User-generated Metadata in Audio-visual Collections

Riste Gligorov
supervised by Guus Schreiber
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands
r.gligorov@vu.nl

ABSTRACT

In recent years, crowdsourcing has gained attention as an alternative method for collecting video annotations. An example is the internet video labeling game *Waisda?* launched by the Netherlands Institute for Sound and Vision. The goal of this PhD research is to investigate the value of the user tags collected with this video labeling game. To this end, we address the following four issues. First, we perform a comparative analysis between user-generated tags and professional annotations in terms of what aspects of videos they describe. Second, we measure how well user tags are suited for fragment retrieval and compare it with fragment search based on other sources like transcripts and professional annotations. Third, as previous research suggested that user tags predominately refer to objects and rarely describe scenes, we will study whether user tags can be successfully exploited to generate scene-level descriptions. Finally, we investigate how tag quality can be characterized and potential methods to improve it.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL—*Information Search and Retrieval*

General Terms

Experimentation, Measurement

Keywords

tagging, video, tag analysis, professionals vs. end-users, games with a purpose, fragment retrieval, tag quality

1. CONTEXT AND RESEARCH PROBLEMS

The central goal of this research work is to investigate the added value of user-generated metadata in professionals environments such as audio-visual archives.

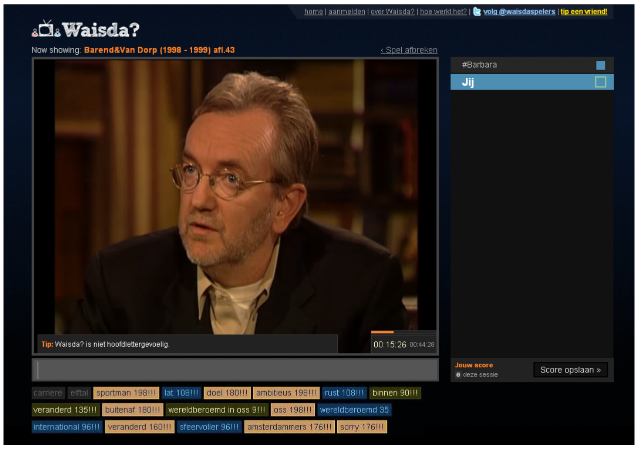
1.1 Audio-visual Collections in the Digital Era

Audio-visual (AV) content collections are undergoing a transformation from archives of analogue materials to very large stores of digital data accessible online. Prerequisite for successful information retrieval and collection management is quality metadata associated with collection items. Traditionally, the task of annotating video programs is strictly in the hands of professional cataloguers who adhere to well-established guidelines in the cataloguing process [1]. Usually, the resulting manually-crafted professional annotations are coarse-grained in a sense that they are referring to the entire program describing the prevalent topics. Fine-grained manual annotation of video fragments, on the other hand, is prohibitive, as the work involved is inevitably tedious, incomplete, and costly [7]. However, a significant portion of user-content requests are targeted at video fragments rather than entire programs; a transaction-log analysis [3] performed in a broadcast archive showed that fragment purchases account for 66% of all purchases. As the program level professional annotations may be inadequate for locating or retrieving of fragments — or any other task that involves fragments rather than entire programs — the need for fine-grained annotations becomes clear. A possible solution to the problem of scarcity of fine-grained video metadata is to harness users' efforts to amass video fragment-level descriptions.

