# An Evaluation of Labelling-Game Data for Video Retrieval

Riste Gligorov[1], Michiel Hildebrand[1], Jacco van Ossenbruggen[1,2], Lora Aroyo[1], and Guus Schreiber[1]

[1] VU University Amsterdam
[2] CWI Amsterdam

**Abstract.** Games with a purpose (GWAPs) are increasingly used in audio-visual collections as a mechanism for annotating videos through tagging. This trend is driven by the assumption that user tags will improve video search. In this paper we study whether this is indeed the case. To this end, we create an evaluation dataset that consists of: (i) a set of videos tagged by users via video labelling game, (ii) a set of queries derived from real-life query logs, and (iii) relevance judgements. Besides user tags from the labelling game, we exploit the existing metadata associated with the videos (textual descriptions and curated in-house tags) and closed captions. Our findings show that search based on user tags alone outperforms search based on all other metadata types. Combining user tags with the other types of metadata yields an increase in search performance of 33%. We also find that the search performance of user tags steadily increases as more tags are collected.

## 1 Introduction

Games with a purpose are a way to make humans solve tasks in an entertaining setting. Video tagging games —a type or GWAPs— could become an attractive alternative (or enhancement) to professional annotators in terms of both price and scale. While user tags are virtually for free and plentiful, professional annotations are costly and scarce. The Institute for Sound and Vision (S&V)[1] launched *Waisda?*[2], a multi-player video labelling game where players describe streaming video by entering tags and score points based on temporal tag agreement. The underlying assumption is that tags are faithful descriptions of the videos when entered independently by at least two players within a given time-frame. From here on we shall refer to such mutually agreed upon tags as *verified* tags.

The archive expects that tags collected with *Waisda?* will improve video search. In this study, we put this hypothesis to the test. Knowing that other types of video metadata will also be present, our first research question is: *RQ1: Can user tags, on their own or in combination with other types of metatada,*

[1] S&V, `http://www.beeldengeluid.nl/`, is the Netherlands national archive.
[2] At the time of writing, *Waisda?* is an ongoing project for three years and the game has seen its second release, `http://woordentikkertje.manbijthond.nl/`

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 50–61, 2013.

*improve video search?* To test the assumption that agreement is a good filter, our second research question is: *RQ2: Does limiting only to verified user tags gives better video search performance than considering all user tags?* When GWAPs are used to tag large video collections generally care must be taken to insure 'fair' distribution of game-time across the collection items. In this sense it is instructive for collection administrators and scheduling algorithms designers to know if search performance deteriorates or stagnates after certain point, or if more tags always give better search performance. Therefore, our last research question deals with search performance change over time: *RQ3: How does the user tag search performance change when tags are added?*

The rest of the paper is structured as follows. After discussing related work, Sect. 3 presents our approach. Section 4 describes the datasets and resources that are used in our study. Section 5 introduces the experimental setup. Finally, Sect. 6 and 7 present the results and conclusions of this study, respectively.

## 2 Related Work

*User annotations for video.* Video annotation is tedious and time-consuming activity. Not surprisingly, there exist various initiatives that aim at collecting video annotations through crowdsourcing. In particular, *LabelMe video* is an online video annotation system that allows users to identify objects and annotate visual features such as motion and shapes [5]. However, this frame-by-frame conceptually low-level annotation remains a tedious task. The willingness of people to participate without compensation is limited at best. To alleviate this, [6,7] employ the crowdsourcing MTurk platform to recruit annotators which are paid for the task. An alternative way to motivate people is to gamify the annotation experience through GWAPs. GWAPs are computer games, in which people, as a side effect of playing, perform tasks computers are unable to perform. The main example of a GWAP is Luis von Ahn's ESP image labeling game [8]. Evaluation shows that these labels can be used to retrieve images with high precision and are almost all considered as good descriptions in a manual assessment. The idea to annotate through GWAP has been applied to video in, for example, the Yahoo! video tag game [9], VideoTag[3] , PopVideo[4] and *Waisda?*. With some slight differences, in each of these games players describe streaming video by assigning free-text tags. Thus, we deem *Waisda?* as a typical representative of video GWAPs.

