Incorporating Crowd Perspectives into Multimedia Retrieval Systems

Vliegendhart, Raynor

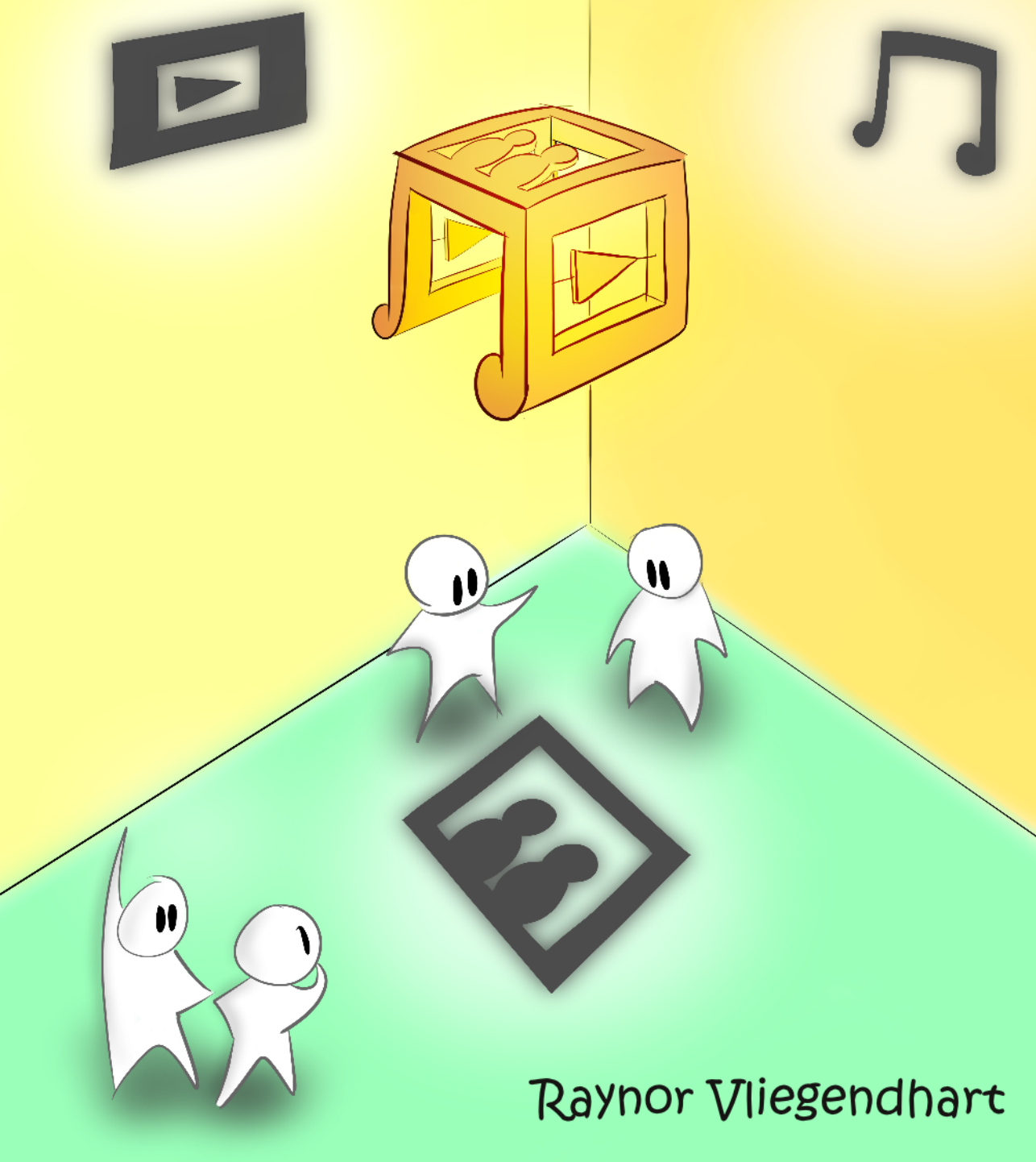
DOI
10.4233/uuid:bfa8a148-0d3a-49bc-a0db-aba16211a90d

Publication date
2017

Document Version
Final published version

Citation (APA)
Vliegendhart, R. (2017). *Incorporating Crowd Perspectives into Multimedia Retrieval Systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:bfa8a148-0d3a-49bc-a0db-aba16211a90d

Incorporating Crowd Perspectives into Multimedia Retrieval Systems



Raynor Vliegendhart

Propositions

accompanying the dissertation

INCORPORATING CROWD PERSPECTIVES INTO MULTIMEDIA RETRIEVAL SYSTEMS

by

Raynor VLIEGENDHART

1. The inability of fully automatic multimedia content analysis and indexing approaches to keep pace with user needs necessitates the incorporation of crowd perspectives into multimedia retrieval systems. (this thesis)
2. Non-linear video access requires more than just topical and affective relevance dimensions. (this thesis)
3. Whether people find a moment in a video to be noteworthy cannot be determined using the video itself. (this thesis)
4. The correct way of collecting realistic video viewing behavior in a crowdsourcing setting is to have the viewer population dictate the video dataset rather than the other way around. (this thesis)
5. The conventional language of mathematics is unfit for clearly and succinctly conveying a message in Computer Science papers.
6. Ecological validity is unsustainable in a purely academic setting.
7. Imposing a page limit on an article's length is detrimental to innovative research.
8. Productivity would skyrocket if all the advances in artificial intelligence would be applied to daily basic tools (such as search and replace in editors), which have been braindead for decades.
9. Campaigns aiming to get more women in STEM fields should not target girls, but their parents instead.
10. Every office should have a shower and every shower should have a whiteboard.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors prof. dr. A. Hanjalic and prof. dr. M. A. Larson.

INCORPORATING CROWD PERSPECTIVES INTO MULTIMEDIA RETRIEVAL SYSTEMS

INCORPORATING CROWD PERSPECTIVES INTO MULTIMEDIA RETRIEVAL SYSTEMS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 20 november 2017 om 12:30 uur

door

Raynor VLIEGENDHART

Master of Science in Computer Science,
Technische Universiteit Delft, Nederland,
geboren te Waddinxveen, Nederland.

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. A. Hanjalic
Prof. dr. M. A. Larson

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. A. Hanjalic,	Technische Universiteit Delft, promotor
Prof. dr. M. A. Larson,	Technische Universiteit Delft, promotor
Onafhankelijke leden:	
Prof. dr. C. M. Jonker,	Technische Universiteit Delft
Prof. dr. F. De Natale,	Università di Trento, Italië
Prof. dr. P. Halvorsen,	Universitetet i Oslo, Noorwegen
Prof. dr. T. Hoßfeld,	Universität Duisburg-Essen, Duitsland
Prof. dr. M. Worring,	Universiteit van Amsterdam
Prof. dr. ir. D. Epema,	Technische Universiteit Delft, reservelid

Keywords: multimedia retrieval, crowdsourcing, video search, user study

Printed by: Ridderprint BV

Front & Back: Raynor Vliegendhart, 'Perspectives on Multimedia'.

ISBN 978-94-6299-787-5

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To all ideas lost to time,
may you be discovered again.

CONTENTS

SUMMARY

The twenty-first century has brought plentiful computational power and bandwidth to the masses and has opened up access to multimedia recording devices for everyone. With these developments, a shift in the landscape of multimedia took place: from traditional one-to-many programming (the paradigm of traditional television) to many-to-many creation of diverse content. Nowadays, everyone can become a content creator and connect with new audiences, which has resulted in an explosion of diverse and available multimedia content. In tandem with this change, user needs have evolved as well. Yet, existing multimedia retrieval systems have been struggling to keep up with what users are looking for.

In this thesis, we argue that a multi-perspective approach is desired in order to cater to a diverse range of user needs. In order to know which perspectives should be taken, we turn to the crowd as a source of information on which perspectives would be actually helpful for serving users of multimedia retrieval systems. The central question underlying the research presented in this thesis is: How can we incorporate these perspectives of the crowd into multimedia retrieval systems?

The first major part of the thesis consists of the development of methodologies for effectively addressing the crowd in crowdsourcing studies. It first introduces the concept of framing. Framing allows people to picture a particular scenario that helps them to understand the task at hand and thus would result in high quality answers. Following the framing methodology, the focus shifts to the refinement of elicitation techniques in order to effectively model the common understanding on a particular topic. The methodologies presented in this first part are shown to be useful in informing the design of new features for a multimedia retrieval system.

The second major part of the thesis builds upon the methodologies developed in the first part and uses them to push the research on non-linear video access, i.e., supporting users in consuming relevant parts of a video, further in two ways. First, in a carefully designed crowdsourcing experiment, user comments referring to specifically mentioned time-points in a video are analyzed to build a crowd-informed typology that captures new dimensions of relevance at the time-code level. The usefulness of this typology is tested through a crowdsourced user study on a simulated search scenario. Second, a methodology is developed for obtaining realistic viewing behaviors through crowdsourcing experiments, which can be used in designing and testing new non-linear video access methods. This methodology stresses the importance of not only properly framing the crowdsourcing task, but also that the crowd and multimedia domain are jointly chosen in order to observe behavior that resembles behavior that participants would normally exhibit outside of the experiment. The methodology is used to demonstrate its ability to capture implicit viewing behavior that can be used to support users in non-linearly accessing videos.

The final contributions of the thesis consist of practical pointers for future work and a set of open research questions pertaining crowdsourcing tasks with an interpretive nature. The practical pointers for future work are fueled by experience gained through the various crowdsourcing campaigns that have been carried out throughout the thesis. Addressing these pointers will help in making crowdsourcing research more effective and reduce the effort needed in carrying out experiments. The set of open research questions are formulated by positioning this thesis in relation to prior related work. These questions serve as a starting point for future research on interpretive crowdsourcing tasks and pursuing them could aid the development of retrieval systems with multiple perspectives on multimedia.

SAMENVATTING

De integratie van perspectieven van de grote menigte in multimediazoeksystemen

De eenentwintigste eeuw heeft de massa een overvloed aan rekenkracht en bandbreedte gebracht en zorgde ervoor, dat apparatuur om multimediaopnamen te maken voor iedereen beschikbaar werd. Met deze ontwikkelingen vond er een verschuiving plaats in het multimedialandschap: van een situatie waarin voornamelijk door een enkeling de programmering voor velen werd bepaald (het paradigma van klassieke televisie) naar één waarin diverse multimediainhoud door velen voor velen gecreëerd wordt. Nu is het voor eenieder mogelijk om contentmaker te worden en in contact te komen met nieuwe publieken, met als gevolg een explosie van veel en divers beschikbaar multimediamateriaal. Gelijk met deze verandering evolueerden de behoeften van gebruikers eveneens. Echter, multimediazoekmachines konden deze veranderingen niet bijbenen en worstelen nog steeds om aan de vraag van gebruikers te voldoen.

In dit proefschrift wordt er beargumenteerd, dat een aanpak vanuit meerdere perspectieven wenselijk is om aan een divers bereik van gebruikersbehoeften te kunnen voldoen. Om erachter te komen welke perspectieven er genomen moeten worden, keren we ons tot de *crowd*, de grote menigte, en beschouwen wij deze als een informatiebron die inzicht geeft welke perspectieven daadwerkelijk van dienst kunnen zijn bij het dienen van gebruikers van multimediazoekmachines. De centrale vraag die ten grondslag ligt aan al het onderzoek in dit proefschrift, luidt dan ook als volgt: Hoe kunnen we deze perspectieven van de grote menigte integreren in multimediazoeksystemen?

Het eerste hoofddeel van dit proefschrift beschrijft de ontwikkeling van methodieken om een menigte in *crowdsourcing*studies—waarbij crowdsourcing een samentrekking is van crowd en outsourcing—effectief te adresseren. In dit deel wordt eerst het *framing*concept geïntroduceerd. Het opstellen van een kader (het *frame*) helpt mensen om zich een bepaald scenario in te beelden, dat ze op weg helpt om de taak die voor hen ligt beter te begrijpen, met als beoogde resultaat dat de antwoorden die zij zullen geven van hoge kwaliteit zullen zijn. Na de introductie van het framingconcept verschuift de aandacht naar het verfijnen van bevragingstechnieken, die op een slimme manier geschikte antwoorden onttrekken, om een gemeenschappelijk begrip over een bepaald onderwerp effectief te kunnen modeleren. Het nut van deze methodieken uit het eerste hoofddeel van dit proefschrift wordt bewezen door hun praktische inzet bij het ontwerpproces van nieuwe functionaliteit voor een bestaand multimediazoeksysteem.

Het tweede hoofddeel van dit proefschrift bouwt voort op de methodieken uit het eerste deel en gebruikt ze voor het bevorderen van onderzoek op het gebied van niet-lineaire toegang tot video's, d.w.z. gebruikers helpen om enkel de relevante delen van een video te laten bekijken, op twee manieren. Ten eerste zijn gebruikersreacties die een expliciet geschreven verwijzing naar een tijdstip in een video bevatten, geanalyseerd met behulp van een zorgvuldig ontworpen crowdsourcingexperiment om tot een door

de menigte geïnformeerde typologie te komen die nieuwe relevantiedimensies op tijdniveau vastlegt. Het nut van deze typologie is getest voor een gesimuleerd zoekscenario door middel van een crowdsourcingstudie. Ten tweede is er een methodiek ontwikkeld voor het verkrijgen van realistisch kijkgedrag via crowdsourcingexperimenten. Het verkregen kijkgedrag kan op zijn beurt worden gebruikt in het ontwerpen en het testen van nieuwe oplossingen voor niet-lineaire videotoegang. Deze methodiek benadrukt, dat alleen het opstellen van een juist kader in een crowdsourcingtaak niet voldoende is, maar dat het daarnaast essentieel is, dat de menigte en het multimediadomein tezamen gekozen worden om kijkgedrag te kunnen waarnemen dat vergelijkbaar is met gedrag dat men ook zou vertonen buiten het experiment om. Er is voor de methodiek aangetoond, dat het in staat is om impliciet gedrag tijdens het kijken van video's vast te leggen en dat zulk gedrag kan worden ingezet om gebruikers te ondersteunen in het niet-lineair bekijken van video's.

De laatste bijdragen van het proefschrift bestaan uit praktische aanwijzingen voor toekomstig onderzoek en een lijst van open onderzoeksvragen die betrekking hebben op crowdsourcingtaken van interpretabele aard. De praktische aanwijzingen voor verder onderzoek komen voort uit de ervaring die is opgedaan tijdens het uitvoeren van uiteenlopende, in dit proefschrift beschreven crowdsourcingexperimenten. Het opvolgen van deze aanwijzingen kan bijdragen aan het effectiever maken van crowdsourcingonderzoek en het verminderen van de vereiste inspanning voor het uitvoeren van experimenten. De lijst van open onderzoeksvragen is opgesteld aan de hand van de positionering van dit proefschrift ten opzichte van eerder gerelateerd werk. Deze vragen dienen als een vertrekpunt voor nader onderzoek naar interpretabele crowdsourcingtaken, waarop de te vinden antwoorden kunnen bijdragen aan de ontwikkeling van zoeksystemen met meerdere perspectieven op multimedia.

I

PRELUDE

1

INTRODUCTION

The multimedia landscape has changed considerably since the turn of the century. With the advent of more personal processing power, the increasing availability of bandwidth and recording devices, a shift has taken place from traditional one-to-many programming (the paradigm of traditional television) to many-to-many creation of diverse content. Whereas previous technological advances in communication, such as the printing press, telephony and television, allowed human communication to transcend space and time, the rise of the Internet allowed for personalized reception of media [27]. With the emergence of Web 2.0 and content hosting services like Flickr and YouTube, it has become possible for everyone to author content and connect with new audiences [20]. Content creation is no longer necessarily tied to considerations of mainstream reception. In other words, it is no longer obligatory to ensure that a mainstream audience is indeed addressed before making a decision to produce content. In a sense, mass communication has been democratized by the Internet, as it is no longer a privilege solely available to larger companies.

This shift in the multimedia landscape is evidenced in changing media consumption patterns. We nowadays spend less time watching televsion and spend more time on the Internet [57]. We now watch whatever we want, whenever we want and wherever we want [69], and are no longer confined to a fixed location and fixed times as was the case with traditional television. The emergence of Web 2.0 has made a vast variety of content covering a broad spectrum of subjects, some highly specialized, available. This variety is a result of the ease with which content can be produced and disseminated. It is now possible for everyone to become a content creator. Creators who focus on niche topics are able to connect with and serve niche audiences.

The fact that everyone can become a creator raises new issues for multimedia retrieval systems. The quantities and diverse topics of content now available pose a challenge for multimedia information retrieval. This new challenge of retrieving the right multimedia item for users in order to accommodate their needs has three parts to it: We can view this challenge from a data angle, a user angle and a technology angle.

From the data angle, we are facing an increase in both variety and volume of multimedia content. By looking at one particular example, we can already see that an abundance of content is being unleashed onto the Web at an astonishing rate. Just on YouTube alone, an extremely popular video sharing platform, hundreds of hours of video content are uploaded every single minute [116], spanning from home videos, video blogs, cat videos to professional-grade, well-edited mini-series, reviews and instructional videos.

From this spectacular increase in variety and volume follows a matching change from the user angle. Because more and different multimedia content can be found on the Internet, users are able to demand finer-grained perspectives that also reflect their points of view. We make the assumption that this potential for demand translates into an actual demand, both present and future. Several sources of evidence indicate that user needs are developing to encompass a wider topical scope and a need for fine-grained access:

1. The physical reality that you can only scan through so much video to get to the parts that you need.

2. What people explicitly write that they are looking for, e.g., on question answering forums [35];

3. What people explicitly write that they are watching, e.g., mentioning time-codes in comments on a video [103] (covered in more detail in Chapter 5);

4. As a special case of (3), people telling other people where to start watching;

We now move to discuss the technology angle. Considering the data and user angles just discussed, we will see that from a technical point of view existing multimedia retrieval systems have not kept up. More background information on retrieval systems will be provided in the next section, but, in general terms, the increased volume of content and the more demanding user needs affect what and how a retrieval system should store information about this content in its index in order to be able to effectively find the content a particular user wants. Without proper index information, a system would be unable to meet its users' demands. Current systems are clearly struggling. For one, the sheer volume of content necessitates that we index multimedia content at a time-code level in a way that is more closely coupled to user needs. While dedicated labeling of content could bring us closer to storing proper metadata, the amount of work required due to sheer volume renders this approach infeasible. Hence, some form of automatic indexing is necessary, but a fully automatic one-size-fits-all approach is not appropiate. We need a closer coupling of indexing features with what users are actually looking for. This raises the question what a system should index. Which segments of a multimedia file are special and what makes that segment special compared to the whole volume of all content? Hence, we need to generate indexing features that differentiate media at a fine-grained level in order to meet the user's need.

In order to move from a one-size-fits-all approach to multimedia information retrieval to a multi-perspective approach, one has to know which perspective to take. Relying on our own ideas of what people are looking for does not give us the wide variety of needs that are presumably in existence. Instead, we need fresh ideas on what people generally look for in multimedia content. One obvious source of a new perspective is the

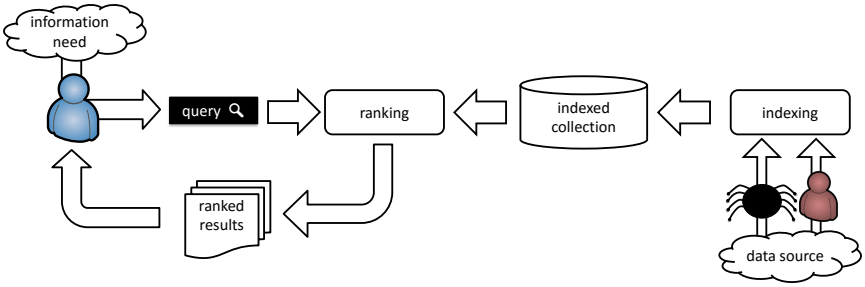


Figure 1.1: Classical high-level overview of a multimedia information retrieval system.

users of multimedia retrieval systems themselves. However, it is not always feasible or easy to consult with these users. So for this reason, in this thesis, we have singled out the crowd as a source of information on which sorts of perspective would be helpful for users and as a way to complement any form of automatic indexing. We consider the crowd as a source of perspectives that will resemble the perspectives of future users of a system if the crowd is carefully addressed.

The central question this thesis concerns itself with is *how can we incorporate the perspectives of the crowd into multimedia retrieval systems?* We will further expand on this question and related concepts in the remainder of this introductory chapter. The contributions of this thesis are methods for obtaining a more productive merger between the use of crowdsourcing, collective intelligence and social computing.

In the remainder of this chapter, we will first provide background information on multimedia retrieval systems (Section 1.1) and elaborate what we mean by "the crowd" in the central question of this thesis (Section 1.2). We will then discuss the philosophy behind the methodology we developed and used in this thesis (Section 1.3) and end the chapter by stating the contributions and the outline of this thesis (Section 1.4).

1.1. MULTIMEDIA RETRIEVAL SYSTEMS

In this section, we provide the necessary background information on multimedia retrieval systems. We first cover the basics of any retrieval system (Section 1.1.1). We then define the notion of non-linear video access, a theme that recurs throughout this thesis, and how the concept differs from how most common multimedia systems treat videos (Section 1.1.2). Finally, we discuss how multimedia retrieval systems can be evaluated, which challenges may arise, and how crowdsourcing can be used to address these challenges (Section 1.1.3).

1.1.1. BASICS OF A RETRIEVAL SYSTEM

An information retrieval system allows a user to search for information in a particular collection of documents. In case of a multimedia retrieval system, the collection consists of multimedia documents, such as video files. The schematic of an information retrieval system is shown in Figure 1.1. A user consults a retrieval system in case he or she has an *information need*. The need for particular information is expressed by the user in the

form of a *query*, which is submitted to the retrieval system. The query typically takes the shape of a sequence of keywords. However, some systems allow the user to express information needs differently, for instance by supplying the system with an example image, for which the system then tries to find images that are similar. Although it is not depicted in the diagram, the retrieval system could perform additional *query processing*, in which the query is analyzed to better infer what the user is looking for (e.g., expanding the query with additional keywords or correcting misspelled words). The query is used by the system to retrieve relevant documents from the *indexed collection* and *rank* the results accordingly.

