Search and Exploration of X-rated Information

WSDM'13 workshop proceedings : February 5, 2013, Rome, Italy

Murdock, V.; Clarke, C.L.A.; Kamps, J.; Karlgren, J.

Search and Exploration of X-rated Information

WSDM'13 Workshop Proceedings

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

February 5, 2013
Rome, Italy
http://sexi2013.org/

Cover image: Michelangelo. The Fall of Man and the Expulsion from the Garden of Eden. 1508-1512. Fresco. Sistine Chapel, Vatican.

Published by: IR Publications, Amsterdam. ISBN 978-90-814485-9-8.

Preface

These proceedings contain the research and position papers of the WSDM'13 Workshop on *Search and Exploration of X-rated Information*, held in Rome, Italy, on February 5, 2013.

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 workshop started from the premise that the research questions surrounding adult content access behaviors are unique, and that interesting and valuable research in this area can be done ethically.

The workshop consisted of three main parts:

- First, keynotes by Maryanne L. Fisher and Susanna Paasonen that help frame the problems, and outline potential solutions.
- Second, paper sessions with five papers selected by the program committee from seven submissions (a 71% acceptance rate). 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 2013

Vanessa Murdock
Charlie Clarke
Jaap Kamps
Jussi Karlgren

Table of Contents

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`.

What We Know about the Sexual Side of Human Nature

Invited Keynote

Maryanne L. Fisher
Department of Psychology
Saint Mary's University
Halifax, Nova Scotia
mlfisher@smu.ca

One of the major and recent shifts in psychological research has been the application of evolutionary principles to human behaviour. Scholars have aligned evolutionary biology with cognitive science, and as a result, have learned far more about human nature than ever before. This discipline, evolutionary psychology, has dramatically changed how psychologists view human sexuality, and with it, how we understand human behaviour more generally. For example, we know much more about sex-specific mating strategies and mate preferences, which may explain the sexual content seen on the Internet.

Internationally, when asked about the characteristics they prefer in mates, men place attractiveness higher on the list than women [e.g., 1]. Women, instead, tend to place more emphasis on resources, or characteristics, such as ambition and intelligence, that relate to the acquisition of resources. The explanation for this difference is that women biologically invest more in children; they have fewer opportunities to conceive and are the sex who carries the developing child. Due to the costs of caring for young children and themselves, it was historically (and arguably is, currently) advantageous to have a mate with resources, which includes the abilities to provide food, shelter, protection and paternal care. In contrast, men are able to have far more children and are faced with the problem of finding fertile, available, mates. Youth is a proxy cue of fertility, in that young women are more likely to be able to conceive than older women, and particular features are reliable signals of youth, such as skin firmness, thinness, and hair color [see 7, 8].

Consequently, evolutionary psychologists predict that, aside from characteristics such as honesty and kindness, the sexes differ in mate preferences. Men will generally seek young, attractive women, and women will typically seek men with resources. Moreover, men will often use a strategy of quantity, and report that they hope to have more sexual partners over the course of their life than women, who often use a strategy of quality and tend report a desire for long-term, committed relationships. As Piazza and Bering [3, p. 1261] review in their study outlining the various ways that evolutionary psychology maps onto Internet content, the specific characteristics that the sexes prefer were shaped by natural selection, such that they allowed individuals to solve recurrent problems of genetic fitness and reproductive success.

In terms of acquiring insights into behaviour, we have many tools at our disposal; we can ask people via surveys or interviews, perform eye-tracking studies, or use the argument that when we see an artifact created by humans, we are also viewing something about human nature. That is, just as the consistent types of information presented in tabloids reveals our evolved human nature, so do the frequency of specific search terms used to find pornography. One very striking benefit of relying on data from sources that are not survey or interview based is that there is less chance for participants to bias the results due to issues such as social desirability or demand characteristics. I argue that for studying sexual behavior, examining specific search terms is an ecologically valid and useful way to gain insight into evolved human behaviour and motivations.

Many scholars have shown that people search for sexual content on the Internet. For example, Spoerri [6] found that of the 100 most visited pages in Wikipedia between September 2006 and January 2007, over 50% were related to entertainment and sexuality. Spink et al. [4] examined queries from the Excite search engine. The table [4, table 3, page 231] listing the 75 most frequently occurring terms of the 531,416 unique queries reveals that "sex" is the third most occurring term, after "and" and "of." Indeed, many of the search terms were sexually relevant.

Although there has been past attempts to use evolutionary psychology to explain the existence of sexual content on the Internet [e.g., 5], what has often been missing is a deep understanding of why the Web contains the information it does. Ogas and Gaddam [2] are a refreshing exception, as they use evolutionary psychology to explain the data they obtained from various search engines, as well as other sources of Internet content such as personal ads, websites and electronic romance novels. For example, they collected 400 million unique queries from Dogpile search engine from July 2009 to July 2010. They report that 55 million of these queries were ones seeking sexual content, and of these, the most common term was "youth" (13.5%).

Using the logic of evolutionary psychology, Ogas and Gaddam's finding clearly reflects men's evolved interest in young women. They further argue that only 20 different categories (or 'interests') accounts for 80% of all sexual content searches, which shows a surprising lack of diversity. From an evolutionary perspective, this homogeneity is reasonable, given that mating-relevant problems (and interests) individuals face tend to be rather similar. In other words, the ways in which we engage with the Internet and seek certain types of content reflects solutions to issues our ancestors faced during human evolution.

I argue that the ways in which the sexual sides of men's

SEXI'13: WSDM Workshop on Search and Exploration of X-rated Information, Febuary 5, 2013, Rome.

and women's human nature differ are predictable and grounded in our evolutionary history. For example, men have been challenged to gain access to mates, to locate fertile women, and to be concerned with issues surrounding paternity certainty (i.e., cuckoldry); all of which are problems faced historically and contemporarily. In contrast, women historically and today face the challenge of finding mates who will invest in them and any children, and are presented with issues surrounding emotional commitment. Thus, men and women have faced distinct mating-related issues, the ramifications of which are observed in the types of information that is found on the Internet.

REFERENCES

[1] D. M. Buss. Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, 12:1–49, 1989.

[2] O. Ogas and S. Gaddam. *A billion wicked thoughts*. New York: Dutton, 2011.

[3] J. Piazza and J. M. Bering. Evolutionary cyberpsychology: Applying an evolutionary framework to internet behavior. *Computers in Human Behavior*, 25: 1258–1269, 2009.

[4] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52:226–234, 2001.

[5] A. Spink, A. Koricich, B. J. Jansen, and C. Cole. Sexual information seeking on web search engines. *CyberPsychology and Behavior*, 7:65–72, 2004.

[6] A. Spoerri. What is popular on wikipedia and why? *First Monday*, 12(4), 2007. `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1765/1645`.

[7] M. Voracek and M. Fisher. Shapely centerfolds? temporal change in body measures. *British Medical Journal*, 325:1447–1448, 2002.

[8] M. Voracek and M. Fisher. Success is all in the measures: Androgenousness, curvaceousness, and starring frequencies in adult media actresses. *Archives of Sexual Behavior*, 35:297–304, 2006.

Ubiquitous Yet Filtered: Porn and the Search

Invited Keynote

Susanna Paasonen
Department of Media Studies
University of Turku
Turku, Finland
suspaa@utu.fi

It is commonly known that search terms for pornographic content top the listings of most popular search terms. It is equally common knowledge that such terms are systematically filtered out when keyword listings are published. In the rhetoric applied by filtering software and search engine filters such as Google's SafeSearch, porn is marked as a risk. In information society discourses and scholarly debates, porn has been similarly associated with danger, addiction and uncontrollability. While the question may seem naive, I am interested in why this is so.

Although porn is recurrently marked off from everyday media use, such erasure is impossible to accomplish as few users are unfamiliar with porn: porn usage is mundane among male and female users of different ages. The needs of the porn industry have driven the development of web technologies and business practices from hosting services to safe credit-card processing, banner advertisements, pop-ups, web promotions and streaming video technology [e.g., 1, 2, 6]. Rather than being an outsider element to be filtered off, porn can, then be seen as part and parcel of the fabric of the web.

The association of porn with risk and filth draws on a history of censorship and regulation as long as the history of modern pornography itself. It can even be argued that without the censorship there would be no pornography per se: for this is a genre exploring displays and acts deemed obscene, indecent or nasty. As Kendrick [3, p. 212] points out, the pornographic has, since the 19th century, been mapped out as something necessitating regulation, secrecy, and limited access. In the process of regulation and censorship, it has been embedded in a discourse of filth, smut, depravity, garbage and sewage. And, as Kuhn [4, p. 23] notes, "in order to maintain its attraction, porn demands strictures, controls, censorship." In other words, the attraction of porn owes considerably to its status as a forbidden fruit. Filtering carried out by agents such as Google feeds into this demarcation and, paradoxically, helps to preserve some of the titillation of porn at the very moment when its ubiquity seems to efface the aura of the forbidden.

Articulations of porn as garbage and filth conflate the genre with moral judgments, which, in the U.S. context, draw on the Judeo-Christian tradition equating extramarital and non-reproductive sex with sin [cf., 5]. Following Warner [8, p. 4], this tends to feed into moralizing where certain moral notions are projected as the general norm. Hierarchies are built between the self and others, the normal and the abnormal, the acceptable and the deplorable by naming certain acts and preferences as shameful. This has little to do with ethical concerns – such as those concerning the ethics of commercial sex work or porn distribution – that need to be dialogical, reflexive and sensitive to context.

