industry actors (male and female).
queries. 716 typical adult queries (e.g. "porn gratis", "porn gallery", etc.).
small-set. 11 terms that are common on adult websites (e.g. "sex", "xxx", "porn", etc.).
tags-en. 2000 most frequent tags (in english) for pornographic videos.
tags-fr. Similar to the previous list, 69 tags in french.

The last attribute is quantitative, and concerns images.

 # images. Number of images in the page.

tree id	size	error	tree id	size	error
0	7	13.7%	5	6	7.1%
1	3	23.0%	6	8	12.4%
2	4	9.7%	7	10	3.5%
3	6	6.2%	8	11	3.5%
4	10	8.0%	9	10	3.5%
global error			**0%**		

Table 1: Decision trees sizes and errors

One could remark that we are not using the number of videos or ads as an attribute. Our experience is that it does not give better results to use these attributes, and they are more difficult to compute than all the attributes above.

4.3 Obtained Decision trees

We obtained 10 decision trees, whose errors on the training dataset are depicted in the Tab. 1. We present in Fig. 2 a graphical view of the decision tree #2 (the one with a 9.7% error). This tree has a small complexity since only

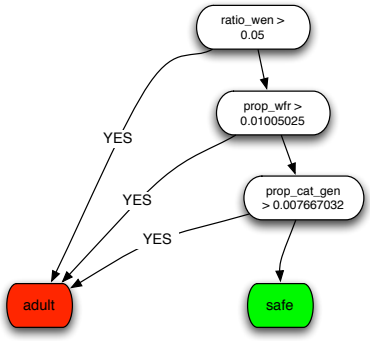

Figure 2: Decision tree #2

3 attributes are used. When a Web page is analyzed, it first computes the proportion of terms from **en-words** that are in the page (noted ratio_wen here). If ratio_wen > 5% then the page is considered as unsafe, otherwise the proportion of terms in the page that are also in **french-words** is computed, and compared to a threshold. The page is filtered depending on the comparison. Last, the attribute **categories-gen** is used.

All decision trees behave similarly, each of them being important to the final decision. Each decision tree has its own (sometimes high) error percentage. It is the aggregation of the decision of all trees through a majority vote that allows to a small global error.

We use a majority vote to aggregate decision from all trees. However, it is possible to tune the filtering power of our method by lowering the threshold (for instance, we can consider that a Web page is unsafe if at least 3 decision trees are in agreement on this outcome). Lowering the threshold will increase the number of false positive, which is less harmful when dealing with unsafe (adult) content.

In section 4.2, we saw that we deal with 36 different attributes. Amongst them, 23 are used by the decision forest. But not all 23 are used for analyzing each web page. 6 attributes were always computed when classifying Web pages from the training dataset: ratio_cat_gen, prop_cat_gen,

frequency	attribute	frequency	attribute
73%	IN-URL	50%	prop_brand
73%	ratio_brand	49%	ratio_tags_fr
70%	ratio_queries	46%	IN-NDD
68%	ratio_wfr	28%	ratio_tags_en
65%	nbr_img	28%	nb_star
63%	ratio_star	24%	nbr_brand
58%	prop_cat_en	24%	nb_wen
53%	nb_wfr	19%	nb_tags_fr
51%	prop_star		

Table 2: Attributes usage frequency

(a)	(b)	<- classified as
821	18	(a): class adult
14	300	(b): class safe

Table 3: Classification results

prop_wfr, prop_small, prop_queries. All these attributes depend on the presence of very specific terms in the textual content of the page. Both ratio and proportion are important, which is natural since it is unlikely that an adult website contains only one category of adult content, or only a few terms relevant to adult topics. Other attributes are used with frequencies presented in the following table.

We learn from Tab. 2 that an attribute such as the domain name is more useful at the blacklist level than at the decision forest level. Indeed, major players from the adult entertainment industry have domain name that are brands and as such that does not contain typical adult terms. However, attributes about the URL are important, mainly because many adult websites are using tags and categories in the URL in order to improve their search engines optimization (SEO). This is for instance the case of *youporn.com*.