RANKING

For each of the documents in the collection, a *relevance score* is computed. This score indicates how well a document matches the query based on some set of *relevance criteria*. A simplified example of computing a relevance score is to count the number of matching query keywords in the document's title. A document whose title contains more keywords from the query would be in this case considered more relevant. In this case, it is assumed that this particular score reflects *topical relevance*. Topical relevance is the criterion that most retrieval systems use by default. That is, the system assumes the user's query describes a particular topic and tries to match that to the documents stored in the system's collection.

After retrieving the initial top k results, a system could potentially improve the list of search results by *reranking* it. That is, it could rank results within this list according to a secondary criterion. For example, a system could diversify the top results [50, 83].

INDEXING

The collection of a retrieval system consists of documents that have been added by an automatic crawling process or manually by a human curator. An automatic crawling process, commonly called a *crawler* or spider, is provided with a data source, typically the World Wide Web, and crawls through this wealth of data to find new documents that could be added to the system's collection. Such a collection is dynamic and updated all the time. Test collections are usually created by a human curator. Most of the collections used within this thesis could be considered to be human-curated.

Documents that are added to a system's collection are *indexed*. That is, characteristic features of a document, which are related to the relevance criteria that the search system supports, are extracted and are used to build and update the retrieval sytem's *index*. The index allows for the system to quickly select relevant candidates from the collection without having to examine every single item.

1.1.2. NON-LINEAR VIDEO ACCESS

One recurring theme in this thesis is *non-linear video access*. Here we discuss non-linear video access with a focus on aspects that are relevant to this thesis. Most multimedia information retrieval systems consider each multimedia document in their collection as the smallest retrieval unit. That is, the results are presented as a list of relevant documents in response to the user's query. However, in the case of video content, one can imagine that for certain information needs, a user might be better served with a list of

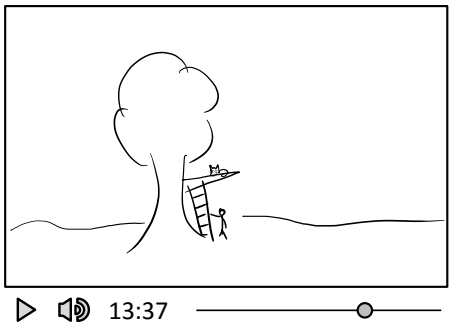


Figure 1.2: Common conventional video players allow the user to jump to any point in the video at any time by using the seek bar. Even though this is a form of non-linear video access in the strictest sense, this thesis considers it to be too trivial as the bare seek bar does not further aid the user in finding relevant content.

video fragments or jump-in points rather than a list of whole videos. This ability to directly access any part of a video is what we call non-linear access.

The term non-linear access is derived from the field of video editing. In the early days of this field, video editing was limited by the linear nature of reels and tapes that were used as the storage medium. It was inherently necessary to sequentially play back the tape's content. If you wanted to broadcast two non-adjacent parts of a recording on a video tape, you would first have to resort to editing and record these two parts to a new tape. With the advent of more computing power and the increasing amount of economically available random access storage, *non-linear video editing* became feasible. Video content could now be stored digitally and any frame could be accessed in any order at any time.

In most multimedia retrieval systems, when a user has selected a video from the results to be played back, he or she can use the video player controls to jump to any point in the video (see Figure 1.2). Strictly speaking, this is a form of non-linear video access. However, we argue that this form is rather trivial and we therefore refine our earlier definition of non-linear video access. In this thesis, whenever we are referring to non-linear video access, we refer to systems that aid the user in finding parts of a video by providing the user timeline-specific information about these parts of the video in some way. This definition excludes systems that allow users to play a video through a conventional video player (as depicted in Figure 1.2) as such systems do not provide additional information about the time-points users can jump to.

Looking at the general pipeline of a retrieval system (Figure 1.1), we can identify at least two ways in which a multimedia retrieval can support non-linear video access:

- Using a smaller unit of retrieval as the basis of the ranked results, e.g., returning a list of jump-in points for videos rather than videos as a whole.
- Enhancing the visualization of a result once a user has selected it from the ranked results list, e.g., providing a more informative time bar in the video player.

Examples of how a system can support the user in accessing video in a non-linear fashion are depicted in Figure 1.3.

Figure 1.3: Examples of how a retrieval system can support non-linear video access. A system could visualize measures of interestingness as a heat map on the seek bar. Potential noteworthy time-points could be derived from user comments and shown on the timeline or below the player. The system could also choose to use noteworthy points as the smallest retrieval unit rather than whole videos, of which a concept is presented later in this thesis (Chapter 5).

Additionally, a system could provide non-linear video across videos in the collection through *anchors*. Anchors are media fragments, e.g., fragments defined by their start and end time, for which users could require additional information [22]. These anchors could be used to retrieve other multimedia fragments and allow the user to explore a video collection in a non-linear fashion [73].

In order to support non-linear access, a multimedia retrieval system needs to extend the collection index with additional information on videos at a time-code level. Two possible, but not mutually exclusive, approaches can be taken here. One approach is media-centric, the other is user-centric. In the media-centric approach, video content of a file can be analyzed when the file is added to the system's collection. This analysis has to be only carried out once. For example, a retrieval system for soccer video content could perform highlight detection based on features that correlate with arousal (e.g., [37]) such as demonstrated in [89]. On the other hand, the user-centric approach requires the system to continuously collect user activity regarding each video document in the collection and use this information to update its index. For example, a system could analyze which parts of a video are viewed more often or commented on by its users to find what is considered noteworthy by them.

1.1.3. EVALUATION OF MULTIMEDIA RETRIEVAL SYSTEMS

How can we measure how good a retrieval system is at retrieving documents? Traditionally, text-based information retrieval systems are evaluated using a *test collection* and a set of *test queries*. By running each test query on the test collection, we can measure the system's *precision*, i.e., the fraction of the returned results that are actually *relevant*, and the system's *recall*, i.e., the ratio of relevant results returned and the total number of rel-

evant items in the test collection [64]. A panel of judges, or assessors, carry out *relevance judgments* in order to determine which documents in the whole collection are relevant for each query. For large collections, instead of exhaustively assessing whether each document is relevant, which is labor intensive and expensive, *pooling* is usually carried out to limit the number of assessments needed: from each of the retrieval system under evaluation, only the top k returned results for each query are assessed [64].

Multimedia retrieval systems as as whole can be evaluated in a similar way. Since relevance assessments can be quite expensive to carry out, researchers often rely on existing test collections or join benchmark initiatives to work collaboratively on a particular multimedia task to alleviate this particular cost [40]. The only limitation of this approach is that it requires the researchers' aim to fit the test collection's nature or to match the benchmark's task.

For most of the work presented in this thesis, no suitable test collection or benchmark could be found, the reason being that most of the work explores new ways of interacting with a retrieval system or uses sources of data generally not found in dataset. Due to this unavailability of test data and benchmarks, custom datasets had to be created for this thesis and alternative evaluation methods were needed such as evaluating aspects of multimedia retrieval systems through user studies. The user studies were carried out on *crowdsourcing* platforms as a way of having access to a large pool of individuals (see next Section for a discussion on crowdsourcing). Through user studies, we could sketch particular retrieval scenarios, pose questions regarding multimedia and gather preferences on alternative retrieval and ranking algorithms by presenting participants, e.g., different results lists.

Crowdsourcing was also a tool in the creation of new datasets in this thesis. Data collected and mined from multimedia platforms, such as YouTube, could be labeled without it being too labor intensive for a small group of individuals by distributing the workload through crowdsourcing. Crowdsourcing also helped in gathering people's interactions with multimedia. These collected multimedia interactions enrich the dataset and are a reflection of understanding what the world is like. We expand further on this notion of common understanding among users in Section 1.3. Collecting this data, such as user interactions, also sheds light on to what extent these interactions would be useful in a deployed multimedia retrieval system as it allows us to evaluate aspects such as how quickly the collected data would converge and whether the collected data is universal.

1.2. THE CROWD AND COLLECTIVE INTELLIGENCE

Recall the central question of this thesis posed at the beginning of the question: How can we incorporate the perspectives of the crowd into multimedia retrieval systems? Here, it is important to make clear what we mean by the crowd. Within this thesis, the term will be used in a quite general fashion to refer any large group of individuals, ranging from users of a specific system to people on social media. We use several other terms in this thesis for which it is also important to clearly define them up-front. The definitions used throughout this thesis are generally adopted from the taxonomy that Quinn and Bederson have proposed [79], which is depicted in Figure 1.4. We will first present the taxonomy as defined by Quinn and Bederson below and refine its definitions afterwards for the purpose of this thesis.

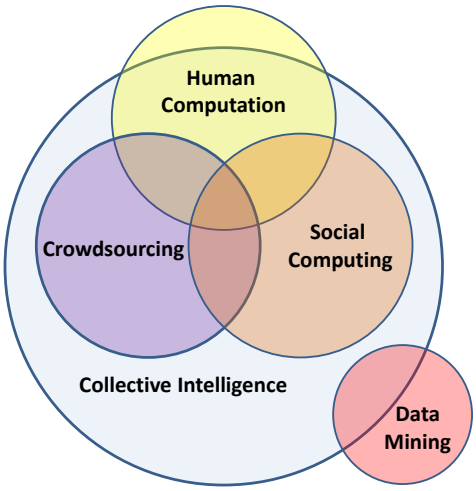


Figure 1.4: The human computation taxonomy as defined by Quinn and Bederson distinguishes between different forms of collective intelligence [79].

Human computation is the central concept in [79] and is defined as a computational process that directs human participation in order to solve a problem that fits the general paradigm of computation and might one day be solvable by computers.

Social computing is about facilitating social interactions between humans. Examples include technologies such as blogs, wikis and online communities. Social computing mainly differentiates itself from human computation with respect to the factor that is driving human behavior. In the case of social computing, this behavior is relatively natural as opposed to it being directed by the system in a human computational process. Even though the purpose is usually not to perform a computation, social computing can result in the evolution of aggregate knowledge [74, 79].

Crowdsourcing is a term coined by Jeff Howe and is derived from the term outsourcing [41]. We find a comprehensive definition of crowdsourcing in [23]: *"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task."* The user in the crowd is rewarded, either economically or socially, for answering the call. The crowdsourcer benefits from the crowd-provided labor, funds, knowledge or experience. In this thesis, we will often use the term *requester* to refer to the crowdsourcer and *worker* to refer to the user.

Data mining is simply the extraction of patterns from data using a specific algorithm. Even though patterns can be mined from data created by humans, Quinn and Bederson argue that they do not consider it to be a form of human computation [79].

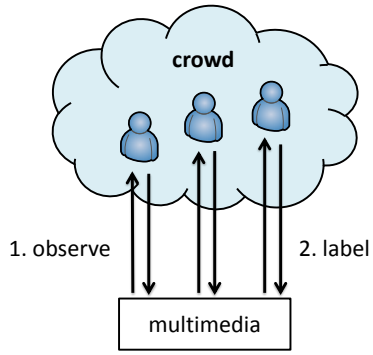


Figure 1.5: Classic view on the crowd for crowdsourcing: A collection of individuals observing and labeling multimedia data.

Collective intelligence is very broadly defined as *"groups of individuals doing things collectively that seem intelligent."*[63] It encompasses everything discussed so far, except in the event when human computation concerns a single individual, a case for which no well-developed examples are known, and in the event when data mining does not concern data created by groups, as depicted in Figure 1.4.

The work in this thesis concerns itself mainly with all things that fall under collective intelligence as defined above. We are not specifically interested in the finer details of the taxonomy as presented by Quinn and Bederson, but we would like to treat crowdsourcing separately. From now on, whenever we refer to collective intelligence, we intend to designate everything that can be considered as collective intelligence, as previously defined, except for crowdsourcing. When referring to crowdsourcing, we will use that term explicitly. The reason for our distinctive use of terms comes from the nature of motivation. Generally, people consume and interact with multimedia for intrinsic reasons. In case of crowdsourcing applied to the multimedia domain, while workers may be intrinsically motivated to look for tasks to work on, the motivation to consume multimedia does not come from within but is initially external. Recall, however, that we do not reserve the word crowd to be used exclusively in the context of crowdsourcing, as stated at the beginning of the section, but instead use it to refer to any general large group of users of any system or platform, which could potentially include workers of a crowdsourcing platform.

1.3. MENTAL MODELS

One of the dominant assumptions held by many is that a limited personal observation is informative. When studying multimedia content, we as researchers make personal observations and are inclined to decide which aspects of multimedia are important. We are tempted to formulate assumptions, but we do not know whether these assumptions hold at large. As single individual researchers, we might be missing the bigger picture. Even when we decide to involve the crowd to rely on their reputed wisdom, the crowd

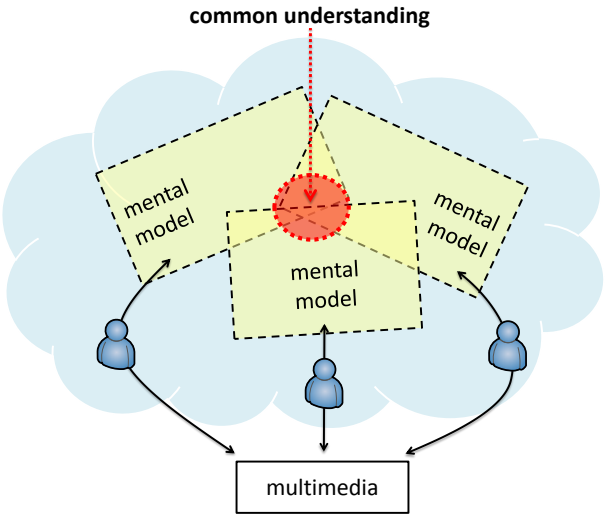


Figure 1.6: This thesis' view on the crowd: The crowd is not treated as merely a collection of individuals, but instead as a collection of mental models used by these individuals that overlap with each other. These mental models are activated when interacting with multimedia content.

is often treated as a bunch of individuals who are good at observing a large corpus of multimedia items and labeling these items in one way or another (Figure 1.5). With such a particular attitude, mismatching views from the crowd are often discarded. As a result, we run the risk of conformism and inadvertently creating a subservient crowd, leading to a situation in which it is hard to discover new perspectives on multimedia. This is why in this thesis, when reaching out to the members of the crowd for their perspective, we do not merely view them as a collection of individuals, but instead as a collection of various *mental models* that these individuals use when they are interacting with multimedia (Figure 1.6).

A mental model is an abstract representation that we have in our head. As Forrester described in [25]:

> *"The mental image of the world around you which you carry in your head is a model. One does not have a city or a government or a country in his head. He has only selected concepts and relationships which he uses to represent the real system."*

As Forrester further describes in his work, we use these mental models when making decisions. These models help us to predict behavior of systems and manage our expectations and act upon these expectations. These mental models also play a role in how we perceive multimedia. In the case of multimedia, mental models help us to interpret the semantically complex content. For example, it is how we view the world that impacts what we could potentially find noteworthy when we are viewing a video stream. When we see a recording of some extraordinary feat that exceeds our expectations, it is precisely because our mental models did not predict our observations.

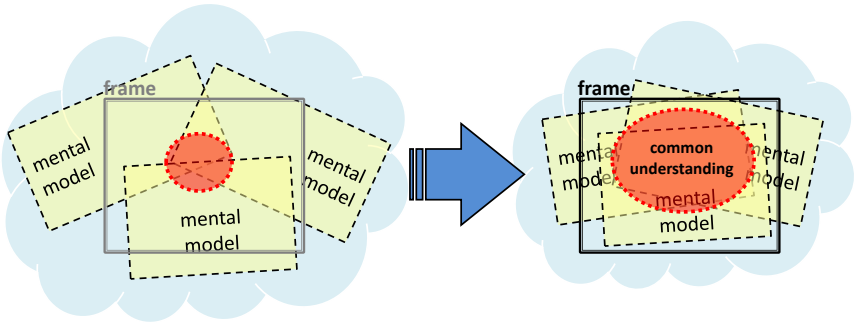


Figure 1.7: The importance of framing: A frame of reference helps to draw someone into a particular, contextual mindset and align or select a fitting mental model. In a sense, it helps to get everyone on the same page.

1.3.1. COMMON UNDERSTANDING

These mental models we employ as individuals when viewing the world are likely to be different from each other, each individual having their own set of assumptions and opinions. However, our mental models do overlap. We share certain expectations of the world around us. It enables us to start a conversation about a particular topic and have a discussion, which would be impossible if our mental models would not overlap. In this thesis, we call the overlap of mental models the *common understanding* on a particular subject (Figure 1.6).

However, it is important to note that mental models are fuzzy, incomplete, and in flux:

> *"The mental model is fuzzy. It is incomplete. It is imprecisely stated. Furthermore, within one individual, a mental model changes with time and even during the flow of a single conversation. The human mind assembles a few relationships to fit the context of a discussion. As the subject shifts so does the model." [25]*

First, this should make us realize that we can only approximate the mental model of a user as people are unable to precisely describe their own mental models [70]. We can only pose carefully formulated questions through which the users' mental models will be reflected in their answers. Second, since the active mental model depends on the context perceived by the human mind, the perceived context influences the degree of common understanding as well. It is therefore important to get everyone on the same page, something that is seemingly effortless in a face-to-face conversation, but much harder in a crowdsourcing setting.

1.3.2. FRAMING

In order to get everyone on the same page, we have devoted a great deal of attention to formulating the task description used in the crowdsourcing campaigns that we carried out throughout this thesis' work. By formulating a *frame*, we are providing a certain context with the purpose of triggering a particular mental model and making sure people are

responding and interacting within that mental model. By framing the task in a certain way, we want to avoid people resorting to a default mental model, whatever that may be. A specific danger of a default mental model is that it is a mental model that reflects adjustment of the user to the projected expectations of the researcher asking the questions rather than a mental model that resembles a real-world scenario. More concretely, we want to avoid thoughts like "I think the researcher meant this with this question". Instead, it should feel like the researcher is not there at all when the user is carrying out a task. The intended effect of framing is depicted in Figure 1.7. By providing an explicit frame, we are providing the context clues that allow people to shift their mental models towards the context.

1.3.3. CAPTURING COMMON UNDERSTANDING

Our goal is a picture of the world that is a common understanding. In order to obtain this picture, we have to go beyond our own individual perspective. This assumption is the starting point of our discussion on how to capture common understanding in this section.

A common understanding on a particular topic includes variety of perspectives, but these perspectives need to be sufficiently similar in order for people to have the idea that they are talking about the same thing. In everyday life, when we try to develop our own mental models for the purposes of talking to other people, and in general effectively interacting with the world around us, we use several approaches, ranging from introspection to active elicitation. We could simply use introspection, but most of the time introspection is only useful for painting an initial picture. Soon, we consult what we consider expert voices to widen our horizon, read or listen closely to explicit expressions formulated by others, actively elicit opinions and views from others on a particular subject, or carefully observe implict interactions of others and their behavior.

These same approaches are also applicable to conducting research for trying to get a picture of the world. A researcher could stick to introspection. For example, a researcher could decide what is important by simply watching multimedia content and making a personal judgment. However, thinking about oneself is not necessarily the best way of thinking about how others think. Instead, a researcher in the domain of multimedia retrieval would be better off applying one of the other four approaches:

1. Expert opinion: Consult some experts in the relevant domain to decide what is important;

2. Explicit expressions: Collect comments or messages on social media that users have posted in response to a multimedia document and find out what they found worth commenting on;

3. Active elicitation: Ask people questions about multimedia that reveal their mental models which determine how they interact with information in the domain;

4. Implicit interactions: Collect the way people interact with multimedia content, e.g., video viewing patterns, and move from there to what is important.

These approaches make it possible for a researcher to observe and collect to a varying degree different views from people involved in a scientific study, more so than mere introspection could. As we move closer to the bottom of the list, the approaches get more sophisticated, since motivating people to respond and interpreting the responses are the key challenges.

The underlying philosophy in this thesis is to shift from introspection to the approaches listed above in order to capture the various views, based on people's various, yet commonly aligned, particular mental models. These mental models are people's viewing angles, i.e., their perspectives, on multimedia and using these approaches we can collect and then incorporate these perspectives in designing and fueling multimedia retrieval systems.

1.4. CONTRIBUTIONS OF THIS THESIS

The contributions of this thesis consists of a collection of methods for obtaining a more productive merger between the use of crowdsourcing, collective intelligence and social computing. The methods were developed to answer the thesis' central question *'How can we incorporate the perspectives of the crowd into multimedia retrieval systems?'*. Underlying these methods are the principles and ideas discussed in the previous sections.

1.4.1. THESIS OUTLINE

This thesis consists of four parts and is organized in a way such that as it progresses through the chapters, collective intelligence and crowdsourcing is used in an increasingly more sophisticated way. After the initial chapter showcasing a classic example of mining collective intelligence and a conventional evaluation using crowdsourcing, the subsequent chapters move on using crowdsourcing in novel ways. In these chapters, crowdsourcing methodologies and new ideas for multimedia retrieval systems are developed in tandem. This joint-development emerged naturally through the need of evaluating and informing the design of these features.

PART I: PRELUDE

The prelude of this thesis, as the name implies, is the part of the thesis that prepares the reader for the parts that follow.

Chapter 1 (this very chapter) provides the necessary background information and the general philosophy of the methodologies developed within this thesis. Its function is to offer sufficient guidance to the reader by setting up the frame of context, describing the motivation and organizational framework of the thesis.

Chapter 2 sets the stage by presenting an example. In this chapter, we present an example of leveraging social computing by mining social networks in order to solve a non-linear video access problem. We start the chapter with an assumption that questions posted on microblog platforms such as Twitter are indicators for finding suitable anchors in broadcast videos. Our assumption is then tested using a classic case of crowdsourced relevance judgment.