One of the most common forms of user-generated metadata are *tags*. With the advent of Web 2.0 the tagging phenomenon witnessed rapid proliferation in many areas; image collections[1], bookmark collections[2], user video collections[3]. Experience has shown that joint efforts of user communities result in massive amounts of tags usually unparalleled — in terms of quantity — by what professionals can provide. Just to put things into perspective, in 200 years of existance the Library of Congress has applied their expert-maintained taxonomy to 20 million books[4], whereas, in only four years, Flickr's users applied their ad-hoc tagging vocabulary to over 25 million photos. This suggests that there is, indeed, a potential in engaging end-users in the video-annotation process. One way of achieving this is through so-called 'games with a purpose' [13]. To this end, the Netherlands Institute for Sound and Vision[5] (S&V) in cooperation with KRO broadcasting[6] launched *Waisda?* in May 2009, a multi-player video labeling game where players describe streaming video by entering tags and score points based on temporal tag agreements . The underlying assumption is that tags are probably valid — trustworthily describe the video fragments — if they are entered indepen-

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] http://www.flickr.com
[2] http://www.delicious.com
[3] http://www.youtube.com
[4] http://www.loc.gov/about/reports
[5] http://portal.beeldengeluid.nl/
[6] http://www.kro.nl/



Tip: Waisda? is niet hoofdlettergevoelig.

Figure 1: Screenshot of the *Waisda?* game interface: the central part of the screen is reserved for the video. Immediately below the tag input field and the list of tags entered already are placed. The coloration of the tags is indication of the number of points the tags scored. On the right, there is the list of players currently in the game.

dently by at least two players within a given timeframe (see figure 1). The motivation behind *Waisda?* is to improve the access to S&V collection [10]. With *Waisda?* the S&V institute aims to collect metadata in a user vocabulary, as previous research [8] suggests that such metadata can help bridge the gap between the search queries and the indexing vocabulary. In addition, it is expected that the resulting time-related metadata of the content within the video can improve support for finding fragments within entire broadcasts [7].

1.2 Project context

This research is part of and funded by the *PrestoPrime*[7] project which brings together various European audiovisual archives (including S&V), research institutions, and industry partners. PrestoPRIME develops practical solutions for the long-term preservation of video and audio broadcasts, and finds ways to increase access by integrating the media archives with European on-line digital libraries. One of the considered ways to increase accessibility to videos is by exploiting user-generated tags collected through the labeling game *Waisda?*.

1.3 Research questions

The overall problem statement for this research is as follows.

> *What potential added value do user-generated video annotations have in professional environment?*

To successfully integrate user-generated tags into AV collections' workflows a better understanding of their characteristics, compared to preexisting professionals annotations, is required. In particular, the terminology that users employ when describing videos and the aspects of the video that they usually describe. The first research question, therefore, is

1. What are the relationships between user-generated tags and professional annotations in terms of what they describe and the vocabulary?

Locating a fragment within a video is an important use case in AV collections for which the fine-grained user tags could potentially provide an added value. Thus, the second research question addresses the usefulness aspect of user tags in terms of locating fragments within a video.

2. Can we improve fragment search within video, with the help of user-generated data?

Previous research [4, 5, 10] has shown that user tags predominately describe objects and rarely refer to topics of a scene. Considering that professional annotations at scene-level are scarce, we investigate whether user tags can be used to deduce what the scene is about. Therefore, the third research question is

3. Can we derive topical description for scenes from user-generated tags?

Lastly, successful integrations of user tags also requires that issues around assessing and improving tag quality to be addressed. The forth and final research question, therefore, is

4. Can the quality of user-generated tags for videos be evaluated and improved?

2. RELATED WORK

In this section we outline some of the related work. It should be noted that this is an incomplete list of the more relevant studies.

2.1 Games With a Purpose

Games with a purpose (or GWAPs) are computer games, in which people, as a side effect of playing, perform tasks computers are unable to perform [13]. The first example of a GWAP was the ESP game [12], designed by Luis von Ahn, which harnesses human abilities to label images. The idea to collect metadata through games with a purpose has been applied to video footage in, for example, the Yahoo! video tag game [11], VideoTag[8], PopVideo[9] and *Waisda?*. Compared to the other video labeling games, *Waisda?* is unique in the sense that it is initiated by an audiovisual institute (S&V) with the purpose to improve the access to their collection [10].

