Current Research on Search and Exploration of X-rated Information

Vanessa Murdock Charles L.A. Clarke[1] Jaap Kamps[2] Jussi Karlgren[3]

[1] University of Waterloo, Canada
[2] University of Amsterdam, The Netherlands
[3] Gavagai, Sweden

ABSTRACT

This paper provides an overview of the work presented at the Workshop on Search and Exploration of X-Rated Information (SEXI) at the Conference on Web Search and Data Mining (WSDM) 2013 in Rome, Italy. The workshop represented a first attempt to study adult content from the perspective of the research communities in Web Search and Data Mining. To this end, five short papers were presented covering different research questions in searching and evaluating adult content on the Web, with two invited talks from experts in adult content from the fields of evolutionary psychology and media studies. The day ended with a panel that included the two invited speakers, and an expert in human trafficking on the Web.

1. INTRODUCTION

The Workshop on Search and Exploration of X-Rated Information was developed based on the observation that while adult content is pervasive on the Web, it is largely ignored in the scientific literature in the Information Retrieval and Data Mining communities. This is a notable omission given the high volume of web traffic devoted to adult content. The adult content industry is a billion dollar industry, that is changing in fundamental ways due to the ease with which people can generate and consume adult content at little or no cost [6, 7, 15].

The neglect of adult content in the research literature on web search and data mining can be partly explained by the assumption that definitions and algorithms developed for non-adult content are sufficiently general that they can be applied to adult content without further study. We propose that this is incorrect, and that 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 usual 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 senses of the query terms. Relevance and personalization are potentially more inextricably linked in an adult context than in non-adult contexts. Furthermore, as proposed by Cunningham [2], it is possible users access adult information primarily through dedicated portals, rather than through web search. That is, most often when they search the Web for adult content, they are searching for a portal rather than a specific item. Search on the portal is frequently done by browsing categories, and less often with a keyword search. This suggests that *enjoyment* might be a fruitful primary principle for the design of access interfaces, instead of building systems solely for the purpose of fulfilling a topical need [2].

Equal in importance to serving adult content when it is requested, is filtering adult content when it is not appropriate to show it. Identifying adult content is often difficult because the terms used to describe the content are frequently euphemistic. Seemingly innocuous terms such as "snake" and "cougar" take on new meaning in an adult context. Even a term such as "swimsuit" whose sense is unambiguous, is satisfied by very different results in an adult context than in a non-adult context.

The goal of SEXI is to provide a first attempt to understand the limitations of the current research in web search and data mining as it relates to adult search intents, and to examine issues such as relevance, diversity, personalization, and query intent. We seek a greater understanding of the particular issues surrounding the access of adult information, specifically user-generated adult content. The agenda of the workshop was designed to encourage discussion among the participants, as this is a new area. The workshop enlisted the perspectives of two communities that have studied Internet pornography extensively over the last 20 years. Finally we included the perspective of industry in a panel discussion at the end of the day.

One result of the workshop was that academic researchers do not face push-back from advisors or university management when doing this type of research. None of the participants had impediments to doing the research, or submitting the papers, as a result of the sensitive nature of the topic. Industrial researchers had some difficulty, mostly because information about how a search engine identifies and filters adult content is proprietary. Industrial researchers working in this area are not at liberty to discuss their findings or methodologies.

Another result of the workshop discussions was that the ethics involved in doing research in this area, such as assembling test collections and relevance judgements, are a barrier to be overcome. From a social perspective, many researchers are not inclined to do this type of research unless there is a

SEXI'13 February 5, 2013, Rome, Italy.

greater social good. From a practical perspective, in terms of assembling test collections, much of the data contains personally identifying information and so cannot be used for research purposes. The issue of having graduate students do user studies and collect relevance judgements appears to be problematic in Computer Science, but this has never been a problem in the social sciences, so this may be a non-issue, if computer science departments follow the same human subjects guidelines that other sciences use when asking people to do research on sensitive topics.

The ClueWeb data set[1] was proposed as a test bed for detecting adult content in a large Web corpus, although it represents a challenge as no ground truth has been established to determine which pages contain adult content.

Finally, a more concrete outcome of the workshop is a test collection assembled by [13]. The collection is a crawl of YouPorn categories. The collection consists of a crawl of 165,000 YouPorn video pages, from which they extracted textual metadata. The data includes the unique video title, the average rating and the ratings count, any categories and tags assigned to the video, and all comments in including comment text, user nickname, and the comment date. It is available at `http://blog.uni-mannheim.de/mschuhma/yp-corpus/`[2]. Contact the authors for more information about the collection.