PART II: FRAMING AND ELICITATION METHODOLOGY

The second part of the thesis covers the core methodologies we have developed and used throughout the crowdsourcing studies as a basis.

We lay down the foundation of our framing methodology for crowdsourcing tasks in **Chapter 3A** and present its first application in **Chapter 3B**. The framing methodology, which helps people to picture a scenario, even one that is outside of their typical daily life, is applied to designing a crowdsourcing user study for evaluating a new feature in a multimedia file-sharing application. The methodology's goal is to obtain high quality answers from the study participants and ecologically valid results and we make the case that crowdsourcing tasks may not have the concept of "wrong answers" (Chapter 3A). Following the discussion of the methodology, we present the design of the feature that should aid new users in discovering content available in the multimedia file-sharing application (Chapter 3B). With the help of domain experts, the user study designed in the preceding chapter is used to evaluate the new feature.

We continue the development of crowdsourcing methodologies in **Chapter 4A and Chapter 4B**. In these twin chapters, we focus on the refinement of the elicitation techniques in order to effectively model the common understanding on a particular topic. In particular, we are interested in discovering new similarity dimensions to inform the design of a feature in a multimedia retrieval system. The new feature targets the problem of near-duplicates in multimedia information retrieval systems. However, without knowing what users consider to be a near-duplicate, the utility of the feature would be limited. By asking the crowd the right questions on a large diverse sample of multimedia files, we are able to discover various views on multimedia that we would not have possibly found with introspection alone (Chapter 4A). Taking some of those views to implement the new feature, we then showcase that crowdsourcing can be employed to test a user interface of an actual software product and show that it is a sufficiently robust method to carry out A/B feature testing (Chapter 4B).

PART III: ADVANCING NON-LINEAR ACCESS TO VIDEO CONTENT

The penultimate part of this thesis builds upon the methodologies developed in the previous part and uses it push the research on non-linear video access further.

In **Chapter 5**, we take all lessons that we have learnt from crowdsourcing and apply them to discover new dimensions of relevance at the the video time-code level. We build a crowd-informed typology that categorizes a particular type of user comments, namely user comments that explicitly refer to a specific point in a video, by mentioning time-codes with the purpose of using these comments to support non-linear video access. The chapter showcases the combination of using explicit expressions and active elicitation in order to arrive at the crowd-informed typology. The new dimensions of relevance stemming from the typology are tested in a carefully framed crowdsourcing user study in which the participant has to picture a certain search scenario.

In **Chapter 6**, we combine the use of active elicitation and implicit interactions. We develop a methodology for collecting realistic implicit information via a crowdsourcing platform that is outside a normal platform in which real implicit behavior usually takes place. In this case, it is not only important that the task is properly framed, but also that the user population and multimedia domain are jointly chosen in order to observe

implicit behavior that resembles behavior that participants would exhibit if they were to view the multimedia content on their own. We demonstrate through a crowdsourcing user study that implicit viewing behavior can be used to support non-linear video access.

PART IV: OUTLOOK
In the final part, we conclude the thesis with an outlook on how interpretive crowdsourcing could be used in future research in **Chapter 7**. Here, we focus on surveying crowdsourcing work that does not confine the allowable responses to a crowdsourcing task to a single set of "correct" answers, but instead frames tasks in a way that are still open to interpretation. We contextualize our work and identify the challenges that lie in using interpretive crowdsourcing.

1.4.2. FULL LIST OF PUBLICATIONS

In the years leading up to this thesis, the following papers have been published:

12. **Raynor Vliegendhart**, Martha Larson, and Alan Hanjalic. Collecting realistic viewing behavior from the crowd for non-linear video access. Under review [100]. **—[Chapter 6]**

11. **Raynor Vliegendhart**, Cynthia C.S. Liem, and Martha Larson. Exploring microblog activity for the prediction of hyperlink anchors in television broadcasts. In *CEUR Workshop Proceedings, no. 1436, 2015. MediaEval 2015 Workshop, Wurzen, Germany, 15-15 September, 2015*. CEUR-WS, 2015 [105]. **—[Chapter 2]**

10. **Raynor Vliegendhart**, Martha Larson, Babak Loni, and Alan Hanjalic. Exploiting the deep-link commentsphere to support non-linear video access. *IEEE Transactions on Multimedia*, 17(8):1372–1384, 2015 [103]. **—[Chapter 5]**

9. Michael Riegler, Mathias Lux, Vincent Charvillat, Axel Carlier, **Raynor Vliegendhart**, and Martha Larson. Videojot: A multifunctional video annotation tool. In *Proceedings of International Conference on Multimedia Retrieval*, page 534. ACM, 2014 [81].

8. **Raynor Vliegendhart**, Babak Loni, Martha Larson, and Alan Hanjalic. How do we deep-link?: Leveraging user-contributed time-links for non-linear video access. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 517–520, New York, NY, USA, 2013. ACM [106].

7. Eelco Dolstra, **Raynor Vliegendhart**, and Johan Pouwelse. Crowdsourcing GUI tests. In *IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST)*, pages 332–341, March 2013 [18].

6. Babak Loni, Maria Menendez, Mihai Georgescu, Luca Galli, Claudio Massari, Ismail Sengor Altingovde, Davide Martinenghi, Mark Melenhorst, **Raynor Vliegendhart**, and Martha Larson. Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 72–77. ACM, 2013 [59].

5. **Raynor Vliegendhart**, Martha Larson, and Alan Hanjalic. LikeLines: collecting timecode-level feedback for web videos through user interactions. In *Proceedings of the 20th ACM international conference on Multimedia*, MM '12, pages 1271–1272, New York, NY, USA, 2012. ACM [99].

4. **Raynor Vliegendhart**, Eelco Dolstra, and Johan Pouwelse. Crowdsourced user interface testing for multimedia applications. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 21–22. ACM, 2012 [98]. **—[Chapter 4B]**

3. **Raynor Vliegendhart**, Martha Larson, and Johan Pouwelse. Discovering user perceptions of semantic similarity in near-duplicate multimedia files. In *Proceedings of the 1st International Workshop on Crowdsourcing Web Search*, pages 54–58. CEUR-WS.org, April 2012 [104]. **—[Chapter 4A]**

2. **Raynor Vliegendhart**, Martha Larson, Christoph Kofler, and Johan Pouwelse. A peer's-eye view: network term clouds in a peer-to-peer system. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1909–1912. ACM, 2011 [102]. **—[Chapter 3B]**

1. **Raynor Vliegendhart**, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse. Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load. In *WSDM'11 Workshop on Crowdsourcing for Search and Data Mining*, February 2011 [101]. **—[Chapter 3A]**

2

EXPLORING MICROBLOG ACTIVITY FOR THE PREDICTION OF HYPERLINK ANCHORS IN TELEVISION BROADCASTS

This chapter showcases the use of social computing and a classic case of crowdsourcing used for collecting relevance judgments. In this chapter, we present a social media based approach to finding anchors in video archives. We use social activity on Twitter to find topics on which people have questions about in order to select suitable anchors. The experiments were carried out on the MediaEval Search and Anchoring in Video Archives Task (SAVA) data set of 2015, consisting of 68 hours of BBC video content broadcasted in 2008. The performance of our relatively simple, but straightforward method seems sufficiently promising to pursue further research.

This chapter is an extension of **Raynor Vliegendhart, Cynthia C.S. Liem, and Martha Larson. Exploring microblog activity for the prediction of hyperlink anchors in television broadcasts. In *CEUR Workshop Proceedings, no. 1436, 2015. MediaEval 2015 Workshop, Wurzen, Germany, 15-15 September, 2015.* CEUR-WS, 2015** [105].

2.1. INTRODUCTION

One of the research questions of the 2015 SAVA task, and the one that is being addressed in this paper, is how to automatically identify anchors for a given set of videos, where anchors are media fragments for which users could require additional information [22].

Microblog platforms, such as Twitter,[1] reflect social activity that takes place around a TV show at the time that it is broadcast. Our approach is based on the idea that users will want to learn further information on segments that discuss topics that trigger questions. We use activity on Twitter to find which topics trigger user questions. The more Twitter questions associated with topics discussed in a certain shot, the greater we consider the likelihood that the corresponding part of the video represents a viable anchor. Our approach is further based on keyphrase mining. We understand keyphrases in the sense of [49], namely, as noun phrases that capture the main content of a document. We make the simplifying assumption that the relationship between questions and shots is reflected in the number of keywords that they share in common. We also assume that the presence of a question mark in a tweet on Twitter indicates the presence of a question.

2.2. METHOD

Our approach to anchor generation in broadcast videos exploits social chatter about topics on the microblogging platform Twitter. The method requires that subtitles and shot boundary information for a video are available. Anchors for a given video are then generated as follows.

For each subtitle $s = (l, t_s, t_e) \in S$ consisting of a line of text l, a start time t_s, and an end time t_e, we extract set of keyphrases K_s using `nltk` [7] and a chunker from [49]. The set of all keyphrases is then denoted as $K = \bigcup_{s \in S} K_s$.

For each shot defined by its boundaries $b = (b_s, b_e) \in B$, consisting of a start time b_s and end time b_e, we introduce the notion of a subset $S_b \subseteq S$ representing all subtitles that start in that shot: $S_b = \{s \mid s = (l, t_s, t_e) \in S \wedge t_s \in [b_s, b_e)\}$. With these definitions set in place, we can now define all keyphrases that occur in a shot b:

$$K_b = \{k \mid s \in S_b \wedge k \in K_s\}.$$

To determine the importance of a keyphrase term $k \in K$, we retrieve tweets containing a question about the keyphrase by sending the following query to Twitter: "k ? `since:`d_s `until:`d_e". In this paper, we use a fixed date range corresponding to the given set of videos for all keyphrases, regardless of the airing date of the video. This means we retrieve questions of social relevance during that general time period, not "the issues of today" in the past. Let $q : K \to \mathbb{N}$ denote the function to count the number of tweets retrieved for a keyphrase $k \in K$. The weight of each keyphrase k is then determined by a weighing function $w : K \to \mathbb{R}$. This function w is of the form $w(k) = f(q(k))$ where f is implementation-dependent and its purpose is to scale the tweet count returned by q.

We then rank shots by the summed weight of all keyphrases appearing in each shot. Let $W : B \to \mathbb{R}$ denote this summation: $W(b) = \sum_{k \in K_b} w(k)$. Then $r : B \to \mathbb{N}$ denotes the function that assigns a rank to a shot defined by its boundaries $b \in B$:

$$r(b) = 1 + \left|\left\{b' \mid b' \in B \wedge W(b') > W(b)\right\}\right|.$$

[1] `https://twitter.com`

After ranking the shots, we simply generate anchors of an arbitrarily chosen fixed length, which is at minimum $T = 30$ seconds, using shot boundaries for alignment. The underlying assumption is that cutting at shot boundaries should result in clean media fragments. Let $a : B \to \mathbb{R}^2$ denote the function that computes the start and time of the anchor derived from a shot $b \in B$. The start time is equal to the start time of the shot itself, i.e., b_s, and the end time is equal to end time of the first shot that ends at least $T = 30$ seconds later, or the end time of the whole video:

$$a(b) = (b_s, \min\left(\begin{array}{c}\{b'_e \mid b' \in B \wedge b_s + T \in [b'_s, b'_e)\} \\ \cup \max\{b'_e \mid b' \in B\}\end{array}\right)).$$

2.3. EXPERIMENTS

2.3.1. DATASET

The dataset used in the SAVA 2015 task is a subset of collection of 4021 hours of video broadcasted by the BBC [22]. This subset consists of a dev set (37 videos, 37 hours) and a test set (33 videos, 31 hours). The experiments presented here make use of manually transcribed subtitles provided by the BBC and use shot boundaries that ship with the SAVA dataset.

2.3.2. SETUP

In this paper, we have tested two weighing functions, w_1 and w_2, each being submitted as a separate run, i.e., experimental condition, for evaluation by the organizers of the SAVA task. The first function takes the popularity of each keyphrase in a shot into account, while the second function only considers the number of different keyphrases. They are defined as follows:

$$w_1(k) = \begin{cases}(\ln \circ q)(k) & \text{if } q(k) > 0 \\ 0 & \text{otherwise}\end{cases}$$

$$w_2(k) = \begin{cases}1 & \text{if } q(k) > 0 \\ 0 & \text{otherwise}\end{cases}$$

Furthermore, the keyphrase extraction used in our method only considered words of length 2 to 40 characters and ignored stopwords (using `nltk`'s default stopwords list) and words that were written in all capital letters. The latter filter was used to ignore words appearing in descriptive subtitles for the hearing impaired, such as "APPLAUSE". This resulted in 37,154 nounphrase candidates for both dev and test set. To reduce the number of queries to be crawled, we pruned the list of keyphrases using the following heuristics:

- The phrase should contain at least one capital letter;
- The phrase may not start or end with a stopword;
- The phrase does not start with a quote;
- The phrase does not contain periods or commas.

The last two heuristics were put in place to deal with tokenization mistakes. Applying these pruning heuristics, we reduced the number of phrases from 37,154 to 5,663.

Table 2.1: Run results averaged over 33 videos in the test set

	Precision @10	Recall	MRR	Unjudged @10	Unjudged @1000
Run w_1	0.55758	0.47496	0.87879	1.21212	6.75758
Run w_2	0.50000	0.43224	0.93939	1.39394	8.39394

Querying Twitter for these 5,663 phrases for the period between 2008-04-01 to 2008-07-31 (corresponding to the original broadcast dates of the dataset) resulted in 66,934 tweets. We did not impose any language or geographic restrictions and issued queries as specified in Section 2.2. In this querying process we used a cut-off point to speed up the crawling process. After collecting more than 150 tweets for a query, the process was stopped. This means that in this experiment $0 \leq q(k) \leq 150$. The software we used to unrestrictedly crawl Twitter is available on GitHub.[2]

2.3.3. RESULTS

Anchors from all submissions made by participants to the 2015 task were pooled and assessed by crowdsourcing workers on Amazon Mechanical Turk [22]. We note that our method itself does not take overlap of anchor segments into account, but that this was addressed by the automatic evaluation of our submission. The results of the two runs are summarized in Table 2.1, showing precision at 10 (P@10), recall and mean reciprocal rank (MRR) averaged over the 33 videos in the test set.

From the results, we can see that run 1, which uses the w_1 weighting function that assigns a higher weight to more popular keyphrases, appears to have a better precision and recall than run 2, which uses the w_2 weighing function that treats all keyphrases equally. Run 2 on the other hand achieves a higher MRR. However, in both runs, the MRR never drops below 0.5, indicating that the first relevant anchor is always amongst the first two results.

When we compare our results to submissions from other participants [28, 87], we see that our run w_1 achieved the highest precision and recall and that run w_2 performed best on MRR. The approaches taken by the other participants rely on natural language processing, multimedia content analysis and information retrieval techniques in order to select suitable anchors given a particular broadcast video and its associated metadata. No further external sources were used by the other participants, indicating the use of social data in our approach is promising.

2.3.4. ANALYSIS

It is of most interest to see when our proposed method performs best and when it does not. For this analysis we will look at the correlation between different performance metrics and video features.

Our first observation is that P@10 and recall appear to be positively correlated, with Pearson's r being 0.42 (with $p < 0.02$) and 0.76 (with $p \ll 0.01$) for run 1 and run 2, respectively. This suggests that some videos are easier and others are harder for our method to get right. We found that video length correlate positively with P@10 (0.56 and 0.54 for

[2] `https://github.com/ShinNoNoir/twitterwebsearch`

the respective runs) and that this correlation is significant ($p < 0.01$, for both runs). No conclusions could be drawn for the correlation between video length and recall, however. The answer to why our method seems to favor longer videos could be found in the following two correlations. First, our method tends to perform better (P@10) when the unpruned list of different extracted keyphrases is longer (correlation of 0.58 ($p < 0.001$) and 0.51 ($p < 0.01$) for run 1 and 2, respectively). Second, we can extract almost undoubtedly more keyphrases from longer videos, as the correlation between these two features is significant (0.83 with $p \ll 0.001$ for both runs). Our final observation is that, interestingly, P@10 also seems to correlate with something related to how we score individual segments, namely the sum of the summed keyphrases weights ($W(b)$) of the first 10 results: 0.44 ($p < 0.02$) and 0.48 ($p < 0.01$) for run 1 and 2, respectively.

2.4. CONCLUSION

We have explored a method for predicting anchors in television broadcasts by measuring interrogative activity on microblogs. Using the simplifying assumption that a shot is important if its subtitles contain keyphrases that appear in questions asked on a microblog platform, our method is able to achieve promising performance. Suggestions for future work include the following. We believe that finetuning the keyphrase extraction process and looking at incorporating tf-idf could help with dealing with generic keyphrases that sometimes are extracted (e.g., “Good evening”). Furthermore, investigating the impact of narrowing and broadening a query’s date range seems interesting to see what type questions are important and how to strike the balance between questions which are ephemeral and questions which are evergreen. Last but not least, we could look into personalization of the method, such as localizing the query to a geographic region.

II

FRAMING AND ELICITATION METHODOLOGY

3A

INVESTIGATING FACTORS INFLUENCING CROWDSOURCING TASKS WITH HIGH IMAGINATIVE LOAD

This chapter lays down the foundation of our framing methodology for crowdsourcing tasks. The methodology was born out of necessity, as we discovered that the task we initially put out on crowdsourcing platforms was different from tasks more commonly found on these platforms and that the task did not resonate well with crowdsourcing workers. The chapter introduces the concept of crowdsourcing tasks with a high "imaginative load", a term we use to designate tasks that requires workers to answer questions from a hypothetical point of view that is beyond their daily experiences. We report useful observations made during the design and test of such a crowdsourcing task and find that workers are able to deliver high quality responses to tasks of this nature. However, it is important that the title given to a task allows workers to formulate accurate expectations of the task. Also important is the inclusion of free-text justification questions that target specific items in a pattern that is not obviously predictable. These findings were supported by a small-scale experiment run on several crowdsourcing platforms.

This chapter is published as **Raynor Vliegendhart, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse. Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load. In *WSDM'11 Workshop on Crowdsourcing for Search and Data Mining*, February 2011** [101].

3A.1. INTRODUCTION

Crowdsourcing platforms increase the ease and speed with which new search functionality can be evaluated from a user perspective. In this paper, we take a closer look at issues that arise when a search-related feature is to be evaluated, but has not yet been implemented in working form into the system. The system in question is a file-sharing system. The evaluation takes place as part of the design cycle and has the purpose of allowing us to decide which of several possible realizations of the feature will be most effective for users of the system.

During the course of designing and testing the evaluation task for the crowdsourcing platform, we realized that our task was rather different in an important respect from other, more conventional, tasks carried out by workers on crowdsourcing platforms. Specifically, we needed the workers to be able to project themselves into the role of a user of the file-sharing system and to provide feedback from the perspective of that role. The projection is necessary for two reasons, first, because the system feature that we are evaluating does not yet exist, and second, because our target group of users are general, mainstream Internet users for whom the mechanics of file sharing is rather a stretch beyond their daily online activities. In an initial exploratory phase, we noticed that there was something "special" about our task. Few workers were choosing to carry out the HITs that we published to the crowdsourcing platform, and the batch completion time was longer than was acceptable given the time constraints of our design and implementation process. Our aim was to increase the number of participants in our HIT and also the rate at which new workers took up our HIT without changing the HIT in such a way that would discourage projection or attract cheaters.

In this paper, we report on this investigations that we undertook in order to design a HIT that would achieve this aim. First, we carry out an exploratory analysis of several experimental HIT designs on Amazon's Mechanical Turk (MTurk) and formulate our findings as a series of observations. Then, we build on these observations, performing a small-scale experiment on several crowdsourcing platforms. The experiment tests two aspects of HIT design (title and free-text justifications) that we found helpful for encouraging workers to undertake projection. We refer to tasks such as our evaluation task that require workers to project beyond tangible reality and beyond their daily experience as "crowdsourcing tasks with high imaginative load". We choose the designation *imaginative load* since we see certain similarities with tasks with a high cognitive load (e.g., they take relatively long, cannot be easily routinized and are difficult to carry out in highly distracting surroundings), but have concluded it is not possible to conflate such tasks with high cognitive load tasks, which would typically require using memory or at least some factual recall effort.

The contribution of this paper is a compilation of considerations that should be taken into account when using crowdsourcing for tasks with a high imaginative load, including suggestions for choices concerning HIT design and crowdsourcing platform that make it easier to design effective HITs for such tasks. Notice that we do not report the results of the evaluation itself in this paper. Rather, we concentrate on conveying to readers the information that we acquired during the design of the evaluation tasks that we anticipate will be helpful in design of further tasks.

The paper is organized as follows. In the next section, we discuss related work (Section 3A.2), then we describe the evaluation task (Section 3A.3). In Section 3A.4, we summarize our observations during the design and test of the task. In Section 3A.5, we report on experiments carried out to investigate the impact of the titles and the verification on the behavior of the workers carrying out our HITs. Finally, in Section 3A.6 we offer a summary of our conclusions.

3A.2. RELATED WORK

In this section, we provide a brief overview of crowdsouring literature using techniques similar to ours. Often a crowdsourcing task will use a qualifying HIT to identify a set of workers who are suited to carry out the main task. In [19, 67, 94], recruitment and screening HITs were used to differentiate between serious workers and cheaters. In [45], methods to prevent workers from taking cognitive shortcuts are investigated. Many more workers completed the qualification HIT than returned to complete an actual HIT, an effect we also observe. Senstivity of workers to titles is mentioned in [94], who notice, as we do, that the selection of HIT titles influence their attraction to workers. In [32], experiments were carried out with different titles, pay rates, whether a bonus should be granted, and if so, whether this fact should be communicated to workers or not. Following the evaluation results, workers gravitate towards HITs with "attractive titles", i.e., titles which are easier to understand. In contrast to HITs that explicitly offer an additional bonus, easier-to-understand titles do not imply a high accuracy per worker. Free-text and open-ended response possibilities are often used to check whether workers had an understanding of the task, as in [94]. We make use of a similar approach, in particular asking for justifications of answers. In this respect, our work is related to that of [67], who conducted a subjective study about political opinions by asking workers to justify their given answers in free-text explanations. Giving an opinion often requires a certain degree of projection, which we equate with imaginative load. Note, however, that our task goes beyond asking mere opinions to asking workers to formulate an opinion about a feature that does not yet exist in a use context that is unfamiliar from their daily experience.