2.2 Evaluation of End-user Tags in Professional Environment

The Steve project [9] was one of first attempts to explore the role of user-generated metadata. In this collaboration of several art museums a collection of artworks was made available to the general public who were asked to tag them. Among other things, the project studied the relationship of the resulting folksonomy to professionally created museum documentation. The results showed that users tag the artworks of art from a perspective different than that of museum documentation: around 86% of tags were not found in museum documentation.

[7] http://www.prestoprime.org/

[8] http://www.videotag.co.uk/

[9] http://www.gwap.com/gwap/gamesPreview/popvideo/

Museum staff also assessed the tags from the steve.museum project on usefulness when used to search for artworks. From the total number of tags, 88.2% were found to be useful. Following the methodology of steve-museum, S&V institute also asked a senior cataloguer to judged a sample of *Waisda?* tags on their usefulness when searching for videos [10]. The sample consisted of the 20 most frequent and the 20 least frequent tags from two television programs. The cataloguer found the majority of the tags to be useful. She also noted that there seems to exist a difference between professional descriptions and end-user tags. While professionals describe the topical subject of the program, the players in *Waisda?* generally tag things that can be directly seen or heard in the video. One of the aims of this research is to investigate the characteristics of the tags and what they describe in the video more methodically, and on a larger scale.

3. APPROACH

In this section we outline the approach to answer the research questions stated above. The specific approach for each of the research questions is described in separate subsection. The order of the subsections is respective to the order in which the research questions are stated in section 1.3.

3.1 Waisda? User Tags vs. Professional Annotations

To answer the first research question 'What are the relationships between user-generated tags and professional annotations in terms of what they describe and the vocabulary used?' we perform two studies. Details can be found in [4, 5].

The first study is a quantitative data analysis of the entire tag collection gathered with *Waisda?* during the first six months after the game was deployed. In order to estimate the lower bound of the fraction of user tags that are meaningful words, we examine the overlap between them and general lexical resources and vocabularies. Furthermore, to determine if users and professionals use different vocabularies when describing videos, we investigate the overlap between all user tags and a typical domain thesaurus used by professionals in the cataloging process. The results from the study show that there is, indeed, a terminological gap between users and professionals; while approximately 89% of all user tags are meaningful words only 8% of them were found in the domain thesaurus used by professionals.

In the second study, we take a combined approach. First, we investigate what do users tend to describe more: things *heard* or things *seen* on screen. To this end, we perform a study on the overlap between the user tags and the audio signal — subtitles for hearing impaired persons — for a sample of episodes. Second, to get a more comprehensive understanding of the types of tags users usually add, we perform a qualitative study of a sample of user tags obtained through the *Waisda?* video tagging game. In particular, each tag from the sample is manually analyzed in the light of the video content it describes and categorized in terms of the Panofsky-Shatford classification framework. The results of the study show that user tags predominately describe objects and rarely refer to topics of a scene. This is in sharp contrast with the professional annotations which exclusively target the topic(s) of the entire video and (lot less frequently) of particular scenes.

3.2 Investigating the Added-value of User Tags for Fragment Search

Fragment search within video is widely recognized as an application scenario of particular business importance in the AV collections world [3]. However, the existing professional annotations generally refer to the entire video and are not tied to a specific time-point, which decreases their usefulness for fragment retrieval; without temporal data one needs to manually locate the fragment of interest. *Waisda?* tags, on the other hand, do refer to fragments and are time-based, with time codes that *deep link* to particular point in the video. Our aim is to investigate their added-value for fragment retrieval. Furthermore, we plan to explore what kind of query tasks (search for *events*, *objects*, etc.) benefit from the user tags.

The methodology that we consider is *quantitative system evaluation* [14]. In order to evaluate effectiveness of information retrieval, this methodology requires collection of "documents" (in our case video fragments), set of queries, and relevance judgments indicating which "documents" in the collection should be returned for each query. We plan to create evaluation dataset from the videos that were part of *Waisda?* game deployment and tagged by players. For such a dataset real-life, authentic user tags would already be available. However, set of fragments, relevance judgments, set of queries and query task types need to be defined. The method of achieving this is going to be part of and one of the contributions of the research.