*Relevance judgements and search.* Designing ground truth in a form of document relevance w.r.t. given topic has been playing central role ever since Cranfield experiments gained prominence[4]. The leading actor in IR benchmarking is TREC[5] which employs substantial manpower in creating the ground truth. For organizations lacking the manpower, crowdsourcing is an alternative; [10,11]

[3] `http://www.videotag.co.uk/`
[4] `http://www.gwap.com/gwap/gamesPreview/popvideo/`
[5] `http://trec.nist.gov/`

showed that this task can be reliable fulfilled by crowd workers. Alternatively, Eickhoff et al. gamified the task resulting in increased reliability and reduced cost [12]. In this study we also rely on the crowd; the relevance assessment is outsourced to targeted fan groups.

Search based on user-generated metadata, in particular folksonomies, has been studied before. Morison compared the search performance of folksonomies from social bookmarking Web sites against search engines and subject directories [13], showing that search engines had the highest precision and recall rates. Folksonomies, however, performed surprisingly well. Geisler and Burns state that YouTube tags provide added value for searching, because 66% of them do not appear in other metadata [14]. Hildebrand et al. proposed and investigated a semi-automatic process of assigning explicit meaning to user tags for video by linking them to concepts from the Linked Open Data cloud [15]. To the best of our knowledge, no work has been done to evaluate the performance of GWAP data for video search. Our study aims to fill this void.

## 3 Approach

In order to assess the added value of user tags for video search we use a quantitative system evaluation methodology [4], for which we need a document collection (i.e. video fragments) that is being tagged by a video labelling game, a set of representative queries with associated relevance judgments. We created this evaluation dataset as follows: (i) select a collection of video fragments tagged by players in *Waisda?*, (ii) select a set of user queries from real-life query logs, and (iii) create relevance judgements. All these steps are described in more detail in Sect. 4.2. We use the dataset in two experiments. In the first, we compare performance of search based on different types of metadata. In the second experiment, we study the search performance of user tags over time. In both experiments, we create a number of systems that use the same probabilistic ranking function BM25 [1]; the only variation is the metadata that they index.

## 4 Datasets and Resources

In this section we describe the datasets and resources that are used in the study.

### 4.1 The MBH Video and Metadata Collection

At the time of writing, *Waisda?* is used to tag fragments from the popular Dutch TV program 'Man Bijt Hond' (MBH) produced by the Dutch broadcaster NCRV. MBH is a humoristic TV show that focuses on trivial, everyday news and ordinary and unknown people. Every episode consists of 7-8 unrelated, self-contained fragments where each fragment topically comes under a recurring heading. Players in *Waisda?* tag these fragments. The entire collection to which we have access has 11,109 fragments from episodes aired in the last 11 years.

In addition to the video fragments, we have access to four types of descriptive metadata that are used as input for search:

***Waisda? Tags.*** We consider the collection of all user tags acquired with *Waisda?* during the first five months, starting from October, 2011. In this period 436,456 different tag entries were assigned to 2,192 video fragments by roughly 24,000 players. The number of unique user tags exceeds 47,000. Each tag entry is associated with the point in time — relative to the beginning of the fragment — when the tag was entered. Additionally, each tag entry is marked as 'verified' or not based on the tag agreement in its temporal neighbourhood. As the game is advertised only in Dutch media and the material being tagged is exclusively in Dutch, the language of almost all tags is Dutch. The average number of tags per video is 199. Approximately 55% of all user tags ($\approx$ 243,000) are 'verified' and the number of unique verified tags is 12,861. The average number of verified tags per video is 111.

***NCRV Tags.*** NCRV, the broadcaster, maintains an in-house collection of tags to facilitate Web access to MBH fragments via search and browsing. In contrast with *Waisda?* tags, NCRV tags are not time-based, meaning they are not linked to a particular time-point in the video, and generally cover only the prevalent topics. The average number of NCRV tags per video is 11. Thus they are usually much scarcer than the game tags.

***NCRV Catalogue Data.*** Along with the curated NCRV tags, each MBH fragment has a short textual description, usually one paragraph, and a title. We consider the collection of all titles and textual descriptions (i.e. catalogue data) as another metadata type that will be used in the study.