3A.3. EVALUATION TASK

Our evaluation task involved assessing the usefulness of a time-evolving term cloud intended to make it possible for users to gain an understanding of the kinds of content that are available within a specific file-sharing system in order to facilitate browsing and search. The term cloud will offer users the possibility to find items within the system, but most importantly it is meant to allow new users unfamiliar with the system to quickly build a mental picture of what kind of content is available via the system. Users should not have to spend extensive time interacting with the system or trying out queries that are frustrating since they do not return results. In order to evaluate whether users have gained an understanding of the content available in the file-sharing system, we test their ability to distinguish five kinds of content available in the system (TV, music, books, movies and software) from five kinds not available in the system (current news, commercials, sports, how to videos and home videos). We compare this ability without the term

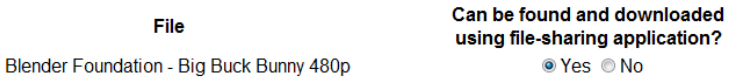


Figure 3A.1: Example question from the evaluation HIT

cloud and with several different different cloud designs. Our HIT asks users to make a series of judgments on whether specific files exist in the file-sharing system. An example judgment is shown in Figure 3A.1.

Their answers to these questions will reflect whether or not users have generalized the information available in the term cloud into a mental picture that correctly represents the type of content in the system.

In order to prime workers to project themselves into the role of users of the file-sharing system and to discourage them from trying to use Internet search to determine which file-sharing system we are discussing and what sorts of files are present in it, we introduce the HIT with a "frame" that sets up an imaginary situation. The frame includes the following text and the diagram in Figure 3A.2: *Jim and his large circle of friends have a huge collection of files that they are sharing with a very popular file-sharing program. The file-sharing program is a make-believe program. Please imagine that it looks something like this sketch:*

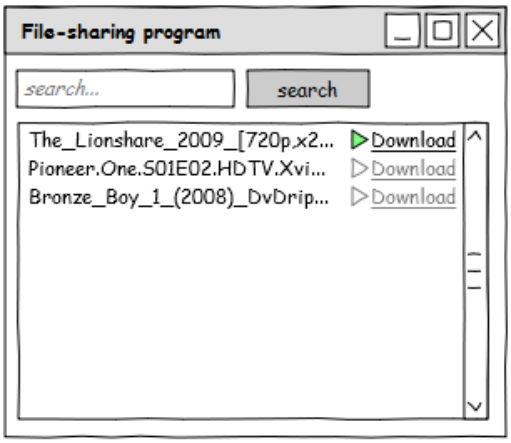


Figure 3A.2: Mockup of a file-sharing program used to introduce (i.e., to "frame") our evaluation HIT

By naming a specific user of the file-sharing system, "Jim", we hope that users will better identify with a user of the file-sharing system, i.e., project themselves into that role.

We then ask for 10 worker judgments like the one in Figure 3A.1. The HIT concludes with three validation questions, i.e., questions that do not ask for information necessary for the task, but rather allow us to judge the way in which the worker is approaching the task and eliminate low quality answers: (1) PrefQ, a personal preference question (multiple choice) *If you could download one of these files, which one would it be?* (2)

PrefEx, a request to explain the personal preference (free-text question) *Why would you choose this particular file for download and viewing?* and (3) AnsEx, a request to justify one of the choices made while answering the 10 evaluation questions (free-text question) *Think again about the file that you chose. Why did you guess that Jim or one of his friends would have this file in their collection?* Note that there is an important difference between PrefEx, which asks workers to give a motivation for their own opinion, and AnsEx, which asks workers to give a motivation from the perspective of the role in which we would like them to project themselves, i.e., a user of the file-sharing system.

We use multiple versions of this HIT, called the "evaluation HIT", in order to collect the information necessary for our study. Most of the cases discussed here are versions of the HIT that do not contain term clouds. We are interested in gauging the user's baseline evaluation answers before exposure to the term cloud. In some cases, we also use a recruitment HIT that establishes a closed pool of qualified workers. In the next section, we discuss observations concerning our HIT made during the design and test process.

3A.4. EXPLORATORY ANALYSIS

This section provides a qualitative discussion of the issues that we encountered during the design and testing process of our evaluation. We relate these issues to the particular nature of our task—its high imaginative load.

Recruitment and worker volume Because the evaluation needed to fit our design and implementation schedule, it was important that our evaluation HITs quickly attracted an adequate volume of workers so that the total number of assignments associated with that HIT (i.e., the batch) completed within reasonable time. We soon noticed that workers from the recruitment HIT did not continue immediately on to carry out an evaluation HIT. We started our first evaluation HIT right after manually handing out qualifications to the 81 workers that completed our recruitment HIT successfully. Since the recruitment HIT took less than 24 hours to complete, we initially assumed that the evaluation HIT would complete within roughly the same amount of time. However, only 10 out of 405 HIT-assignments offered were completed the next day. A second recruitment yielded 79 new qualified workers, but only one of them took up the main evaluation HIT within 24 hours. We conjectured that this slow uptake was due to the mismatch in expectations raised by the recruitment HIT. The recruitment HIT was titled "Like movies and music? Earn qualification with a background survey and two short questions". It contained a list of relatively easy to answer background questions, but only one question containing titles as in Figure 3A.1. In short, it did not reflect the focus of the main HIT. Workers were possibly misled to believe the main HIT would be more related to music and movies and did not expect to receive questions like Figure 3A.1 in the main HIT. There are two possible interfering factors affecting the volume of workers: reward level, which we did our best to optimize before publishing this HIT, and total number of assignments available to workers. During a previous crowdsourcing project, e-mails from workers suggested that HIT popularity is related to offering a large volume of assignments and keeping them in steady supply. Because our recruitment HIT asks for free-text answers that must be individually judged, it is not possible to automate the assignment of qualifications in our evaluation, and for this reason the slow worker uptake was a real concern. We decided to publish an "open" evaluation HIT, i.e., one that

did not require workers to earn a qualification, and were surprised that the quality of the responses to the free-text validation questions remained very stable. Apparently, our HIT has an aspect of its design that discourages workers who are not serious and makes recruitment less necessary.

Matching strategies. Because our evaluation task is attempting to gather information about people's mental pictures and not about the external world, there are no "correct answers" to the task questions. We could enlarge our HIT with questions for which the answer is known – a popular method for quality control – but the workers' ability to answer the control questions is not guaranteed to reflect the quality of their evaluation answers. For our task, it is more important to control for the strategy the worker is using to answer the question. In particular, we need the workers to be projecting themselves into the role of the user of the file-sharing application and not applying a strategy that reflects an external source of information (such as making use of general Internet search). A particular danger in the case of the evaluation HIT is that workers will try to apply a matching strategy using the information given in the "frame" of the HIT. In other words, it is possible that workers answer the evaluation questions by literally comparing the filenames in the example in Figure 3A.2 or the terms in the term cloud (described in Section 3A.3, but not pictured) to the filenames in Figure 3A.1. Reading the explanations of why the workers thought that certain files were in the file-sharing system (i.e., the answer to AnsEx), it was clear that a few of the workers would base their decision on literal matches (e.g., one answers "cloud contains DVDRIP"). However, the majority were attempting to generalize the situation and make a decision on the basis of what kind of media enjoy overall popularity (in the case which does not include the term cloud) or what general categories of content are represented in the term cloud (e.g., one answers, "With the cloud screens showing words like programming and microsoft, I think this file should be available in the collection").

3A.5. FURTHER INVESTIGATION

We carried out a small-scale experiment run on several crowdsourcing platforms in order to further investigate the impact of title choice and of the validation questions on the quality of the workers' responses. Each version of the HIT was made available to workers with a total of 50 assignments (5 sets of 10 different filenames to be judged) paying US$0.10 each. Results are reported in Table 3A.1 in terms of batch statistics: number of assignments that we rejected due to obvious non-serious workers (e.g., blank text boxes), total number of workers participating, effective hourly rate, run time needed to complete the batch and median time between arrivals of new workers to work on the HIT-assignments.

We experimented with three titles. Title A ("Jim, his friends and a make-believe file-sharing program"), which emphasized the imaginative nature of our HIT by including reference to "make believe". Title B ("Jim, his friends and digital stuff to download"), deemphasized the fact that the HIT involved file sharing, terminology we thought might seem overly technical to workers. Title C ("Jim, his friends and interesting stuff to download"), which attempted to make the HIT generally attractive to a wide audience. We also experimented with omitting our validation question in order to understand which ones were important for maintaining high quality answers. We ran a version of our HIT

Table 3A.1: Batch statistics for the five experimental conditions (varying title and validation questions) on Mechanical Turk

	Title A	Title B	Title C	Only AnsEx	No PrefEx
#Rejected assignments	0	0	2	0	0
#Workers	25	22	19	17	20
Effective hourly rate	$2.54	$2.08	$1.76	$3.13	$1.51
Run time	50h28m	13h45m	19h13m	15h55m	20h45m
Median arrival interval	67m36s	24m05s	18m41s	17m22s	35m06s

which only asked for an explanation of the answer to the evaluation questions ("Only AnsEx") as well as a HIT that asked for a personal preference, but did not ask for that personal preference to be justified ("No PrefEx"). For completeness, we include a list of the limitations of this experiment, necessarily imposed by its small scale and short duration: We were able to control for temporal variation by starting each version of the HIT at approximately the same time on consecutive weekdays. We did not control for differences among weekdays or for the effects of holidays (for example, Title A ran the day before the Thanksgiving holiday in the US and we are careful not to read too much into its significantly longer runtime). We did not control for workers becoming acclimated to us as a requester and thereby more inclined to do our HITs. We simply checked that the number of workers that participated in multiple conditions remained limited (2–5). In this way, we know that our results are not dominated by workers who are developing strategies on how to approach the task from one HIT version to the next.

The following generalizations emerge from our investigation. First, all HITs yielded serious results—in only two cases did we reject an assignment completed by a worker due to blatant cheating. Second, the generally attractive title (Title C) seemed to attract workers at a better rate, but needed a longer total run time than Title B. Only requiring an explanation of the answer and not of personal opinion attracted workers quickly and also improved the total running time. However, here we noticed that we attracted two types of workers: first, workers who were taking the HIT seriously, spending relatively long to complete it and giving thoughtful answers to AnsEx and, second, workers who approached AnsEx with a "quick and dirty" strategy. Either these workers realized that the same answer was more or less applicable to all 5 sets of ten filenames and copied and pasted the same answer for each HIT-assignment that they completed or they fell into trivial non-specific observations, such as "That's what people share". In order to understand this effect, it is important to note that the wording of AnsEx was necessarily affected by the removal of the personal preference question from the "Only AnsEx" condition. It was no longer possible to ask for an explanation concerning the file that the user had picked. Instead of the original wording, the question was changed to "Think about the files that you thought were available for download. Why did you guess Jim and his friends would have these files in their collection?" This relatively small change meant that the question no longer targeted one specific file—the generality of the question apparently was enough to encourage non-serious workers to apply cut and paste strategies. Interestingly, the workers that answered the AnsEx question seriously in the "Only AnsEx" version of the HIT gave more elaborate answers than the workers doing the

version of the HIT that required them to answer multiple validation questions. Also interesting was that the "No PrefEx" condition, which omitted the question requiring workers to justify their personal interests, yielded thoughtful answers on the AnsEx question, suggesting that the PrefEx question is not necessary. We would like to note that because the number of workers was relatively small, a single worker with a particular style (e.g., tending to apply a matching strategy) could have an inordinately large influence on the outcome of the experiment. If it is not possible to completely control for worker style, it appears important to use a quite large pool of workers in order to ensure the generality of results.

We ran the same set of experiments on other available crowdsourcing platforms to make a cross-platform comparison. Gambit and Give Work did not yield any judgments at all. This finding was largely independent of the financial reward offered. We conjecture that the lack of uptake may be due to technical limitations (mobile device, etc.) or a consequence of a different culture of HITs on these platforms. Samasource seems to be a very difficult platform to use. There were several negative observations to be made with our current experiment setup: Largely independent of title or question style we notice a very high share of uncreative copy and paste answers. Additionally there seem to be issues with their worker identification system as we have multiple submissions from different worker ids, that were issued from the same IP address and contained identical copy & paste answers. The very impressive exception to this trend was one worker from Nairobi who provided extremely detailed, informed and well-written answers.

3A.6. CONCLUSIONS

We conclude that "high imaginative load" tasks can be successfully run on MTurk. The key appears to be a combination of signaling to workers the unique nature of the task, possibly quite different than tasks they generally choose, and at the same time making each HIT-assignment require a highly individualized free-text justification response.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme (FP7) under grant agreement № 216444 (NoE PetaMedia).

3B

A PEER'S-EYE VIEW: NETWORK TERM CLOUDS IN A PEER-TO-PEER SYSTEM

This is a companion chapter to the previous chapter in which the framing methodology is put to use. Here, the methodology is used as part of an evaluation of a new browsing feature for a real-world retrieval system. We investigate term clouds that represent the content available in a peer-to-peer (P2P) network. Such network term clouds are non-trivial to generate in distributed settings. Our term cloud generator was implemented and released in Tribler—a widely-used, server-free P2P system—to support users in understanding the sorts of content available. Our evaluation and analysis focuses on three aspects of the clouds: coverage, usefulness *and* accumulation speed. *A live experiment demonstrates that individual peers accumulate substantial network-level information, indicating good coverage of the overall content of the system. The results of a user study carried out on a crowdsourcing platform confirm the usefulness of clouds, showing that they succeed in conveying to users information on the type of content available in the network. An analysis of five example peers reveals that accumulation speeds of terms at new peers can support the development of a semantically diverse term set quickly after a cold start. This work represents the first investigation of term clouds in a live, 100% server-free P2P setting.*

This chapter is published as **Raynor Vliegendhart, Martha Larson, Christoph Kofler, and Johan Pouwelse. A peer's-eye view: network term clouds in a peer-to-peer system. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1909–1912. ACM, 2011** [102].

3B.1. INTRODUCTION

New users encountering a search system can search more effectively if they have appropriate expectations of the sort of content that can be found in the system. Tribler is a real-world peer-to-peer (P2P) file-sharing system (downloadable from `http://www.tribler.org`) that offers a search functionality [77]. We developed and implemented a term cloud generator in order to promote successful searches by providing users with an impression of the types of content available in the system. Informal observation of user interaction patterns suggests that users having more experience with the Tribler system formulate a greater number of successful queries. The term clouds are intended to provide a quicker substitute for system interaction experience. If the clouds support users in understanding which information needs Tribler can fulfill, it can be expected that their queries better match the content of the system, leading, in the long term, to higher satisfaction and better user retention rates.

The term clouds are generated using the frequency counts of terms extracted from the names of files within the network that are accumulated at an individual peer by way of the underlying process used to exchange information among peers. This paper investigates the question of whether effective term clouds reflecting overall network content can be created in a distributed environment. We focus on three aspects: *coverage, usefulness* and *accumulation speed*. Note that this focus excludes investigation of cloud animation. Here, we simply state that animation switches cloud views at regular intervals to give the user the impression of the scope and dynamic development of the content of the system. Analysis of use pattern statistics and long-term impact on the uptake of Tribler are also left for future work.

In a completely distributed environment such as Tribler, building a network term cloud is non-trivial. Within the network, content is stored not on a central server with a 'bird's-eye' view, but rather at the individual peers. An individual peer can receive information about content at other peers only by communicating with its direct neighbors. In other words, in an environment that is 100% server free, the only view of the content collection that is available is a 'peer's-eye' view. In order for term clouds to be useful, the communication between peers must provide fast and high-coverage information about the content in the network. The key contribution of this paper is to demonstrate that a server-free architecture does not prevent peers from generating clouds that provide a global overview and are helpful for users.

After presenting background and related work, we report results of a live discovery experiment investigating cloud *coverage*, i.e., how well 'peer's-eye' clouds reflect network-level content. Then, we investigate the *usefulness* of the clouds, i.e., their ability to convey an impression of the content of the network to users, with a user study. Finally, we examine the *accumulation speed* of the clouds with a qualitative analysis of the cold start phase of example peers that reflects the experience of new users entering the network. We finish with our conclusion and outlook.

3B.2. BACKGROUND AND RELATED WORK

The basic motivation for using term clouds to communicate the content of the system to users derives from work on tag clouds, which shows that clouds support browsing

and discovery [88]. Our term clouds contain mixed uni- and bi-gram terms. This design choice is based on studies showing user preference for bigram or mixed clouds [46], which suggest that longer terms are easier for users to interpret. Since our aim is to investigate the viability of peer's-eye clouds in a P2P setting, we do not focus on the specific benefits derived from individual design characteristics of clouds, e.g., the use of terms vs. tags or the benefits of mixing terms in clouds.

Understanding our network term cloud requires careful distinction between 100% server free P2P systems and solutions with central dependencies. The presence of a central component in the design of a P2P system allows a significant simplification of the search set up, e.g., search in Napster using a central file index. In such a system, generation of term clouds is trivial. However, in a fully decentralized, i.e., 100% server free, P2P system such as Tribler, there is no central point that can, e.g., aggregate term frequencies and peers are required to propagate or search for information throughout the network. Depending on its network topology, a P2P system can use different communication protocols based on, e.g., flooding [60]. Gossip-based algorithms [15] provide the relative advantage of scalability and Tribler uses its own specific gossip protocol called Buddycast [78]. Through the exchange of periodic Buddycast messages, a peer discovers other peers and new content. Each message contains a list of live peers (divided into peers similar to the sending peer and peers that have been selected randomly), a download profile of the sending peer and a list of selected content hashes known by the sending peer. When a peer finds a peer with a similar download profile or an unknown content hash, it can connect to that peer or request the metadata of the file corresponding to that hash. Note that although the similarity relationship between connected peers is a distinguishing characteristic of Tribler, here, we exclude it from consideration in order to ensure our results achieve better generalization beyond Tribler to server-free P2P in general. If a peer downloads a file, it retrieves the file's content using the BitTorrent protocol [14].

Work closely related to our own is limited, and arguably effectively restricted to a single research effort, [31]. This work proposes an architecture for aggregation and representation of information resources that enables tagging in a P2P network with a Distributed Hash Table (DHT) topology. A DHT is a structured P2P network in which nodes and values are assigned a key through a hashing function and can be found using key-based routing. The basic challenge, that of information aggregation in a distributed environment, faced in [31] is shared with our work. However, our work differs in that Tribler is an unstructured P2P network, with greater flexibility and lower security risk. Further, here, we focus on term discovery, while [31] investigates maintaining frequency approximations of previously-known tags.

3B.3. NETWORK TERM CLOUDS

Network term clouds are created by extracting terms from the filenames and accumulating raw counts. Peers acquire filenames from neighbors through torrent files and in turn pass these torrent files along again. The collecting of torrent files is driven by the Buddycast protocol, which allows peers to discover new content as described in the previous section. A torrent file contains metadata required for the BitTorrent protocol to download an actual file [14]. The metadata includes the filename, size and integrity

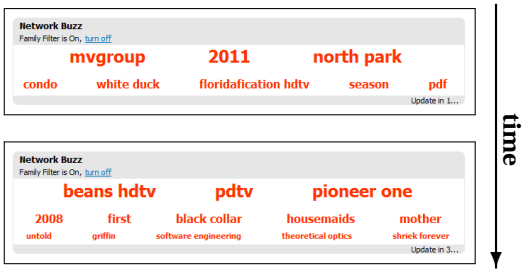


Figure 3B.1: Network term cloud: peer's-eye view (two snapshots of the animated cloud in Tribler)

hashes. From a filename both unigram and bigram terms are extracted. The unigram terms are obtained by tokenizing the filename using non-alphanumeric characters as delimiters. English stopwords and unigrams shorter than three characters are ignored. Bigram terms are constructed by joining the first two unigrams extracted from a filename, based on the assumption that the most important unigrams are at the beginning of a filename. The extracted unigram and bigram terms are displayed together in a single mixed cloud.

The network term cloud in Tribler is illustrated in Figure 3B.1, which pictures two frames of the cloud, which is animated. Each frame of the animated term cloud shows for five seconds a random sample of 13 terms from high, medium and low frequency levels represented on three different tiers. The design decision to include three levels of frequency enhances the user's impression of the network content as evolving, since new terms are low frequency and would be completely excluded from the cloud, had frequency been the sole criterion for inclusion in the cloud. We restrict ourselves here mentioning the tiers, but do not investigate them further in the current work. The network term cloud can be observed live by downloading and installing Tribler; as an alternative we provide a demonstration video: `http://youtu.be/hZeQlf5V8tA`.

3B.4. LIVE TERM DISCOVERY EXPERIMENT

The first aspect of the network term cloud we investigate is *coverage*. We carry out a live experiment within the Tribler network whose aim is to discover whether peer's-eye views of the network are mutually exclusive or whether the coverage of terms accumulated at a peer is substantial enough to support the generation of a cloud representing the overall content of the network. We acquired filenames at each of a pool of 30 peers under our control during their normal operation within the P2P system over an extended period of time (ca. 671 hours) and extracted terms from them. The 30 peers were started in succession on a single machine, joined the network, and only executed Buddycast to discover new peers and new content. Our peers did not initiate any downloads and did not build up a download profile. In this way, we ensured that they only connected to random and not semantically similar peers, as mentioned above. During the whole experiment they did not leave the system.