An alternative is to creating an evaluation dataset is using an existing one. There are various video retrieval evaluation initiatives like TRECVID[10] that provide datasets for testing content-based retrieval techniques. The advantages of reusing such a dataset are (i) the set of fragments is defined, (ii) the set of queries and the relevance judgments are defined. The drawbacks are that *Waisda?* like games have not been run yet on this dataset and setting up such a game and attracting sufficient users for this material could turn out to be hard. Also we are limited to the query task types defined by the dataset. Our position is that the effort required to collect user tags for the material outweighs the advantages offered by this approach.

3.3 Deriving Topical Descriptions From User Tags

Video fragments or scenes are often multivalent in terms of their meaning and as such can be observed as a mixture of topics. User tags collected through *Waisda?*, which are mostly referring to objects, can be seen as instantiations of these topics. The challenge that we are facing is to derive topical descriptions from the user tags. The unstructured nature of the tags, — only weak temporal ordering of tags within video exists, based on the tag entry time — makes statistical approaches excellent candidates for this task. The research we have done so far puts statistical *topic models* at the top of the list. Topic models [6, 2] are a type of statistical models for discovering abstract 'topics' in collection of documents. One of the most common topic models currently in use is the Latent Dirichlet Allocationc (LDA). The idea behind LDA is to model documents as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms. Specifically, it is assumed that K

[10] http://trecvid.nist.gov/

topics are associated with a collection, and that each document exhibits these topics with different proportions. Furthermore, LDA assumes that words are exchangeable within each document, i.e., their order does not affect their probability under the model. In other words, each document is treated as a 'bag of words'. We believe that the assumptions underlying the LDA model are valid in and applicable to our context as well. *Videos* and *scenes within videos*, much like documents, have many layers of meaning and can be viewed as mixture of topics. The high-level professional *topical* descriptions, on the other hand, can be considered as the analog to topics in LDA. The low-level user tags are referring to things heard or seen on screen and as such can be viewed as instantiations of the particular topics the video/scene is about. The unstructured and unordered nature of user tags — in the borders of a particular shot/scene — fits the 'bag of words' metaphor quite well. Regardless of the choice of the method the procedure we envision will proceed in four steps.

1. *Video segmentation.* Videos are segmented into shots (possibly scenes) using state of the art video analysis tools.
2. *Tag-to-Shot association.* Tags are associated to the shots (scenes) derived in the first step. In this process the temporal information associated to each tag will be used and special attention will be payed to user tags assigned close to the detected boundaries.
3. *Model training.* Statistical model is trained. The collection of topics used by the professional catalogers from the Sound and Vision Institute will be used as a list of abstract topics. We consider this set to be a representable and relevant for two reasons. (1) The videos that will be considered will originate from their catalogue. (2) Every single topical description in their catalog comes from this set. As a training corpus we shall use the Sound and Vision cataloque: each cataloque entry is annotated with topical descriptors and is associated to contextual documents that contain textual description of the program's content. Additional corpora may be used as well.
4. *Topics inference.* Using the trained model, for each shot (scene), the most probable topics will be inferred from the tags associated to it.

We plan to evaluate the effect that the inferred topical descriptions will have on fragment retrieval. For this we will use the test data set from the fragment retrieval study described in section 3.2.

3.4 User Tag Quality Metrics

Our supposition is that any characterization of quality of the user tags is largely determined by how different institutions will use them. In other words, it is immaterial to talk about tag quality without a specific application scenario in mind. Fragment search within video is widely recognized as an application scenario of particular business importance in the AV collections world. For these reasons, our characterization(s) of tag quality will be limited to this scenario. Some of the aspects we plan to investigate in the context of the retrieval scenario are tag frequency and discriminative power of tags, correlation between reputation of players and tag relevance, semantic ambiguity of tags, overlap with transcripts, etc. The exact specifics of this study will be determined by the outcome of the previous studies. Therefore, we plan to make the choice of research methodology and the type of output (e.g. quantitative metrics or set of recommendations) this study will provide after the fragment retrieval study is performed.

