Human Computation (2017) 4:1:47-70

ISSN: 2330-8001, DOI: 10.15346/hc.v4i1.3

Topical Video Search: Analysing Video Concept Annotation through Crowdsourcing Games

RISTE GLIGOROV, VU UNIVERSITY AMSTERDAM

MICHIEL HILDEBRAND, VU UNIVERSITY AMSTERDAM

JACCO VAN OSSENBRUGGEN, VU UNIVERSITY AMSTERDAM

LORA AROYO, VU UNIVERSITY AMSTERDAM

GUUS SCHREIBER, VU UNIVERSITY AMSTERDAM

ABSTRACT

Games with a purpose (GWAPs) are increasingly used in audio-visual collections as a mechanism for annotating videos through tagging. One such GWAP is *Waisda?*, a video labeling game where players tag streaming video and win points by reaching consensus on tags with other players. The open-ended and unconstrained manner of tagging in the fast-paced setting of the game has fundamental impact on the resulting tags. Consequently, the *Waisda?* tags predominately describe visual objects and rarely refer to the topics of the videos. In a previous study [ECIR 2013, 50-61], Gligorov at al. showed that the *Waisda?* tags are effective in finding video segments that depict a specific entity (person, object, etc.) of interest. This study evaluates the performance of the game tags for retrieval of videos that are about a given topic. To this end, we setup a Cranfield-style experiment for which we use an evaluation dataset that consists of: (i) a collection of videos tagged in *Waisda?*, (ii) a set of queries derived from real-life query logs, and (iii) relevance judgements. While we reuse the collection of videos and the set of queries compiled by Gligorov at al., we designed the set of relevance judgments specifically for this study. The novelty aspects of this paper are as follows. Our results demonstrate that the raw, unprocessed game tags are not well suited for retrieving videos based on topic. We perform a qualitative analysis of the search results which reveals that this is mainly caused by the presence of general tags which refer to visual objects unrelated to the topics of the video. Thus, we characterize the quality of the game tags as topical annotations in order to detect and filter out the non-topical ones. We explore several features of the game tags which could serve as an indication of their quality as topical descriptors. Our results show that after filtering, the game tags perform equally well compared to the manually crafted metadata when it comes to accessing the videos based on topic. An important consequence of this finding is that tagging games can provide a cost-effective alternative in situations when manual annotation by professionals is too costly.

1. INTRODUCTION

In the past decade, audio-visual (AV) content collections have been undergoing a transformation from archives of analogue materials to very large stores of digital data accessible online[1]. As the AV collection items become accessible to the wider internet audience the lack of adequate annotations is highlighted: users cannot find what they are looking for because annotations are either not present or too expert-centric — created from the perspective of the catalogers (Oomen et al., 2009). Video tagging games are an attempt to alleviate this problem. By engaging the internet community to tag their videos, the AV collection owners can benefit in at least two important ways. First, there is a clear cost-benefit, collecting annotations in this way is usually cheaper than using professional documentalists. Second, the game tags can help bridge the *terminological gap* in search if the *searchers* and *annotators* originate from the same community and use similar terminology when searching and tagging.

A successful online video tagging game is *Waisda?* (Hildebrand et al., 2013). The game was launched in 2009 by the Netherlands Institute for Sound and Vision (S&V), one of the largest AV archives in Europe[2]. *Waisda?* is an ESP game (von Ahn et al., 2003) applied to audio-visual content. It is a multi-player 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. The first pilot of the game run until January 2010 and produced over 420,000 user tags. Considering the acquired experience and insights, S&V launched a second version of the game[3] which features several improvements, albeit the basic idea of the temporal tag agreement is preserved. Figure 1 shows the game page of *Waisda?*. It contains a video player and below it a text-entry field. When the player enters the game page the video automatically starts playing, the text-entry field receives focus, and the player can start entering tags. The right side of the page contains the score board. It consists of the current score of the player, the current rank and a listing with all tags entered by the player.

The second version of *Waisda?*[4] is a mature, production-grade, customizible GWAP with diverse

[1] A good example is PrestoPRIME, http://www.prestoprime.org/, which was an European Union funded project about digitisation and preservation of the AV heritage. The consortium included the major national AV archives in Europe.

[2] S&V archives over 70% of the Dutch AV heritage. The collection contains more than 750.000 hours of television, radio, music and film from the beginning in 1898 until today.

[3] The source code is published as open-source under the GPL licence and can be downloaded from https://github.com/beeldengeluid/waisda.

[4] At the time of writing the second version of the game has been discontinued. S&V is deploying (still under development) a third installment of the game available at http://waisda.beeldengeluid.nl/. S&V intends to integrate *Waisda?* more tightly in their internal workflows and to use it to collect tags for the items in their online collection http://in.beeldengeluid.nl/.

usage. The second version of *Waisda?* is included[5] as a showcase in the Europeana[6] internet portal. *Waisda?* has also been deployed and incorporated in Spotvogel[7]. In the former the internet users tag archival footage from the European cultural heritage and in the latter they identify occurrences of wildlife in footage. These are two relatively different domains yet *Waisda?* was fitted to them seamlessly. This is due to the simple design of the game and the unconstrained way the tags are entered — the players get a small set of instructions and then are free to enter whatever they please. This is a double-edged sword. On the one hand, the entry barrier for the players is low and the target audience is wide as no specific skills are needed. On the other hand, the open-ended way the tags are entered may be overwhelming in the fast-pased setting of the game and players may succumb to entering low quality tags which lack descriptive power just to reach consensus with other players and win points. In this study we perform a qualitative analysis to investigate this issue.

The cost-effectiveness of *Waisda?* versus manual annotation by professionals comes with a caveat: successful deployment of a GWAP is no small feat. Attracting new players and keeping them engaged over time is vital for success and requires continuous publicising efforts. The empirical evidence collected from both *Waisda?* pilots supports this claim. On both occasions the bulk of the tags was accumulated during periods when there was an active campaign for promoting the game. Targeting the fanbase of the TV series via various channels (e.g. TV series website, social media, etc.) has proved to be a successful strategy for attracting new players. Player's engagement is sustained with in-game mechanisms such as leaderboards to honour the best players and motivate the others, and with time-limited contests offering awards for the top performers. Planning and executing activities like these is certainly costly, however in the case where the growth rate of AV collections is ever increasing, these costs will be outweighed by the costs of manual annotation. This is one of the lessons learned from the *Waisda?* project.

The second version of the game has amassed more than 710,000 user tags. The number of unique tags is 71,448 and each of the top five most-tagged videos has more than 3,600 tags ascribed to it which amounts to an average tag density of more than 13 tags per second of video time. This is a substantially higher number than the number of tags assigned by professional catalogers, typically 10-20 tags per video[8]. However, are all these user tags of any use? Does this overwhelming quantity implies quality? The expectation of S&V is that the tags collected with *Waisda?* will improve video search especially since the tags are integrated in the S&V's internal workflows. A previous study demonstrated that the *Waisda?* tags can indeed be successfully exploited for retrieving video fragments that feature *visual appearances* of objects of interest (persons, objects, etc.) (Gligorov

[5]http://labs.europeana.eu/apps/waisda-floss

[6]Europeana is an internet portal that provides multi-lingual access to millions of books, paintings, films, and archival records that are part of the European cultural heritage. More than 2,000 institutions across Europe have contributed to the project.

[7]*Spotvogel* (http://spotvogel.vroegevogels.vara.nl/) which translates to *mocking bird* from Dutch, was a *Waisda?* deployment by the Dutch broadcaster VARA (http://www.vara.nl/) and the S&V institute with the aim of collecting user tags for the footage from the Dutch television program *Vroege Vogels* (http://vroegevogels.vara.nl/). The game run for 6 months in 2013 and was nominated for Dutch Game Awards 2013 http://www.dutchgameawards.nl/2013/spotvogel/, note that the link is in Dutch.

[8]The average and median length of the videos is 2.9 and 3.2 minutes, respectively. Mode detail is provided in section 4.2.2.