This is an important point since pornography itself is a highly evasive denominator, there being no unanimity over its definitions. With web distribution, indexing, tagging and metadata facilitating content searches, different porn niches, fringes, subcategories and subgenres have become increasingly articulate. And as porn production has equally widened to encompass operations involving a range of distinct aesthetical, ethical and economical underpinnings, the field of the pornographic has become extensively fragmented. See Paasonen [7]. At the same time, filters have been critiqued for conflating sex education, erotic fiction and information resources for sexual minorities with pornography, equally filtering all. This further adds to the obscurity of pornography as a point of reference.

As browsers and search engines developed in the U.S. maintain a hegemonic market status, there is a risk of particularly North American notions of obscenity becoming generalized as the global norm. It remains crucial to question received notions of obscenity, indecency and pornography, are we to resist the conflation of sexuality with notions of risk, harm and filth, to account for the diversity within the pornographic and to better understand the genre.

SEXI'13: WSDM Workshop on Search and Exploration of X-rated Information, Febuary 5, 2013, Rome.

REFERENCES

[1] J.-A. Filippo. Pornography on the web. In D. Gauntlett, editor, *Web.Studies: Rewiring Media Studies for the Digital Age*, pages 122–129. London: Arnold, 2000.

[2] J. A. Johnson. To catch a curious clicker: a social network analysis of the online pornography industry. In K. Boyle, editor, *Everyday Pornography*, pages 147–163. London: Routledge, 2010.

[3] W. Kendrick. *The Secret Museum: Pornography in Modern Culture.* Berkeley: University of California Press, second edition edition, 1996.

[4] A. Kuhn. *The Power of the Image: Essays on Representation and Sexuality.* London: Routledge, second edition edition, 1985.

[5] T. Laqueur. *Solitary Sex: A Cultural History of Masturbation.* New York: Zone Books, 2003.

[6] J. Lillie and J. McCreadie. Cyberporn, sexuality, and the net apparatus. *Convergence*, 10:43–65, 2004.

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

[8] M. Warner. *The Trouble with Normal: Sex, Politics, and the Ethics of Queer Life*. Cambridge: Harvard University Press, 2000.

Adult Query Classification for Web Search and Recommendation.

Aleksandr Chuklin, Alisa Lavrentyeva
Yandex, Russia
{chuklin, amneziya}@yandex-team.ru

ABSTRACT

The Internet is a rich source of various adult content. Though, the web search engine must provide the user with X-rated content only in case the user's query has an adult intent. In this paper we suggest a three-class distribution of search queries with respect to their adultness and present an adult content classificator based on language models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Classification, Filtering

Keywords

query classification, web search, web search suggestion, adult queries

1. INTRODUCTION AND RELATED WORK

The Internet today is a powerful source of all kinds of information, that can satisfy almost any users' interest. As a result, the dark side of WWW is a large amount of adult content, that can be freely accessed. However, there are a number of cases, when this type of content is not desired to be retrieved or should be treated in a particular way. So there stands a problem of building a reliable adult content classifier. In this work we consider a problem of classification of user search queries. Every query given to the search engine should be classified for having or not having the adult intent, i.e. if the user is expecting to see adult content in the search result. Unlike the common two-class adult content classification (see e.g.[4]) we mark out three classes of queries: *black* (explicit adult intent), *gray* (ambiguous queries, adult intent possible) and *white* queries. The related research referring to adult content filtering mainly focuses on multimedia content ([6, 1]), complete web document classification ([4, 5]) and document text classification. The most simple approach to text filtering is building *black*- and *white*-lists of queries or queries' keywords is useful but cannot be effectively extended or handle with misspells. There are a number of statistical methods, employed to preform text classification.

In [8] a dynamic Markov compression for texts extracted form html-documents is considered. A complex classifier based on bag-of-words algorithms, feature selections and incremental feedback is presented in [3]. A statistical language model approach with a fixed document structure is presented in [7]. In this paper we will describe a language model based algorithm that preforms query text classification.

2. METHOD

We started with collecting human judgements for the queries. Overall we have 1889 queries judged: 1118 marked as *white*, 132 *gray* and 639 *black*. Queries were sampled from the search traffic and then enriched by some well-known *black* words. In order to automatically classify the queries into classes developed a set of features and pick a machine-learning algorithm. As a learning algorithm we used a Gradient Boosting Decision Trees (GBDT) [2] which can be considered as a state-of-the art machine learning algorithm.

2.1 Features.

We used features of two sorts, namely language features and SERP[1] features. **SERP features** are built using the documents returned by a web search engine. We assume that for each document returned we have pre-computed the value of "adultness" of this document. We do not aim to estimate the document "adultness" level in the current work. Instead we are going to infer the query class (*white*, *gray*, *black*) using these "adultness" levels as well as other signals. We assume that "adultness" of the document is a real-valued variable normalized to $[0, 1]$ (0 means *clean* document, 1 means *highly adult* document). In Yandex search engine adult query classification is run after receiving the document list but before presenting it to the user. We used the following SERP features: number of "adult" documents ($adultness > 0$) in TOP-N, "adultness" of the first document, number of documents with "adultness" within intervals $[0.25, 0.5)$, $[0.5, 0.75)$, $[0.75, 1.0)$, $[1.0, +\infty)$ in TOP-N (4 features), average value and variance of "adultness" in TOP-N (2 features). In the current work we used $N = 30$ as if one asked a human to classify query based on a SERP he would unlikely examine lower documents.

Language features are calculated using only the query submitted by the user. Along with simple features using manually collected *black*- and *white*- lists we present a novel set of features based on language models. A language model is a function that for each sequence of words outputs the probability of this sequence to be a valid phrase in some "lan-

[1]SERP is the short for Search Engine Result Page

Table 1: Results of classification.

Query Class	Precision	Recall	F1
black	95.2 %	97.3 %	96.2 %
gray	64.0 %	22.4 %	33.2 %
white	99.4 %	99.9 %	99.7 %

Table 2: Classification without language models.

Query Class	Precision	Recall	F1
black	95.5 %	97.3 %	96.4 %
gray	65.7 %	16.1 %	25.8 %
white	99.4 %	99.9 %	99.7 %

guage". For this task we used a bigram model with simple back-off smoothing (i.e. falling back to unigram model when the bigram is not present in a dataset) and some heuristic to choose a frequency value for out-of-vocabulary words. In the current work we randomly sampled a subset of queries submitted to Yandex during the year 2011. In total we have 116'511'310 queries. We then applied a high-precision classifier to build a corpora of *black* and *white* queries. The classifier was used by a previous revision of the adult query detector that did not make use of language models (some SERP-based features were used). We tuned this algorithm to reach 90% precision (by reducing recall). In total we had 3 corpora (all queries, *black* queries, *white* queries) and were able to build a language model for each corpus separately. One might argue that high-precision *adult* query classifier could be constructed by using only *black*-list of words. However, we argue that a language model is able to learn some new information. For example, if we have a query "AAA BBB XXX" where "XXX" word is present in a *black*-list and "AAA", "BBB" do not, our language model will probably assign a high enough probability to the query "AAA BBB" which can be further classified as *gray*. Finally, for each query we have 3 probabilities $P_{general}$, P_{black}, P_{white} which are transformed to the language features by comparing them to each other: $P_{general} - P_{black}$, $P_{general} - P_{white}$, $P_{black} - P_{white}$, $\log(1 + P_{general}) - \log(1 + P_{black})$, etc.

2.2 Results

We take all these features and use them as an input for GBDT machine learning algorithm. In order to evaluate its performance we adopted a stratified 10-fold cross-validation test. The results are presented in Table 1. We can see that the task of detecting *black* and *white* queries could be solved with high precision and recall, while classifying *gray* queries is a relatively difficult task and we should concentrate our future effort on solving this problem. For the comparison we also show the results of a classifier that does not use language features, i.e. uses only SERP features (Table 2). This is particularly useful when we are going to apply our algorithm to a new market without having large enough corpus of queries. We can see that it shows almost identical performance for all classes but *gray*, which means that language models are particularly useful for detecting *gray* queries.

3. DISCUSSION

APPLICATION TO WEB SEARCH. For the web search it is important to avoid showing undesired adult content to the users who do not want to see it. At the same time, we do not want to degrade relevance and loose search traffic by filtering adult documents for all users. As we have 3 classes of queries we can use the following algorithm:

- If the query was classified as *black* do not perform any filtering
- If the query was classified as *white* remove all documents with "adultness" higher than α
- If the query was classified as *gray* remove all documents with "adultness" higher than β

Instead of removing documents one may want to use reordering, but we leave it for further discussion. One of the interesting questions for future work would be the algorithm to choose the thresholds α and β to use in a production setup. We argue, that α should be higher, because for *white* queries "adult" documents are generally unlikely to make into the TOP. At the same time we want to carefully clean up SERPs for *gray* queries as they may contain many controversial documents. Another possible improvement of the algorithm could be usage of continuous query adultness level instead of 3 classes. By using continuous scale we could assign varying filtering threshold (i.e. α, β).

APPLICATION TO QUERY SUGGESTION. Web Suggest is a search engine's tool that provides a user with query suggestions on a typed prefix. Assuming we have no information about the users' age, general interests and the typed prefix is neutral, than we should not provide suggestions that have explicit adult intent — *black* queries. The *gray* queries are still possible. On the other hand, if the input is already a *black* query (for example, contains explicit adult words), we can provide all sort of suggestions.

As the future work we want to study the phenomenon of *gray* queries in more detail and increase the quality of classification. Another important direction is the problem of adult filtering evaluation. What are the optimal values for the filtering thresholds? Shall we perform some document reordering along with / instead of filtering? When rolling out new filtering versions to production we could see that users who got used to access their favorite adult resources by submitting *gray* queries will generally learn that from now on they need to add some words to their queries to make them explicitly "adult". This phenomenon of user adaptation might also be interesting to investigate.

4. REFERENCES

[1] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *ICPR*. IEEE, 2008.

[2] J. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Statist*, 29(5), 2001.