4. CONTRIBUTIONS

Contribution from the study described in section 3.1.

- A method to analyze quantitatively the overlap between user and professional terminology exploited in video annotation.
- A method of qualitative analysis of samples of user tags to establish the facets—*who*, *what*, *where*, and *when*—of videos typically described by users and the level of specificity—*abstract*, *generic*, and *specific*—of tags.

Contribution from the study described in section 3.2.

- A method to design a dataset —including fragments, queries, and relevance judgments — for evaluating the added value of user tags in terms of fragment retrieval.
- Evaluation dataset which will contain fragments, user tags from *Waisda?* collection, queries, and relevance judgments.
- Quantitative evaluation of the added value of user tags in terms of fragment retrieval.

Contribution from the study described in section 3.3

- Algorithm for deriving topical descriptions for scenes obtained from user tags.
- Quantitative analysis of the added value of the derived topical description for fragment retrieval.

Contribution from the study described in section 3.4

- Quality metrics or guidelines for accessing the quality of user tags with respect to fragment retrieval.

4.1 Acknowledgments

We like to thank Johan Oomen, Maarten Brinkerink and Lotte Belice Baltussen from the Netherlands Institute for Sound and Vision for initiating and guiding the *Waisda?* project. We also like to thank Q42 for the development of *Waisda?* and making the collected data available. This research is partially supported by the PrestoPRIME project, funded by the European Commission under ICT FP7 (Seventh Framework Programme, Contract No. 231161).

5. REFERENCES

[1] Edmondson, R.: Audiovisual Archiving: Philosophy and Principles, UNESCO, Paris, France, 2004.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] J. Carmichael, M. Larson, J. Marlow, E. Newman, P. Clough, J. Oomen, and S. Sav. Multimodal indexing of electronic audio-visual documents: a case study for cultural heritage data. In *CBMI 2008*, 2008.

[4] R. Gligorov, L. B. Baltussen, J. R. van Ossenbruggen, L. Aroyo, M. Brinkerink, J. Oomen, and A. van Ees. Towards integration of end-user tags with professional annotations. In *Web Science 2010*, 2010.

[5] R. Gligorov, M. Hildebrand, J. Van Ossenbruggen, G. Schreiber, and L. Aroyo. On the role of user-generated metadata in audio visual collections. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*, pages 145 – 151. ACM Press, June 2011.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[7] B. Huurnink, C. G. M. Snoek, M. de Rijke, and A. W. M. Smeulders. Today's and tomorrow's retrieval practice in the audiovisual archive. In *ACM International Conference on Image and Video Retrieval*, 2010.

[8] C. Jorgensen. Image access, the semantic gap, and social tagging as a paradigm shift. *Proceedings 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin*, 2007.

[9] T. Leason and steve.museum. Steve: The art museum social tagging project: A report on the tag con tributor experience. In *Museums and the Web 2009: Proceedings*, Toronto, Canada, March 2009.

[10] J. Oomen, L. Belice Baltussen, S. Limonard, A. van Ees, M. Brinkerink, L. Aroyo, J. Vervaart, K. Asaf, and R. Gligorov. Emerging practices in the cultural heritage domain - social tagging of audiovisual heritage. The Web Science Trust, April 2010.

[11] R. van Zwol, L. Garcia, G. Ramirez, B. Sigurbjornsson, and M. Labad. Video tag game. In *WWW 2008*, April 2008.

[12] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04*, New York, NY, USA, 2004. ACM.

[13] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[14] E. Voorhees. The philosophy of information retrieval evaluation. In *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer-Verlag.