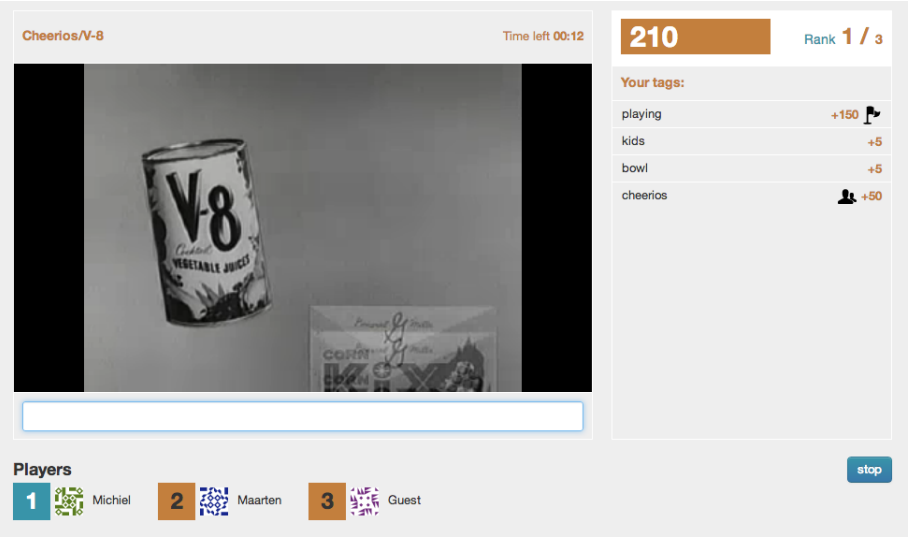

Figure 1. ***Screenshot of the Waisda? game page with public domain video clip from Prelinger Archives.***

et al., 2013). The results showed that *Waisda?* tags excel at this kind of *visual instance search*, outperforming the closed captions and annotations created by professionals.

In this study we wish to understand whether the fast pace of *Waisda?* forces players to have tunnel vision about the content of the video and tag predominately non-topical aspects. Previous analysis performed by a senior cataloger from S&V ruled the game tags to be limited in scope, referring mainly to things seen or heard on the screen (Baltussen, 2009). However, we hypothesize that players do describe topics and topical tags are present albeit buried under the myriad of non-topical tags. The first research question, therefore is

RQ1 Do players enter tags which describe the topics of the videos?

In the context of topical video search, the presence of non-topical tags, which do not refer to the topics covered in the video, can lead to false positives. For example, if a video v has a tag t ascribed to it and t does not refer to the topics covered in v. If a searcher is interested in videos that are about t and uses t as a search term the video v will be incorrectly retrieved. Thus, it is important to know whether a tag is non-topical and ignore it in the retrieval process to eliminate its negative influence on the search results. Our second research question, therefore is

RQ2 Can the access to videos based on topic be improved by detecting and filtering out the non-topical game tags?

The rest of the paper is structured as follows. In Section 2 we discuss related work. Section 3 outlines our approach. In Section 4.1 and 4.2 we describe the collection of video fragments and the evaluation dataset, respectively. Section 5 presents the findings with respect to the effectiveness of the user tags for topical search. This sets the baseline for the remainder of the study. In Section 6

we outline a number of tag filters and evaluate their effectiveness for eliminating non-topical tags. Section 7 presents the conclusions from the study.

2. RELATED WORK

User annotations for search. Search based on user-generated metadata, in particular folksonomies, has been studied before. Morrison compared Web search performance of folksonomies from social bookmarking Web sites against search engines and subject directories (Morrison, 2008), showing that search engines had the highest precision and recall rates. Folksonomies, however, performed surprisingly well. In fact, user tags show promise to alleviate the vocabulary mismatch problem for search: bridging the gap between user queries and metadata used for retrieval (Furnas et al., 1987). Indeed, Geisler and Burns state that YouTube tags provide added value for search, because 66% of them do not appear in other metadata (Geisler et al., 2007). Heymann et al. investigated a large-scale sample of forty million bookmarks from the social bookmarking site del.icio.us and found that in 20% of the cases user tags do not occur in the page text, backlink page text, or forward link page text of the pages they annotate. Studies in (Bischoff et al., 2004a; Halvey et al., 2007; Rorissa, 2010; Yanbe et al., 2007) investigate this phenomenon across multiple domains and multimedia resource types and identify the gaps between the tag space and the querying vocabulary. The common conclusion is that user tags can improve search by bridging the vocabulary gap.

Studies reported in (Bischoff et al., 2004b; Sun et al., 2010; Marshall, 2009) take a more critical stance. They conclude that while overall user tags improve search, not all tags are suitable for retrieval. In fact, Marshall (Marshall, 2009) even suggests that tags may be less effective descriptors for image retrieval, classification, and description than other forms for descriptive metadata such as title and narrative captions. This hints that a characterization of the quality of the tags is needed to filter out the tags that are not suited for retrieval. This is one of the aspects that our study addresses.

Another line of research is exploiting the tripartite structure ($Users \times Tags \times Resources$) of folksonomies to improve search (Hotho et al., 2006; Bao et al., 2007). Alternatively, the semantics of tags can be grounded in some lexical sources and the grounded tags utilized for improving search. For example, 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 (Hildebrand et al., 2012).

Quality and Refinement of Annotations. There is a substantial body of research into the refinement and quality assessment of annotations of still images. Lee at al. propose a tag refinement technique that aims at differentiating noisy tag assignments from correct tag assignments (Lee et al., 2010). Each tag is assigned a probability of being noisy based on the visual similarity of the images and tag co-occurrence statistics. Tags with a probability above a threshold are discarded as noisy.

In (Truong et al., 2012; Li, Snoek, & Worring, 2012) neighbour voting schemes for determining the tag relevance are explored. In this approach, a tag is considered more relevant to the image it is ascribed to, also known as the seed image, if the tag is also used to annotate the neighbouring images. The neighbourhood relation is defined in terms of the visual similarity among images. Lee at al. expand the approach by not only considering the visually similar images, but the dissimilar images as well, thus providing negative examples (Lee et al., 2012). Kennedy at al. exploits visual similarity among images in a sense that tags ascribed to images by the creators of the image are

used as seed annotations and also attached to visually similar images (Kennedy et al., 2009). Zhao at al. propose a data-driven method to automatically determine the relatedness between a tag and the image's visual content taking into consideration the tag co-occurrence and the visual similarity among images (Zhao et al., 2010).

Probabilistic methods that exploit random walk based techniques have also been explored (Wang et al., 2006; Liu et al., 2009; Li, Tang, Li & Zhao, 2012). These methods produce a ranking of the tags according to their relevance with respect to the image with which they are associated. The tag relevance estimations are computed as the stationary or the convergence probabilities of a random walk processes. Notable example of this approach is the PageRank algorithm (Ding et al., 2009; Koschützki et al., 2008; Junker et al., 2008). Another group of methods exploit background knowledge (such as the lexical database Wordnet and a massive corpus indexed by Google) to perform the refinement of the image annotations (Jin et al., 2010; Wang et al., 2007). The semantic relations encoded in Wordnet and the semantic similarity quantified by the Google-based measures like the Normalized Google Distance (Cilibrasi et al., 2007) provide contextual evidence for the relationship among the annotations. This evidence is then used as an input for machine learning algorithms which give the final word for the quality of the annotations.

Games With a Purpose (GWAPs) GWAPs are a human-based computation technique in which humans solve tasks, too difficult for computers, in a game setup which provides entertainment for the players. The predecessor of all GWAPs is the ESP game designed by von Ahn and Dabbish (von Ahn et al., 2003), which harnesses human abilities to label images. The general idea of the ESP game is that two players who can see the same image try to come up with matching tags. The players are paired up randomly without any means of communication. Therefore, when players agree on a tag it is a strong indication that the tag is a valid descriptor of the image. In (von Ahn et al., 2008) von Ahn and Dabbish generalize the design principles of the ESP game into a conceptual framework for designing GWAPs. Given a problem which is hard or impossible to solve by computer the framework provides high-level guidelines and principles for transforming the problem into GWAP that solves it. Following the ESP game, many GWAPs were designed that tackle problems from various domains. *Peekaboom* is a GWAP in which players identify and locate objects in images (von Ahn et al., 2006). The output of the game is precise location information and other useful information which can be used to train computer vision algorithms. Several GWAPs have been designed for collecting or validation of common sense knowledge. *Verbosity* is a GWAP (von Ahn, Kedia et al., 2006) which addresses the problem of creating a database of 'common-sense facts', statements about the world known to most people. Similarly, *Common Consensus* is a GWAP that aims at collecting a database of human *goals* (Havasi et al., 2007). *Top10* and *Pirate & Ghost*(Chang et al., 2010), on the other hand, are GWAPs that deal with the problem of verification of the common sense assertions of the form *concept* $\rightarrow$ *relation* $\rightarrow$ *concept*. Another area where GWAPs are applied is collecting visual data for real world locations. *EyeSpy* is a GWAP in which players contribute photos or textual data about geographic locations (Bell et al., 2009). The collected data is subjected to in-game validation by other players and once validated can be used for navigation tasks. *PhotoCity* is another GWAP where players contribute photos of real world locations in urban areas (Tuite et al., 2010, 2011). The ultimate goal is to use the collected photos to build detailed 3D models of real world places varying over time. GWAPs have also been applied in the music and art domains. *TagATune* (Law et al., 2007, 2009) and *Listen Game* (Turnbull et al., 2007) are