In Figure 3B.2, discovery of terms from filenames is illustrated over time for the first four hours of the experiment. Each line represents the intersection of the number of terms discovered by a given number of peers, with the bottom line showing terms dis-

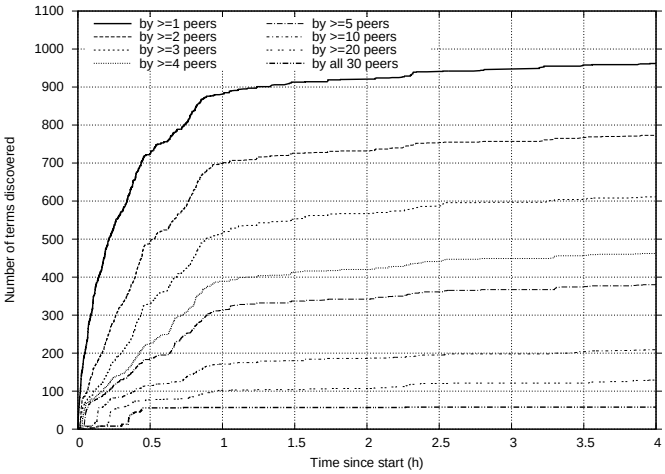


Figure 3B.2: Peer-level discovery of terms over time

covered by all peers. This figure yields several important insights. First, peers are far from discovering mutually exclusive term sets in the network, but rather a large overlap can be seen between the term sets that accumulate at the individual peers. Recall, that it is not possible to collect a complete global view of network content and that we approximate this view using the pool of peers under our control. In particular, we assume that terms discovered by at least one of the peers are representative of the overall content of the system. Figure 3B.2 shows that peers in our pool discover a substantial proportion of the global term set. Second, the system does not reach a steady state, but rather the number of terms discovered by a single peer keeps growing and the other peers never converge with respect to the composition of their term sets. The growth reflects the constant entry of new content into the system. Although only the first four hours are pictured in Figure 3B.2, the parallel growth trend is displayed during the entire collection interval of ca. 671 hours. Note that the flatness of the lowest lines in Figure 3B.2 before ca. 0.4 hours can be attributed to the sequential startup of the peers. In order to get better insight on the early startup phase of peers, we will return later to investigate term accumulation immediately after the cold start of a peer.

We carried out an additional analysis to investigate the nature of those terms that are discovered by some peers, and thus assumed to belong to the global view, but not by others. In particular, we are interested in determining whether individual peers are likely to miss high frequency terms, under the assumption that such terms are the most important for characterizing the network-level collection. Figure 3B.3 plots the probability that a term will fail to be discovered by one of the peers in our pool against the global frequency of that term estimated using the peer pool at two time-points in the life of a peer: a relatively immature (4 hours) and a mature (24 hours) stage. The exact times were chosen according to Tribler-specific considerations: four hours is the smallest resolution with which we can observe peers entering and leaving the network and 24 hours is the point at which Tribler considers the startup phase to have ended and switches messaging to a lower rate. We use Figure 3B.3 to draw important general con-

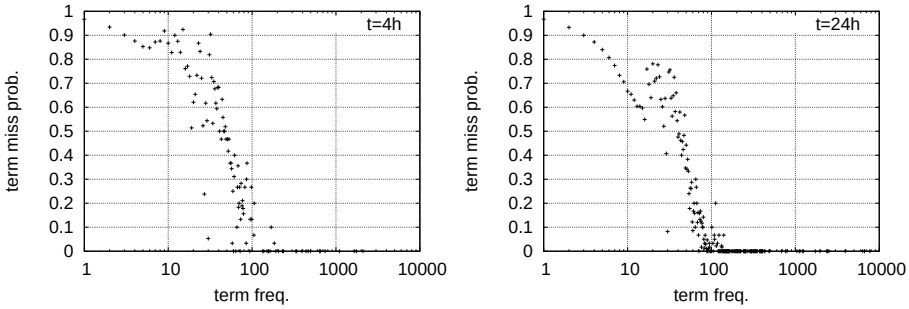


Figure 3B.3: Term discovery failure vs. term frequency

clusions. First, it can be seen that the terms the most likely to be missed are in general low frequency terms. For high frequency terms, the miss probability approaches zero. Second, although the mature peer has fewer terms with high miss probabilities, the difference with the immature peer is not staggering, suggesting that the discovery process is already effective early in the life of a peer. In sum, the 'peer's-eye' view does not vary radically from peer to peer and does well in approximating the 'bird's-eye' view, generally missing a relatively restricted set of lower frequency terms.

3B.5. USER STUDY

The second aspect of the network term cloud we investigate is *usefulness*. We performed a user study to investigate the question of whether term clouds can act as a surrogate for user experience with Tribler. We measure understanding of the collection in terms of the ability of a user to classify a file as either available or not available in Tribler. Informal observation of user interaction patterns provides evidence that a preponderance of Tribler queries correspond to known item search needs. We assume that the ability of a user to predict the availability of a file in Tribler reflects the ability to formulate successful searches.

We carried out experiments using Amazon Mechanical Turk (`http://www.mturk.com`), a crowdsourcing platform providing access to a pool of workers. Details of the study design and set up, including the crowdsourcing quality control mechanism to ensure serious user study participants, are described in [101]. The study investigates two conditions: `no-cloud`, in which the subject is presented with a mockup and a basic description of the system, and `with-cloud`, in which the subject is additionally presented with a series of five term clouds, such as they would be viewed (in animated sequence) in the system. The clouds were selected randomly using the set of terms that had been discovered after four hours by a typical peer chosen from our peer pool.

The file list we ask the subjects of the user study to classify consists of 100 filenames, half drawn from the Tribler system and half fake. The fake filenames were generated to represent types of content that were chosen by a panel of 'expert' users with extensive Tribler experience. They correspond to five categories clearly not represented in the Tribler system: 'home videos', 'news' and 'how-to videos', 'commercials' and 'sports'. The real filenames represented types of potentially findable content: 'TV', 'movies', 'music', 'software' and 'books'. The basis of the fake filenames were titles of existing videos from

Table 3B.1: User filename prediction (%correct)

	no-cloud	with-cloud
real	57.0%	63.1%
TV	66.3%	63.8%
movies	59.5%	69.5%†
music	59.5%	65.5%
software	50.5%	63.5%†
books	45.0%	55.0%†
fake	51.9%	52.1%
home videos	53.5%	49.5%
news	61.5%	61.5%
how-to	55.5%	61.0%
commercials	51.7%	42.5%
sports	33.8%	45.5%†
all	54.5%	57.6%†

† *Statistically significant improvement, Pearson's χ^2 test ($\alpha = 0.05$)*

popular websites and titles of news items. The titles were modified to resemble the filenames of the real files, such that the difference between the two was disguised. This was done by adding plausible group names, format extensions and format specifications to the title.

In total, 184 workers took part in the user study, making a total of 4000 classification decisions divided over the 100 filenames. Subjects were first offered the `no-cloud` and then the `with-cloud` condition; they could participate in one or both conditions—offering this option effectively gave us access to more subjects. The filenames in each condition were mutually exclusive, but chosen to be comparable. Accuracy in the `no-cloud` condition was 54.5% and rose to 57.6% in the `with-cloud` condition (Table 3B.1), a significant improvement according to Pearson's χ^2 test ($\alpha = 0.05$). The above-random performance in the `no-cloud` condition reflects assumptions that the users make about the content of the system from their prior file-sharing experience and the short description of the task. Out of 92 workers, 28 explicitly referred to the cloud when they were asked to justify one of their classification decisions, confirming that the clouds were consulted. Statistically significant improvements of the `with-cloud` over the `no-cloud` condition were observed in four categories: 'movies', 'software', 'books' and 'sports'. These results suggest that users do indeed gain an impression of the system content via the term cloud. The effect measured here is subtle, however, combined with aspects not yet studied here (e.g., use of cloud to support browsing, volume of user clicks) has a potential to increase user satisfaction with the system. Gains in specific categories less conventionally associated with file sharing (e.g., books) make term clouds particularly valuable for our P2P system.

3B.6. COLD START ANALYSIS

The final aspect of the network term cloud we investigate is *accumulation speed*. We examine the cold start phase of five example peers from our pool with the aim of gaining an impression of the potential of very young clouds to be helpful to users, an issue assumed to be important for retention of new Tribler users. Our analysis procedure is based on the insight that it is not necessary for two clouds to be exactly identical in order to be

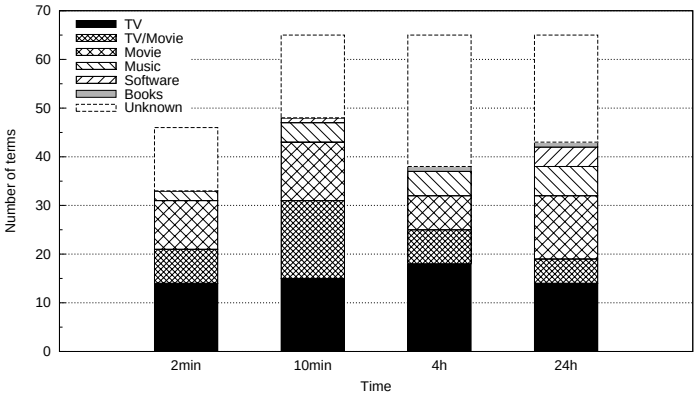


Figure 3B.4: Semantic diversity of term clouds

equally useful to the user. For this reason, we concentrate not on the identities of cloud terms, but rather on their semantic diversity. We take diversity to reflect the ability of the clouds to convey an impression of the variety and scope of the content available in the network. We assume that certain terms in the cloud trigger users to infer the presence of certain types of content in the system. In particular, we focus on the categories of content known to exist in the system and used in the user study. For each of the terms in example clouds from our five peers, we make a best guess on which category it might reflect. In the case of TV vs. movies, it is often difficult to make a single best guess and in this case we label the term with a combined TV/movie category. Figure 3B.4 shows the distribution of the terms over the categories in the clouds at four points along the life of the peer (2min, 10min, 4h and 24h). Between 2min and 10min clouds become mature enough to contain a full set of 13 terms. Although not all categories are present in the youngest (2min) clouds, the diversity is still good. Further, clouds at peers older than 10min are not radically more diverse, suggesting that very young clouds can be just as effective as mature clouds.

3B.7. CONCLUSION AND OUTLOOK

We have demonstrated that information aggregated at a single peer within a distributed system is adequate to support generation of term clouds that provide a user with useful information about the content of the system. Since the Tribler client attracts heavy use—it has been downloaded more than 800,000 times within the last five years—even a small or modest improvement in users' understanding of the system has the potential to lead to a large impact in terms of improved query success and better user experience. Further, communication of an impression of the global content of a P2P network to users is critical as P2P moves into new domains, since it helps users to quickly shake outdated assumptions about the nature of the items being shared by file sharing. Future work will involve analysis of the click behavior of users interacting with the term cloud and will shed light on the details of cloud use, particularly on the ability of clouds to support browsing and discovery and to improve the retention of new users of the system.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme under grant agreement № 216444 (PetaMedia).

4A

DISCOVERING USER PERCEPTIONS OF SEMANTIC SIMILARITY IN NEAR-DUPLICATE MULTIMEDIA FILES

The focus of this chapter is on elicitation techniques for obtaining useful responses from the crowd. The motivation behind this chapter's work is to fuel the design of a new feature for presenting search results in a retrieval system. Evaluation of this feature is presented in the next chapter. Here, we address the problem of discovering new notions of user-perceived similarity between near-duplicate multimedia files. We focus on file-sharing, since in this setting, users have a well-developed understanding of the available content, but what constitutes a near-duplicate is nonetheless nontrivial. We elicited judgments of semantic similarity by implementing triadic elicitation as a crowdsourcing task and ran it on Amazon Mechanical Turk. We categorized the judgments and arrived at 44 different dimensions of semantic similarity perceived by users. These discovered dimensions can be used for clustering items in search result lists. The challenge in performing elicitations in this way is to ensure that workers are encouraged to answer seriously and remain engaged.

This chapter is published as **Raynor Vliegendhart, Martha Larson, and Johan Pouwelse. Discovering user perceptions of semantic similarity in near-duplicate multimedia files. In *Proceedings of the 1st International Workshop on Crowdsourcing Web Search*, pages 54–58. CEUR-WS.org, April 2012** [104].

4A.1. INTRODUCTION

Crowdsourcing platforms make it possible to elicit semantic judgments from users. Crowdsourcing can be particularly helpful in cases in which human interpretations are not immediately self evident. In this paper, we report on a crowdsourcing experiment designed to elicit human judgments on semantic similarity between near duplicate multimedia files. We use crowdsourcing for this application because it allows us to easily collect a large number of human similarity judgments. The major challenge we address is designing the crowdsourcing task, which we ran on Amazon Mechanical Turk, to ensure that the workers from whom we elicit judgments are both serious and engaged.

Multimedia content is semantically complex. This complexity means that it is difficult to make reliable assumptions about the dimensions of semantic similarity along which multimedia items can resemble each other, i.e., be considered near duplicates. Knowledge of such dimensions is important for designing retrieval systems. We plan ultimately to use this knowledge to inform the development of algorithms that organize search result lists. In order to simplify the problem of semantic similarity, we focus on a particular area of search, namely, search within file-sharing systems. We choose file-sharing, because it is a rich, real-world use scenario in which user information needs are relatively well constrained and users have a widely-shared and well-developed understanding of the characteristics of the items that they are looking for.

Our investigation is focused on dimensions of semantic similarity that go beyond what is depicted in the visual channel of the video. In this way, our work differs from other work on multimedia near duplicates that puts its main emphasis on visual content [6]. Specifically, we define a notion of near duplicate multimedia items that is related to the reasons for which users are searching for them. By using a definition of near duplicates that is related to the function or purpose that multimedia items fulfill for users, we conjecture that we will be able arrive at a set of semantic similarities that will reflect user search goals and in this way be highly suited for use in multimedia retrieval results lists.

The paper is organized as follows. After presenting background and related work in Section 4A.2, we describe the crowdsourcing experiment by which we elicit human judgments in Section 4A.3. The construction of the dataset used in the experiment is given in Section 4A.4. Direct results of the experiment and the derived similarity dimensions are discussed in Section 4A.5. We finish with conclusions in Section 4A.6.

4A.2. BACKGROUND AND RELATED WORK

4A.2.1. NEAR-DUPLICATES IN SEARCH RESULTS

Well-organized search results provide an easy means for users to overview search results lists. A simple, straightforward method of organization groups together similar results and represents each group with a concise surrogate, e.g., a single representative item. Users can then scan a shorter list of groups, rather than a longer list of individual result items. Hiding near duplicate items in the interface is a specific realization of near-duplicate elimination, which has been suggested in order to make video retrieval more efficient for users [111]. Algorithms that can identify near duplicates can be used to group items in the interface. One of the challenges in designing such algorithms is be-

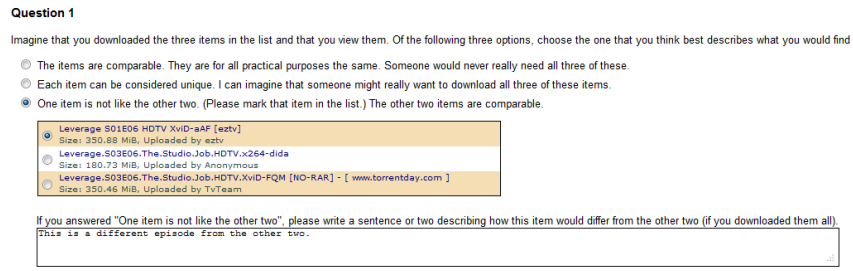


Figure 4A.1: One of the triads of files and the corresponding question as presented to the workers.

ing able to base them on similarity between items as it is perceived by users. Clustering items with regard to general overall similarity is a possibility. However, this approach is problematic since items are similar in many different ways at the same time [39]. Instead, our approach, and the ultimate aim of our work, is to develop near-duplicate clustering algorithms that are informed by user-perceptions of dimensions of semantic similarity between items. We assume that these algorithms stand to benefit if they draw on a set of possible dimensions of semantic similarity that is as large as possible.

Our work uses a definition of near duplicates based on the function they fulfill for the user:

> *Functional near-duplicate multimedia items are items that fulfill the same purpose for the user. Once the user has one of these items, there is no additional need for another.*

In [111], one video is deemed to be a near duplicate of another if a user would clearly identify them as essentially the same. However, this definition is not as broad as ours, since only the visual channel is considered.

Our work is related to [12], which consults users to find whether particular semantic differences make important contributions to their perceptions of near duplicates. Our work differs because we are interested in discovering new dimensions of semantic similarity rather than testing an assumed list of similarity dimensions.

4A.2.2. ELICITING JUDGMENTS OF SEMANTIC SIMILARITY

We are interested in gathering human judgments on semantic interpretation, which involves the acquisition of new knowledge on human perception of similarity. Any thoughtful answer given by a human is of potential interest to us. No serious answer can be considered wrong.

The technique we use, triadic elicitation, is adopted from psychology [26], where it is used for knowledge acquisition. Given three elements, a subject is asked to specify *in what important way two of them are alike but different from the third* [48]. Two reasons make triadic eliciation well suited for our purposes. First, being presented with three elements, workers have to abstract away from small differences between any two specific items, which encourages them to identify those similarities that are essential. Second, the triadic method is found to be cognitively more complex than the dyadic method [9],

supporting our goal of creating an engaging crowdsourcing task by adding a cognitive challenge.

A crowdsourcing task that involves the elicitation of semantic judgments differs from other tasks in which the correctness of answers can be verified. In this way, our task resembles the one designed in [92], which collects viewer-reported judgments. Instead of verifying answers directly, we design our task to control quality by encouraging workers to be serious and engaged. We adopt the approach of [92] of using a pilot HIT to recruit serious workers. In order to increase worker engagement, we also adopt the approach of [21], which observes that open-ended questions are more enjoyable and challenging.

4A.3. CROWDSOURCING TASK

The goal of our crowdsourcing task is to elicit the various notions of similarity perceived by users of a file-sharing system. This task provides input for a card sort, which we carry out as a next step (Section 4A.5.2) in order to derive a small set of semantic similarity dimensions from the large set of user-perceived similarities we collect via crowdsourcing.

The crowdsourcing task aims to achieve workers' seriousness and engagement with judicious design decisions. Our task design places particular focus on ensuring task credibility. For example, the title and description of the pilot makes clear the purpose of the task, i.e., research, and that the workers should not expect a high volume of work offered. Further, we strive to ensure that workers are confident that they understand what is required of them. We explain functional similarity in practical terms, using easy-to-understand phrases such as "comparable", "like", and "for all practical purposes the same". We also give consideration to task awareness by including questions in the recruitment task designed to determine basic familiarity with file-sharing and interest level in the task.

4A.3.1. TASK DESCRIPTION

The task consists of a question, illustrated by Figure 4A.1, that is repeated three times, once for three different triads of files. For each presented triad, we ask the workers to imagine that they have downloaded all three files and to compare the files to each other on a functional level. The file information shown to the workers is taken from a real file-sharing system (see the description of the dataset in Section 4A.4) and are displayed as in a real-world system, with filename, file size and uploader. The worker is not given the option to view the actual files, reflecting the common real file-sharing scenario in which the user does not have the resources (e.g., the time) to download and compare all items when scrolling through the search results.

The first section of the question is used to determine whether it is possible to define a two-way contrast between the three files. We use this section to eliminate cases in which files are perceived to be all the same or all different. This is following the advice on when not to use triadic elicitation that is given in [82]. Specifically, we avoid forcing a contrast in cases where it does not make sense.

The following triad is an example of a case in which a two-way contrast should not be forced:

Despicable Me The Game
VA-Despicable Me (Music From The Motion Picture)
Despicable Me 2010 1080p

These files all bear the same title. If workers were forced to identify a two-way contrast, we would risk eliciting differences that are not on the functional level, e.g., "the second filename starts with a V while the other two start with a D". Avoiding nonsense questions also enhances the credibility of our task.

In order to ensure that the workers follow our definition of functional similarity in their judgment, we elaborately define the use-case of the three files in the all-same and all-different options. We specify that the three files are the same when someone would never need all of them. Similarly, the three files can be considered to be all different from each other if the worker can think of an opposite situation where someone would want to download all three files. Note that emphasizing the functional perspective of similarity guides workers away from only matching strings and towards considering the similarity of the underlying multimedia items. Also, we intend the elaborate description to discourage workers to take the easy way out, i.e., selecting one of the first two options and thereby not having to contrast files.

Workers move on to the second section only if they report it is possible to make a two-way contrast. Here they are asked to indicate which element of the triad differs from the remaining two and to specify the difference by answering a free-text question.

4A.3.2. TASK SETUP

We ran two batches of Human Intelligence Tasks (HITs) on Amazon Mechanical Turk on January 5th, 2011: a recruitment HIT and the main HIT. The recruitment HIT consisted of the same questions as the regular main HIT (Section 4A.3.1) using three triads and included an additional survey. In the survey, workers had to tell whether they liked the HIT and if they wanted to do more HITs of this type. If the latter was the case, they had to supply general demographic information and report their affinity with file-sharing and online media consumption.

The three triads, listed below, were selected from the portion of the dataset (Section 4A.4) reserved for validation. We selected examples for which at least one answer was deemed uncontroversially wrong and the others acceptable.

- Acceptable to consider all different or to consider two the same and one different:

  Desperate Housewives s03e17 [nosubs]
  Desperate Housewives s03e18 [portugese subs]
  Desperate Housewives s03e17 [portugese subs]

  Here, we disallowed the option of considering all files to be comparable. For instance, someone downloading the third file would also want to have the second file as these represent two consecutive episodes from a television series.

- Acceptable to consider all different:

  Black Eyed Peas - Rock that body
  Black Eyed Peas - Time of my life
  Black Eyed Peas - Alive

Here, we disallowed the option of considering all files to be comparable as one might actually want to download all three files. For the same reason, we also disallowed the option of considering two the same and one different.