***Captions.*** Closed captions are textual versions of the dialogue in films and television programs for the hearing impaired, usually displayed at the bottom of the screen. Each dialogue excerpt is accompanied with time-points — relative to the beginning of the video — when the dialogue excerpt appears on and disappears from the screen. We use captions obtained from S&V that cover most of the MBH episodes aired in 2010 and 2011 which amounts to a total of 897 fragments.

### 4.2 Evaluation Dataset

In this section we describe the creation of the three separate components of our evaluation dataset: set of video fragments, set of queries, and relevance judgements.

***Video Fragment Subset.*** The set of fragments for our experiment is selected from the MBH fragments tagged in *Waisda?*. Not all metadata types described above are available for every single fragment. To do a fair comparison of the search performance of various metadata types, we use only a subset. The filtering criterion is as follows: we include only the fragments that have at least one *Waisda?* tag and NCRV tag ascribed to them, and for which captions files are

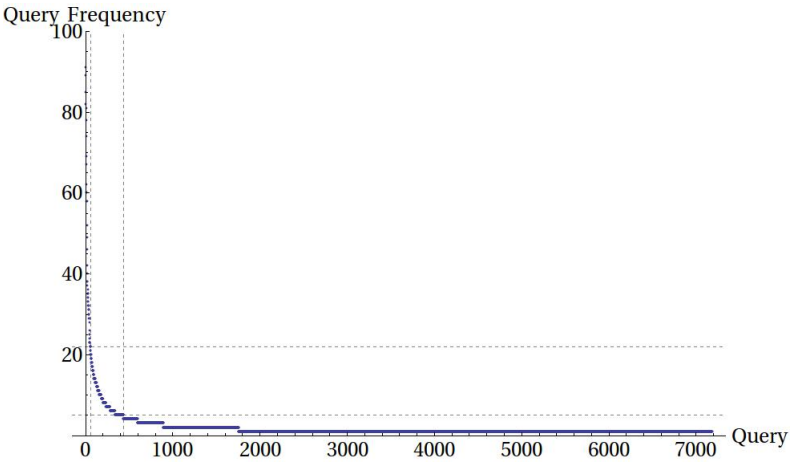



**Fig. 1.** Query frequency distribution. Horizontal dashed lines represent the "appeared 5 times" and "appeared 22 times" thresholds when observing from bottom to top. Vertical lines divide the area under the curve in three equal parts.

available. This results in a collection of 197 fragments. The accumulative duration of our test collection is almost 11 hours of video material, with an average fragment length of approximately 3.3 minutes and a median of 3.6 minutes. The duration of the shortest and the longest fragment in our collection is 0.5 and 8.6 minutes, respectively. The total number of user tags, verified user tags, and NCRV tags ascribed to the the videos of these collection is 107,531, 80,805, and 2,066, respectively. Thus, the average number of user tags, verified user tags, and NCRV tags per fragment is 545, 410, and 10 respectively.

***Query Set.*** To measure the information retrieval performance we use real-life user queries. NCRV provided us with one month of query logs from the MBH web site. The logs contain 15,219 queries posed by internet users to the site's search engine asking for video fragments. Figure 1 shows the query frequency distribution. As seen, the query frequency follows a power law; aside from few frequent ones most of the queries appear infrequently. In fact, only 6% of the queries appear at least 5 times (points under or on the lower horizontal dashed line in Fig. 1).

Out of the complete set of 15,219 user queries we select, in two steps, a subset of 50 queries to include in the study. First, we partition the query set into three classes: a high, mid and low frequency class. The borders of the classes are chosen so that the area under the curve in Fig. 1 for each class is one third of the area. Queries appearing more then 22 times form the high-frequency class, between 5 and 22 form the middle-frequency class, and queries appearing less than 5 times form the low-frequency class.