GWAPs where the aim is to collect annotations for music pieces. *Artigo*[9] is a game platform that offers six GWAPS for annotating artworks (Wieser et al., 2013). Three of the games *Artigo game*, *Artigo Taboo*, and *TagATag* are variations of the ESP game. *TagATag* includes a more challenging aspect where players are tagging pairs consisting of an image and a tag. The resulting tags describe relationships between the members of the pair(the image and the tag). The other three game of the platform, *Karido*, *Artigo-Quiz*, and *Combino* are designed to complement the data collected with the first three ESP-like games. In *Karido* (Steinmayr et al., 2011) the process of annotation is carried out as a guessing game where one player is trying to guess *goal image* out of set of images. Second player tries to help the first player by providing description of the image. If the first player correctly guesses the goal image then the description is valid annotation with high probability. The aim of the *Combino* game is to collect semantically more complex multi-word tags (Stoerkle, 2012). Semantic web is another area for application of GWAPs. *OntoGame* is a series of GWAPs that aims to cover the complete Semantic Web life-cycle: building and maintaining ontologies, alignment of ontologies, and semantic annotation of data (Siorpaes, Hepp, 2008). Building and maintaining ontologies is carried out with the *OntoPronto* GWAP (Thaler et al., 2012), alignment of ontologies is achieved with the *SpotTheLink* GWAP (Thaler et al., 2011), and semantic annotation of resources is done with the *OntoTube* and the *OntoBay* GWAPs. *GiveALink Slider* and *Great Minds Think Alike* are two more GWAPs that address the problem of collecting reliable semantic annotations (Weng et al., 2011). Another area where GWAPs are applied is linguistics. *Jinx* is a GWAP that tackles the problem of word sense disambiguation (Seemakurty et al., 2010). *Borsa Parole* and *Poker parole* are GWAPs that aim at collecting linguistic data (François et al., 2013). In *Borsa Parole* word phrases and their characteristics are collected from the user community. *Poker parole* aims at collecting meta-data about the data collected in *Borsa Parole* i.e. players form conjectures about the word phrases and their characteristics. Other notable GWAPs are *Odd Leaf Out* which addresses the problem of misclassified leaf images (Hansen et al., 2011) and *Polarity* which deals with collecting attributes and attribute values for resources (Law, Settles et al., 2011). Pearl and Steyvers in (Pearl, Steyvers., 2010) outline a GWAP-based methodology for identifying emotions, intentions, and attitudes in text. Kneissl in (Kneißl, 2014) studies how GWAPs and more generally crowdsourcing can be used in the field of e-learning.

Important area of research is on the limitations of ESP-like games. Robertson at al. demonstrated that the ESP game in its original design encourages players to enter 'obvious' or predictable tags (Robertson et al., 2009). Weber at al. built a probabilistic language model which using only the already assigned tags for the image and without any knowledge about the image was able to predict with high probability[10] what the new tags added by players will be. Jain and Parkes provided a game-theoretic explanation for this observation (Jain et al., 2013). Their game-theoretic model of the ESP game indicated that from the players' point of view it is more beneficial if they focus on more obvious (low-effort) tags.

[9] *Artigo*, http://artigo.org, is a game platform for artwork annotation as well as artwork semantic search engine in German, English, and French founded in 2008. According to (Bry et al., 2015), thanks to a good media coverage over several years, the German version of ARTigo has a sufficient number of regular players.

[10] Even without any understanding of the actual image, the probability of agreement with randomly assigned human partner on a label was 69% for all images, and 81% for images which have at least one tag assigned to them.

3. APPROACH

To answer the research questions stated above we use a quantitative system evaluation methodology (Voorhees, 2012) which requires an evaluation dataset that consists of three components: a document collection (in our setting video fragments tagged by players in *Waisda?*), a set of representative queries, and relevance judgements. To evaluate the performance of a search system, the query set is run against the system and the retrieved results are graded w.r.t the relevance judgements[11]. The search performance or the search system is then quantified by search performance metrics such as Mean Average Precision (MAP), Recall, Precision, etc. Central to this evaluation methodology is the creation of the relevance judgements which plays the role of *gold standard*. In the making of the gold standard we exploit in-house annotations created by the broadcaster which describe the main topics in the video. More precisely, we deem a video to be relevant for a query if the query concurs with at least one the topics (annotations) of the video (see Sec. 4.2.3).

Our study is divided in two parts. In the first part we address the first research question: do players enter topical tags and in effect how suited are the game tags for topical search. To this end, we create four search systems, each exploiting different metadata. We run each system against the evaluation dataset and carry out comparative analysis of their retrieval performance. The first and the second system use all game tags and *verified*[12] only game tags, respectively. The third system is our baseline and exploits the in-house catalog metadata — it is an approximation of the present search functionality. The forth system combines all game tags and the in-house catalog metadata. By comparing the first and the second system against the baseline system, we deduce how suited the game tags for topical search are on their own, and see if they could replace the need for the current expensive cataloguing practice. By comparing the forth system with the baseline, we infer whether the game tags provide added value on top of the existing catalog metadata. To get a better qualitative insight of what happens under the hood, we carry out an analysis of example false/true positives. We classify the game tags according to the Panofsky-Shatford scheme (Hollink et al., 2004) to establish which parts of the video content are described by the tags. The aim is to derive insights about users' tagging practices that are associated with ascribing topical tags.

In the second part we address the second research question: we investigate several ways to detect and filter out non-topical tags with the goal of improving topical search performance. In particular, we take tag features such as TF-IDF score and player's reputation and derive a binary yes/no decision as to whether to retain the game tag or drop it. Each filtering method yields a subset of all game tags. To judge how well each of the filtering methods work we build a search system for each of them which uses the corresponding subset of all game tags as input for search. Moreover, to see how well the filtering methods work together with the existing catalog metadata we build a search system for each of them which uses the corresponding subset of all game tags and the catalog metadata as input for search. The search performances of these systems are compared to the baseline system from the first part of the study (see Sect. 6 for more details).

When comparing the performance metrics of any two systems, to assess whether the difference is statistically significant we use the student's paired t-test at 0.01 level of significance as suggested

[11] Relevance judgments give binary (yes/no) answer as to search result is relevant to the query. Thus, the retrieval process was fully automatized.

[12] Tags entered by at least two players within a time interval of 10 seconds.

by (Smucker et al., 2007).

4. EXPERIMENTAL DATA

In this section we describe the data and the evaluation dataset used in this study. More precisely, Section 4.1 outlines the metadata that we exploit for search and Section 4.2 describes in more detail the evaluation dataset.

4.1. The MBH Video and Metadata Collection

In the second pilot, *Waisda?* was used to tag fragments from the popular Dutch TV program 'Man Bijt Hond' (MBH, English: 'Man Bites Dog') produced by the Dutch broadcaster NCRV[13]. MBH is a humoristic TV show that focuses on trivial, everyday news and ordinary, 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 video 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 we use as input for search in this study:

***Waisda?* game 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 whether the tag was entered by at least two players within a time interval of 10 seconds. 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 catalog 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. catalog 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.

[13] http://www.ncrv.nl/

4.2. Evaluation Dataset

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

4.2.1. Query set

We will reuse the same set of queries from a previous evaluation of the *Waisda?* tags for *visual* search (Gligorov et al., 2013). Here by visual search we mean *keyword* search where the goal is to retrieve videos that visually depict the object/artifact of interest. The query set consists of fifty queries that were sampled from user query logs of the TV series' web site. The sampling procedure involved grouping the queries into three classes based on their frequency in the logs: *high*, *mid*, and *low* frequency class. Each of the three classes were then filtered to remove bias and to ensure fair comparison of the retrieval performance of the metadata types. The top-ranked 12, 19, and 19, queries, w.r.t. frequency, from the high, mid, and low frequency class, respectively, comprise the final query set. The exact details of the sampling procedure can be found in (Gligorov et al., 2013). An example of a high-frequency query is *Mandy* which is the main character in one of the rubrics in the MBH show. On the other hand, an example of a low-frequency query is *Friesland* which is a region in the Netherlands and the filming location of one of the rubrics in the MBH show.

An objection may be raised for reusing the same queries from the visual search study, where the goal is to retrieve videos that *depict the artefact* specified by the query, for topical search, where the goal is to retrieve the videos that are *about the topic* specified by the query. However, most queries can refer both to an object/artefact appearing in the video or to one of the topics the video is about. For example, consider the query *horse*. A video may depict a horse as part of the scenery and this would qualify it as a relevant result for visual search where we might be looking for stock footage containing horses. Another video may be about equines and therefore be a relevant result for topical search on the topic *horses*. In fact, the query selection procedure used in the visual search study (Gligorov et al., 2013) was oblivious about the searcher's intent behind the queries, be it visual, topical or something else. Only after the queries were selected a visual search interpretation was assigned to them and the gold standard was created accordingly. In the same manner, in this study we give the queries a topical interpretation which is embodied in the gold standard (see Sect. 4.2.3 below). Figure 2 shows the number of topically relevant videos per query as judged by the gold standard. As seen, for every query there is positive number of relevant videos which means all queries can be interpreted topically.