- Acceptable to consider all same or to consider two the same and one different:

  The Sorcerers Apprentice 2010 BluRay MKV x264 (8 GB)
  The Sorcerers Apprentice CAM XVID-NDN (700 MB)
  The Sorcerers Apprentice CAM XVID-NDN (717 MB)

  Here, we disallowed the option of considering all files different. For instance, someone downloading the second file would not also download the third file as these represent the same movie of comparable quality.

The key idea here is to check whether the workers understood the task and are taking it seriously, while at the same time not to exclude people who do not share a a similar view onto the world as us. To this end, we aim to choose the least controversial cases and also admit more than one acceptable answers.

We deemed the recruitment HIT to be completed successfully if the following conditions were met:

- No unacceptable answers (listed above) were given in comparing files in each triad.
- The answer to the free-text question provided evidence that the worker generalized beyond the filename, i.e., they compared the files on a functional level.
- All questions regarding demographic background were answered.

Workers who completed the recruitment HIT, who expressed interest in our HIT, and who also gave answers that demonstrated affinity with file sharing, were admitted to the main HIT.

The recruitment HIT and the main HIT ran concurrently. This allowed workers who received a qualification to continue without delay. The reward for both HITs was $0.10. The recruitment HIT was open to 200 workers and the main HIT allowed for 3 workers per task and consisted of 500 tasks in total. Each task contained 2 triads from the test set and 1 triad from the validation set. Since our validation set (Section 4A.4) is smaller than our test set, the validation triads were recycled and used multiple times. The order of the questions was randomized to ensure the position of the validation question was not fixed.

4A.4. DATASET

We created a test dataset of a 1000 triads based on popular content on The Pirate Bay (TPB),[1] a site that indexes content that can be downloaded using the BitTorrent [14] file-sharing system. We fetched the top 100 popular content page on December 14, 2010. From this page and further queried pages, we only scraped content metadata, e.g., filename, file size and uploader. We did not download any actual content for the creation of our dataset.

[1] http://thepiratebay.com

Table 4A.1: Dimensions of semantic similarity discovered by categorizing crowdsourced judgments

Different movie vs. TV show	Different movie
Normal cut vs. extended cut	Movie vs. trailer
Cartoon vs. movie	Comic vs. movie
Movie vs. book	Audiobook vs. movie
Game vs. corresponding movie	Sequels (movies)
Commentary document vs. movie	Soundtrack vs. corresponding movie
Movie/TV show vs. unrelated audio album	Movie vs. wallpaper
Different episode	Complete season vs. individual episodes
Episodes from different season	Graphic novel vs. TV episode
Multiple episodes vs. full season	Different realization of same legend/story
Different songs	Different albums
Song vs. album	Collection vs. album
Album vs. remix	Event capture vs. song
Explicit version	Bonus track included
Song vs. collection of songs+videos	Event capture vs. unrelated movie
Language of subtitles	Different language
Mobile vs. normal version	Quality and/or source
Different codec/container (MP4 audio vs. MP3)	Different game
Crack vs. game	Software versions
Different game, same series	Different application
Addon vs. main application	Documentation (pdf) vs. software
List (text document) vs. unrelated item	Safe vs. X-Rated

Users looking for a particular file normally formulate a query based on their idea of the file they want to download. Borrowing this approach, we constructed a query for each of the items from the retrieved top 100 list. The queries were constructed automatically by taking the first two terms of a filename, ignoring stop words and terms containing digits. This resulted in 75 unique derived queries.

The 75 queries were issued to TPB on January 3, 2011. Each query resulted in between 4 and 1000 hits (median 335) and in total 32,773 filenames were obtained. We randomly selected 1000 triads for our test dataset. All files in a triad correspond to a single query. Using the same set of queries and retrieved filenames, we manually crafted a set of 28 triads for our validation set. For each of the triads in the validation set, we determined the acceptable answers.

4A.5. RESULTS

4A.5.1. CROWDSOURCING TASK

Our crowdsourcing task appeared to be attractive and finished quickly. The main HIT was completed within 36 hours. During the run of the recruitment HIT, we handed out qualifications to 14 workers. This number proved to be more than sufficient and caused us to decide to stop the recruitment HIT prematurely. The total work offered by the main HIT was completed by eight of these qualified workers. Half of the workers were eager and worked on a large volume of assignments (between 224 and 489 each). A quick look at the results did not raise any suspicions that the workers were under-performing compared to their work on the recruitment HIT. We therefore decided not to use the validation questions to reject work. However, we were still curious as to whether the

eager workers were answering the repeating validation questions consistently. The repeated answers allowed us to confirm that the large volume workers were serious and not sloppy. In fact, the highest volume worker had perfect consistency.

The workers produced free-text judgments for 308 of the 1000 test triads. The other 692 triads consisted of files that were considered either all different or all similar. Workers fully agreed on which file differed from the other two for 68 of the 308 triads. Only two judgments out of the three given judgments agreed which file was different for 93 triads. For the remaining 147 triads no agreement was reached. Note that whether an agreement was reached is not of direct importance to us since we are mainly interested in just the justifications for the workers' answers, which we use to discover the new dimensions of semantic similarity.

4A.5.2. CARD SORTING THE HUMAN JUDGMENTS

We applied a standard card sorting technique [82] to categorize the explanations for the semantic similarity judgments that the workers provided in the free-text question. Each judgment was printed on a small piece of paper and similar judgments were grouped together into piles. Piles were iteratively merged until all piles were distinct and further merging was no longer possible. Each pile was given a category name reflecting the basic distinction described by the explanations. To list three examples: the pile containing explanations like

> "The third item is a Hindi language version of the movie."
>
> "This is a Spanish version of the movie represented by the other two"

was labeled as *different language*; the pile containing

> "This is the complete season. The other 2 are the same single episode in the season."
>
> "This is the full season 5 while the other two are episode 12 of season 5"

was labeled *complete season vs. individual episodes*; and the pile containing

> "This is a discography while the two are movies"
>
> "This is the soundtrack of the movie while the other two are the movie."

was labeled *soundtrack vs. corresponding movie*.

The list of categories resulting from the card sort is listed in Table 4A.1. We found 44 similarity dimensions, many more than we had anticipated prior to the crowdsourcing experiment. The large number of unexpected dimensions we discovered support the conclusion that the user perception of semantic similarity among near duplicates is not trivial. For example, the "commentary document versus movie" dimension, which arose from a triad consisting of two versions of a motion picture and a text document that explained the movie, was particularly surprising, but nonetheless important for the file-sharing setting.

Generalizing our findings in Table 4A.1, we can see that most dimensions are based on different instantiations of particular content (e.g., quality and extended cuts), on the serial nature of content (e.g., episodic), or on the notion of collections (e.g., seasons and albums). These findings and generalizations will serve to inform the design of algorithms for the detection of near duplicates in results lists in future work.

4A.6. CONCLUSION

In this work, we have described a crowdsourcing experiment that discovers user-perceived dimensions of semantic similarity among near duplicates. Launching an interesting task with the focus on engagement and encouraging serious workers, we have been able to quickly acquire a wealth of different dimensions of semantic similarity, which we otherwise could not have thought of. Our future work will involve expanding this experiment to encompass a larger number of workers and other multimedia search settings. Our experiment opens up the perspective that crowdsourcing can be used to gain a more sophisticated understanding of user perceptions of semantic similarity among multimedia near-duplicate items.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme under grant agreements № 216444 (PetaMedia) and № 287704 (CUbRIK).

4B

CROWDSOURCED USER INTERFACE TESTING FOR MULTIMEDIA APPLICATIONS

Whereas the previous chapter used crowdsourcing to collect different perspectives on semantic similarity for designing a user interface feature in a retrieval system, this companion chapter uses crowdsourcing to test that feature's implementation. Normally, evaluation of an application user interface is carried out in a conventional lab environment, but this form of evaluation is a costly and time-consuming process. In this chapter, we show that it is feasible to carry out A/B tests for a multimedia application through Amazon's crowdsourcing platform Mechanical Turk involving hundreds of workers at low costs. We let workers test user interfaces within a remote virtual machine that is embedded within the crowdsourcing task interface and we show that technical issues that arise in this approach can be overcome.

This chapter is published as **Raynor Vliegendhart, Eelco Dolstra, and Johan Pouwelse. Crowdsourced user interface testing for multimedia applications. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 21–22. ACM, 2012** [98]. An extension of this work containing comprehensive details on the test framework's prototype is published as **Eelco Dolstra, Raynor Vliegendhart, and Johan Pouwelse. Crowdsourcing GUI tests. In *IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST)*, pages 332–341, March 2013** [18].

4B.1. INTRODUCTION

Conducting an experiment to test an application's user interface is a costly and time-consuming process. A lab setting is needed in which the experimenter has full control over the environment and technical setup and in which the participants can be instructed. This generally means the experimenter can only accomodate a small number of user subjects at a time due to limited capacity. Furthermore, in order to draw statistically significant conclusions, a large number of participants is needed. Hence, conventional usability studies do not scale well.

In this paper, we show that it is technically feasible to conduct a large scale usability study on Amazon's Mechanical Turk crowdsourcing platform (`http://www.mturk.com`) involving hundreds of participants at low costs. In our approach, we face the challenge of no longer having full control over the experiment. While we can maintain control over the technical setup (OS, browser, etc.) that is running the application under test, we cannot control the environment in which the worker is performing the task. We show, however, that we can design usability studies to accommodate for this lack of control.

While user interfaces of web pages are already being evaluated on Mechanical Turk using services like TryMyUI (`http://www.trymyui.com`), our work allows any application's user interface to be tested by workers from a crowdsourcing platform. We have implemented a prototype that presents Mechanical Turk workers a display of a virtual machine (VM) embedded within the web page of the Human Intelligence Task (HIT). This VM runs the graphical user interface (GUI) under test on a server operated by the experimenter, ensuring we have full control over the technical setup. Workers can interact with the GUI using the keyboard and mouse and are asked to execute a series of steps as described by the HIT (Figure 4B.1). These steps are visually recorded by the VM. The resulting video can be used by developers for analysis, e.g., in case workers reported any problems.

To test whether our prototype can be used for usability studies, we ran A/B tests [51] for variants of Tribler, a multimedia sharing application [118]. In this usability study, we evaluated whether an experimental user interface feature inspired by previous work [104] would help users in finding multimedia content faster.

Implementation details of our prototype are outside the scope of this paper and are reserved for future publications. Some of the collected statistics that we present here are from a larger study which this work is part of. The focus of the larger study is not limited to usability studies, but also includes semi-automated continuous (e.g., periodic) testing.

The structure of this paper is as follows. We first describe technical factors that impact the HIT design and affect the testing of user interfaces (Section 4B.2). We then discuss the design and the results of our usability study (Section 4B.3). Finally, we summarize our findings (Section 4B.4).

4B.2. TECHNICAL FACTORS

In a conventional lab setting, the experimenter has the opportunity to eliminate any environmental factors that may have an impact on the results of an experiment. In our approach, the experimenter only has full control over the technical setup running

the user interface that needs to be tested. We cannot control for technical factors that play a role when workers are connecting to one of the remote virtual machines, such as: *a*) Bandwidth of the worker's connection; *b*) Latency of the worker's connection; and *c*) Screen resolution of the worker's display. We collected this information on each worker's technical setup as well as each worker's location in our larger evaluation study on crowdsourced GUI testing. The study involved 398 unique workers submitting 700 assignments from 32 different countries.

We found that our workers generally had a fairly slow Internet connection. Their connections had a median download speed of 48 KiB/s and had an average ping of 260 milliseconds. This factor has an effect on the task completion time, which needs to be accounted for when completion time is a key element in the usability study.

We also found that our workers were using low resolution displays. The majority of the workers had a 1024x768 (25.3%), 1366x768 (20.7%) or 1280x800 (11.8%) screen. To accommodate for these screens, the display of the VM should also be small. If it is too large, the worker cannot see both the embedded VM and the HIT's instructions simultaneously, which negatively impacts the worker's workflow. We therefore chose to use a resolution of 640x480 for the usability study which we describe in the next section.

4B.3. USABILITY STUDY

One aspect of usability is efficiency, e.g., how quickly can a user carry out a specific task. If our crowdsourced user interface testing approach were to be feasible, it is required that task completion times measured during experiments are reliable and are not influenced by external factors. However, as we have seen, there is a large variance in worker connection speeds (Section 4B.2) which could cause a large variance in the task completion time.

To test whether we can detect significant differences in task completion times, we focused on A/B testing for the usability study. Workers were instructed to issue several specific queries to find and download multimedia content in Tribler. The VM server presented connecting workers either variant *A* or variant *B* of the Tribler application. The application was instrumented to log the time between each query and download action. We modified variant *B* to include an artificial delay of 2 seconds in displaying search results, expecting that each query-download task would take 2 seconds longer when compared to variant *A*.

We launched a HIT of 100 assignments (and thus 100 workers), which took 28h58m to complete and costed a total of US$25. This yielded 354 and 330 measurements for variant *A* (normal) and *B* (delayed), respectively. The median interval between searching and downloading was 19.6s for variant *A* and 21.7s for variant *B*, conforming to the artificial 2 second delay that was introduced in variant *B*. The same was not observed for the arithmetic mean due to extreme outliers in variant *A* (max: 748.9s), but discarding the 25% highest measurements to account for skew resulted in a statistically significant difference between the two trimmed arithmetic means (Student t-test, $P = 0.049$). We thus conclude that the variance on connection speeds is not an issue.

Following this conclusion, we repeated the experiment to evaluate a new feature. This time, variant *B* contained an experimental Tribler feature called "bundling". This feature groups related search results together based on one of a few different notions

of similarity inspired by earlier work [104] such as filename or size, but we found no statistical difference in task completion time. The second HIT took 28h38m to complete. Thus the total runtime of both HITs was less than three days and costed in total only US$50.

4B.4. CONCLUSIONS

In this paper, we have shown that it is feasible to carry out A/B tests for a multimedia application on Mechanical Turk. While technical factors impact the design of the HIT and the usability study, we can account for them. Using our approach, we have been able to involve hundreds of workers to evaluate an experimental user interface feature within a few days at reasonably low costs.

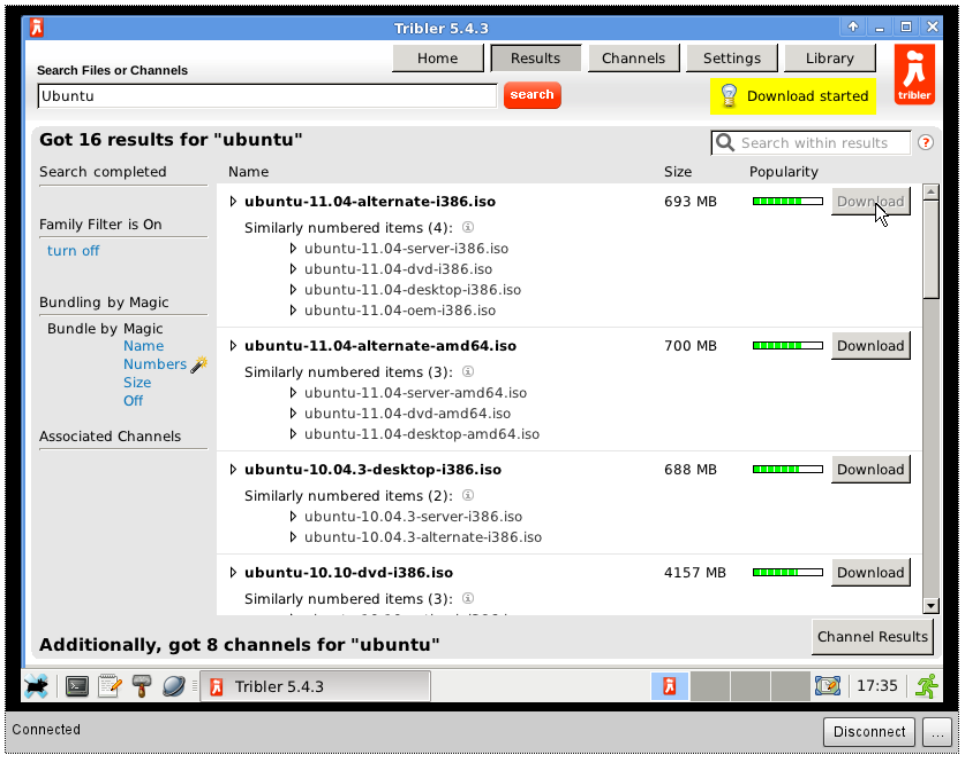
ACKNOWLEDGMENTS

We wish to thank Martha Larson for her advice on crowdsourcing and her comments on the design of the A/B test; Niels Zeilemaker for discussions and fixing bugs in Tribler; and of course all workers who participated in our HITs. This research was partially supported by: NWO-JACQUARD project 638.001.208, *PDS: Pull Deployment of Services*, as well as the NIRICT LaQuSo Build Farm project.

Test a Graphical User Interface

The goal of this task is to perform a list of actions to test software. Below you see the display of a computer running some software. The task is **to perform the following steps *precisely* and report whether they succeed**. If you don't succeed in any step, **report what went wrong** in the form at the bottom.

Virtual machine display



Step 3 / 8: Click on the **Download** button next to the top result. This should start the download.

Did you succeed? ○Yes ○No

Figure 4B.1: An example of a GUI testing HIT as it appears in a worker's web browser. The steps are shown below the embedded VM's display.

III

ADVANCING NON-LINEAR ACCESS TO VIDEO CONTENT

5

EXPLOITING THE DEEP-LINK COMMENTSPHERE TO SUPPORT NON-LINEAR VIDEO ACCESS

This chapter builds upon the foundation that the previous chapters have laid down. With this foundation in place, the chapter employs the crowdsourcing methodologies from past chapters for developing a crowd-informed typology of explicit expressions in the form of user comments that could improve search results in a retrieval systems. Specifically, we investigate the usefulness of deep links. *Deep links are time-coded comments with which viewers express their reactions to the content at specific time-points of a video that they find noteworthy. The rationale underlying our work is that deep links can open up an interesting new perspective on the relevance of a video, namely focusing on individual video segments, in addition to the existing ones that typically concern a video as a whole. In this perspective, deep-link comments provide non-linear access to videos via their time-codes, which can match alternate dimensions of user needs that extend beyond topical and affective relevance. We explore the different types of deep-link comments and develop a* Viewer Expressive Reaction Variety *(VERV) typology that captures how viewers deep-link on YouTube. We validate this typology through a user study on Amazon Mechanical Turk to show that it is a typology human annotators can agree upon. We then demonstrate through experiments that deep-link comments can automatically be classified into VERV categories and show the potential of our proposed usage of deep-link comments for video search through a user study.*

This chapter is published as **Raynor Vliegendhart, Martha Larson, Babak Loni, and Alan Hanjalic. Exploiting the deep-link commentsphere to support non-linear video access.** ***IEEE Transactions on Multimedia,* 17(8):1372–1384, 2015** [103].

5.1. INTRODUCTION

A conventional video search engine, such as the YouTube engine, imposes two constraints that limit its ability to respond to users' queries. First, search-engine internal representations of videos conventionally admit only topic-based matches with queries. In other words, the videos in the search result list are selected for their topical relevance to users' information needs, but other dimensions of relevance are not taken into account. Second, a conventional video search engine can only respond to queries by returning a whole video. It cannot return a results list that contains jump-in points (i.e., time-points within the video at which the user should start watching), even if such points would be the most useful response for the user. In other words, the search engine cannot address cases in which the user need would be better satisfied by a specific video segment, rather than the whole video.

This paper presents an approach that tackles both of these constraints at once, enabling *non-linear video access*, i.e., video search engines that return time-codes matching alternate dimensions of user needs. The added value of this approach is clear from a description of a simple example. Results from a conventional search engine may match the subject that the user is interested in, but they fall short of being optimal because they are too long to watch, and the user may quickly find them boring. Our approach is able, in this case, to return specific time-points within the videos that are particularly informative, surprising or amusing. Another example, depicted in Figure 5.1, illustrates the search engine that we envision, which returns results in the form of jump-in points (i.e., time-codes). In the example in Figure 5.1, the user is searching for information in videos related to BMW-type automobiles. The search engine first finds a set of videos that are topically related to BMWs (i.e., YouTube's current functionality), and on the basis of this set creates a reranked list of jump-in points for the viewer (i.e., the contribution of the deep-link categorization presented in this paper). Before reranking (cf. top panel), the jump-in points are ranked by a baseline condition (comment popularity), and the top of the list can be seen to contain deep links associated with exclamations and inquiries. After reranking (cf. bottom panel), the jump-in points at the top of the list are associated with factual information to match the user's need. We point out that in a different case the underlying information need of the user might be different and our approach would allow reranking according to other dimensions. Anticipating the findings of this paper, the example makes plausible that other dimensions might relate to factors such as amusement and surprise.

Our approach is based on the novel insight that the deep-link commentsphere can be exploited to enable non-linear video access. We consider the deep-link commentsphere to be the totality of time-coded comments, or deep links, that have been contributed by viewers of Web videos. The commentsphere can be thought of as an online information space, akin to the blogosphere, which represents the totality of all blogs. An example of a deep-link comment is:

> *"The way they come in at **4:00**.....LOVE IT :D"*
> — A user's reaction to a team's victory celebration

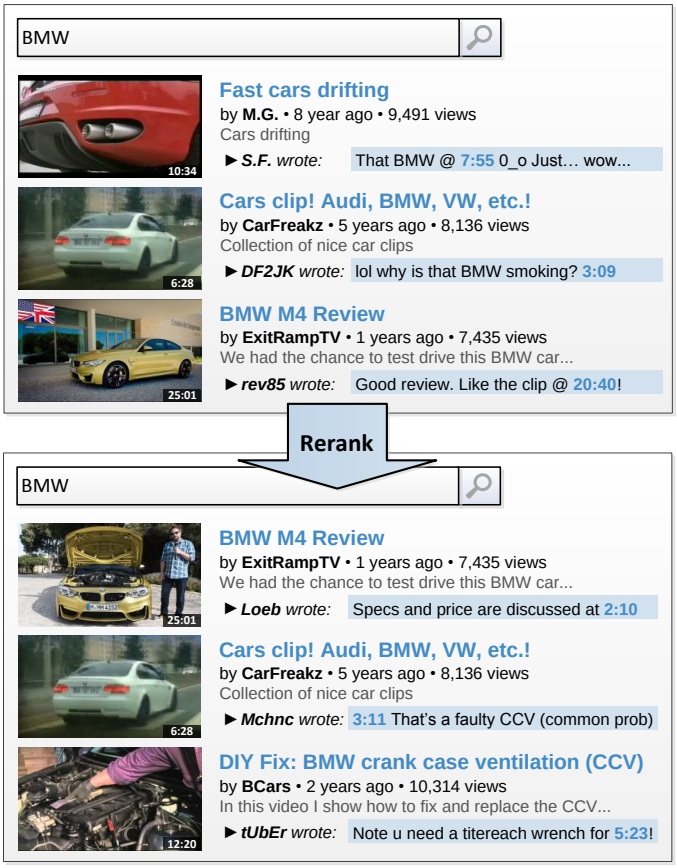


Figure 5.1: Excerpts of a video search results list containing time-code level results, before and after a reranking process that promoted results of a relevant deep-link type. The example illustrates the variety of systems that we envision in our work. Such a system would exploit deep links in order to offer users non-linear video access.