[3] J. M. G. Hidalgo, E. P. Sanz, F. C. Garcia, and M. B. R. and. Web content filtering. 2009.

[4] W. H. Ho and P. A. Watters. Statistical and structural approaches to filtering internet pornography. In *IEEE SMC*, 2004.

[5] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank. Recognition of pornographic web pages by classifying texts and images. *TPAMI*.

[6] F. Jiao, W. Gao, L. Duan, and G. Cui. Detecting adult image using multiple features. In *ICII*, 2001.

[7] B. Medlock. A language model approach to spam filtering.

[8] I. Santos, P. Galán-García, A. Santamaría-Ibirika, B. Alonso-Isla, I. Alabau-Sarasola, and P. Bringas. Adult content filtering through compression-based text classification. In *CISIS*, 2012.

Learning from the Internet Porn Industry: What Porn Sites May Tell Us about Pornography Location Behaviors

Sally Jo Cunningham
University of Waikato
Computer Science Department
Hamilton, New Zealand
+64 7 838 4402
sallyjo@waikato.ac.nz

ABSTRACT
This paper suggests directions for identifying user preferences for organization and access of pornography, by analyzing existing popular collections. These insights can then inform design of improved access mechanism.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Collection, System Issues, User Issues

General Terms
Design, Human Factors.

Keywords
Collection organization, browsing, collection access points, user feedback into collection design.

1. INTRODUCTION
The largest and most popular digital image and video collections on the Internet are, unarguably, pornography websites. As pornography websites proliferated, site design has become more sophisticated; sites attempt to differentiate themselves from competitors by collection organization as well as by collection contents, and by encouraging user feedback and interaction in the form of comments and ratings. This paper suggests that an initial step in understanding user preferences in accessing pornography is to recognize that one significant behavior is to search for a site to frequent (in contrast to conducting general web searches for individual pornographic items). We can then leverage on the expertise of pornography site designers for an initial investigation of preferences in user behaviors in locating pornography, by finding commonalities in the site design of popular pornography sites. This position paper presents an overview of initial results from an examination of 30 interfaces and organization of the highest-traffic and most frequently bookmarked pornography sites.

2. DATA GATHERING
This investigation focuses on high traffic and heavily bookmarked pornography websites, under the assumption that it is primarily (though not exclusively) user preferences for the access mechanisms of these sites that draw the high traffic, rather than novel content at these sites. The "Adult" section of the Alexa website was consulted to draw up an initial list of the 30 highest-traffic sites (http://www.alexa.com/topsites/category/Top/Adult; Alexa is a web analytics service that provides website traffic rankings in a number of categories).

To capture heavily bookmarked sites, the social bookmarking service Delicious.com was consulted for sites tagged with "pornography" and "porn" and the results sorted by frequency of tagging. 11 of the Alexa sites were also heavily tagged in Delicious.com. These 30 sites were then visited and snapshots taken of the home page (to identify search and browsing functionalities provided). Snapshots were also taken of a representative proportion of the browsing hierarchies.

3. SUPPORTING BROWSING
Browsing is the primary access mechanism for pornographic video and image sites. The home page of most sites is image heavy—including tens or even hundreds of images—encouraging the user to begin the document search by browsing rather than by searching. Browsing is further supported by extensive classification schemes—several sites surveyed included schemes with over 400 categories. To make browsing easier, the structures are overwhelmingly broad but shallow—so that the user does not become lost in long hierarchies. Further, most sites place the entire browsing structure on the site's top level; this allows the user to visually explore the entire browsing structure at once.

Why emphasize browsing over search? Categories and individual documents (primarily videos) are represented by thumbnail images or animations, which are themselves pornographic. Browsing is thus an additional mechanism for consuming pornography, and exposes the collection user to additional pornographic options that s/he may not have initially considered, as browsing progresses. The additional benefit of broad-but-shallow browsing hierarchies is that the extent of the collection (in terms of number of categories) is visually maximized—again, supporting the user in visually scanning as many parts of the collection as possible, bringing potentially unknown but interesting categories to the user's attention.

4. RECOGNIZING NICHE INTERESTS AND INTEREST NICHES
Pornography sites recognize that their users may be turned away by content outside their interest or fetish. The classic example is that of male heterosexual and homosexual content: most sites recognize that male viewers exclusively view one or the other,

Conference'10, Month 1–2, 2010, City, State, Country.

and men interested in heterosexual content are likely to leave the site if presented with male homosexual images. To complicate this further, heterosexual male viewers frequently seek out lesbian imagery, although men interested in homosexual content do not. Successful porn sites recognize these viewing preferences by clearly distinguishing these categories of content in the browsing structures.

The classifications themselves are not fixed or standard, beyond a few broad categories (eg, "lesbian"); each site appears to develop its own organization scheme. The schemes themselves vary widely, in structure, complexity (eg, number of categories), and terminology. The main organizing principle is to cluster videos / images by fetish, where 'fetish' is loosely defined as the defining characteristic of the sexual act(s) depicted. While all of the sites examined in this study included a set of broad categories, some also chose to provide narrow categories as well, apparently to appeal to users with highly specific pornography preferences. A next step in this research is to analyze the commonalities and differences in terminology and structure for these 30 websites, and to investigate whether there is any correlation between the site organizational scheme and site popularity.

5. USERS INFORM SITE DESIGN

Following on from point 5, the success of a site is clearly contingent on the site manager understanding the video/image information needs of the site's potential users, ensuring that the site includes desirable categories of content and that the content is accurately classified into the correct category. While it is difficult to develop a general model of information behavior in accessing pornography, site managers can (and do) create models of the information behavior of own site's users. This site management activity is so common that a pornography-specific web analytics service exists: Sextracker (www.sextracker.com) tracks site traffic at subscribing soft and hardcore pornography sites, and produces both high level statistics useful to users (eg, the 'Top Ten Sites' list), and extensive and fine-grained reports for individual site managers. These reports are updated at least daily, allowing site managers to constantly tweak their sites to downplay or eliminate unpopular content, and to provide higher profile access to more popular material. The availability of near real-time information on user activities is considered crucial for keeping a site high in the rankings.

Research in this area could benefit hugely from access to detailed site reports, to build profiles of search and browsing behavior as well as gaining insight into content consumption preferences. It is difficult, however, to contact site owners and gain permission to access the site reports (beyond the difficulties in gaining ethics consent from the researcher's home institution). For these reasons, the main insights into search behavior to date come from analysis of search engine logs (eg, Spink et al 2002, Spink et al 2004), and focus exclusively on search behavior rather than browsing, re-finding, or other information seeking behaviors.

6. SUPPORTING RETURNING USERS

Frequent visitors to a document collection want to be made aware of new material added to the collection, and particularly new material within their specific interests. Pornography websites draw attention to recent additions by setting up galleries clearly labeled to indicate when the content was added to the site (eg, "Today's Galleries", "Yesterday's Galleries", etc.)—encouraging users to visit the site frequently, and also supporting irregular site visitors by pointing them to the freshest content. Conventional 'serious' document collections (eg, academic digital libraries, historic or cultural document collections, etc), by contrast, frequently 'bury' new content by adding new documents to the collection without identifying their accession date—making it difficult for users to distinguish newly acquired documents.

Many of the sites provide little or no support facilities for re-finding videos / images located in earlier site visits (eg, personal site usage history, tagging of 'favorites'). It is unclear whether the weak support for re-finding is a consequence of lack of user demand (perhaps linked to a preference for novel, rather than familiar, videos/images), a reluctance on the part of users to create the necessary login accounts on the websites, a large churn rate among users seeking novel sites as well as novel videos/images, or some other factor. The recently emerged Pinterest-like sites for pornography (Pornterest and Snatchly) offer the capacity for re-finding across many sites, thereby allowing users to curate their own pornography collections. Examining the use of these sites over time will provide some insight into the importance of re-finding to pornography site users.

7. SUPPORTING SOCIAL INTERACTION

Most sites provide some support for interaction and feedback—at the very least, the ability to rate site documents. More complex social media – style interactions are currently supported through pornography – focused copies of general social media sites: for example, Fantasti.cc is described as "a sort of Facebook for porn", and Pornsifter.com and Socialporn.com are intended to be the Digg for pornography. If these sites survive and gain a reasonable user base, they will be a valuable source of insights into pornography information behavior.

8. CONCLUSIONS

The focus in this project is on pornography information seeking and behavior on pornography sites, rather than on behavior over general web search engines. At this point, more questions are raised than answered. Emerging pornography – specific versions of general social media sites may provide additional resources for understanding how people find and use pornography.

9. REFERENCES

[1] Spink, Amanda, et al. "From e-sex to e-commerce: Web search changes." *Computer* 35.3 (2002): 107-109.

[2] Spink, Amanda, H. Cenk Ozmutlu, and Daniel P. Lorence. "Web searching for sexual information: An exploratory study." *Information processing & management* 40.1 (2004): 113-123.

[3] Stevenson, S. The dirtiest comments in the world: why do people comment on porn videos? *Slate Magazine*, August 24 2012, http://www.slate.com/articles/technology/the_browser/2012/08/fantasticc_the_people_who_comment_on_porn_and_the_weird_things_they_say_.html.

[4] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

Sex, Privacy and Ontologies

Adriel Dean-Hall
DRC School of Computer Science
University of Waterloo
Canada
adeanhal@uwaterloo.ca

Robert Warren
Math and Statistics
Carleton University
Ottawa, Canada
rwarren@math.carleton.ca

ABSTRACT

Personal profiling has long had negative connotations because of its historical association with societal discrimination. Here we re-visit the topic with an ontology driven approach to personal profiling that explicitly describes preferences and appearances. We argue that explicit methods are superior to vendor-side inferences and suggest that privacy can be maintained by both exchanging preferences independently from identity and only sharing preferences relevant to the transaction. Furthermore this method is an opportunity for additional sales through the support of anonymous 'drive by' shopping that preserve privacy. We close by reviewing the computational advantages of accurate profiling and how the ontology can be applied to complex real world situations.