4.2.2. Video fragment set

The set of fragments for this particular experiment is selected from the MBH fragments tagged in *Waisda?*, described in more detail in Section 4.1. For a fair comparison of the search performance, we select a subset from the entire collection of fragments. The selection criterion is as follows: only fragments that have at least one verified *Waisda?* tag ascribed to them are considered. The resulting collection contains 2,562 fragments with accumulative duration of almost 123 hours of video material. The average fragment length is approximately 2.9 minutes and the median is 3.2 minutes. The duration of the shortest and the longest fragment in our collection is 0.1 and 25 minutes, respectively. The total number of user tags, verified user tags, and NCRV tags ascribed to

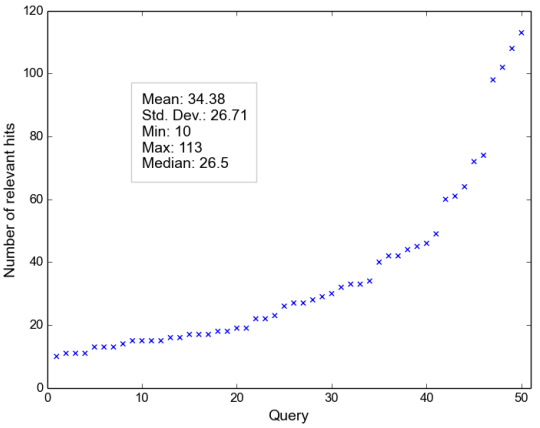

Figure 2. ***Number of relevant hits per query in increasing order.***

the videos of this collection is 591,468, 355,522, and 28,248, respectively. Thus, the average number of user tags, verified user tags, and NCRV tags per fragment is 231, 139, and 11 respectively.

4.2.3. Ground Truth

As we said earlier, NCRV tags are in-house tags that describe the topics of the fragment with which they are associated. We consider the NCRV tags as the ground truth about what topics are covered by the fragments. With this in mind, given a query q and a fragment f, we deem f to be topically relevant for q if there is an NCRV tag that is equal with q, a synonym of q, or a hypernym of q. Y is a hypernym of X if every X is a Y, for example *canine* is a hypernym of *dog*. Hyperonymy is a transitive relation (Miller, 1995). More formally, the topical relevance relation is defined as follows

$$\begin{aligned} Topical\_Relevance \quad = \quad & \{(q,f) \mid \exists t \; (t \in NCRV(f) \; \wedge \\ & (lower(q) = lower(t) \; \vee \\ & synonym(lower(q), lower(t)) \; \vee \\ & hypernym(lower(t), lower(q)))) \} \end{aligned} \tag{1}$$

where $NCRV(f)$ is the set of all NCRV tags associated with the fragment f, $lower(\cdot)$ is the lower case string function, $synonym(w_1, w_2)$ is a binary predicate which is true iff w_1 and w_2 are synonyms, and $hypernym(w_1, w_2)$ is a binary predicate which is true iff w_1 is a hypernym of w_2. Figure 2 shows the number of relevant fragments in our collection for the queries in the dataset along with additional descriptive statistics. As seen, the median and the average number of relevant hits for a query is approximately 27 and 34, respectively. The lowest number of relevant hits for a query is 10 whereas the highest is 113.

The NCRV tags were primarily created to be used for searching and browsing on the TV-show website by the online community. The tags are displayed prominently on the website's UI which enhances the *social proof* effect (Floeck et al., 2011; Golder et al., 2006), thus many of the tags are

incorporated in the terminology of online community. In fact, our chosen query set belongs to the intersection between the NCRV tags and the online community's search terminology as witnessed by Fig. 2. As implied by the definition ((1)) and seen in Fig. 2, for every query in the query set there is at least one video annotated with the same NCRV tag. This is sufficient for our narrower aim to determine the relevance for this particular query set and not for *any* query in general. In other words, the mismatch between the NCRV tags and online community's terminology has no impact on the derived relevance judgements for our query set.

Note that while the NCRV tags are currently used by the broadcaster on the TV-show website, there is no guarantee that they, when used as described above, form a complete and accurate ground truth for topical search. We choose this set-up because the alternative scenario, creating a dedicated topical search ground truth for a limited subset ourselves, would have resulted in a much smaller data set.

We considered variation of definition (1) where we defined the topical relevance relation only by considering case-insensitive string comparison (omitting synonyms and hyponyms). Even with the modified definition the general conclusions from Sections 5.1 and 6 below remained the same; while the retrieval metrics of the systems (given by Tables 1 and 5) varied slightly the ordering did not change and the differences remained statistically significant.

5. EVALUATION OF THE QUALITY OF USER TAGS FOR TOPICAL VIDEO SEARCH

In this section we will evaluate the effectiveness of the *Waisda?* tags for topical video search. The results obtained here will serve as a starting point for comparison in the subsequent sections where we will try to improve the retrieval effectiveness of the *Waisda?* tags by filtering those that are irrelevant as topical descriptors.

5.1. Game tags vs. catalog data

To address the first research question stated above we created four search engines. Each of them utilizes the same state-of-the-art probabilistic ranking function BM25 and the only variation among them is the metadata they index and use as input for search. Consequently, differences in retrieval performance are attributed solely to the data. We evaluted search engines that index:

1. SE_{user} all *Waisda?* tags
2. SE_{vuser} only verified *Waisda?* tags
3. $SE_{catalog}$ NCRV catalog data
4. $SE_{user+catalog}$ catalog data and all *Waisda?* tags

We did not consider using captions for search because we only have them available for a small subset of MBH fragments. If we would have used them in the study, we would have had to settle for much smaller evaluation collection of fragments. Naturally, we did not use the NCRV tags either, since we used them as ground truth for topical relevance.

As said, the combinations of metadata types that are indexed by the various systems are strategically chosen so that the resulting performance metrics from the evaluation dataset will provide answers to the first research question. We compare the performance of SE_{user} against $SE_{catalog}$ to evaluate the

System	MAP	Precision	Recall
SE_{user}	$0.131^{\approx\uparrow}$	$0.16^{\approx\downarrow}$	$0.589^{\approx\uparrow}$
SE_{vuser}	$0.081^{\downarrow\approx}$	$0.193^{\uparrow\approx}$	$0.286^{\downarrow\approx}$
$SE_{catalog}$	$0.168^{\uparrow\uparrow}$	$0.438^{\uparrow\uparrow}$	$0.291^{\downarrow\uparrow}$
$SE_{user+catalog}$	$0.151^{\uparrow\uparrow}$	$0.17^{\uparrow\downarrow}$	$0.654^{\uparrow\uparrow}$

***Table 1.** MAP/Precision/Recall scores for the search engines. ↑, ↓, 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.*

performance of the game tags alone when compared to the catalog metadata. By comparing SE_{user} against $SE_{user+catalog}$ we see if the performance of the catalog metadata can be further improved by adding the game tags. Furthermore, we check whether the performance of the game tags can be improved by just using the verified tags (SE_{user} versus SE_{vuser}).

5.2. Results

Table 1 summarizes our findings. The search performance of game tags (SE_{user}) is 28% below that of the NCRV catalog data ($SE_{catalog}$). Combining the game tags with the catalog metadata does not help either: $SE_{catalog}$ is better than $SE_{user+catalog}$ by 11%. The good news is that SE_{user} outperforms $SE_{catalog}$ on recall: the poor MAP score is largely due to the low precision, only 0.16.

The results are even worse for the verified user tags. While they yield only marginally better search precision, their recall is far worse (see Table 1). Consequently, SE_{user} outperforms SE_{vuser} by 62% which suggests that considering all tags is better for topical search than limiting the scope only to verified tags.

In conclusion, the good recall of the game tags does not outweight the low precision. The latter is caused by *Waisda?* tags that match the query but are associated with fragments that are not considered topically relevant. Below, in Section 6, we attempt to detect and filter out these tags to improve the search performance.

5.2.1. A Closer Look on Verified Tags

The results presented in the previous section suggest that verified tags leave much to be desired when it comes to topical search. To get a better qualitative insight of what happens behind the scenes we carry out an analysis of samples of the results returned by the system that indexes the verified *Waisda?* tags. Our hypothesis is that the tags referring to objects that are visually depicted in the videos are the leading reason for the poor retrieval performance. To test this we analyse a subset of the *false positives* and a subset the *true positives* returned by SE_{vuser}. For each query we consider all returned videos— either true or false positive — for which captions are available. The subject of the analysis is the tag that caused a given video to be returned for a given query.