Because this is a new area, the workshop generated many more questions than answers. Among the research questions identified were the following:

- Are users searching for adult content more or less sensitive to non-relevant results?
- Does changing the definition of relevance and diversity change the research questions?
- Do people use search engines to find porn? Or do they rely on trusted sites?
- People watch things they would prefer not to do. How does this affect personalization?
- Is the browsing need better served by the tail distribution than the information need?
- Is adult content search a scenario for slow search: browsing, coming back to the same thing?

The rest of the report summarizes the keynote addresses, the papers that were presented, and the panel discussion.

2. KEYNOTE: MARYANNE L. FISHER

Dr. Fisher is an Associate Professor in the Department of Psychology at Saint Mary's University in Halifax, Nova Scotia, where she is also a member of the interdisciplinary Women and Gender Studies Program. Her primary areas of research interest include women's competition for men and understanding human universals in popular culture. She is the lead editor of M. L. Fisher, J. R. Garcia, and R. S. Chang, editors. *Evolutions Empress: Darwinian Perspectives on the Nature of Women*. Oxford University Press, 2013, writing a book about women's same sex competition, and is currently planning a book about variable in women's sexuality.

[1]http://boston.lti.cs.cmu.edu/clueweb09 visited January 2013

[2]visited January 2013

2.1 What We Know About the Sexual Side of Human Nature

Maryanne Fisher structured her talk [4] around the question of what we sexually desire, and why, based on the premise that what we desire is determined by evolution and biology. She related our sexual desires to terms we use when searching, to describe what we want, and proposed that the way we describe what we want is a result of mate selection over the course of human history. She presented the evolutionary psychology view that our motivations, awareness and behavior are the result of finding solutions to problems humans faced over time. Those with effective solutions had an advantage in survival and reproduction.

She presented findings in research on mating strategies, notable among them that men have lower standards for short-term mating, and women have higher standards for physical attractiveness for long-term mating. Furthermore, across cultures, both men and women seek honest and kindness in their mates, and men place more emphasis on physical attractiveness, while women place more emphasis on characteristics related to accruing resources. In terms of preferences of men and women, Fisher notes that these are biologically driven, and they follow common stereotypes (such as men preferring younger women, and women preferring taller men).

Although our preferences are biologically driven, that doesn't tell the whole story. Imprinting early in life accounts for preferences, and some degree of variability in preferences. Furthermore, many of these associations are learned, and indicate a response to a visual cue (such as small feet). Men have been found to be more visual, and their physical arousal is tied to psychological arousal, whereas women's physical arousal is separate from her psychological arousal.

All of the background information about human sexuality was in preparation for a discussion of adult content search behaviors on the Internet. She presented several query log studies, to characterize what people search for, and show a correlation between search terms and what we know from evolutionary psychology. For example, of unique queries to Dogpile having to do with sexual content, the most common term was "youth". While many searches reflect the curiosity of the user, the majority of searches reflect either our evolutionary history, or our capacity for imprinting.

She finished with a study of titles of Harlequin Romance novels, in which the professions of the male protagonist were mentioned, to determine which stereotypes of men were most appealing to women. She ranked the top 20 professions. Computer scientists will be happy to note that although "Programmer," "Researcher," and "Scientist" did not make the list, "professor" was listed in the top 20, below "Fireman," "Pirate," and "Viking." She finished with a note that most psychology studies are done on atypical samples (people from the western hemisphere who are relatively educated and wealthy), whereas the Internet has the potential to reveal a much less biased view of human behavior due to its ubiquity.

3. KEYNOTE: SUSANNA PAASONEN

Susanna Paasonen is professor of media studies at University of Turku, Finland. Specialized in Internet research, cultural studies and studies of sexuality, she is most recently the author of S. Paasonen. *Carnal Resonance: Af-*

fect and Online Pornography. MIT Press, 2011, coeditor of M. Liljeström and S. Paasonen, editors. *Working with Affect in Feminist Readings Disturbing Differences.* Routledge, 2010, and S. Paasonen, K. Nikunen, and L. Saarenmaa, editors. *Pornification: Sex and Sexuality in Media Culture.* Berg Publishers, 2008.

3.1 Ubiquitous Yet Filtered: Porn and Search

Susanna Paasonen began her keynote [11] with Rule #34: If it exists, there is porn of it.[3] Her talk set out some basic facts: porn is easily accessible, everyone knows it exists, a majority of Internet users access porn and most do not pay for that access. These starting points, she argued, should inform our understanding of Internet porn. Porn is distributed by niche channels rather than mainstream channels, but these niche channels are easily accessible and known even by people who do not access them.