1. INTRODUCTION

In this paper, we consider which additions to user profiles are helpful to support suggestion, matching and classical information retrieval needs in shopping, dating, and erotica contexts. The novel approach of communicating preferences without identity is explored, as well as the relationships between the notions of sex and gender within recommendation systems. This research is a continuation of previous research efforts on data extraction [15] and recommendation systems [4], using user gender and personal appearance preferences.

Keeping track of gender has historically only been important for a limited number of purposes: a) it is a high selectivity identifier for identification purposes, b) it permitted the automation of expected social conventions and salutations and c) allowed persons to be pre-qualified according to a marketeer's sales plan.

A strong desire for personal privacy is now preventing this information from being widely shared because preferences and identity have historically been unified. Concurrently, social mores have changed in that alternative interpersonal relationships are being recognized as mainstream and the classical definition of gender as male or female is being questioned. It follows that these changes will drive the creation of new tools which are used to seek entertainment, relationships and goods.

Far from classifying individuals in a box, we seek only to provide models accurate enough for a software agent to adjust its retrieval algorithms. We note that human preferences are notoriously fickle as well as the difficulty in getting a person to express their wants and desires when they themselves are unsure. The intent is not to turn the individual into a product, but to facilitate communications with the information retrieval agent and permit proper content negotiation with information providers. A number of research problems remain in the privacy-preserving processing of the information, but this research is a first step in documenting them.

This paper is organized as follows: we first review the previous work done in customizing recommender and IR systems using personal preferences and then motivate our research based on a number of use cases that occur in the IR field. An ontology of personal characteristics, gender and ethnicity is also presented as an experimental reference for assisting IR for positive and negative feedback. Finally, we report on some experimental results in the retrieval of erotica materials and how the use of a separate preference data structure improves retrieval performance. We conclude on the extension of the ontology to other IR and matching problems.

2. PREVIOUS WORK

The use of recommender systems and profiled IR systems is not new and several other research approaches have been attempted in the past. Approaches using mobile software agents were proposed early on to perform distributed information retrieval by Brewington et al. [2] while other researchers, such as Pipanmaekaporn[12] focused on learning a user's interest based on a single relevancy class using research paper collections.

Daoud [3] and Middleton [11] furthered this approach by utilizing web server logs to determine user interest which was then classified within topic ontologies based on the Open Mozilla Directory.

Schiaffino and Amandi [13] also supported the creation of user profiles through the use of demographic databases and

SEXI 2013 Rome, Italy

user questionnaires. Gasparini [6] used an ontology to report the needs, demographics and languages spoken by a user in order to customize the material for presentation. Sutterer et al. [14] made use of OWL [1] ontologies in an attempt to provide better context to retrieval situations.

Katifori [9] and Ghosh [7] also made attempts at building integratable user profiles using semantic web technologies. Ghosh [7] makes use of a simple "Shopping List" property and Sutterer [14] attempted to create context for preferences based on locations. One of the reoccurring issues is the lack of support to record preferences within personal profiles.

The relevance track that was part of the TREC 2012 conference explored the performance of recommender systems. For these systems, preference data was provided as a list of 50 venues, along with positive and negative feedback data supplied by the users. Out of 23 recommender systems submitted to TREC 2012, 14 systems performed better than simple baseline systems that did not incorporate user profile data and simply made general, non-user specific, suggestions [4].

In the following research, we present a modified approach to information retrieval using a novel ontology titled Appearances [2] that can support both preferences and generic identity information. Terms are provided for physical appearance traits, body measurements as well as sexual, romantic and entertainment preferences and aversions. This is also done in a manner that allows for the independent use of different aspects of the ontology without having to produce an actual identity, which is similar to the approach used by Gulyás and Imrel [8] for anonymizing social network applications.

3. SPECIFIC USES CASES

Figure 1 is a high-level representation of what we believe the appropriate use of the ontology should be. Here the profile information is completely stored on user's computer. Previously, most if not all user preferences would have been stored as part of a user account on the server side, which would require the user to register with the web-site before usage. Previous research [5] has shown that the effort required for registration is an effective deterrent to the buying behaviour.

Instead we propose for most of the profiling information to be stored on the user's client. The profile agent will communicate which aspects of the profile it could make use of to improve retrieval, these portions on the profile can then be dispensed to the server after the user has authorized it. The user can decide to withhold certain parts from the server or never include them in the profile. The profile can be sent to the server irrespective of whether the user has registered with the merchant. For privacy reasons, the profiling information can also be dispensed anonymously.

This is not unlike the mobile agent information retrieval paradigm of Brewington et al. [2], with the caveat that ontological structures are exchanged instead of computer code. To the best of our knowledge, providing partial elements of a personal profile to the seller without identity information in a novel contribution that has not been studied before. We now review several use cases where we think that this approach may be useful and a solution to specific corner cases within each one.

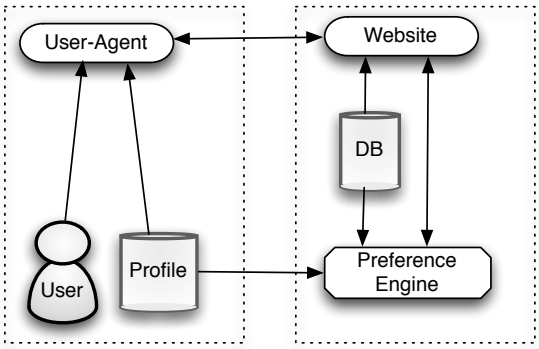


Figure 1: A user's profile will be stored locally and portions of the profile are used by the provided to improve information retrieval.

3.1 Shopping

Online shopping is an application that has traditionally made use of 1) short-term, cookie based preference modelling based on advertising click through, 2) the relationship between products "people who bought A also bought B", and 3) proprietary long-term preferences modelling for registered users. Method 1 has been reported as having conversion rates as high as 10% in trade magazines for large retailers, while method 2 is known to increase sales by providing the users with a set of products likely to be of interest. In both cases, these systems have been known to make recommendations that were not only ineffective, but actively detrimental to the relationship between vendor and customer, for example by recommending a book on anal sex as a Father's Day reading list and as complementary to an evangelist book[10, 16].

The third option of requiring customer registration is known to be a hindrance since it requires effort from the user to enter his personal information, additionally some users will not be comfortable sharing all the requested information. Vendors have a vested interest in getting the user to register since demographic and marketing information can be derived from logistic information such as billing address. This type of profiling or recommendation system works best for existing customers who have already expended effort on registering.

Lastly, an existing problem with recommendation and profiling systems is their inability to handle contexts, such as a difference between the person buying and the person receiving the gift. While florists and some online bookstores have had to tackle this problem, they do so primarily by separating shipping and billing addresses and focusing on calendar

[1] www.w3.org/TR/owl-ref/

[2] http://rdf.muninn-project.org/ontologies/appearances.html

events, such as Mothers' day, instead of the customer's profile.

A re-occurring use case for third party purchases has to do with a husband shopping for lingerie for his wife. What makes this particular case so interesting is that the product is completely removed from the shoppers' experience and any profiling information likely to be available: the sex, gender, body measurement (which vary across country and sex) preferences are unusable and in some cases the husband is ignorant of their spouse's size. This particular use case has spawned several web-sites dedicated only to this problem which is dealt with as a presentation problem instead of a recommendation problem.

Besides inducing large errors in 3rd-party recommendation systems that monitor browsing behaviour (purchased goods drive the objective function), these specific cases represent lost sale opportunities in that the user is fighting the recommender system while attempting to search for a relevant gift.[3] The next-generation use case is one person shopping for another person whose lifestyle or culture are completely orthogonal to their own, and who are looking for gifts that are appropriate without having a complete understanding of the world of the gift receiver.

3.2 Dating

As one of the primary human drives, dating is an application where personalization and profiling are key. Furthermore, it is a matching problem in that the preferences of both potential matches must be taken into account concurrently. The removal of geographic restrictions and limited online anonymity have enabled the creation of new alternative communities for dating, such as academic or military singles dating. Currently the popular online classified ad web-site Craigslist has no less than 21 different types of relationships listed, ranging from traditional marriages to polyamourous relationships.

Interestingly, the mass customization of dating communities seems to rely primarily on the re-branding of the same back-end systems and / or the restriction of the sex field on the site registration form. We performed a short survey of alexa.com's directory of gay, lesbian and alternative dating sites, classifying them by their treatment of gender. We also performed the same classification for the dating websites found in the first 20 search results for the Google queries "gay dating" and "lesbian dating".

The results in Table 1 tabulate the number of dating sites according to their treatment of a person's gender. The first class, "Generic", identifies dating sites that are simply targeted advertisements for larger, brand name dating sites. The "Hard-coded" class contains the dating sites that use commercial off the shell dating website software, with only one gender as a choice. Lastly, the "Choice" class lists websites where any kind of gender differentiation opportunity is provided to the user.

Source	Generic	Hard-coded	Choice
Alexa	2	4	4
Google 'Gay Dating'	5	6	3
Google 'Lesbian Dating'	3	7	2

Table 1: Gender customization of different dating websites

This survey is by no means comprehensive, but is valuable in identifying the lack of support for the self identification of gender. In all three surveys, the majority of dating websites treat gender as a binary choice with no attempt at differentiation, even through this additional profiling would make matching easier.

What is interesting is that the marketing documentation of the websites makes it clear that the website operators are aware of the target community and its terminology. This understanding has not been ported to the search and matching functions of the website since its profiling system is incapable of recording the data.