First, we classify each tag in the samples into three categories based on the content component — audio, visual, or both — it refers to (note that the classification was carried out by a single rater). Tags that refer to concepts that are visually depicted but are not mentioned in the dialog are classified

	Number	Average TF-IDF
Only visual	361 (67%)	22.302
Visual and in captions	19 (3%)	55.793
Only in captions	161 (30%)	49.319

Table 2. False positives analysis.

as *only visual*. Tags that are mentioned only in the dialog are classified as *only in captions*. The third category contains the tags that refer to concepts that are both visually depicted and present in the dialog. Second, for each of the categories we compute the average TF-IDF (Term Frequency - Inverse Document Frequency) for the tags in that category. TF-IDF is a numerical statistic which reflects how important a term is to a document in a collection or corpus (Jones, 1972; Salton et al., 1988). In our context, the TF-IDF score of a tag reflects the relevance of the tag for the associated fragment where the fragments are represented as bags of all tags ascribed to them. The average TF-IDF score of a category is an indicator of the average relevance of its members relative to the other categories. Lastly, to establish what aspects of the video content the tags from our samples are describing we classify them according to the Panofsky-Shatford model (Hollink et al., 2004). This model divides the descriptions into three levels: *general* (generic things in the video), *specific* (specific things), and *abstract* (symbolic things). Each of the levels is further broken down into four facets: *who*, *what*, *where*, and *when* producing the Panofsky-Shatford 3x4 matrix. There are alternative tag classification schemes (Sigurbjörnsson et al., 2008), however we picked Panofsky-Shatford since it provides insight about the relation of the tag and video content it describes; the role of the concept, denoted by the tag, in the video content is captured by the model, e.g. if the tag *Amsterdam* is classified in *Where Specific* this would signify that this is the location of the scene in the video.

Sampling procedure. The tags for our analysis are selected as follows. For each query we consider all returned videos for which captions are available. The sample of tags consists of all verified tags ascribed to these videos. The reason we restrict only to videos with captions is that in the course of the analysis we check for presence in captions, as described above. The results of the analysis are given in continuation.

5.2.2. False positives

We start by analysing the false positives returned by the system that indexes the verified tags, SE_{vuser}. The results of the analysis are summarized in Table 2. The total number of analysed instances is 541. As suspected, the majority of the false positives, about 67%, are caused by tags referring to concepts that are only visually depicted. Around 30% of the false positives are caused by tags that appear only in the audio component of the content. The remaining 3% are caused by tags present both in audio and visual part of the content. Interestingly, it seems that the tags which are present in the audio and refer to a concept that appears visually are less likely to yield false positives. Their presence in both the audio and visual component is a strong indication that they are denoting salient aspects of the content. This is witnessed even more by the fact that this category has the highest average TF-IDF score which measures the importance of a term for a document. In our case the "document" is the bag of tags associated with the fragment.

	Number	Average TF-IDF
Only visual	20 (27%)	50.195
Visual and in captions	33 (45%)	96.182
Only in captions	21 (28%)	60.185

Table 3. True positives analysis.

	True positives			**False positives**		
	Abstract	General	Specific	Abstract	General	Specific
Who		7	20		31	1
What	4	34		37	377	
Where		3	6		63	68
When						1

Table 4. Classification of positives.

5.2.3. True positives

In this section we continue our analysis with the true positives returned by the system SE_{vuser}. The results are summarised in Table 3. The total number of analysed instances is 74. The figures are following the same trend as the figures for the false positives analysis documented in Table 2. The tags that are present in the captions and represent concepts depicted visually make the category that yields the highest number of true positives, 45%. Around 28% of the true positives are result of tags which are present only in the captions. The remaining 27% are yielded by tags that denote concepts depicted only visually. Again, the category of tags found both in the audio and visual component have the highest TF-IDF score.

We also classified the positives using the Panofsky-Shatford model (Hollink et al., 2004). More precisely, we classified the tags that led to the hit with respect to the returned video fragment. The results from the classification are summarized in Table 4. It is interesting to note the disproportionally large number of false positives compared to the number of true positives for the *What* and *Where* facet. In fact, the set of false positives in the *What* facet significantly overlaps with the *only visual* false positives set from Table 2. For most part these are objects appearing in the foreground and the background of the scenery. The case of the *Where* facet is the more interesting one. The false positives in the *General Where* facet are mostly caused by tags which refer to the dialog in situations where the actors are talking about generic places e.g. their current whereabouts like the *farm*. The false positives in the *Specific Where* facet almost all originate from one query, namely *Amsterdam*. In fact, our collection features a series of videos where a TV crew from Amsterdam travels to other places in the Netherlands and interviews ordinary people. At the beginning of every interview the crew presents themselves at which point they mention that they come from Amsterdam. It is a signature motif in the series and is usually picked up by the players in *Waisda?*.

After analysing the true and false positives, the general conclusion is that the verified tags usually refer to the more “obvious", more noticeable, aspects of the content which is in agreement with the conclusions from (Robertson et al., 2009; Jain et al., 2013). For example, moving objects in the

	System	MAP	Precision	Recall
Baseline	$SE_{catalog}$	$0.168^{\approx\uparrow\uparrow}$	$0.438^{\approx\uparrow\uparrow}$	$0.291^{\approx\downarrow\uparrow}$
	SE_{user}	$0.131^{\downarrow\approx\uparrow}$	$0.16^{\downarrow\approx\downarrow}$	$0.589^{\uparrow\approx\uparrow}$
	SE_{vuser}	$0.081^{\downarrow\downarrow\approx}$	$0.193^{\downarrow\uparrow\approx}$	$0.286^{\downarrow\downarrow\approx}$
Tag filters	$SE_{F_{tfidf,80}}$	$0.171^{\uparrow\uparrow\uparrow}$	$0.221^{\downarrow\uparrow\uparrow}$	$0.523^{\uparrow\downarrow\uparrow}$
	$SE_{F_{lda,30}}$	$0.149^{\downarrow\uparrow\uparrow}$	$0.22^{\downarrow\uparrow\uparrow}$	$0.454^{\uparrow\downarrow\uparrow}$
	$SE_{F_{tfidf,80+catalog}}$	$0.199^{\uparrow\uparrow\uparrow}$	$0.233^{\downarrow\uparrow\uparrow}$	$0.603^{\uparrow\uparrow\uparrow}$
	$SE_{F_{lda,30+catalog}}$	$0.174^{\uparrow\uparrow\uparrow}$	$0.258^{\downarrow\uparrow\uparrow}$	$0.53^{\uparrow\downarrow\uparrow}$

***Table 5.** MAP/Precision/Recall scores for the search engines. Notational convention: the system that indexes the data obtained by applying the filter F is denoted by SE_F; only the best performing filters for each filtering approach are shown. $\uparrow$, $\downarrow$, and $\approx$ indicate if a score is significantly better, worse, or statistically indistinguishable from the scores of $SE_{catalog}$, SE_{user}, and SE_{vuser}, respectively.*

background or in the foreground, prominent stationary objects, or words from dialog are among the things the verified tags denote. This is hardly surprising, after all these are things that are easy to reach consensus on and ultimately that is the goal of the game from the perspective of the players.

6. IMPROVING THE QUALITY OF USER TAGS FOR TOPICAL VIDEO SEARCH

Previously, in Sect. 5 we investigated how well the tags collected with *Waisda?* are performing with respect to topical search. The general conclusion was that when it comes to using *Waisda?* tags to retrieve fragments that are about a given topic the search performance is unsatisfactory. This conclusion came with the following caveat: while the search recall is relatively high ($\approx 59\%$), the search precision is rather low ($\approx 16\%$). What this means is that a significant portion of topics covered by the fragments are in fact entered as tags by the *Waisda?* players hence the high recall. Moreover, there are also many tags that do not refer to the topics covered by the fragments and when these tags result in a hit the overall search precision goes down. This being said, should the non-topical tags be detected and filtered out from the collection, that would result in increased precision and unchanged recall. In this section we describe the tag filtering methods, from this point on called *tag filters*, that are investigated in this study. We also present how well each of the filters eliminates the non-topical tags. The evaluation metrics are summarized in Table 5. For reference we copied the metrics for the systems SE_{user}, SE_{vuser}, and $SE_{catalog}$ from Sect. 5.2. We use the following notational convention in the table: the system that indexes the data obtained by applying the filter F is denoted by SE_F and the system that indexes the data obtained by applying the filter F and the catalog data is denoted by $SE_{F+catalog}$. The system that we are trying to beat is $SE_{catalog}$ since it is the best performing one and an approximation of the present search functionality.