Today, the distribution of porn over the Internet is fueled by rapidly evolving technology which has enabled low-budget production of content by amateurs, often made available for free. The business of the adult content industry is shifting from the production of porn to its distribution. In view of this, the need to stand out among the many other distributors of adult content has created an attention economy. This has resulted in a rapid and constantly changing diversification of offerings, new subcultures appearing, changing, and rapidly moving along to new labels.

The creation of subcultures around a specific offering of porn has the effect of community building: finding others and identifying shared preferences among them. This is sometimes enabled by social features on the adult content portal, but not always. This sense of a community focused on a single subculture of porn may be a driver for search and access activity on the portal.

In spite of adult content being everywhere, and popular, and part of our natural state of being, it is treated as if it is something taboo and dangerous. As an example, she presented a cover of a Time Magazine issue devoted to the topic of cyberporn, from the early 1990's which shows the face of an innocent child with a shocked expression on his face. The article asks the questions "How pervasive is it?" and "Can we protect our kids?"

Her talk walked through statistics about the Internet pornography industry, relating to its prevalence, along with several examples of search interfaces in which the primary goal is to protect people from viewing adult content. She discussed how sexual content is characterized as "objectionable" and equated with "hate content" by search engines. As an extreme example, websites that show samples of web engine query streams in real time attempt to filter out adult terms, and then apologize and issue a warning that although the list is filtered, there may be adult terms among the queries. She points out that this is ridiculous, because the people viewing the query stream are from the same population as the people who issued the queries, and that in any case the sensibilities of the filtering (and the attendant warning) are culture-centric, based on the specifics of sexual morality of the U.S., where the major search engines are based.

In general the finding was that sexual content is frequently linked with *harm*, and technology surrounding sexual content is designed to conform with the sensibilities of a concerned community of guardians, but that those objections are fueled by concern for what other people might be accessing, viewing, or doing, not concern for the convenience, comfort, or well-being of the consumers themselves. Largely, the concerns may be related to the "community-building" aspect for specific sub-cultures noted above and the fact that easy access is normalizing the act of viewing porn. That is, one aspect of understanding attempts at limiting or hindering Internet porn consumption is that it is driven by people are unsettled by the habits of others object to the normalisation process rather than the content itself.

[3]http://knowyourmeme.com/memes/rule-34 visited January 2013

4. ACCEPTED PAPERS

As this is a new research area in the WSDM community, we solicited short papers. We accepted a total of 5 papers.

Chuklin and Lavrentyeva [1] discussed the classification of queries with or without adult query intent, proposing a three way classifier, labeling queries as black (explicit adult intent), gray (ambiguous intent, both adult and non-adult intent possible), and white (no adult intent).

Cunningham [2] proposed that users searching for porn do so within adult content portals, rather than directly from the search engine. She did a preliminary study of 30 adult content portals to examine what tasks are supported by the user interfaces, in order to determine user preferences in information access. She presents several observations about how the portals manage and are informed by user interactions and preferences. She concluded with a suggestion that the use case for accessing collections of enjoyable material such as music, videos and porn needs to include the value of the interaction with the collection itself.

Dean-Hall and Warren [3] looked at which attributes in a user profile are most important for personalization, and they examine the trade-off between personalization and privacy, which is particularly important in adult content sites. They propose an ontology for user preferences which is stored on the user's client, in contrast to the typical set-up where the content is stored at the server.

Schuhmacher et al. [13] created a collection of textual meta-data from the YouPorn[4] website, described in the Introduction. They investigated whether YouPorn user nicknames could be used as a feature for predicting the type of content the user was interested in.

Steiner et al. [14] investigated video artifacts that are the result of conversion of a VHS tape to digital, and proposed that the presence of these artifacts can be used to identify a particular genre of porn.

5. PANEL DISCUSSION

The panel discussion featured the two keynote speakers as well as Rane Johnson-Stempson from Microsoft Research. Rane serves on several White House committees on human trafficking, and offered an expertise in identifying harmful content on the Web. She is primarily interested in questions of using technology for a social good, and making sure it is not abused. She pointed out that there was very little technical literature relating to adult content, with the exception of some work related to fighting human trafficking on the Web.

The panel started with a discussion of the ethical considerations in doing research in this area. Rane started off by dis-

[4]http://www.youporn.com visited January 2013

cussing an agenda promoted by the Obama administration to provide support for research against human trafficking. Because the law is strict with regard to child pornography, there is very limited academic approval to access data that can be used for developing methods to fight human trafficking.

