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.