On a platform such as YouTube, these deep-link comments occur in the wild, and no special action on the part of the user is required to create a deep link. Simply mentioning the time-code using one of the supported formats (i.e., "0m00s" or "0:00") is sufficient and the platform will automatically turn the time-codes into a link that leads to the corresponding point in the video when clicked [114]. Our envisioned video retrieval engine illustrated in Figure 5.1 directly uses and selects suitable deep-link comments that have been contributed by users on YouTube to create a result list of jump-in points that are relevant to the user's information need.

Note, however, that the application of deep links is not limited to the use case of retrieving jump-in points. Deep links could potentially also be used to offer richer forms of navigation after a user has selected a video for viewing. We will return for further discussion of this point in Section 5.7.

The development and testing of the approach presented in this paper is made possible by crowdsourcing. Through crowdsourcing, it is possible to address a large pool of people from an online community or platform, i.e., the crowd, and obtain input, resources or other services. Here, we use crowdsourcing to gather opinions on deep-link comments in order to understand the deep-linking phenomenon on YouTube, and discover new relevance criteria. Crowdsourcing is also used to evaluate the ability of the approach to improve non-linear video access, with a focus on the reranking application illustrated in Figure 5.1.

The paper is structured as follows. First, we continue our discussion on the novel value of deep links in Section 5.2 and state our key contributions, as well as the five research questions addressed in this paper. We then cover related work in Section 5.3. In Section 5.4, we develop and validate a typology of deep-link comments through a series of user studies carried out on a commercial crowdsourcing platform. We train automatic text-based classifiers in Section 5.5 to test the suitability of deep links for practical use. Using these classifiers, we investigate the potential of reranking jump-in points (cf. Figure 5.1) in Section 5.6 through an additional crowdsourcing user study. The paper concludes with a discussion that summarizes our findings and provides pointers for future work (Section 5.7).

5.2. KEY CONTRIBUTIONS AND NOVELTY

The specific value of deep links for non-linear video access has two aspects. First, the presence of a deep-link comment at a time-point in a video allows us to make an important background assumption for our work. This assumption is that the deep-linked moment is more *noteworthy* than other moments in the video, since it triggered a user to add a comment. We understand noteworthiness as a general indicator that a specific moment is worth an investment of viewing effort, independently of the specific interests of individual viewers. We do not assume that moments lacking deep links are not noteworthy, but rather that the precision of the result list will be higher, if we focus on moments that have already been commented. Choosing this narrow focus does not necessarily limit our output, since new deep links can be exploited by the video search engine as soon as they are added by users.

Second, and more specifically to the contribution of this paper, deep links allow us to make a significant departure from indexing methods that analyze the video itself, and, in particular, from work on non-linear video access involving content analysis. Such methods provide users with time-codes relevant to their queries by detecting visual semantic objects [90] or by detecting highlights within a video [37]. Although content analysis has a clear contribution to make towards moving forward the state of the art of video search engines, here, we take another perspective. Specifically, we note that aspects of a video important for users are often inherent in the audience reception, i.e., the reaction of the people watching the video, and thus not derivable from the video itself. For example, the 'newsworthiness' of a news broadcast is not an intrinsic property, but instead is function of the response of the audience [85]. Here, we do not look at news broadcasts, but rather online social video, specifically at videos on YouTube.

On the basis of our observations concerning these two aspects, we introduce the concept of *Viewer Expressive Reaction* (VER), i.e., aspects of audience response that are spe-

cific to particular time points in the video. By analyzing the textual content of YouTube comments containing deep links, i.e., time-codes that have been added by viewers, we show that VER occurs in multiple varieties, and demonstrate that text-based classifiers are capable of differentiating between these varieties. The different kinds of VER (referred to as *VER Varieties* or VERVs) encompass, but extend beyond topic and affect to include new dimensions of relevance. Similarly, in relevant related research, increasing attention has been recently devoted to video search engines that extend topical relevance, with additional dimensions, such as affect [37] or intent [35]. Our unique contribution is the introduction of new dimensions of relevance based on our VERV typology at the time-code level, rather than at the video level.

The contributions of this paper are threefold: introduction of a VERV typology for video deep-link comments, demonstration that text-based classification can be applied to sort deep-link comments by VERV, and demonstration of the impact of reranking video jump-in points with respect to VERVs. Note that in this paper, we do not address the inference of relevance dimensions from queries. Rather, we focus on two key scenarios necessary to provide a proof-of-concept of the use of deep-link comments for non-linear video access. The work presented here builds on and significantly extends our previous work on the topic of deep links, initially presented in [106]. Specifically, the creation of the VERV typology is explained in detail, we introduce improved classifiers, and we carry out an extensive validation of the usefulness of deep links for reranking jump-in point results.

In this paper, we build a case for our approach exploiting deep links for non-linear video access by addressing five research questions. The first two are devoted to establishing a practical typology of Viewer Expressive Reaction Varieties. This was accomplished by carrying out user studies out on a commercial crowdsourcing platform, which analyzed a large number of deep-link comments (Section 5.4).

RQ1 Which types of deep links exist?

RQ2 Are they human-interpretable?

The next question investigates the potential of deep links for practical use (Section 5.5).

RQ3 Is it possible to classify deep links automatically with a text-based classifier?

The final two questions investigate the potential of applying VERVs to rerank a list of jump-in points drawn from videos that are topically related to a user query. This investigation is again carried out with a crowdsourcing user study (Section 5.6).

RQ4 Do we see evidence that people notice a difference when video search reranking based on a given deep-link category is applied?

RQ5 Do people report such reranking to be useful?

Before addressing these research questions, the next section covers related work and discusses how our approach goes beyond existing techniques.

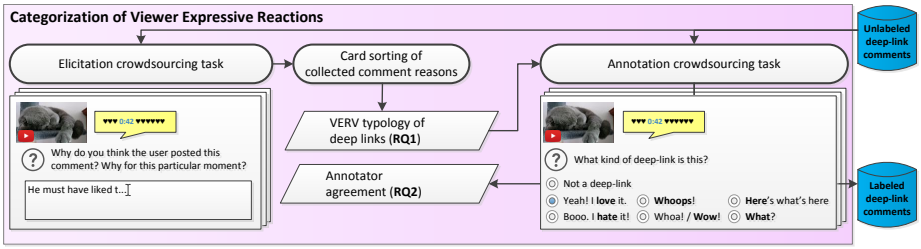


Figure 5.2: Overview of our approach for categorizing deep-link comments. Different kinds of deep links are discovered through a crowdsourcing experiment and card-sorted into a typology. The typology is validated through a crowdsourced annotation process, which also gives us a set of labeled deep-link comments for training classifiers.

5.3. RELATED WORK

In this section, we put our work in context by focusing on two different aspects. First, we take a look at existing relevance criteria for non-linear video access (Section 5.3.1). Second, we discuss how previous research has made use of user comments for information retrieval purposes (Section 5.3.2).

5.3.1. RELEVANCE CRITERIA FOR NON-LINEAR VIDEO ACCESS

Non-linear access to video content has been researched extensively in the past. The study by Yang and Marchionini [112] revealed, among other insights, that users would like to have access to scene or shot level information in order to find a small piece of a video more easily, rather than having to go through the whole video to find that piece. Past research that contributes results in this direction includes work on high-level temporal video segmentation, aiming at automatically discovering the boundaries of semantic (topical) segments potentially interesting as retrieval units [36]. More recent research in this direction has addressed the problem of generating pointers to different video segments, like those based on visually depicted semantic concepts [90] or affective peaks [37], as already mentioned in the previous section. More general approaches targeting the detection of interesting jump-in points in video deployed implicit relevance criteria, like user playback behavior [13] [99]. However, none of the methods discussed so far considered the notion of noteworthiness or studied reactions that are explicitly expressed by the viewer for detecting noteworthy jump-in points. We believe an approach exploiting deep-link comments is a promising one to provide insights in this respect.

5.3.2. USER COMMENTS FOR RETRIEVAL TASKS

User comments have proved to be useful for various information retrieval tasks. In the survey by Potthast et al. [76], the authors define three tasks for which user comments can be used: filtering, ranking, and summarization. Orthogonally, the authors also define two usage paradigms: comment targeting (i.e., the comments themselves are the retrieval target) and comment exploiting (i.e., the commented items are the target of retrieval). The most interesting insight expressed by [76] is that, even though comments contain a limited amount of information, users want to find some kind of "surprise" in these comments, such as "complementary information but also a joke". We wish to ex-

tend this insight from conventional comments to deep-link comments. Hsu et al. studied how an online community perceives the relative quality of its user comments with the aim of predicting the relative order of comments, such that potentially useful yet unrated comments get the chance of accumulating ratings [42]. Chelaru et al. focused on using social feedback, including general user comments, to better rank YouTube videos [11]. In further work, Siersdorfer et al. [86] have noted the correspondence of the presence of links to quality comments on Slashdot and related contexts. In our work, we go beyond the study of comments in general, to look specifically at the contribution of comments with deep links.

Very few studies, however, investigate user comments that refer to specific time-points of a video. Laiola Guimarães et al. discuss the deep-linking phenomenon on YouTube [54] and point out its shortcomings from an user interaction point of view. The authors argue that, since the deep-link comments appear somewhere below the video player (buried under other comments) and are not synchronized with the video playback, the phenomenon does not reproduce commenting-while-watching activity. They therefore propose and evaluate a video commenting paradigm through a user interface that allows users to directly create time-linked comments in a similar fashion to a Japanese social video platform [68]. Wakamiya et al. [108] studied the use of comments tied to a specific time interval and spatial region for the purpose of extracting and retrieving scenes from video clips. The comments that were studied were created in a laboratory setting on a limited scale. The described system required users to explicitly specify intervals, similar to the idea presented in [54], and cannot benefit from time-link comments that were generated casually by users. Kordumova et al. extract terms from social commentary on Twitter that discuss live broadcasts to present jump-in points for the recorded versions, using the time the comment was posted as an implicit time-code [52].

The study by Madden et al. [62] is the closest to our work. The authors propose a classification scheme for classifying YouTube comments. The scheme consists of a large number of categories and sub-categories. The authors offer only guidelines for classifying YouTube user comments into their proposed comment classes, but did not test whether the comments could be automatically classified by a machine. Our typology presented here differs from the one by Madden et al. as it is specific to deep-link comments that are associated with specific time-points of a video.

5.4. CATEGORIZATION OF VIEWER EXPRESSIVE REACTIONS

In this section, we address two research questions. First, *RQ1*, i.e., what kind of deep-link comments exist? Second, *RQ2*, i.e., can they be grouped into intuitive categories? We apply a method similar to the social-Web mining approach taken in [35] to find answer to *RQ1*. We crawled YouTube for deep-link comments for a set of videos from an existing dataset (Section 5.4.1). We then asked a large pool of workers on a crowdsourcing platform to analyse the crawled deep-link comments and provide explanations of why people posted them (Section 5.4.2). Through an iterative card-sorting process of the responses, we arrived at a typology built on a set of abstract dimensions (Section 5.4.3). Finally, we ran a second crowdsourcing experiment, but this time to annotate our dataset of deep-link comments. The annotation process also serves to answer *RQ2* (Section 5.4.4). Our approach is summarized in Figure 5.2.

Table 5.1: Number of videos per YouTube category in the annotated deep-link comments set

music	454	sports	96	howto	14
entertainment	388	autos	48	games	8
comedy	214	news	36	tech	8
film	178	animals	22	education	4
people	115	travel	18	shows	3

5.4.1. COLLECTING THE VIDEO DEEP-LINK COMMENTS DATASET

The dataset used for addressing our first two research questions consists of 1,707 YouTube videos taken from the publicly available MSRA-MM dataset [56] and their associated user comments and video metadata. These were collected through Web scraping since the YouTube API only provides access to the 1,000 most recent comments for any video. These videos were selected because they fulfilled three criteria. First and second, a video had to have "survived" since the MSRA-MM dataset release (June 2009) up to the time of our crawl of their associated comments and metadata (October 2012) and also had to have accumulated at least one deep-link comment by that time. Third, a video had to plausibly be in the English language to ensure most of the crawled comments would be in English as well. For the third criterion, we heuristically assumed that the language of the description was the overall language of the video. The language of the video description was determined through use of a language identification tool [61].

For each video, we randomly sampled three deep-link comments from all the deep-link comments associated with it. The sampling procedure was chosen to yield a deep-link comment set that was sufficiently large, yet still manageable for conducting crowdsourcing experiments. Through a large crowdsourcing study, annotations were collected for a total of 3,659 deep-link comments (Section 5.4.4). Table 5.1 shows the distribution of YouTube categories of the corresponding videos for which annotated deep-link comments exist.

5.4.2. COLLECTING DEEP-LINK MOTIVATIONS FROM THE CROWD

In order to discover why users post deep-link comments and derive the different types of deep links in existence, we make use of crowdsourcing. Crowdsourcing allows us to review deep links at a larger scale in two different ways. First, it allows us to process more deep-link comments than we could possibly do exclusively by hand. Second, it allows us to elicit input from a more diverse pool of human judges. This technique has been successfully applied to collect user perceptions of multimedia, e.g., in [104] it was used to create classes of user-perceived multimedia similarity.

We ran our crowdsourcing experiments on Amazon Mechanical Turk (AMT, `www.mturk.com`), a large commercial crowdsourcing platform. We published the microtask, commonly called a Human Intelligence Task (HIT) on AMT, for our elicitation experiment under the title: "Why watch this video moment? Give reasons for which people create video time-links." This HIT presented workers with a single video and three deep-link comments. The workers were informed that the comments had been written by a viewer of the video. They were instructed to watch the video around the time-point mentioned in the comment and provide a description of the reason for which they thought

Table 5.2: Examples of reasons provided by workers in the elicitation crowdsourcing task

Comment	*at 2:15 Zac has his mike on the stand when he clearly took it off earlier[...]*
Reason	"this comment is mentioning the technical errors in the movie [...]"
Comment	*luv the pic 1:38 thru 1:40 nice video*
Reason	"He must have liked the pics of Jennifer Aniston, [...]"
Comment	*what the name of this song there is from 0:37 to 0:44[...]*
Reason	"He was curious about the song title. [...]"

that the person posted the comment. We applied two criteria to ensure high quality work: First, we explicitly stated the quality requirement in the HIT, which specified that workers must provide answers using two to four sentences. Second, each answer was inspected by hand and we discarded those that were grammatically not reasonably well formed. Note that with this HIT, we are concerned with *discovering* reasons why people create deep-link comments. In other words, the goal is not to be exhaustive, but instead to understand the main trends, which we explore further in the next section through a card sorting process (Section 5.4.3).

For the elicitation HIT, we took 33 videos from our dataset and sampled three comments from each of the videos. Each video and its three comments was reviewed by three different workers. Each worker received a compensation of US$0.11 per task. In total, 17 different workers participated in the experiment. Examples of reasons provided by these workers are listed in Table 5.2.

5.4.3. VERV: A VARIETY OF VIEWER EXPRESSIVE REACTIONS

In order to capture the variety of viewer expressive reactions, we applied a card sorting technique [82] to the responses we collected in the crowdsourcing experiment. By iteratively forming new groups, we ended up with a typology of six different classes that characterize the reactions expressed in deep-link comments. The six classes were chosen with the aim of covering the main trends of deep-link comments observed in the previous HIT. To fulfill this aim, we only added a new class to our typology if we could also find a contrasting class. In addition, we chose the labels of the classes to be short, intuitive colloquial phrases that would capture the spirit of the comment and could easily have been typed instead of the comment by the original user who posted the comment. The six classes, which we refer to as Viewer Expressive Reaction Varieties (VERVs), are listed in Table 5.3 and will be discussed in the remainder of this section.

Quite a few reasons that workers listed boiled down to the original person expressing a liking towards a certain point in the video. We labeled this class of comments as "Yeah! I love it." (`love`). Comments expressing the opposite, although less frequent, were also found, leading to the class "Booo. I hate it!" (`hate`). These two classes cover the personal taste of the original writer of the comment and map to positive and negative valence, respectively.

Table 5.3: Viewer Expressive Reaction Variety (VERV) typology of deep-link comments

VERV class	ID	Description
Yeah! I love it.	`love`	Personal, positive reaction to linked time-point
Booo. I hate it!	`hate`	Personal, negative reaction to linked time-point
Whoops!	`whoops`	Reacting to a mistake or blooper
Whoa! / Wow!	`wow`	Reacting to something uncommon that is surprising and impressive
Here's what's here	`here`	Describing or providing neutral information
What?	`what`	Asking for information

Another set of reasons reflected surprise in viewers. In these type of comments, the viewers reacted to something uncommon and impressive that happened in the video and that exceeded conventional expectations, thus transcending the personal nature of the `love` and `hate` classes. We categorized these comments as "Whoa! / Wow!" (`wow`). The class of comments we positioned to oppose the `wow` class is "Whoops!" (`whoops`). The latter class comprise comments that, instead of surprising the viewer by exceeding conventional expectations in a positive way, actually surprise the viewer by something that should not happen, i.e., mistakes or bloopers.

Finally, some of the comments were of a neutral descriptive nature and only declared what could be seen or heard in the video around a certain time-point. This gave rise to the "Here's what's here" (`here`) class. Complementing this class of comments is the "What?" (`what`) class, which comprises the interrogative comments. Comments in this class were posted by viewers because they were curious and therefore asked questions about a certain point in the video.

In addition to these comment types, the card sorting process also revealed comments that were crawled, but were not true expressive reactions by viewers to a certain point in the video. Instead, they were comments of a promotional nature (e.g., advertisements and chain letters) or contained substrings that share the same appearance as denotations of time-points, but instead referred to something else, such as the time of day, e.g., "3:15 PM", or scripture verses, e.g., "Jeremiah 13:23". These types of comments are not reactions to the video content at a time-point. For this reason, we consider them to be separate from our typology and refer to them as non-VER comments.

5.4.4. VALIDATION OF THE VERV TYPOLOGY

We ran a second crowdsourcing experiment to accomplish two goals. First, it is to find out whether our typology of deep-link comments is intuitive and could easily be understood by others. That is, we check whether human judges, when given Table 5.3, can reliably judge the VERV class of deep-link comments. This would also provide evidence that our typology serves to explain how users create a deep link, i.e., it covers most of the

Table 5.4: Distribution of VERV classes in the annotated dataset

love	934	wow	183
here	873	hate	72
what	560	whoops	53

possible types of reactions. Second, the crowdsourcing experiment allows us to generate ground truth annotations for our dataset. The ground truth is used to understand the distribution of VERV deep-link comments and as the basis for our deep-link comment classification experiments (Section 5.5).

The AMT HIT used for the validation and annotation experiment was titled "Watch short snippets of YouTube videos (and support research on video search engines)". The workers were presented a single video and three selected comments for the video. They were instructed to read the comments and watch parts of the video around the time-points mentioned in the comments prior to answering a series of questions for each comment. For each comment, workers had to judge first whether a comment contained an actual deep link and an actual Viewer Expressive Reaction at all. This question allowed us to isolate comments like advertisements, but also, as discussed in the previous section, isolate comments that happen to contain substrings that resemble time-points but are not meant to designate a time-code of the video. If a comment indeed contained a true deep link and a viewer's expression, the worker was then asked to pick one of the VERV classes, as listed in Table 5.3, that best corresponded to the spirit of the comment. The task also contained an additional set of questions with the goal of collecting more information about the comments and checking the consistency of workers, but they are not further discussed in this paper.

The crowdsourcing experiment consisted of a campaign of five staggered batches of tasks and took fifteen days to complete in February, 2013. The reward of a single task was US$0.09 and each task was performed by five different workers. While most comments were annotated by five workers, some were by less, since some workers failed quality control. Ignoring those whose work failed quality control, 263 workers participated in the experiment. In total, they annotated 3,659 deep-link comments. The ground truth was formed by majority vote. Only comments for which workers reached an agreement were used in our classification experiments (Section 5.5). Workers reached an agreement on 3,552 comments (97.1%) when deciding whether it contained a true deep link and a true expressive reaction. Of these 3,552 comments, 3,110 comments were labeled to contain a true deep link. When deciding on the VERV-class, workers reached a consensus for 2,675 of the 3,110 comments (86.0%). The distribution of the VERV classes is shown in Table 5.4.