6.1. Latent Dirichlet allocation-based filtering

Topic models (Hofmann, 1999; Blei et al., 2004) are a type of statistical models for discovering abstract 'topics' in a collection of documents. One of the most common topic models currently in use is the Latent Dirichlet Allocation (LDA). The idea behind LDA is to model documents as

arising from multiple topics, where a topic is defined to be a probability distribution over a fixed vocabulary of terms — the set of all unique words in the collection. Specifically, it is assumed that there exists a fixed set of K topics associated with the 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, much like documents, have many layers of meaning and can be viewed as mixture of topics. User tags collected through *Waisda?* can be seen as instantiations of these topics. The unstructured nature of the tags, — only weak temporal ordering of tags within video exists, based on the tag entry time — fits the 'bag of words' metaphor quite well. According to LDA the probability that a tag t appears in video v is given by

$$P(t \mid v) = \sum_{i=1}^{K} P(t \mid z_i) P(z_i \mid v) \tag{2}$$

where $P(t \mid z_i)$ is the probability of the tag t for the topic z_i and $P(z_i \mid v)$ is the probability of picking a tag from the topic z_i in v i.e. the proportion of z_i exhibited by v. We use the probability $P(\cdot \mid v)$ given by (2) to rank the tags ascribed to v in descending order. The filtering is carried out by taking the top k tags for v. We denote the tags filters defined in this was by $F_{lda,k}$. In the experiment, we vary the value of k in the set $\{5k \mid k \in \mathbb{N},\ 2 \leq k \leq 20\}$, i.e. all integers from 10 to 100 with increments of 5. The best results are achieved for $k = 30$ and are shown in Table 5. As we can see, $S_{F_{lda,30}}$ outperforms that system SE_{user} that indexes all user tags by 14%. However, the retrieval performance of $S_{F_{lda,30}}$ falls short when compared with the systems $SE_{catalog}$ and $S_{F_{tfidf,80}}$ which outperform it by 28% and 31%, respectively. The combination of the catalog data and LDA tag filtering yields only a marginal improvement in performance: $SE_{F_{lda,30}+catalog}$ outperforms $SE_{F+catalog}$ only by 3% with respect to the MAP score. A potential explanation for the poor performance of the LDA based filtering is the size of the corpus. Our tag/fragment collection is not large enough for the LDA model to provide a good estimation of the underlying topic structure. In fact, this was our suspicion all along however we decided to include LDA for the sake of completeness — since TF-IDF and LDA are among the most commonly used methods for measuring word to document relevance and topic inference.

6.2. *TF-IDF-rank Take Top k*

We saw in Sect. 5.2.1 that the TF-IDF score of a tag is rather indicative as to whether the tag is topical or non-topical. Referring back to Sect. 5.2.1, for the analyzed sample of positives, the average TF-IDF score of the true positives was higher than the average TF-IDF score of the false positives. This suggests that the TF-IDF measure favours topical tags. Assuming the correctness of this hypothesis, we exploit this measure to filter out the potentially non-topical game tags. In particular, for each fragment in the collection we rank the tags associated with the fragment based on their TF-IDF[14] scores in descending order: the tag with the highest TF-IDF score is at the top. Then the filtering is performed by taking only the top k tags for every video. We denote the tag filters defined in this way by $F_{tfidf,k}$. In the experiment, we vary the value of k in the set $\{5k \mid k \in \mathbb{N}, 2 \leq k \leq 20\}$, i.e. all integers from 10 to 100 with increments of 5. Figure 3 presents the search performance

[14]The TF-IDF measure is computed over a corpus of documents. In this particular case the "documents" are the bag of tags associated with the fragments.

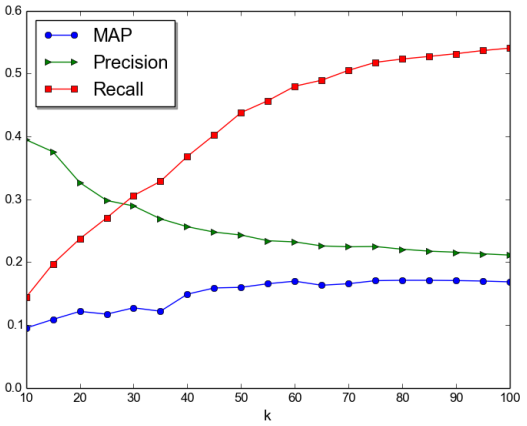

Figure 3. ***Search performance statistics for the systems*** $S_{F_{tfidf,k}}$.

statistics as the value of k varies. Not surprisingly, as the value of k increases the search precision and the search recall decrease and increase, respectively. Moreover, the MAP score steadily increases until k equals 80, where the maximal value is reached, and then it starts to decrease. As shown in Table 5, the MAP score of $S_{F_{tfidf,80}}$ is 0.171 which means it significantly outperforms SE_{user}, the systems that indexes all user tags. Indeed, the increase in search performance is 31%. What is more important is that $S_{F_{tfidf,80}}$ slightly outperforms even $SE_{catalog}$, which until now was the best performing one and the system that we are trying to beat. This suggests that TF-IDF score is indeed a good indication of the quality of the tags as topical descriptors and can be used to filter out the non-topical tags. Furthermore, the combination of catalog metadata and TF-IDF filtering proves to be beneficial: $SE_{tfidf,80+catalog}$ outperforms $SE_{catalog}$ by 18%. The improvement is caused by the fact that the TF-IDF ranked tags increase the search recall of the catalog metadata by factor of 2; the recall of $SE_{tfidf,80+catalog}$ is higher than the recall of $SE_{catalog}$ by 107%.

6.3. Honorable mentions

Besides the two filters described above, we tried two other filtering methods. Alas, they did not perform well and for space considerations we only briefly describe them here.

Network Analysis-based filtering Network analysis tools and mechanisms (Hanneman et al., 2005; Easley, Kleinberg, 2010) are increasingly used to study folksonomies and social tagging related phenomena (Ji-Lung, 2011; Wu, 2008; Shen et al., 2005). The crux of these approaches is to represent the domain knowledge as a network (graph) and apply network analysis tools to investigate the phenomenon of interest. In our particular case, we exploit network analysis to detect and filter out non-topical tags. The general idea is to build a network for each video fragment that captures the semantic connectedness among the tags associated with that fragment. Once the network is build we exploit network centrality measures to rank the tags according to their importance. The intuition is the more central a given tag is the higher its connectedness with the other tags is

and therefore that tag has higher importance as content descriptor. We considered three centrality measures: pagerank, weighted degree centrality and eigenvector centrality.

Player reputation-based filtering Player's reputation can be exploited to reduce the influence of the tags ascribed by 'bad' players. We define the reputation of a player as the ratio of the number verified tags entered by the player to the number of all tags entered by the player. We consider a verified tag to be a positive evidence for the player's reliability, therefore a higher fraction of verified tags implies higher player's reputation. Instead of computing the ratio directly, we estimate the value by taking the lower bound of Wilson score confidence interval for a Bernoulli parameter (Agresti,Coull, 1998; Wilson, 1927). The latter approach is more robust in cases where the number of observations (evidence) is small. One limitation that we faced was the fact that most of the players ($\approx 99\%$) are anonymous meaning they only have one recorded session in which they played one or more games. This means that even if the same person played two different sessions as anonymous player there is no way to reliably correlate the sessions. In effect, the amount of evidence we can collect for the players is limited which led to incorrect estimation of the players' reputation. Moreover, considering (Robertson et al., 2009; Jain et al., 2013) our reputation estimation method assigns higher reputation to players which settle on low-effort tags. Consequently, tags favored by this filtering method will tend to be more 'obvious'.

7. CONCLUSIONS AND DISCUSSION

In this paper we studied to what extend players enter tags which are valid topical descriptors of the video material. Another aim was to derive insights about players' tagging practices that are associated with ascribing topical tags. The study was carried out with the focus on topical search.

In Sec. 5 we evaluated the search performance of the entire unprocessed collection of the user tags. The general conclusion is that the search performance of the raw, unprocessed, user tags for retrieving videos based on topic leaves much to be desired. While the search recall of the user tags is relatively satisfactory the precision is rather poor. Our analysis showed that a significant portion of the topics are indeed captured by the user tags. However, there are also many user tags that do not pertain to the subject (*semantics*) of the video, but refer to the more *syntactic* aspects such as what is *seen* or *heard*. It is the latter group that is responsible for the false positives and thereby hurting the search precision. Therefore, if user tags are to be used for topical search, a preprocessing step is required that will identify and filter out the non-topical user tags.

The quality of the tags as topical descriptions was addressed in Sec. 6 where we looked into several ways we can detect and filter out the non-topical user tags. While the different methods that we studied performed with various success, the conclusion is that the game tags can be successfully exploited for topical video search provided there is a filtering process that would reduce or eliminate the effect of the non-topical tags. Our results show that after TF-IDF-based filtering game tags can emulate the retrieval performance of the best performing system that utilizes manually crafted metadata for search. Moreover, combining TF-IDF filtered game tags with the manually crafted metadata yields an improvement of retrieval performance by 18%. The improvement is attributed to the increased retrieval recall stemming from the game tags. An important consequence of this result is that tagging games provide a cost-effective alternative for AV collection owners that do not possess the required manpower to manually annotate their material.