Second, for each class we perform filtering. Namely, a query is skipped whenever it meets one of the following criteria: (i) it equals with the title of one of

the MBH recurring headings or it contains one of the words 'man', 'bijt', and 'hond' from the series title. (ii) if no video was found for the query using independently at least two of the metadata types described in Sect. 4.1. After the filtering, we are left with 12, 78, and 49 queries from the high-frequency, middle-frequency, and low-frequency class, respectively. The top 12, top 19, and top 19 queries from the high-frequency, middle-frequency, and low-frequency class, respectively, comprise the final query set.

***Relevance Judgements.*** In order to collect relevance judgements for the query set and the fragment collection we performed an on-line user experiment. To this end, we deployed a web application which was used by the participants to carry out the evaluation. For each participant the workflow proceeds as follows. Whenever a participant accesses the web application she is presented with a welcome page which contains a description of the task she is required to perform. Before starting with the evaluation, the participants need to fill out a questionnaire that aims at assessing their familiarity with *Waisda?*, the MBH TV series and the MBH website. Then the participants proceed to the evaluation page (see Fig. 2) which plays a randomly assigned fragment and lists the complete query set. During the evaluation process, the participants watch the fragment and indicate which of the concepts denoted by the queries are shown in it. *We asked users to judge a fragment to be relevant for a query if it depicts the concept denoted by the query.* Each participant is asked to evaluate at least five fragments.


**Fig. 2.** Screenshot of the evaluation page. At the top, a video player is placed which displays the fragment. The list of queries is rendered at the bottom.

**Participants.** The participants in the experiment were recruited mainly from the *Waisda?* online community and MBH series fanbase by distributing a call for participation through the major social networking services Facebook[6] and Twitter[7]. The posted messages and tweets contained a link to our web application. 107 participants started the experiment, 83 of them evaluated at least one fragment and 25 participants evaluated more than 5 fragments. Judging from the questionnaire data, the level of familiarity of participants with MBH series almost uniformly ranges from 'never seen it' to 'watch it regularly'. Surprisingly, the participants who never visited the MBH website or visited it only few times are the vast majority. Also for familiarity with *Waisda?*; the participants who never played or played only few times are the overwhelming majority.

**Participant's (Dis)agreement.** From the entire collection of 197 video fragments, 134 or them are evaluated by 2 distinct participants. The rest of the fragments, 63 in total, are evaluated by 3 distinct participants. When consolidating the relevance judgements from different participants we use majority voting; the side — either 'relevant' or 'not relevant' — that gets more votes wins. In case of a tie, we take the side of 'relevant' i.e. we deem the fragment to be relevant for the query. We justify this decision with the following reasoning. The notion of relevance in our particular case is defined in terms of depiction of the concept denoted by the query in the fragment. Our queries are not abstract concepts and there is very little room for different interpretations among the participants. Thus, we believe if one participant rated a query 'not relevant' and another 'relevant' for a given fragment it is most probable that the first participant simply missed it. The consolidated evaluation set is publicly available in the online appendix A[8]. The overlap among the participants in terms of evaluated videos is too small to reliably measure the inter-rater agreement with measures such as Krippendorff's alpha. However, we found that the probability of a rater rating 'relevant' is 9.5% and the probability of disagreement between raters is 10.1%.

## 5 Experiments

To answer the research questions formulated in Sect. 1 we use a quantitative system evaluation. Namely, we implement a number of search engines and run them against the evaluation dataset described in Sect. 4.2. In all experiments we evaluate the performance of the various search engines using the mean average precision (MAP) measure. The number of results returned by the systems is low enough (not more than 30) for the users to be willing to inspect them all. Thus, we deem that it is important that all results are good not just the top ones. This intuition is captured by MAP. To assess if the difference in performance is statistically significant we use the student's paired t-test at 0.01 level of significance as suggested by [2].