In the last category of dating websites, some support was provided for the limited self identification of gender, primarily through the use of the 'Male Trans Female' and 'Female Trans Male' terms. The only other case was the crude use of sexual positions as a proxy for gender and this approach may not be appropriate for all demographics.

Personal preferences in romantic and / or sexual relationships are among the most complex, and can be at times contradictory. This limited support for gender terms within dating sites does not provide adequate support for the profiles specific to the target community. The ability to process a complex personal profile would allow matching engines to locate not only members of the users' preferred community but improve its matching algorithms based on its understanding of that community.

Furthermore, some of these preferences can be awkward to enumerate in public and can incur a certain public stigma. Examples can include a preference for persons with a specific hair colour, or an aversion to persons from certain cultural backgrounds. In these cases, profile preferences can be used as a social lubricant by avoiding unnecessary rejections and ensuring that incompatible matches are never introduced to one another.

Lastly, we note that terms, labels and nomenclature represent a tremendous opportunity for additional profiling by extrapolating additional information from the specific class of terms used within a personal profile. As an example, the terms "man, gentleman, dude, bro, boy, lad", all have been used by men of any age to describe themselves where each implies a set of demographic, social and temporal properties that can populate a personal profile. The key problem in their use, which we will not tackle here, lies in documenting those properties both in the user's semantics and its relationship to the merchant's semantics.

[3]The authors note that the systems that generate and display online ads went to great lengths to try and find a relevant advertisement after several days of online searches on gender, appearances and interpersonal relationships. Results ranged from comical to insightful, but it is obvious that further work on profiling (including a "please-ignore-this-search" button) are needed.

3.3 Erotica

Beitzel et al. [1] estimate that as much as 7% of queries in the 2006 AOL query log were pornography related. Sex being one of the primary human drives, it is no surprise that erotica searches are an important class of search problem.

Creating preferences sets for erotica is a complex endeavour in that there is a strong element of fantasy to the preferences and aversions. There does exist a probabilistic relationship between romantic, sexual and entertainment preferences, but its complexity is too high to easily infer one from the other.

We can, however, let the user specify their entertainment preferences using the ontological constructs for gender, ethnicity and physical appearances. Anecdotal review of erotica web-site folksonomies in Section 6 has identified Appearance, Gender and Race/Nationality as the most selective categories within collections, which makes vendor selection possible for the user and query pre-processing possible for the vendor.

Currently, erotic material is an adult-oriented product that requires special considerations (as with alcohol or medication) and which is a social taboo. These factors therefore dictate that these customers will likely be concerned about privacy and desire a discrete shopping experience. An overly aggressive attempt at getting the customer to register with a vendor will likely fail; the ontological preference terms therefore permits the customer to locate the entertainment that he desires without being driven away too early.

Without doubt, the use cases listed above and the suggested use of ontological terms for describing preferences or aversions will likely engender a new generation of spammers attempting to disguise their intentions. Vendor reputation and statistical normalization methods will therefore be required to perform expected quality control on vendor results and the suggested products.

4. APPEARANCES ONTOLOGY DESCRIPTION

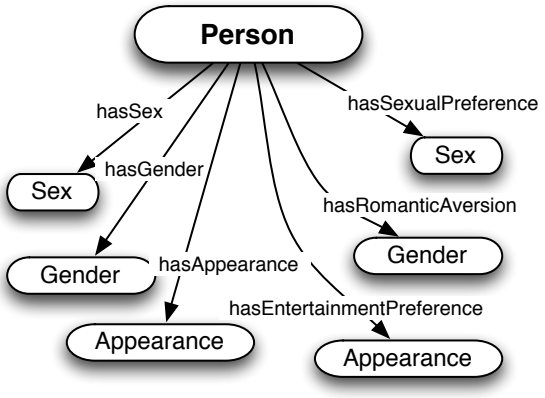


Figure 2: The profile contains attributes, preferences, and aversions for different situations.

User profiling and preferences remain an ongoing problem within e-commerce systems as there is a lack of standardization. What systems do exist are black box systems dedicated to specific tasks whose models are not portable to related modelling questions. Furthermore, what standards do exist vary according to the institution that published them and their intended audience, making data transformation problematic. Figure 2 shows an overview of the profile information which contains information such as the person's gender and entertainment preference.

We make use of OWL based ontological data structures because they have built in equivalence and inheritance for both properties and classes. This ensures that whenever preferences are communicated, a super-class of the desired property is available as a fall-back if the specific semantics desired is unavailable. Furthermore, the data structures are independent of any software package preventing lock-in and the use of terms and properties can describe profile data independently of identity which permits both anonymous "window shopping" and shopping with a third party's preferences.

In the following subsection, we review the aspects of the Appearances ontologies and how it represents a persons' attributes, preferences and aversions.

4.1 Gender, Sex and Orientation

The Appearances ontology was originally meant to deal with ambiguous gender references in text and was expanded to deal with soldiers' personal description. In the coming paragraphs we describe the working of the ontology and its application to information retrieval.

An ongoing problem for ontology design is that a number of social and linguistic conventions are in everyday use while being logically wrong or ambiguous. Examples include the use of sex and gender interchangeably, the use of contradictory genders ("Sarah was a airman") and the ambiguity of perception.

The ontology provides two sexes, which within the ontology is grounded to XX and XY phenotypes and three genders Male, Female, Transsexual[4]. The terms have no restriction on any combination of Gender, Sex or Relationship which gives the end user full descriptive power. Specifically, sexual, romantic and legal relationships can be separated and impose no combinatorial restrictions.

As the ontology has its roots in the processing of war records that straddle the Edwardian and Victorian Eras, a second set of terms are provided which are suffixed Simple. These terms are a replica of the generic terms previously enumerated, but include cardinality and disjoint restrictions that enforce gender assumptions held within official historical records. Hence, instance SimpleGenderM is the same as SimpleSexM and SexISO5218-1 while being disjoint with SimpleGenderF. This permits assertions that the wife of soldier on an 1915 form must be a woman and the mother of their child. Not all cases fit within this model, a number of women do

[4]We note that there exists a great deal more, but we provide only the most obvious ones here.

serve as soldiers as both men or women, but it provides a model that accounts for the majority of cases and that can locate exceptions worthy of study within a database.

4.2 Observed versus self-reported profile properties

One of the more useful aspects of an ontology as opposed to other schema based solutions is its ability to make use of sub-properties and sub-classes. In cases where a perfect match can not be obtained between the merchant and users' systems, it is always possible to obtain partial information. A typical example could be the `hasAppearance` property that can be branched into `hasAppearanceObserved` or `hasAppearanceSelfReported`: different vendors may choose the property as observed by an authority versus a self-reported property, but have a documented alternative to the `hasAppearance` property if the specific property they want is unavailable.

Similarly, an interesting element of the linked open data model is that the parts of ontologies can be separated and used independently without loss of semantic meaning. The utility of this is evident for "window shopping" e-commerce applications where anonymous preferences can be enumerated by the user agent without necessarily providing identity information.

4.3 Eyes, Skin and Hair

The ontology provides several different reference standards for identifying the colour of hair, eyes and skin. Because not all standards have the same degree of specificity or precision and equivalences are not always available from one standard's term to another. In some cases we are able to provide properties that indicate the inclusivity of a term within another through the use of skos:broader and skos:narrower properties. An example of this is the use of the Martin-Schultz eye colour scale which provides several different terms for shades of blue eyes that all map to the single BLUE eye colour for motor vehicle license terms.

Other standards such as the Von Luschan skin colour terms are known to be ambiguous, while some of the hair colour references in the US Federal Bureau of Investigation such as PINK or BALD have no equivalent within the Fischer-Saller hair colour scales.

There does exist a series of statistical relationships between macroetchnicity and skin, hair and eyes: it is not unreasonable to expect a person from Japan to have black hair. However because of the difficulty in reconciling statistical relationships into logical relationships, these are not currently recorded in the ontology.

Several standards are referenced within the ontology to improve interoperability with the caveat that not all known standards are directly interchangeable. Appearances makes use of three levels of properties to record the relationships between the standards: owl:sameAs for terms that are completely similar, skos:related for terms that can easily be confused for one another (eg: grey eyes versus blue eyes) and the skos:broader and skos:narrower properties for terms that can encompass a series of other terms (eg: light blue, dark blue versus blue). Whenever possible, common language labels have been provided and related to these terms.

4.4 Modelling preferences and aversions

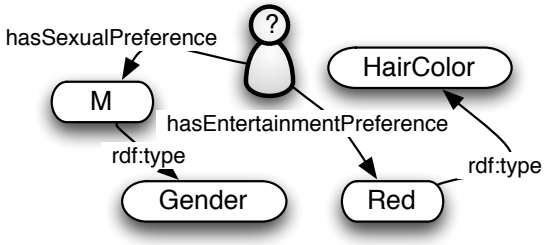


Figure 3: Enumeration of anonymous preferences

Figure 3 is a representation of the anonymous profile information for a person's sexual and entertainment preferences in the context of the information retrieval of erotica, in this case a person interested in males with red hair. An ongoing discussion is whether merchants will accept to process anonymous preferences, as some merchants wish to establish user accounts immediately in order to capture the customer.

Research suggests that online users shun complexity of user account registration and favour web-sites that provide effective search functions [5]. Notwithstanding, this is likely to lead to a second generation of spam and spam sites, with the caveat that the encoding will allow complex normalization.

We choose the term aversion to represent the antonym of a preference. As with preference, these properties do not represent values or moral judgments, but affinities towards certain concepts which recommender systems should use with a certain amount of flexibility. They are not meant to represent absolute requirements: as an example of food preferences one can record a preference for oranges and an aversion to mushrooms but not a deadly allergy to nuts.