Successfully deploying a GWAP is no small feat. Atracting new players and keeping them engaged over time is vital for success and requires continuous publicising efforts. The experience gained from the *Waisda?* project showed that targeting the fanbase of the TV series being tagged in the game is an effective method of attracting new players. Player's engagement can be sustained with in-game motivational mechanisms such as leaderboards, and with time-limited contests offering awards for the top performers. Planning and executing such activities is costly, however the cost is independent of the size of the video collection. The cost of manual annotation, on the other hand, increases linearly with the size of the collection and will eventually outweigh the *Waisda?*-related costs.

What makes video annotation a difficult task is the fact that video is a medium that is extremely rich in meaning. The taggers can get overwhelmed by the complex interplay of objects and events especially in the fast-paced game setting. Our qualitative analysis showed that significant portion of the *Waisda?* tags refer to more noticeable aspects of the content such as moving objects in the background or in the foreground, or prominent stationary objects. This is hardly surprising, after all these are things that are easy to reach consensus on and ultimately that is the goal of the game from the perspective of the players. We also observed that the tags which are present in the audio and refer to a concept that appears visually are more likely to be topical descriptors. Their presence in both the audio and visual component is an indication that they are denoting salient aspects of the content. Therefore, if the goal of the AV collection owners is collecting topical tags, this insight can be operationalized in the game by instructing the players to tag things that are both on screen and in the audio.

Waisda? is a production grade open-source crowd-sourcing tool which is relatively easy to set-up and after the suitable processing the collected tags can be exploited for retrieval. We believe that this is an important step toward making the AV heritage more accessible on the Web and in general making the Web more connected.

8. 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.

9. REFERENCES

Ding, Y, Yan, E, Frazho, A. R, and Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11), 2229-2243. https://doi.org/10.1002/asi.v60:11

Floeck, F, Putzke, J, Steinfels, S, Fischbach, K, and Schoder, D. (2011). Imitation and Quality of Tags in Social Bookmarking Systems - Collective Intelligence Leading to Folksonomies. In *Advances in Intelligent and Soft Computing* (pp. 75-91). doi:10.1007/978-3-642-14481-3_7

Furnas, G. W, Landauer, T. K, Gomez, L. M, and Dumais, S. T. (1987). The Vocabulary Problem in Human-system Communication. In *Communications of the ACM*, 30(11), 964-971. https://doi.org/10.1145/32206.32212

Geisler, G and Burns, S. (2007). Tagging video: conventions and strategies of the YouTube community. In *JCDL* (pp. 480-480). https://doi.org/10.1145/1255175.1255279

Gligorov, R, Hildebrand, M, van Ossenbruggen, J, Aroyo, L, and Schreiber, G. (2013). An Evaluation of Labelling-Game Data for Video Retrieval. In *ECIR* (pp. 50-61). doi:10.1007/978-3-642-36973-5_5

Golder, S. A and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of information science*, 32(2), 198-208. https://doi.org/10.1177/0165551506062337

Halvey, M. J and Keane, M. T. (2007). Analysis of Online Video Search and Sharing. In *HT. ACM* (pp. 217-226). https://doi.org/10.1145/1286240.1286301

Hanneman, R. A and Riddle, M. (2005). Introduction to social network methods. University of California, Riverside.

Hildebrand, M, Brinkerink, M, Gligorov, R, Steenbergen, M. V, Huijkman, J, and Oomen, J. (2013). Waisda?: video labeling game.. In *ACM Multimedia* (pp. 823-826). http://doi.acm.org/10.1145/2502081.2502221

Hildebrand, M and van Ossenbruggen, J. (2012). Linking user generated video annotations to the web of data. In *Advances in Multimedia Modeling* (pp. 693-704). doi:10.1007/978-3-642-27355-1_74

Hofmann, T. (1999). Probabilistic latent semantic indexing. *SIGIR*, (pp. 50-57). https://doi.org/10.1145/312624.312649

Hotho, A, Jäschke, R, Schmitz, C, and Stumme, G. (2006). Information Retrieval in Folksonomies: Search and Ranking. In *ESWC* (pp. 411-426). doi:10.1007/11762256_31

Ji-Lung, H. (2011). Network Analysis of Tagging Structure. In *Proceedings of the American Society for Information Science and Technology*, 48(1), 1-4. doi:10.1002/meet.2011.14504801233

Jin, Yohan and Khan, Latifur and Prabhakaran, B. (2010). Knowledge Based Image Annotation Refinement. In *Journal of Signal Processing Systems*, 58(3), 387-406. https://doi.org/10.1007/s11265-009-0391-y

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. In *Document retrieval systems*, (pp. 132-142).

Junker, B. H and Schreiber, F. (2008). Analysis of biological networks. Vol. 2. John Wiley & Sons.

Kennedy, L, Slaney, M, and Weinberger, K. (2009). Reliable Tags Using Image Similarity: Mining Specificity and Expertise from Large-scale Multimedia Databases. In *WSMC* (pp. 17-24). https://doi.org/10.1145/1631135.1631139

Koschützki, D and Schreiber, F. (2008). Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. In *Gene Regulation and Systems Biology*, 2(2), 193-201.

Lee, S, De Neve, W, and Ro, Y. M. (2010). Tag Refinement in an Image Folksonomy Using Visual Similarity and Tag Co-occurrence Statistics. In *Journal of Image Communication*, 25(10), 761-773. https://doi.org/10.1016/j.image.2010.10.002

Lee, S, De Neve, W, and Ro, Y. M. (2012). Towards Data-driven Estimation of Image Tag Relevance Using Visually Similar and Dissimilar Folksonomy Images. In *SAM* (pp. 3-8). https://doi.org/10.1145/2390876.2390880

Li, M, Tang, J, Li, H, and Zhao, C. (2012). Tag Ranking by Propagating Relevance over Tag and Image Graphs. In *ICIMCS* (pp. 153-156). doi:10.1145/2382336.2382380

Li, X, Snoek, C. G, and Worring, M. (2008). Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *MIR* (pp. 180-187). https://doi.org/10.1145/1460096.1460126

Liu, D, Hua, X.-S, Yang, L, Wang, M, and Zhang, H.-J. (2009). Tag Ranking. In *WWW* (pp. 351-360). https://doi.org/10.1145/1526709.1526757

Marshall, C. C. (2009). No Bull, No Spin: A Comparison of Tags with Other Forms of User Metadata. In *JCDL* (pp. 241-250). https://doi.org/10.1145/1555400.1555438

Miller, G. A. (1995). WordNet: A Lexical Database for English. In *ACM*, 38(1), 39-41. https://doi.org/10.1145/219717.219748

Morrison, P. J. (2008). Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. In *Information Processing and Management: an International Journal* 44(4), 1562-1579. https://doi.org/10.1016/j.ipm.2007.12.010

Oomen, J, Gligorov, R, and Hildebrand, M. (2014). Waisda?: making videos findable through crowdsourced annotations. In *Crowdsourcing our Cultural Heritage*. Ashgate.

Panofsky, E. (1972). Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Row.

Rorissa, A. (2010). A comparative study of Flickr tags and index terms in a general image collection. In *JASIST*, 61(11), 2230-2242. doi:10.1002/asi.21401

Salton, G and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing & Management*, 24(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Shatford, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. In *Cataloging & Classification Quarterly*, 6(3), 39-62.

doi:10.1300/J104v06n03_04

Shen, K and Wu, L. (2005). Folksonomy as a Complex Network. In *CoRR* abs/cs/0509072.