[6] `http://www.facebook.com/`
[7] `https://twitter.com/`
[8] All online appendixes are available at `http://tinyurl.com/9tsd47r`

### 5.1 Experiment 1

In this experiment we address the first and the second research question. To this end, we retrieve fragments for the set of queries using 12 search engines. Each of the search engines utilizes the same state-of-the-art probabilistic ranking function BM25 and the only variation among them is the data they index. Consequently, differences in retrieval performance are attributed solely to the data. We implement search engines that index:

1. $SE_{user}$ all *Waisda?* tags
2. $SE_{vuser}$ only verified *Waisda?* tags
3. $SE_{ncrv}$ all NCRV tags
4. $SE_{catalog}$ NCRV catalogue data
5. $SE_{caps}$ all captions
6. $SE_{caps+user}$ all captions and all *Waisda?* tags
7. $SE_{caps+catalog}$ all captions and all catalogue data
8. $SE_{ncrv+caps}$ all captions and all NCRV tags
9. $SE_{ncrv+user}$ all NCRV tags and all *Waisda?* tags
10. $SE^{baseline}_{ncrv+catalog}$ all NCRV tags and catalog data
11. $SE_{all-user}$ all metadata except *Waisda?* tags
12. $SE_{all}$ all metadata types including *Waisda?* tags

$SE^{baseline}_{ncrv+catalog}$ is an approximation of the search functionality offered on the web site dedicated to MBH series. We use it as a baseline for comparing the search performance of the other search engines. By comparing the performance of $SE_{user}$ and $SE_{vuser}$ we are able to see if using all tags as opposed to only verified tags is detrimental or beneficial for fragment search (*RQ2*). Furthermore, comparing the performance of $SE_{user}$ and systems 3 through 12 will reveal how well user tags are doing — on their own and in combination — compared to other types of metadata (*RQ1*).

### 5.2 Experiment 2

In this experiment we address the third research question. We retrieve fragments for the set of queries using two collections of search engines. The first collection consists of search engines that index snapshots[9] of all user tags taken periodically once a week. Identically, the second collection consists of search engines that index snapshots of the verified tags taken at the same time points as the snapshots from the first collection. As with experiment 1, all search engines use the same probabilistic ranking function BM25 and the only variation among them is the data that they index. Examining the performance of search engines within a collection reveals how tag search performance changes over time. By examining the performance of search engines across collections we learn how all tags perform compared to verified tags.

[9] A snapshot contains all user tags up to a given point in time.

## 6 Results

In this section we present the results of our experiments.

### 6.1 Experiment 1

The results for this experiment are summarized in Table 1. As seen, considering only verified tags yields worse search performance than considering all tags. Intuitively, verified tags should yield higher precision but lower recall than all tags. Indeed, the average search precision of verified tags (0.59) across the queries is higher than the average search precision of all tags (0.49). However, search based on all tags yields more relevant results — the average search recall of all tags (0.42) is higher than the averages search recall of the verified tags (0.28). In fact, for 36 queries the non-verified tags yielded relevant results —on average 4—not found by verified tags[10]. It seems the *tag verification* criterion is too conservative in a sense that it filters out tags that are in fact useful for search.

Search based on user tags ($SE_{user}$) significantly outperforms search based on other metadata types alone. Indeed, search based on user tags is approximately 69% more successful than search based on the in-house NCRV tags ($SE_{ncrv}$). We believe this is attributed to the fact that NCRV tags are relatively scarce and cover mainly prevalent topics. In this sense, user tags are complementary to the NCRV tags and the combination of both is mutually beneficial. Indeed, the search engine that indexes both user tags and NCRV tags, $S_{ncrv+user}$, yields a performance increase of 20% and 90% over search engines $S_{user}$ and $S_{ncrv}$, respectively.

Furthermore, search based on solely on user tags yields better performance from our baseline search engine, $SE^{baseline}_{ncrv+catalog}$. Indeed, the MAP scores of $S_{user}$ and $SE^{baseline}_{ncrv+catalog}$ indicate a performance increase of 46%.

Comparison of the MAP scores of $SE_{user}$ and $SE_{caps}$ indicates that user tags outperform captions by approximately 39%. This can be explained by the fact that captions only cover the audio portion of the video content, whereas user tags cover both audio and visual. In fact, previous work [3] suggested that players tend to describe more things that appear visually in a video. Combination of captions and user tags proves to be beneficial: $SE_{caps+user}$ outperforms $SE_{caps}$ and $S_{user}$ by 64% and 13%, respectively.