Also note that between the set of preferences (P_s) and the set of aversions (A_s), there may exist an unknown region (U_s) within the universe S that is $U_s = S - P_s - A_s$. This has to do with both preferences and aversions not being a complete element set of possibilities.

Figure 4 is a limited representation of the taxonomy that links the different sub-classes of appearances, different measuring standards, and actual measurements.

Appearances makes use of a top level property that records all other aspects of the person appearances, including measurements, hair, eye and skin colour. All of these are accessible according to multiple formats. There are relationships that are encoded within the ontology that allow for the translation of equivalent or similar terms across different standards. This allows an agent and server to negotiate transactions even through both are not using standards that are not completely compatible. One item that created some difficulty is that information that is expected to be true statistically, such as age causing grey and white hair is not

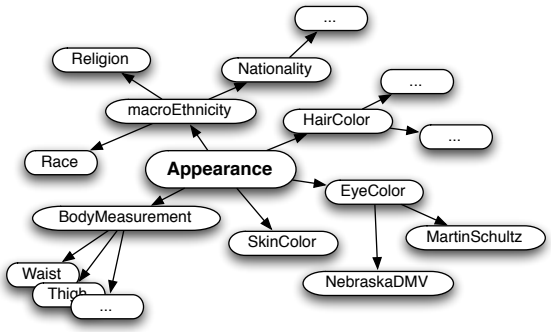


Figure 4: The ontology of appearances has multiple classes with sub-classes and classification standards.

easily encoded into an ontological framework.

4.5 Body Measurements

Body measurements are provided as part of the appearance class for both men and women. Their use-case for online shopping is clear and we use generic clothes fitting measurements, which are historically linked to sex. For this reason, all measurement properties are related to the sex ontological terms for men or/and women instead of their gender. As measurement units are currently problematic in ontologies, separate properties for both decimal inches and meters are provided.

Manufacturer's clothing and shoe sizes have been omitted as a number of different conflicting standards exists across the world for men and women and equivalences are not always available or consistent over time. Thus, we only provide a basic set of body measurement properties as a starting point. In future work, it would be interesting to add the ability to translate clothing size across standards. This would be useful in the use case of Section 3.1 where a person is purchasing a gift for a spouse from one vendor based on the label size of a current garment from a second vendor.

5. IR OF EROTICA MATERIAL

In order to estimate how including attributes, such as gender, into an IR system would improve results we calculated which subset of documents would have to be searched if the system knew a user's attribute preference and the expected speedup a system could have given this information.

Site	Num Documents
Site 1	472,283
Site 2	4,785,909
Site 3	76,531
Site 4	637,650
Site 5	5,441,078
Site 6	2,762

Table 2: Number of documents for each site.

We looked at the categorization of documents (videos) on six pornography sites. These sites were chosen from a list of the top sites in the "Adult" category of Alexa.com. They were all general interest sites, rather than sites that focused on only a certain category of documents, e.g., a site that only contains Japanese oriented documents. It is, however, interesting to see that genre specific categories exist, these sites are, effectively, limiting their results to a specific attribute that could have been communicated in a profile. All of the sites chosen also exposed how many documents were in each of the categories on the site. The number of documents in each site can be seen in table 2.

For each of the top 50 categories on each site, we labelled whether the category was relevant to each of four attributes: gender, ethnicity, age, and hair colour. For example, the category "college girl" is relevant to the age group and the female gender group. Most categories are only relevant to one of the four attributes. These four attributes were chosen as the attributes that had many relevant corresponding categories in our data-set. Each of the four attributes can have certain values. The possible values chosen are values associated with the top 50 categories rather than every possible value.

We then looked at how search performance would improve with knowledge about user preferences from profiles. For example, given the query "swimsuit brunette", a search system that does not consider profile data would have to search all documents for this query. However, with the user preferences from the profile, we could know that, for example, the user likes documents where the hair colour attribute is "brunette". With this knowledge we can not only find more relevant documents but also improve the time it takes to perform the query. This query would only have to be run on the subset of documents in the "brunette" category. The documents are categorized so we know, in advance, all possible values for all four attributes so we would be able to precompute which documents belong to which attribute values. This would enable us to speedup the search for documents where an attribute is specified. Note that not all categories have a value for each of the four attributes. For example, some categories do not specify anything about hair colour.

In order to estimate the speedup improvement we use calculations for the fraction of documents that fall within each attribute. The speedup is the inverse of the fraction of documents that would have to be searched. Each of the four attributes can have one of several values. In order to calculate the advantage of having preference information about an attribute we take the mean of the fraction of documents across all of the attribute values. This is done for each of the four attributes for each of the six sites. So, for hair colour, this would be the mean of the fraction of documents that would have to be searched over the blond, brunette, and redhead preferences.

6. EXPERIMENTAL RESULTS

Figure 5 shows the observed speed up of the system when certain parts of the profile are exposed to us. Again, these collections are general, rather than category specific, we expect that similar results will be found for other general collections. We calculated the average speed up given each of the four factors for each site. For example, if we knew the desired ethnicity from the profile, on site 1 there is a

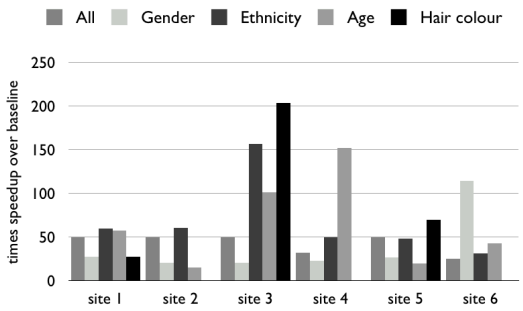


Figure 5: Observed speedup, over baseline of searching the entire collection, across the six sites.

60 times speed up to search time. This increase in speed comes from the fact that we only need to look at the subset of documents that are relevant to a given ethnicity and we have pre-computed which documents are related to which ethnicity.

As a comparison to the expected speed up for each factor we look at the expected speed up if the profile contained which specific category the user preferred. The "All" expected speed ups are the mean expected speed ups over all categories (the categories given on the site rather than the attributes). Note however that for the four attributes more than one category might be included in a particular attributes. For example, if there is a preference for the "male" gender there may be multiple categories that satisfy this preference, which will generally lead to less of an expected speed up than if the preference specific a single category. The "All" category is provided as a means to compare random category selection to given the selection of attributes. These speedups could be realized by picking categories at random and recording whether a user has a preference for that category in the profile.

Factor	Speed up
All	43
Gender	37
Ethnicity	68
Age	65
Hair Colour	50

Table 3: Observed speed up for attributes.

The differences in the expected speed ups between sites is seen because each site has a different number of categories relevant to each factor and a different number of documents within each category. The average expected speed ups for each factor can be seen in Table 3.

7. CONCLUSION

In this paper, we reported on current issues in the use of user preferences for e-commerce application including information retrieval engines. A novel ontology describing a person's preferences and appearances is described and its applications to multiple use cases presented. Finally its application to the problem of the retrieval of erotic material was reported with a speed improvement that can be achieved in a manner that permits customers to preserve privacy. Using an ontology, like the one describes, to record profile data gives the flexibility needed to describe a variety of a user preferences. At the same time it allows for anonymity and privacy to be preserved.

8. REFERENCES

[1] S. M. Beitzel, E. C. Jensen, O. Frieder, D. D. Lewis, A. Chowdhury, and A. Kolcz. Improving automatic query classification via semi-supervised learning. In *ICDM*, pages 42–49. IEEE Computer Society, 2005.

[2] B. Brewington, R. Gray, K. Moizumi, D. Kotz, G. Cybenko, and D. Rus. Mobile agents for distributed information retrieval. In *Intelligent Information Agents, chapter 15*, pages 355–395. Springer-Verlag, 1999.

[3] M. Daoud, L. Tamine-Lechani, M. Boughanem, and B. Chebaro. A session based personalized search using an ontological user profile. In S. Y. Shin and S. Ossowski, editors, *SAC*, pages 1732–1736. ACM, 2009.

[4] A. Dean-Hall, C. L. Clarke, J. Kamps, P. Thomas, and E. Voorhees. Overview of the trec 2012 contextual suggestion track. In *Online Proceedings of TREC 2012*, Gaithersburg, MD, USA, 2012. National Institute of Standards and Technology. To appear.

[5] N. Farag, M. Smith, and M. Krishnan. The consumer online purchase decision: A model of consideration set formation and buyer conversion rate across market leaders and market followers. In *AMCIS*, page 37. Association for Information Systems, 2003.

[6] I. Gasparini, M. S. Pimenta, J. P. M. de Oliveira, and A. Bouzeghoub. Combining ontologies and scenarios for context-aware e-learning environments. In J. C. Anacleto, R. P. de Mattos Fortes, and C. J. Costa, editors, *SIGDOC*, pages 229–236. ACM, 2010.

[7] R. Ghosh and M. Dekhil. Mashups for semantic user profiles. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *WWW*, pages 1229–1230. ACM, 2008.

[8] G. G. Gulyás and S. Imre1. ï£ijanalysis of identity separation against a passive clique-based de-anonymization attack. *INFOCOMMUNICATIONS JOURNAL*, 3(4):11–20, December 2011.

[9] A. Katifori, M. Golemati, C. Vassilakis, G. Lepouras, and C. Halatsis. Creating an ontology for the user profile: Method and applications. In C. Rolland, O. Pastor, and J.-L. Cavarero, editors, *RCIS*, pages 407–412, 2007.

[10] S. K. Lam, D. Frankowski, and J. Riedl. Do you trust your recommendations? an exploration of security and privacy issues in recommender systems. In *Proceedings of the 2006 international conference on Emerging Trends in Information and Communication Security*, ETRICS'06, pages 14–29, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] S. E. Middleton, D. D. Roure, and N. Shadbolt. Capturing knowledge of user preferences: ontologies in recommender systems. In *K-CAP*, pages 100–107. ACM, 2001.