Sigurbjörnsson, B and van Zwol, R. (2008). Flickr Tag Recommendation Based on Collective Knowledge. In *WWW* (pp. 327-336). https://doi.org/10.1145/1367497.1367542

Smucker, M. D, Allan, J, and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *CIKM* (pp. 623-632). https://doi.org/10.1145/1321440.1321528

Sun, A and Bhowmick, S. S. (2010). Quantifying Tag Representativeness of Visual Content of Social Images. In *MM* (pp. 471-480). https://doi.org/10.1145/1873951.1874029

Truong, B. Q, Sun, A, and Bhowmick, S. S. (2012). Content is Still King: The Effect of Neighbor Voting Schemes on Tag Relevancefor Social Image Retrieval. In *ICMR*, article no. 9, 8 pages. https://doi.org/10.1145/2324796.2324808

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In *CLEF* (pp. 355-370). doi:10.1007/3-540-45691-0_34

Wang, C, Jing, F, Zhang, L, and Zhang, H.-J. (2006). Image Annotation Refinement Using Random Walk with Restarts. In *MM* (pp. 647-650). https://doi.org/10.1145/1180639.1180774

Wang, Y and Gong, S. (2007). Refining Image Annotation Using Contextual Relations Between Words. In *CIVR* (pp. 425-432). https://doi.org/10.1145/1282280.1282343

Wu, C. (2008). Analysis of Tags as a Social Network. In *CSSE* (pp. 651-654). https://doi.org/10.1109/CSSE.2008.1268

Yanbe, Y, Jatowt, A, Nakamura, S, and Tanaka, K. (2007). Can Social Bookmarking Enhance Search in the Web?. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 107-116). https://doi.org/10.1145/1255175.1255198

Zhao, Y, Zha, Z.-J, Li, S, and Wu, X. (2010). Which Tags Are Related to Visual Content?. In *MMM* (pp. 669-675). doi:10.1007/978-3-642-11301-7_67

Blei, D. M., Ng, A. Y., & Jordan, M. I.. (2004). Latent dirichlet allocation. In *J. Mach. Learn. Res.*, 993-1022.

von Ahn, Luis, & Dabbish, Laura (2003). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 319-326). https://doi.org/10.1145/985692.985733

Baltussen, B. B. (n.d.). Waisda? Video Labeling Game: Evaluation Report. Retrieved May 4, 2017, from http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/

Bischoff, Kerstin and Firan, Claudiu S. and Nejdl, Wolfgang and Paiu, Raluca (2004a). Bridging the Gap Between Tagging and Querying Vocabularies: Analyses and Applications for Enhancing Multimedia IR. In *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2-3), 97-109. https://doi.org/10.1016/j.websem.2010.04.004

Bischoff, Kerstin and Firan, Claudiu S. and Nejdl, Wolfgang and Paiu, Raluca (2004b). Can all tags be used for search?. In *Proceeding of the 17th ACM conference on Information and knowledge management* (p./pp. 193–202), New York, NY, USA: ACM. ISBN: 978-1-59593-991-3

Bao, Shenghua and Xue, Guirong and Wu, Xiaoyuan and Yu, Yong and Fei, Ben and Su, Zhong (2007). Optimizing Web Search Using Social Annotations. In *WWW*, (pp. 97-109). https://doi.org/10.1145/1242572.1242640

Cilibrasi, Rudi L. and Vitanyi, Paul M. B. (2007). The Google Similarity Distance. In *IEEE Trans. on Knowl. and Data Eng.*, 19(3), 370-383. doi:10.1109/TKDE.2007.48

von Ahn, Luis, & Dabbish, Laura (2008). Designing Games with a Purpose. In *Communications of the ACM*, 51(8), 58-67. https://doi.org/10.1145/1378704.1378719

von Ahn, Luis and Liu, Ruoran and Blum, Manuel (2006). Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 55-64). doi:10.1145/1124772.1124782

von Ahn, Luis and Kedia, Mihir and Blum, Manuel (2006). Verbosity: A Game for Collecting Common-sense Facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 55-64). doi:10.1145/1124772.1124784

Havasi, Catherine and Lieberman, Henry (2007). Common Sense and Intelligent User Interfaces. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (pp. 7-7). doi:10.1145/1216295.1216301

Tao-Hsuan Chang, Cheng-wei Chan, Jane Yung-jen Hsu (2010). Human Computation Game for Commonsense Data Verification. At *AAAI Fall Symposium: Commonsense Knowledge*, article no. 6, num. pages 2.

Bell, Marek and Reeves, Stuart and Brown, Barry and Sherwood, Scott and MacMillan, Donny and Ferguson, John and Chalmers, Matthew (2009). EyeSpy: Supporting Navigation Through Play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 123-132). doi:10.1145/1518701.1518723

Tuite, Kathleen and Snavely, Noah and Hsiao, Dun-Yu and Smith, Adam M. and Popović, Zoran (2010). Reconstructing the World in 3D: Bringing Games with a Purpose Outdoors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 232-239). doi:10.1145/1822348.1822379

Tuite, Kathleen and Snavely, Noah and Hsiao, Dun-yu and Tabing, Nadine and Popovic, Zoran (2011). PhotoCity: Training Experts at Large-scale Image Acquisition Through a Competitive Game. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games* (pp. 1383-1392). doi:10.1145/1978942.1979146

Law, E. and von Ahn, L. and Dannenberg, R. and Crawford, M. (2007). TagATune: a game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 361-364).

Law, Edith and von Ahn, Luis (2009). Input-agreement: A New Mechanism for Collecting Data Using Human Computation Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1197-1206). doi:10.1145/1518701.1518881

Turnbull, Douglas and Liu, Ruoran and Barrington, Luke and Lanckriet, Gert (2007). A game-based approach for collecting semantic annotations of music. In *ISMIR* article no. 1, numpages 4.

Christoph Wieser and François Bry and Alexandre Bérard and Richard Lagrange (2013). ARTigo: Building an Artwork Search Engine With Games and Higher-Order Latent Semantic Analysis. In *Proc. of Disco 2013, Workshop on Human Computation and Machine Learning in Games at HComp* (pp. 1-6).

Bartholomäus Steinmayr and Christoph Wieser and Fabian Kneißl and François Bry (2011). Karido: A GWAP for telling artworks apart. In *CGAMES* (pp. 193-200). doi:10.1109/CGAMES.2011.6000338

Florian Störkle (2012). Combino - A GWAP for Generating Combined Tags. *Bachelor thesis*, Institute of Computer Science, LMU, Munich.

Siorpaes, Katharina and Hepp, Martin (2008). Games with a Purpose for the Semantic Web. In *IEEE Intelligent Systems*, 23(3), 50-60. doi:10.1109/MIS.2008.45

Thaler, Stefan and Simperl, Elena and Wölger, Stephan (2012). An experiment in comparing human-computation techniques. In *IEEE Intelligent Systems*, 16(5), 52-58.

Stefan Thaler and Elena Paslaru Bontas Simperl and Katharina Siorpaes (2011). SpotTheLink: A Game for Ontology Alignment. In *Conference on Professional Knowledge Management: From Knowledge to Action* (pp. 246-253).

Weng, Lilian and Schifanella, Rossano and Menczer, Filippo (2011). Design of Social Games for Collecting Reliable Semantic Annotations. In *CGAMES*, (pp. 185-192). doi:10.1109/CGAMES.2011.6000337

Seemakurty, Nitin and Chu, Jonathan and von Ahn, Luis and Tomasic, Anthony (2010). Word Sense Disambiguation via Human Computation. In *HCOMP*, (pp. 60-63). doi:10.1145/1837885.1837905

François Bry and Fabian Kneissl and Thomas Krefeld and Stephan Lücke and Christoph Wieser (2013). ARTigo: Building an Artwork Search Engine With Games and Higher-Order Latent Semantic Analysis. In *Proc. of Disco 2013, Workshop on Human Computation and Machine Learning in Games at HComp*, (pp. 1-6).

Derek L. Hansen and David W. Jacobs and Darcy Lewis and Arijit Biswas and Jennifer Preece and Dana Rotman and Eric Stevens (2011). Odd Leaf Out: Improving Visual Recognition with Games. In *PASSAT*, (pp. 87-94). doi:10.1109/PASSAT/SocialCom.2011.225

E. Law and B. Settles and A. Snook and H. Surana and L. von Ahn and T. Mitchell (2011). Human Computation for Attribute and Attribute Value Acquisition. In *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization*, article no. 6, numpages 2.

Pearl, Lisa and Steyvers, Mark (2010). Identifying Emotions, Intentions, and Attitudes in Text Using a Game with a Purpose. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 71-79).

Fabian Kneißl (2014). Crowdsourcing for linguistic field research and e-learning. *Doctoral thesis*, Ludwig-Maximilians-Universität München.

Stephen Robertson and Milan Vojnovic and Ingmar Weber (2009). Rethinking the ESP game. In *Proceedings of CHI* (pp. 3937-3942). doi:10.1145/1520340.1520597

Jain, Shaili and Parkes, David C. (2013). A Game-theoretic Analysis of the ESP Game. In *ACM Trans. Econ. Comput.*, 1(1), 3:1-3:35. doi:10.1145/2399187.2399190

François Bry and Corina Schemainda and Clemens Schefels (2015). A Gaming Ecosystem Crowdsourcing Deep Semantic Annotations. *Research report*.

Laura Hollink and Guus Schreiber and Bob J. Wielinga and Marcel Worring (2004). Classification of user image descriptions. In *Int. J. Hum.-Comput. Stud.*, 61(5), 601-626. http://dx.doi.org/10.1016/j.ijhcs.2004.03.002

David, Easley and Jon, Kleinberg (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. New York, NY, USA. Cambridge University Press

Agresti, Alan and Coull, Brent A. (1998). Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. In *The American Statistician*, 52, 119-126.

Wilson, Edwin B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. In *Journal of the American Statistical Association* 22, 209-212.