Lastly, the search engine that indexes all available types of metadata, $SE_{all}$, performs best. This is to a large extend due to the contribution of user tags. Indeed, $SE_{all}$ outperforms the search engine that indexes all metadata types except for user tags, $SE_{all-user}$, by 33%. Obviously, the said difference can only be attributed to the effect of the user tags. Interestingly, search based on user tags alone outperforms by 5% $SE_{all-user}$, which is the best performing search engine that does not index user tags.

[10] More detailed figures can be found the online appendix B. We omit them here due to lack of space.

**Table 1.** Results for experiment 1: MAP scores for the search engines — MAP score for given search engine is given immediately bellow. ↑, ↓, and ≈ indicate if a score is significantly better, worse, or statistically indistinguishable from the MAP scores of $SE_{user}$ and $SE_{vuser}$, in that order.

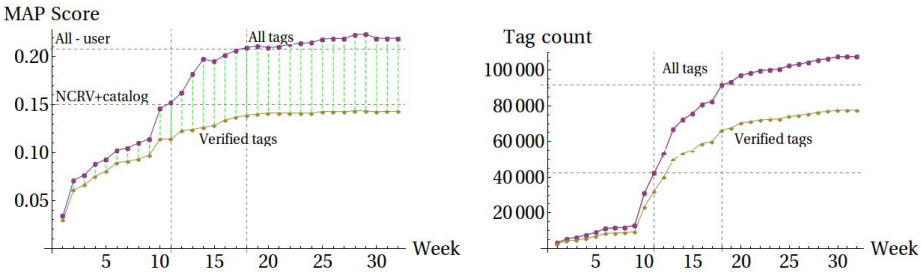
| $SE_{user}$ | $SE_{vuser}$ | $SE_{ncrv}$ | $SE_{catalog}$ | $SE_{caps}$ | $SE_{caps+user}$ |
|---|---|---|---|---|---|
| $0.219^{\approx\uparrow}$ | $0.143^{\downarrow\approx}$ | $0.138^{\downarrow\downarrow}$ | $0.077^{\downarrow\downarrow}$ | $0.157^{\downarrow\uparrow}$ | $0.247^{\uparrow\uparrow}$ |

| $SE_{caps+catalog}$ | $SE_{ncrv+caps}$ | $SE_{ncrv+user}$ | $SE^{baseline}_{ncrv+catalog}$ | $SE_{all-user}$ | $SE_{all}$ |
|---|---|---|---|---|---|
| $0.183^{\downarrow\uparrow}$ | $0.201^{\downarrow\uparrow}$ | $0.263^{\uparrow\uparrow}$ | $0.150^{\downarrow\uparrow}$ | $0.208^{\downarrow\uparrow}$ | $0.276^{\uparrow\uparrow}$ |

### 6.2 Experiment 2

In this section we present the results from our second experiment which addresses the third research question. Figure 3(a) shows the MAP scores of the search engines indexing the weekly snapshots of all tags and only the verified tags. Figure 3(b), on the other hand, shows how the number of all tags and verified tags increased over time. Looking at Fig. 3(a) we conclude that most of the time the search performance for both the verified tags and all tags is monotonically increasing with the number of tags. In other words, the more tags we amass, the better our effectiveness in searching fragments becomes. Furthermore, looking at the pairwise search performance differences between the search engines that index the weekly snapshots of all tags and verified tags (vertical dashed lines between plots in Fig. 3(a)), we conclude that using all tags for search opposed to only verified ones yields consistently better results. In fact, search performance improvements are statistically significant for every single pair.

The performance of search based on all user tags surpasses our baseline, $SE^{baseline}_{ncrv+catalog}$, around the $11^{th}$ week after 42,271 tags have been collected (Fig. 3(b)). Beyond that point the said difference in performance steadily increases as more tags are collected. With $SE_{all-user}$, which is the best performing search




(a) The MAP scores of the user tags over time. Horizontal lines represent MAP scores of $SE^{baseline}_{ncrv+catalog}$ and $SE_{all-user}$

(b) The total number of tags over time

**Fig. 3.** MAP scores and tag count over time

engine that does not index user tags, this happens a bit later. In particular, after the $18^{th}$ week and 91,508 collected tags, $SE_{user}$ starts to outperform $SE_{all-user}$. Thus, there is a point somewhere between the $18^{th}$ and $19^{th}$ week when the collected user tags outperform all search engines that do not index tags.

