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 a different dimension of descriptions, 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 show 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.

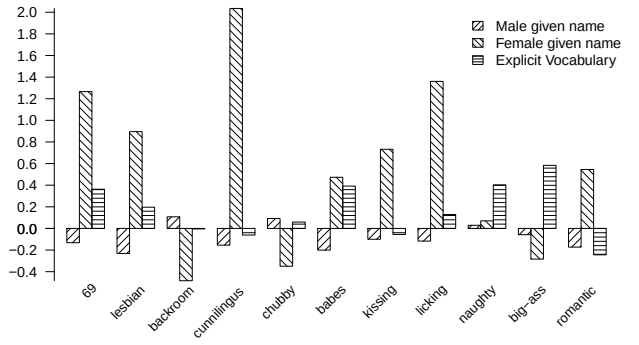

Figure 1: Share of selected categories/tags per nickname type

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 amount of female user names for the tags *69*, *lesbian*, *cunnilingus*, *kissing*, *licking* and *romantic*, while they are fewer represented compared to the other groups 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 rarer in *romantic*.

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.

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

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