[12] L. Pipanmaekaporn and Y. Li. Mining a data reasoning model for personalized text classification. *IEEE Intelligent Informatics Bulletin*, 12(1):17–24, 2011.

[13] S. N. Schiaffino and A. Amandi. Intelligent user profiling. In M. Bramer, editor, *Artificial Intelligence: An International Perspective*, volume 5640 of *Lecture Notes in Computer Science*, pages 193–216. Springer, 2009.

[14] M. Sutterer, O. Drögehorn, and K. David. User profile selection by means of ontology reasoning. In D. Collange, T. Atmaca, M. D. Logothetis, H. Mannaert, J. T. Yu, and C. Dini, editors, *AICT*, pages 299–304. IEEE Computer Society, 2008.

[15] R. H. Warren. Creating specialized ontologies using wikipedia: The muninn experience. In *Proceedings of Wikipedia Academy: Research and Free Knowledge. (WPAC2012)*, Berlin, Germany, June 2012. Wikimedia Deutschland.

[16] A. S. Weigend. Analyzing customer behavior at amazon.com. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 5–5, New York, NY, USA, 2003. ACM.

Exploring YouPorn Categories, Tags, and Nicknames for Pleasant Recommendations

Michael Schuhmacher, Cäcilia Zirn, Johanna Völker
Data and Web Science Group
Universität Mannheim, Germany
{michael, caecilia, johanna}@informatik.uni-mannheim.de

ABSTRACT

YouPorn is one of the largest providers of adult content on the web. Being free of charge, the video portal allows users - besides watching - to upload, categorize, and comment on pornographic videos. With this position paper, we point out the challenges of analyzing the textual data offered with the videos. We report on first experiments and problems with our *YouPorn dataset*, which we extracted from the non-graphical content of the YP website. To gain some insights, we performed association rule mining on the video categories and tags, and investigated preferences of users based on their nickname. Hoping that future research will be able to build upon our initial experiences, we make the ready-to-use *YP dataset* publicly available.

1. INTRODUCTION

Sex sells, and it is not astonishing that adult content is widespread on the internet. According to Alexa Internet Inc., XVideos, PornHub, and YouPorn are ranked among the hundred most popular websites. Recommender systems as well as data mining techniques could help guiding the user through the large amount of available media to the content most relevant to her/him. In this paper, we focus on the user-created, textual meta information of a video for this purpose. We report on first experiences with the YouPorn (YP) data set, a large corpus of nicknames, comments, video tags and categories which we constructed by crawling the publicly available pages of *youporn.com*. We try to outline the key challenges of working with these user-generated contents in the pornography domain, illustrating our findings by the preliminary results of our experiments with this data. Our experiments were motivated by the following research questions: Would it be feasible to recommend videos to users only knowing their nicknames? Can we recommend additional tags (or categories) for a video based on the ones already assigned to it?
The preliminary work presented in the following is still far from giving comprehensive answers to these questions, but it helps getting a grip on this rather unusual type of data. We hope that our experiences and the availability of our data set will inspire and facilitate future research on mining adult content.

2. THE YP DATA SET

The YP data set consists of textual content extracted from 165,402 single HTML video pages from *youporn.com*. We fetched the HTML content with a custom Python screen-scrapping program, effectively retrieving all, as of Oct 2012, available video pages. We used regular expressions to extract the following features and include them in our YP corpus: The unique video title, the average rating and the ratings count, all categories and tags assigned, and all comments including comment text, nickname, and date of commenting. The corpus with all features listed is publicly available to encourage further research by third parties.[1] In the following, we describe the main aspects of the corpus data.

Video Categories and Tags: Any user can assign categories of a fixed vocabulary to a video. The categories are used for the website's main menu categorization. In addition, users can freely create and assign tags which allows for more fine grained differentiations.[2] Of the 165,402 pages, around 50% of pages have at least one category/tag. The maximum numbers of categories/tags we found is 19, the average is at 7.6. The most frequently used category is *amateur* (19,122 videos), followed by *blowjob* (18,964) and the tags *hardcore* (12,868) and *cumshot* (12,061). The categories and tags are not mutually exclusive, e.g. *European* and *Turk* both exist. Furthermore, the categories/tags cover different description dimensions, e.g. *3some* and *Wife* (actor information) co-occurs with *Blow* and *Doggy* (sexual techniques).

User Comments and Nicknames: At *YouPorn*, users can comment on videos and leave a self-selected nickname. 62% of the pages have at least one comment, these having on average 8.8 comments. The distribution of the 910,000 comments shows a nearly steady decline for the page count with respect to the number of comments per video page. The comments themselves are rather brief, with an average word count of 11.2 per comment. An example for a - comparably - meaningful comment is "I kissed my girlfriend like that - she slapped me in my face" from nickname "burning face".

Rank	Nickname	#	Rank	Nickname	#
1	lol	4,911	57	sex	776
2	me	4,898	68	Au Cindy	671
3	xxx	3,597	129	Bill 69	474
8	john	1,603	189	Cunnilinguo	341
14	Con-naisseur	1,393	234	love	303
41	Camille Crimson	966	659	Fred Flintstone	125
45	:)	923	1,288	OldschoolRobert	67

Table 1: Selected nicknames ordered by frequency, i.e. number of comments

The nickname for each comment can be freely chosen and neither has to be registered nor unique.[3] The nicknames in the corpus are thus only plain strings. We identified about 305,000 unique strings of which selected, frequently used nicknames are given in Table 1.

[1] `http://blog.uni-mannheim.de/mschuhma/yp-corpus/`
[2] Since end of Nov 2012, YP does no longer offer user-created tags.
[3] Since end of Nov 2012, YP offers comments with registered nicknames.

3. EXPERIMENTS

For gaining a better understanding of the YP data set, we conducted some experiments in an exploratory manner, as reported below.

Recommendation of Videos by nickname: We assume that the choice of nickname is - besides situational influences in that very moment - largely based on the user's personality. While one person might call her/himself Dragon Slayer or Jabba the Hutt, another one would rather prefer using her/his real given name or point out the size of her/his body parts. This arises the question whether in return it is possible to draw conclusions which videos a user prefers depending on the choice of her/his nickname.

In [1] and [2], the authors analyze nick names in Internet Relay Chat (IRC) and a web forum on eating disorders, manually classifying the nicknames they found. The name category scheme defined by [1] is rather focused on the meaning of the names, using categories like "famous names", "objects", "sex-related nick names". In contrast, the scheme used in [2] is linguistically oriented (e.g. "commonly known names", "nouns and phrases" or "adjectives").

Inspired by the above mentioned category schemes, we categorized the nicknames found in the YP corpus. Aiming at a completely automatic labeling process, we simply matched the nickname strings with predefined lists. Out of the 973,963 comments, we categorized 69,819 comments as showing with a male given name[4] in the nickname, 22,791 with a female given name[4], and 13,409 containing explicit vocabulary (this includes vulgar or pornographic language)[5]. Each name was classified into exactly one category. In case it fit into several categories, explicit content was prioritized; in case of ties with male and female given name, we did not give any label. We have to point out that the category female given name does not necessarily mean the comment was entered by a woman, it could for instance refer to a porn actress.

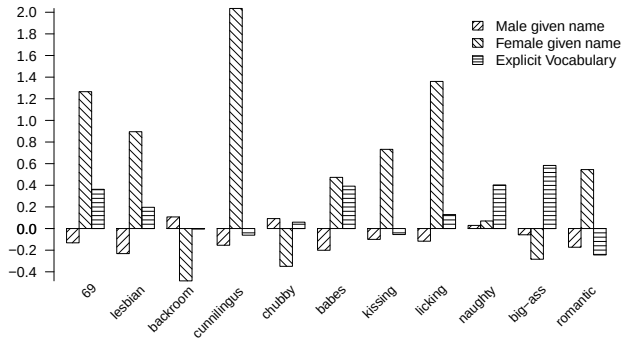


Figure 1: Share of selected categories/tags per nickname type

As a recommender system would rely on measurable differences between these three groups, we analyzed them by comparing their video category/tag distributions. While the majority of the video labels are used equally often amongst the three groups, like for *blowjob*, we also found some differences in terms of preferred labels. Fig. 1 shows the relative difference between the share of a specific category/tag within one nickname group compared to the overall share of this category/tag for all comments. The most salient observation is the increased number of female user names for the tags *69*, *lesbian*, *cunnilingus*, *kissing*, *licking*, and *romantic*, while they are comparably underrepresented for *backroom*, *chubby*, and *big-ass*. Explicit vocabulary user names, however, can be found with augmented occurrences for the tags *naughty* and *big-ass*, and appear less frequently with *romantic*.

[4] http://www.andythenamebender.com

[5] http://www.variety.com/graphics/photos/_storypics/TV-BAD-WORDS.pdf

Recommendation of Tags: Besides the video recommendation for users, improving the completeness and quality of the category/tag assignments, as well as query expansion, with related categories/tags might also be of interest. We therefore analyzed the relationship between the categories/tags by mining for association rules. We used a minimum support value of 5 and a lower confidence bound of 50%. While not accessing the large number of rules with measures of rule relevance, we could identify some interesting rules as reported in Table 2, highlighting the potential to suggest additional tags and thus completing the categorization of the videos. Though including the nickname categories of Section 3, we could not identify any interesting rules containing them.