4.4 Experiments

We saw in Sec. 4.3 that we obtain very promising results on the training dataset. However, it does not means that the filter will be efficient on new data.

We tested our filter on a new sample of 1153 Web pages. 839 pages contain adult content, and 314 are safe pages. The results are summarized in Tab. 3. These are satisfying results. The miss rate is (all the following numbers are rounded values) 2.15%, the accuracy is 97.22%, the recall is 97.85% and the precision is 98.32%.

Note that we have tested here only the decision forest of the filter. We did not used the blacklist update mechanism, nor the TLD and disclaimer detection. If we add these additional mechanisms, the accuracy is higher.

It is interesting to understand why some adult websites are considered as safe by our filter (false negatives). We made 3 additional tests, with specific types of adult websites. What we learned from these tests is that some adult video streaming website contains images and videos without explicit textual content. This is for instance the case of the website *beeg*, a youporn-like website. Since our filter is looking only at textual content, it cannot detect such website. In an index where almost all adult websites have been removed, even this kind of website will not appear easily in the search results: they don't have textual content so they are hardly relevant for textual queries, and their pagerank is very low since most of the sites that linked to them have been removed from the index.

A more ambiguous example is the one of discussion forums. While some of these forums are very explicit (this is for instance the case of "swingers forums"), others are forums of classical websites with a "sexuality" section. On those very specific websites, our filter is performing poorly. This is however a minority of the pages containing adult content, and the content of these forums is mildly explicit (no images, no videos).

5. CONCLUSION

In this paper, we presented a method that builds a safe index for search engines. Our experiments show that the method is efficient. Using only textual content proves sufficient in almost all cases with an accuracy of 97.22%.

Aknowledgements. The authors would like to thank Qwant for funding parts of this research.

6. REFERENCES

[1] G. Amato, P. Bolettieri, G. Costa, F. La Torre, and F. Martinelli. Detection of images with adult content for parental control on mobile devices? In *Proc. of the 6th International Conference on Mobile Technology, Application & Systems*, page 35. ACM, 2009.

[2] Y. Chan, R. Harvey, and D. Smith. Building systems to block pornography. In *Challenge of Image Retrieval, BCS Electronic Workshops in Computing series*, pages 34–40, 1999.

[3] M. Hammami, Y. Chahir, and L. Chen. Webguard: A web filtering engine combining textual, structural, and visual content-based analysis. *Knowledge and Data Engineering, IEEE Transactions on*, 18(2):272–284, 2006.

[4] M. Hammami, D. Tsishkou, and L. Chen. Adult content web filtering and face detection using data-mining based kin-color model. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 403–406. IEEE, 2004.

[5] C. Jansohn, A. Ulges, and T. M. Breuel. Detecting pornographic video content by combining image features with motion information. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 601–604. ACM, 2009.

[6] N. Park and Y. Kim. Harmful adult multimedia contents filtering method in mobile rfid service environment. In *Computational Collective Intelligence. Technologies and Applications*, pages 193–202. Springer, 2010.

[7] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[8] R. Quinlan. Data mining tools see5 and c5. 0. 2004.

[9] H. A. Rowley, Y. Jing, and S. Baluja. Large scale image-based adult-content filtering. In *VISAPP (1)*, pages 290–296. Citeseer, 2006.

[10] C. Sabina, J. Wolak, and D. Finkelhor. The nature and dynamics of internet pornography exposure for youth. *CyberPsychology & Behavior*, 11(6):691–693, 2008.

Author Index

Clarke, Charles L. A. 7

Gowda N, Thamme 10

Jie Zhou, Andrew 10

Kamps, Jaap 7
Karlgren, Jussi 7

Largillier, Thomas 14
Luo, Jiyun 10

Manjunatha, Harshavardhan 10
Mattmann, Chris 10
McGibbney, Lewis John 10
Murdock, Vanessa 7

Peyronnet, Guillaume 14
Peyronnet, Sylvain 14

Yang, Grace 10