Maryanne pointed out that there are no ethical issues in psychology in studying adult content. Academics take care to discuss everything in scientific ways, in a way that is free of value judgements. Students helping with data collection must be older than eighteen years, and must give consent. Many collaborators and students are female. Dealing with this data is very well possible when working within the policies of the ethics boards. Social sciences and psychology study sensitive topics regularly.

Suzanne added that teaching classes on pornography has some challenges. Showing porn in class is possible if it serves a function, and has educational value. The largest problem is getting access to representatives from the adult industry. In general, the adult entertainment community are very reluctant to talk to outsiders such as journalists and academics scholars. It is very difficult to go beyond the self-promotion of industry, in order to do a deeper study..

Rane pointed out that there is an increased awareness of human rights and doing research that benefits the social good movement in computer science. There are many safeguards that could be put into place. In terms of enlisting students and faculty to work on this topic, the White House sent out a call for proposals to work on methods for combatting human trafficking. They were expecting a handful of responses and received 30. There is plenty of interest on the part of faculty and students to support research in this area.

Maryanne added that looking at pornography is one of the least harmful topics to watch. There are far more disturbing topics for people to be exposed to. For example, people find infidelity is a more disturbing topic.

The panel discussed cultural differences in terms of how adult content is perceived and valued. For example, Susanna is free to present any information when speaking in Finland, but when presenting in India where kissing in public is taboo, she tailored her talk and her examples. In terms of research, there are many valuable contributions to be made in this field, and not all of them require looking at objectionable material. In terms of the benefits or detriments to the women's movement, there is a need to position the research carefully to make it contribute to a positive outcome.

Acknowledgments.

We would like to thank ACM and WSDM for hosting this workshop, the WSDM workshop chair Sebastiano Vigna, and especially the team of local organizers.

We thank the *Trattoria Morgana* at the Via Mecenate in Rome for hosting a memorable workshop diner with the workshop attendees and other WSDM participants interested in the topic, with more informal discussion that continued far into the Roman night.

Final thanks are due to the paper authors, the invited speakers Maryanne Fisher, Susanna Paasonen and Rane Johnson-Stempson as well as the participants for their passionate presentations and discussions in the workshop.

Details about the workshop including the presentations and slides are online at `http://sexi2013.org/`.

References

[1] A. Chuklin and A. Lavrentyeva. Adult query classification for web search and recommendation. In Murdock et al. [9], pages 15–16.

[2] S. J. Cunningham. Learning from the internet porn industry: What porn sites may tell us about pornography location behaviors. In Murdock et al. [9], pages 17–18.

[3] A. Dean-Hall and R. Warren. Sex, privacy and ontologies. In Murdock et al. [9], pages 19–26.

[4] M. L. Fisher. What we know about the sexual side of human nature. In Murdock et al. [9], pages 11–12.

[5] M. L. Fisher, J. R. Garcia, and R. S. Chang, editors. *Evolutions Empress: Darwinian Perspectives on the Nature of Women*. Oxford University Press, 2013.

[6] S. Friess. Porn industry sweats recession, piracy. *AOL News*, January 9, 2011. URL `http://www.aolnews.com/2011/01/09/porn-industry-facing-hard-times-in-struggling-economy/`.

[7] B. Fritz. Tough times in the porn industry. *The Los Angeles Times*, August 10, 2009. URL `http://www.latimes.com/news/local/la-fi-ct-porn10-2009aug10,0,3867866,full.story`.

[8] M. Liljeström and S. Paasonen, editors. *Working with Affect in Feminist Readings Disturbing Differences*. Routledge, 2010.

[9] 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.

[10] S. Paasonen. *Carnal Resonance: Affect and Online Pornography*. MIT Press, 2011.

[11] S. Paasonen. Ubiquitous yet filtered: Porn and the search. In Murdock et al. [9], pages 13–14.

[12] S. Paasonen, K. Nikunen, and L. Saarenmaa, editors. *Pornification: Sex and Sexuality in Media Culture*. Berg Publishers, 2008.

[13] M. Schuhmacher, C. Zirn, and J. Völker. Exploring youporn categories, tags, and nicknames for pleasant recommendations. In Murdock et al. [9], pages 27–28.

[14] T. Steiner, S. V. Hooland, R. Verborgh, J. Tennis, and R. V. de Walle. Identifying vhs recording artifacts in the age of online video platforms. In Murdock et al. [9], pages 29–30.

[15] L. Theroux. Louis theroux on porn: The decline of an industry. *BBC News Magazine*, June 8, 2012. URL `http://www.bbc.co.uk/news/magazine-18352421`.