It is also interesting to note that the precision and recall of search based on all tags are monotonically non-decreasing with the number of tags for each query in our set. We did not include the actual figures and numbers in this paper due to lack of space. However, the results for search precision and recall can be found in online appendices C and D, respectively.

## 7 Conclusions and Future Work

In this paper we have studied the added value of user tags for video search. For this reason we have created a publicly available evaluation dataset that consists of real-life user queries, a video fragment collection, and relevance judgements.

Search based solely on user tags outperforms search based on other types of metadata such as in-house (NCRV) tags or captions. Thus if any of the other metadata types are unavailable or costly to acquire, relying only on sufficient user tags for search could yield equal or even better results. In our dataset, combining user tags with other metadata types is beneficial for search. In fact, the search engine that exploits all available metadata performs best, to large part due to the contribution of the user tags—the observed performance improvement is 33%.

Exploiting only verified user tags for search gives poorer performance than search based on all user tags. While search based on verified tags yields higher precision, it also has lower recall compared to all user tags. In fact, for most of the queries non-verified tags provided relevant results that were not found by the verified tags. This proves that considering only verified tags is too conservative filtering criterion resulting in discarding non-verified user tags that are valid video descriptors and thus useful for search.

Search performance steadily increases as more user tags are collected. This is true for both verified and all tags. Moreover, search based on all tags consistently outperforms search based only on verified tags. When the average number of tags is slightly more than 2 tags per second, the search using all tags outperforms all search engines that are not indexing user tags. Such an estimate could be used as an indicator whether a video has been tagged enough.

In the future, we will study whether certain tag features such as reputation of the tag author and provenance can be used to detect and exclude non-useful non-verified tags thereby increasing the search precision without sacrificing the recall.

**Acknowledgements.** We thank Q42 and Johan Oomen, Maarten Brinkerink, Lotte Belice Baltussen and Erwin Verbruggen from the Netherlands Institute for Sound and Vision for running the Waisda pilots, Carole Grootenboer from NCRV for collecting the query logs and the video metadata.

This research was partially supported by the PrestoPRIME project, funded by the European Commission under ICT FP7 Contract 231161.

## References


1. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3(4), 333–389 (2009)
2. Smucker, M., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proc. of CIKM, pp. 623–632 (2007)
3. Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., Aroyo., L.: On the role of user-generated metadata in audio visual collections. In: Proc. of K-CAP, pp. 145–152 (2011)
4. Voorhees, E.M.: The Philosophy of Information Retrieval Evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 355–370. Springer, Heidelberg (2002)
5. Yuen, J., Russell, B., Liu, C., Torralba, A.: LabelMe video: Building a Video Database with Human Annotations (2009)
6. Vondrick, C., Ramanan, D., Patterson, D.: Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 610–623. Springer, Heidelberg (2010)
7. Soleymani, M.: Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In: Proc. of ACM SIGIR (2010)
8. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proc. of SIGCHI, pp. 319–326 (2004)
9. van Zwol, R., Garcia, L., Ramirez, G., Sigurbjornsson, B., Labad, M.: Video tag game. In: Proc. of WWW (April 2008)
10. Kazai, G.: In Search of Quality in Crowdsourcing for Search Engine Evaluation. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 165–176. Springer, Heidelberg (2011)
11. Alonso, O., Baeza-Yates, R.: Design and Implementation of Relevance Assessments Using Crowdsourcing. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 153–164. Springer, Heidelberg (2011)
12. Eickhoff, C., Harris, C.G., de Vries, A.P., Srinivasan, P.: Quality through Flow and Immersion:Gamifying Crowdsourced Relevance Assessments. In: Proc. of SIGIR 2012 (2012)
13. Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. Information Processing & Management. Elsevier (2008)
14. Geisler, G., Burns, S.: Tagging video: Conventions and strategies of the YouTube community. In: Proc. of JCDL (2007)
15. Hildebrand, M., van Ossenbruggen, J.: Linking User Generated Video Annotations to the Web of Data. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 693–704. Springer, Heidelberg (2012)