Conclusion	Premise	Supp in #	Conf in %
female-friendly	kissing, romantic	2,170	91.7
drunk	reality, russian-students	1,113	100.0
british	stockings, senior	315	100.0
nerd	glasses, ponytail	7	100.0

Table 2: Selected association rules between video categories/ tags with absolute support and confidence

4. CONCLUSION

Processing textual data associated with pornographic media requires methods which can effectively deal with a number of problems. The user-generated data obtained from these sites tends to use highly colloquial language or slang, that is not covered by any common lexical resource (e.g. WordNet). Privacy issues and the fact that very little of the actual data is publicly accessible makes it hard to get information about individual users.

Nevertheless, we believe that mining such data is likely to yield interesting conclusions about both a site's contents and its users. In this paper we reported on preliminary experiments using a data set which we created by crawling the publicly accessible contents of *youporn.com*. Though we could not distinguish between disjunct preference profiles for the three nickname categories, we revealed existing differences in their commenting behavior. In order to analyze these differences in more depth, we plan on extending our nickname labeling process with a broad covering, yet fine-grained category scheme, capturing for example nicknames like Jabba the Hutt as a movie character or Camille Crimson as a porn actress. For result interpretation, we head towards a collaboration with sociologists, which we believe is recommendable also for other descriptive analysis of our data set. Furthermore, we will investigate the benefit of organizing categories/tags in a hierarchical scheme for content retrieval.

Acknowledgements. Johanna Völker is financed by a Margarete-von-Wrangell scholarship of the European Social Fund (ESF) and the Ministry of Science, Research and the Arts Baden-Württemberg.

5. REFERENCES

[1] H. Bechar-Israeli. From "Bonehead" to "cLoNehEAd": Nicknames, Play and Identity on Internet Relay Chat. *J. Computer-Mediated Communication*, 1(2), 1995.

[2] W. Stommel. *Mein Nick bin ich!* Nicknames in a German Forum on Eating Disorders. *J. Computer-Mediated Communication*, 13(1):141–162, 2007.

Identifying VHS Recording Artifacts in the Age of Online Video Platforms

Thomas Steiner
Univ. Politècnica de Catalunya
Department LSI
Barcelona, Spain
tsteiner@lsi.upc.edu

Seth van Hooland
Université Libre de Bruxelles
Information and C.S. Dept.
Brussels, Belgium
svhooland@ulb.ac.be

Ruben Verborgh
Ghent University
iMinds – Multimedia Lab
Ghent, Belgium
ruben.verborgh@ugent.be

Joseph Tennis
Information School
University of Washington
Washington, D.C., USA
jtennis@uw.edu

Rik Van de Walle
Ghent University
iMinds – Multimedia Lab
Ghent, Belgium
rik.vandewalle@ugent.be

ABSTRACT

In this position paper, we describe how analogue recording artifacts stemming from digitalized VHS tapes such as grainy noises, ghosting, or synchronization issues can be identified at Web-scale via crowdsourcing in order to identify adult content digitalized by amateurs.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Amateur Video Digitalization, VHS, Online Video Platforms

1. INTRODUCTION

Online adult video is one of the fastest growing Internet industries as recent statistics of a large meta search engine for adult content show[1]. Since its launch in 2006, the search engine has indexed the amount of overall 735,000 videos at a growth rate of 22,000 videos per month with overall 93 billion views. Over this period, 158 million user ratings were collected. It becomes evident that efficient search, recommendation, and navigation capabilities are required in order to use adult video platforms in a meaningful way. Online adult video platforms typically allow their users (i) to search for content based on full-text query terms that are matched against textual descriptions of the video like its title or description, or (ii) to browse the archive of a platform by category or channel, usually based on video tags. Users are presented a top-n ranked list of videos that match a given category or query term, ranked by criteria such as *relevancy*, *view count*, *user rating*, or *upload date*. The default ranking criterion normally is *relevancy*—a platform-specific *black box* concept. Advanced and frequently returning power-users may prefer more transparent and traceable ranking criteria such as the popularity-based *view count* and *user rating*, or the LIFO (last in, first out) ranking criterion *upload date*.

In this position paper, we suggest a computer vision-based approach to automatically identify VHS adult content that has been digitalized in a non-professional manner. This type of niche adult content is characterized by analogue recording artifacts stemming from VHS tapes. Common issues include ghosting, brightness and color channel interferences, chaotic line shift at the end of frames (Figure 1a), and wide horizontal noise strips (Figure 1b).

[1] `http://www.pornwatchers.com/content/statistics11-2012/`

SEXI '13 Rome, Italy

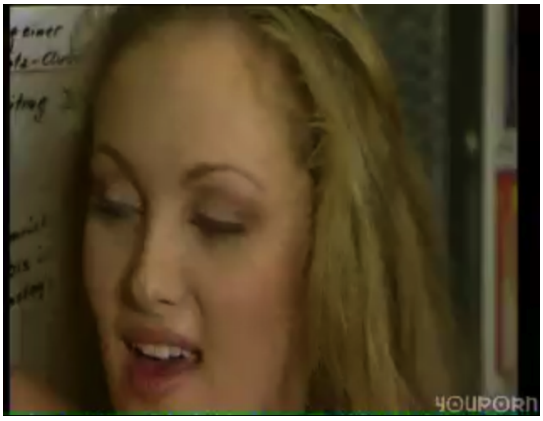
2. PROBLEM STATEMENT

The publication of online content produced by *amateurs*, or non-professionals, has received a substantial amount of attention. This position paper raises the question: to what extent can the identification of content as VHS ADULT CONTENT DIGITALIZED BY AMATEURS offer a useful parameter? Exploiting the fact that an individual invested time and resources for the digitalization of content from a VHS tape can hold a unique value both for information retrieval and research purposes. Especially in the context of the long tail of niche content, automatically identifying VHS ADULT CONTENT DIGITALIZED BY AMATEURS can help identify more quickly unique content items.

Uploaders of this type of content occasionally add tags such as "vintage" or "retro" but these practices are not standardized and sparse. On the aforementioned adult content platform, out of overall 735,000 videos, there were 23.427 tagged as "vintage", 95 as "vhs", and only 50 as "vintage" *and* "vhs". Automated means to aggregate this type of content are needed. In this paper, we propose a scalable, crowdsourced way to identify adult content digitalized by amateurs.

3. PROPOSED METHODOLOGY

In [5], we have introduced a generic crowdsourcing framework for the automatic and scalable annotation of HTML5 video. The term *crowdsourcing* was first coined by Jeff Howe in an article in the magazine Wired [1]. It is a *portmanteau*



(a) Chaotic line shift at the bottom of frames (green)

(b) Wide horizontal noise strip distortions

Figure 1: Typical VHS artifacts and distortions after amateurish digitalization.

of "crowd" and "outsourcing". Howe writes: *"The new pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R&D"*. The difference to outsourcing is that the crowd is undefined by design. For our specific use case, any adult video platform user could be part of that crowd.

While a user watches a video, the framework in the background unobtrusively annotates it, *e.g.*, as demonstrated in the concrete case in [5], to extract events. The annotation framework being generic, we can imagine a video denoising algorithm as presented by Yang in [8] being applied to a video that is currently played to detect if it suffers from VHS artifacts. Over time, *individual* users watching low quality digitalized videos create enough signals to eventually filter out the corpus of content digitalized by amateurs.

4. RELATED WORK

The plethora of online videos can effectively be tackled with the driving force behind it: an enormous community of users. The aim is to make the annotation task as easy and as less time-consuming as possible, in order to avoid disturbing a user's experience. Soleymani and Larson describe the use of crowdsourcing for annotating the effective response to video [3]. They discuss the design of such a crowdsourcing task and list best practices to employ crowdsourcing. The trade-off between the required effort versus the accuracy and the cost of annotating has been described by Vondrick *et al.* [6]. The quality of annotations generated by a crowdsourcing process has been assessed by Nowak and Rüger [2]. Welinder and Perona [7] devise a model that includes the degree of uncertainty and a measure of the annotators' ability. The usefulness of annotations also depends on their envisioned functional value, *i.e.*, what purpose they should serve in the application.

5. FUTURE WORK AND CONCLUSION

Given the streaming nature of online video, our approach inherits the speed and accuracy challenges described in [4]. The solution here is to work with lower resolution versions of the video files in the background. In order to evaluate the accuracy of the generated VHS artifacts annotations, A/B tests with different video resolutions can be used.

In this position paper, we have presented a crowdsourced, scalable approach to detect VHS digitalization artifacts, where users by watching videos do useful work such as detecting VHS artifacts as a by-product of viewing, and thus over time allowing video platforms to identify this type of niche content.

6. REFERENCES

[1] Howe, J.: The Rise of Crowdsourcing. Wired 14(6) (2006), `http://www.wired.com/wired/archive/14.06/-crowds.html`

[2] Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proc. of the Int. Conf. on Multimedia Information Retrieval. pp. 557–566. MIR '10, ACM, New York, NY, USA (2010)

[3] Soleymani, M., Larson, M.: Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In: Proc. of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010). ACM SIGIR, ACM (Jul 2010)

[4] Steiner, T., Verborgh, R., Gabarró Vallés, J., Hausenblas, M., Troncy, R., Van de Walle, R.: Enabling on-the-fly video shot detection on YouTube. In: Proceedings of the 21st International Conference on World Wide Web. ACM (Apr 2012)

[5] Steiner, T., Verborgh, R., Van de Walle, R., et al.: Crowdsourcing Event Detection in YouTube Videos. In: Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) (Oct 2011)

[6] Vondrick, C., Ramanan, D., Patterson, D.: Efficiently scaling up video annotation with crowdsourced marketplaces. In: Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol. 6314, pp. 610–623. Springer (2010)

[7] Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, San Francisco, CA, USA (Jun 2010)

[8] Yang, C.: Video Noise Reduction based on Motion Complexity Classification. In: Asia-Pacific Conference on Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. vol. 2, pp. 176–179 (Nov 2009)

Author Index

ISBN 978-90-814485-9-8