We use Randolph's free-marginal multirater kappa [80] to further assess the inter-annotator agreement of the workers. This statistical measure is suitable when annotators are not aware of the marginal distribution of each of categories. We cannot, however, compute Randolph's kappa directly since not every comment has received the same number of annotations due to quality control. We address this issue by randomly selecting three out of at most five annotators per comment and compute the average kappa $\bar{\kappa}_{\text{free}}$ by repeating this process 100,000 times. Applying this method to the VERV annota-

Table 5.5: Annotator confusion matrix

Majority vote	Percentage of all annotations received						
	love	here	what	wow	hate	whoops	
love	27.7%	2.6%	0.2%	3.2%	0.2%	0.4%	34.4%
here	2.4%	25.7%	0.6%	1.9%	0.7%	1.3%	32.6%
what	0.4%	0.6%	19.7%	0.6%	0.2%	0.2%	21.7%
wow	0.9%	0.8%	0.1%	4.7%	0.1%	0.2%	6.8%
hate	0.1%	0.3%	>0.0%	0.1%	1.9%	0.1%	2.6%
whoops	0.1%	0.3%	>0.0%	>0.0%	0.1%	1.3%	1.9%
	31.7%	30.4%	20.7%	10.4%	3.2%	3.5%	100%

tions, we can report a kappa of $\bar{\kappa}_{\text{free}}$=0.5214, which is generally considered to be moderate agreement.

Looking at the sources of disagreement, we can see that workers tend to conflate `love` and `wow` (Table 5.5). Apparently, the distinction between whether a comment expresses a positive personal reaction or whether it expresses surprise is sometimes hard to make. More surprisingly, we can see some confusion between `love` and `here`. As we will see later when we discuss automatic classification (Section 5.5.3), this source of confusion will also affect classifiers, but not necessarily in a negative way.

Given that a consensus could be reached for a large portion of the comments and a moderate inter-annotator agreement was found, we conclude that human judges can reliably classify deep-link comments. We therefore consider our VERV typology to be valid and that the typology is successful in capturing major trends.

5.5. AUTOMATIC CLASSIFICATION OF DEEP-LINK COMMENTS

In this section, we address our second research question *RQ3*, i.e., can machines automatically classify deep-link comments according to the VERV categorization scheme that was derived in Section 5.4? We approach this question by splitting the classification problem in two. First, does the comment contain an actual Viewer Expressive Reaction (VER/non-VER)? As discussed in the previous section, some comments may be spam or contain substrings that resemble video time-codes only on the surface level. Solving this classification problem will act as a filter, and decouples it from the second, main problem: Given that a comment contains a Viewer Expressive Reaction, can we categorize the comment into one of the six VERV classes discovered in Section 5.4? We represent deep-link comments in our classification experiments as standard bags of words using different techniques for tokenization, filtering, and weighing. The different feature combinations that we test are described in Section 5.5.1. We test the different features using a support vector machine (SVM) with two different kernels and a naive Bayes classifier [110]. Both multi-class and one-vs-all approaches are considered. Further details of the experimental setup can be found in Section 5.5.2. Finally, the results of exploring the different features combinations and the evaluation of the classifiers are presented in Section 5.5.3.

5.5.1. FEATURES FOR CLASSIFICATION

We represent deep-link comments in our dataset as bags of words. We focus on only using lowercased unigrams as previous experiments on a development dataset showed us that they generally outperform mixtures of unigrams and bigrams [106]. In previous classification experiments, we employed simple tokenization (splitting on whitespace, quotes and interpunction) and used unigram frequency as a weight for each feature dimension. Here, we will investigate several approaches for tokenizing comments and weighing each feature. For weighing each unigram term, we compare binary weights, frequency weights, and tf-idf [64]. Additionally, we investigate the impact of removing hapaxes, i.e., terms that only occur once throughout the dataset.

Under the hypothesis that some VERV classes such as `what` are sensitive to punctuation, we consider three different methods for tokenizing comments. The first method simply splits comments on whitespace and thus preserves punctuation as part of other terms. The second method ignores punctuation and only considers sequences of alphanumerical characters as unigrams. The third method preserves punctuation as separate tokens and maps them to the following punctuation classes: `question` for one or more question marks; `exclamation` for one or more exclamation points; `interrobang` for a subsequence of question marks and exclamation points; `punctuation` for everything else. Furthermore, all three methods merge time-codes into a single `timecode` token such that all time-code strings are treated equally on a textual level.

5.5.2. EXPERIMENTAL SETUP

Dataset and ground-truth: We use the annotated dataset as described in Section 5.4.1 (consisting of 3,552 comments with VER/non-VER labels and 2,675 comments with VERV class labels) and split it into three subsets: 60% training set, 20% development set, and 20% test set.

Experiments: We run two sets of classification experiments. In the first set, we focus on classifying whether a deep-link comment truly contains a Viewer Expressive Reaction (VER/non-VER). In the second set, we focus on classifying a comment into one of the six VERV classes. In both sets of experiments, we first test the different approaches to weighing features on the development set with unigrams obtained through simple tokenization, i.e., whitespace as delimiter. The best approach is then used for comparing the different tokenization methods and the impact of removing hapaxes. In the set of VERV experiments, we test the different features with multiclass classifiers. In the final test on the test set, we run the experiments both with binary classifiers (one-versus-all) and multiclass classifiers. Additionally, we test the impact of merging the `love` and `wow` classes based on our findings of human confusion between the two classes.

Evaluation: The experiments are carried out using two different classification algorithms provided by the WEKA toolkit [34]: naive Bayes and SVM. We evaluate two different kernels for the SVM: the default linear kernel and the histogram intersection kernel. Results of the classification experiments are reported using precision, recall and F_1 scores for binary classifiers. For multi-class classifiers, we report precision and recall for each of the classes, as well as the weighted F-measure (WFM). In order to compare the classification performance between the training set and the development set, 10-fold cross-validation was used on the training set.

Table 5.6: VER/non-VER classification results on the test set

	Baseline	Naive Bayes	SVM Linear	SVM Hist
Precision	0.879	0.965	0.958	0.956
Recall	1.000	0.917	0.981	0.984
F_1	0.936	0.940	0.969	0.970

Term weighting: tf-idf; Tokenization: whitespace as delimiter, separate punctuation tokens; Hapaxes: removed.

Baseline: We compare all classification results to a dominant class classifier (ZeroR in WEKA) in each of the set of experiments. Remember that the objective of our research question *RQ3* is to find out whether automatic methods can distinguish between different VERV classes. For the VER/non-VER experiments, the dominant class is VER. For the VERV experiments, the dominant class is `love` (Table 5.4).

5.5.3. RESULTS

When determining whether a deep-link comment truly contains a Viewer Expressive Reaction, the choice of term weights appears to be not important. All classifiers achieve an F_1 score of well over 0.9 with minor differences between the weighing schemes. We therefore choose to evaluate the different tokenization and filtering methods using the commonly used tf-idf. Similarly, different tokenization and filtering methods show minor differences in performance. The best performance was achieved when punctuation was preserved as separate tokens and hapaxes were removed. Results of this setup on the test set are summarized in Table 5.6.

The impact of term weighting is more noticeable when classifying the VERV class of a comment. While there seems to be no difference between frequency and tf-idf term weighting, the linear and histogram intersection SVM classifiers benefit from it compared to the simple binary term weights, while the naive Bayes classifier takes a hit in performance. Based on the SVM classifiers performing best, we pursue the common tf-idf weighing approach in evaluating tokenization methods. Here, we find again that the combination of treating punctuation as separate tokens and removing hapaxes gives the best performance. Classes that benefit the most from this tokenization method are `here` and `what` at the expense of the other VERV classes.

In all further experiments, we use the tokenizer that outputs separate tokens for punctuation and removes hapaxes in order to simplify the feature extraction process. Note, however, that for specific problem instances, e.g., classifying a single VERV class, one might want to extract features that are optimized for that specific task.

The results of the multi-class classification experiments are summarized in Table 5.7. From this table, we can see that it is indeed possible for machines to perform VERV classification at a basic level, as the trained classifiers easily outperform the dominant class baseline. Since the naive Bayes classifier was overall the weakest classifier, the analysis below mainly reflects the results obtained with the SVM classifiers. The top three most occurring classes (`love`, `here`, `what`) are the easiest to classify. The `wow` class can be classified with moderate precision, but suffers from low recall. Not unexpectedly, the least common VERV classes (`hate`, `whoops`) are the hardest to classify. For example, none of the classifiers were able to predict any of the `whoops` instances correctly during training

Table 5.7: VERV multi-class classification results on the test set

	Precision / Recall / F_1		
	Naive Bayes	SVM Linear	SVM Hist
`love`	.561 .746 .640	.657 .714 .684	.653 .714 .682
`here`	.697 .354 .469	.650 .667 .658	.662 .662 .662
`what`	.722 .686 .704	.835 .843 .839	.777 .853 .813
`wow`	.273 .444 .338	.615 .296 .400	.600 .333 .429
`hate`	.079 .273 .122	.357 .455 .400	.333 .273 .300
`whoops`	.182 .133 .154	.500 .133 .211	.500 .067 .118
WFM	0.551	0.671	0.663

(Baseline WFM: 0.178)

Table 5.8: VERV one-versus-all classification results on the test set

	Precision / Recall / F_1		
	Naive Bayes	SVM Linear	SVM Hist
`love`	.529 .789 .633	.720 .681 .700	.702 .676 .689
`here`	.626 .446 .521	.699 .477 .567	.699 .523 .598
`what`	.653 .755 .700	.818 .794 .806	.822 .814 .818
`wow`	.215 .630 .321	.500 .259 .341	.529 .333 .409
`hate`	.065 .364 .110	.333 .182 .235	.667 .182 .286
`whoops`	.176 .200 .188	.000 .000 .000	.000 .000 .000
`love+wow`	.569 .816 .671	.681 .745 .712	.664 .726 .694

and evaluation on the development set, i.e., the precision and recall for this class were 0. The same pattern is reflected in the low performance of the binary classifiers on the test set (Table 5.8).

When investigating the impact of merging the classes `love` and `wow` as directed by our findings of the human annotator confusion, we noticed that the overall performance (WFM) of the multi-class classifiers only improved when evaluating it using the development set (e.g., from 0.638 to 0.691 for the linear SVM), but actually degraded when tested on the test set. We do see an improvement in the case of binary classifiers, as shown by Table 5.8.

The confusion matrix of one of the classifiers is shown in Table 5.9. A large portion of confusion is between the `love` and `here` VERV classes, similar to the confusion among human annotators. An interesting instance of a misclassified comment is the following: "*funny pic at 1:35.*" This comment has been labeled as `here` by human annotators, while the classifier classified it as `love` instead. Interestingly, the notion of "funniness" is not always considered to be a personal opinion by annotators, but some see it simply as a piece of neutral information about a point in the video. Exploring the dataset further reveals that there is less agreement for comments that clearly have multiple interpretations or consist of multiple sentences. For example, a comment like "*At 3:04, Zac touch the nose of Vanessa... it's so cute!!*" contains both a neutral descriptive part and a positive personal reaction.

Table 5.9: Linear multi-class SVM confusion matrix for the test set

	Predicted					
True	love	here	what	wow	hate	whoops
love	132	44	6	1	2	0
here	47	130	7	3	7	1
what	3	11	86	1	0	1
wow	11	5	3	8	0	0
hate	0	3	0	0	5	0
whoops	5	7	1	0	0	2

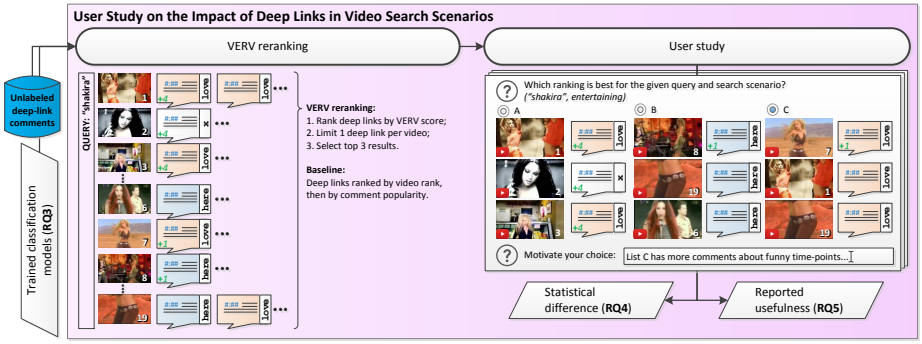


Figure 5.3: Overview of our evaluation of the VERV typology in online video search scenarios. We rerank a results lists with respect to a particular VERV class using previously trained classifiers. The reranked lists (together with the original ranking) are then evaluated by the participants of a user study carried out on a crowdsourcing platform.

In sum, we see that the automatic classifier performs reasonably well, and that many mistakes it makes concern cases that are also ambiguous for humans. We conclude that VERV classification shows promise of being robust enough for use in video search systems. We investigate this promise in more detail in the next section.

5.6. USER STUDY ON THE IMPACT OF DEEP LINKS IN VIDEO SEARCH SCENARIOS

This section addresses our final two research questions. First, *RQ4*, i.e., do we see evidence that people notice a difference when video search results with deep-link comments are being reranked according to the comment's VERV class compared to a challenging popularity baseline that preserves the topical relevance of the videos and selects the most liked deep links? Second, *RQ5*, i.e., do people report the VERV reranking to be useful? We use classifiers trained in the previous section to rerank results (Section 5.6.1). The original and reranked lists have been used in a user study on Amazon Mechanical Turk in which the participants were asked select the best list of search results under given conditions and motivate their choice (Section 5.6.2). The conclusions on these results are presented in Section 5.6.3. Our approach is summarized in Figure 5.3.

5.6.1. RANKED RESULTS DATASET

The dataset used for this user study is an extension of the dataset described in Section 5.4.1. Originally, more videos and user comments were crawled than were actually annotated due to the sheer size of the set. In total, 91,123 deep-link comments were used here.

The original MSRA-MM dataset describes for each video the query that was used to retrieve the video and its original rank in the results list. We use this query information as a basis for simulating a retrieval engine. Some of the queries, however, were unclear. For the set of queries associated with the deep-link comments' videos, we had each query labeled by two other researchers as either clear or unclear. A researcher had to mark the query as unclear if it was difficult to imagine the information need of the user submitting the query. For example, the query "video" would be considered to be too generic. A query would only be included in the final set of queries if both researchers agreed upon the query being clear and the query would have at least 5 videos with deep-link comments associated with them. The second requirement was included to make sure that the set contained a sufficient number of deep-link comments from a diverse set of videos to reduce the risk of producing identical reranked lists for the user study. In total, 7 researchers agreed on 84 out of 143 queries being clear. Of these 84 queries, 45 queries had at least 5 videos and were used in the user study.

5.6.2. CROWDSOURCING USER STUDY

Our crowdsourcing user study investigates whether people notice a difference between differently ranked results lists. In this study, we follow the method presented in [101] and instruct workers to project themselves into the role of another person who likes watching online videos. This person is said to be using a video search engine that provides him search results that have been enriched with deep-link comments. The task explains the general idea behind this search engine and then asks the worker to imagine one of the possible search scenarios: (1) the person wants to be entertained today (`ENT`), or (2) he wants to be informed (`INF`). The two scenarios are based on two VERV classes (`love`, `here`). We limit ourselves to these two scenarios for two reasons. First, these two VERV classes translate best into a search scenario that would be easy for participants to understand and to imagine. Second, the classifiers for these two VERV classes are most reliable.

The worker is then presented with a single query and three lists of search results. Each list consists of the top three deep-link results like the example shown in Figure 5.1. Note that the figures in this paper contain video thumbnails for illustrative purposes, but in the user study, thumbnails were hidden from the workers to put the focus on the deep-link comments. The workers are asked to pick the best results list that suits the given scenario described at the beginning of the task and motivate their answers (cf. Figure 5.3). Afterwards, we also ask the worker whether it was difficult to complete the task, and if so, why. This helped us to identify any systematic issues with the judgements.

The three search results lists presented to the worker are ranked as follows. First, one of the lists is ranked according to the original ranking in the MSRA-MM dataset. Here, we rank the deep-link comments first by the rank of the corresponding video and then rank by popularity, i.e., the number of likes the comment has received minus the num-

Table 5.10: Search results list preference for the query "shakira"

	search results list R			
scenario S	`orig`	`love`	`here`	
`ENT`	11	16	3	30
`INF`	14	4	12	30
	25	20	15	60

$$\chi^2 = 12.96, \quad p = 0.002$$

ber of dislikes. We refer to this ranking as `orig`. Note that we consider this ranking to be a challenging baseline that preserves both the topical relevance of the full video and selects deep links which are well received by the YouTube community. Second, one of the lists is ranked according to the `love` VERV class. For this list, we rank deep-link comments based on the likelihood of belonging to the `love` class using a previously trained classifier. The hypothesis is that this reranking corresponds to the targeted scenario in which the person wants to be entertained. Third, one of the list is ranked according to the `here` VERV class in a similar manner. The hypothesis here is that this reranking corresponds to the scenario in which the person wants to be informed. To ensure a diverse set of results, we only allow one deep-link comment per video in each of the reranked lists. These three ranking methods were chosen to be simple, yet to allow for contrastive experimental conditions. For completeness, we mention that reranking introduced substantial differences between the lists. For all but one of the 45 queries, all three lists were different both with respect to the videos and also the deep links. In the exception case, two lists listed the same videos, but the deep links were different.

5.6.3. EXPERIMENTAL RESULTS

The crowdsourcing campaign published on AMT was set to have each task to be completed by 30 different workers. The campaign consisted of 90 different tasks, one for each pair of query (45) and search scenario (2). Workers were rewarded US$0.13 per task. In total, 321 workers participated in the campaign and it took them one month to complete all 2,700 assignments. The results of the experiment can be essentially captured in a series of contingency tables. The contingency tables count how often a particular ranking (R) was selected given a certain search scenario (S). We can compute such a table for the full experiment, but also for each query separately (e.g., Table 5.10). For these tables we can compute the Pearson's χ^2 statistic to test whether the variables R and S are independent (H_0) or dependent (H_1).

Considering all queries collectively, we can reject the null hypothesis H_0 that R and S are independent under a significance level of $\alpha = 0.05$. This observation leads us to the main conclusion of the user study, namely that people generally perceived a difference between the differently ranked results under different use scenarios. In short, the study provides evidence that VERV reranking has the potential to make a productive contribution to video search approaches that exploit deep links, such as the one illustrated in Figure 5.1.

Table 5.11: Search results list preference for the query "nba crossover"

	search results list R		
scenario S	`+orig`	`-orig`	
`ENT`	18	12	30
`INF`	8	22	30
	26	34	60

$$\chi^2 = 5.498, \quad p = 0.019$$

Table 5.12: Search results list preference for the query "rihanna"

	search results list R		
scenario S	`+love`	`-love`	
`ENT`	21	9	30
`INF`	13	17	30
	34	26	60

$$\chi^2 = 3.326, \quad p = 0.068$$

In order to better understand this conclusion, we now turn to dissecting it in detail. Since considering all queries collectively does not give insight into the impact of VERV reranking in specific cases, we first discuss the results in terms of individual queries. If we look at each query in isolation, we find that we can reject H_0 for 7 out of 45 queries. Note that when we consider each query individually, the significance is calculated here over a smaller amount of data, making this test a stringent one. We present this calculation in detail for the query for which the effect was strongest (Table 5.10) for the purpose of illustrating how it was carried out.

Although this analysis is interesting, it does not yet tell us which VERV ranking (R) contributed to the difference under which scenario (S). To gain insight into this point, we need to carry out a one-versus-rest comparison. We do this by keeping the column of one ranking, merging the other two columns, and computing the χ^2 statistic for the new contingency table. This process is illustrated for two different queries in the case of the original ranking `orig` (Table 5.11) and of the `love` VERV ranking (Table 5.12).

We carry out a one-versus-rest comparison for each query and each reranking of the seven queries that shows a significant impact of VERV reranking according to our test. We found that in four cases, a single reranking contributed to statistical significance: once `love` was strongly preferred in the `ENT` scenario, once `orig` in the `ENT` scenario (i.e., Table 5.11), and twice `here` in the `INF` scenario. For the other three queries, the significance could be attributed to two different rerankings. In one case, `here` was clearly preferred in the `INF` scenario and `love` in the `ENT` scenario. The other two cases were similar, except it was the `orig` reranking users preferred in the `INF` scenario. From this analysis, we can see that whenever the `love` and `here` rerankings are strongly preferred, it happens under the `ENT` and `INF` scenarios, respectively. This observation is consistent with our hypotheses concerning the correspondence of VERV reranking to search scenarios (cf. Section 5.6.2). Also note that the baseline `orig` cannot be linked specifically to either scenario.

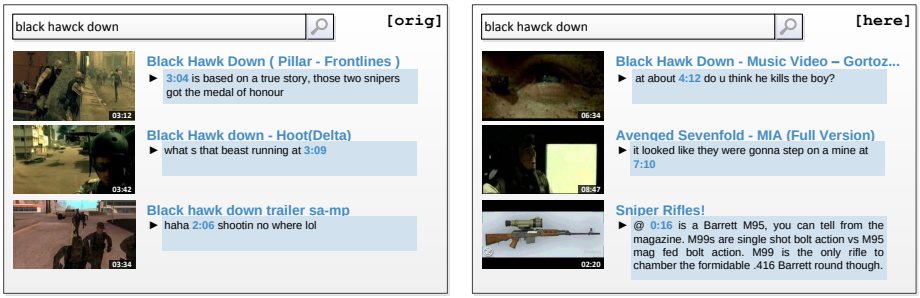


Figure 5.4: Topical vs. VERV relevance. Each of these two results lists presented in the user study for the query "black hawck down" contains an informative comment. When asked to pick the most informative list, participants often preferred the original list on the left because it is more topically relevant. This example shows that VERV relevance at the deep-link level needs to be considered in conjunction with topical relevance at the video level.

Next, we turn back to consideration of the overall contribution of the individual VERV rerankings to users' perceptions of results lists. To this end, we carry out the one-versus-rest comparison just discussed on the full set of 45 queries. This analysis allows us to discover, overall, which VERV reranking contributed most strongly to the user's perception that the lists were different for a given scenario. We find that for the `INF` scenario, there was no clear overall VERV that was responsible. However, for the `ENT` scenario it was clear that the VERV reranking `love` was clearly preferred.

Finally, we dive into how the individual examples were perceived from the point of view of the user study participants. Here, the comments that people gave during the user study about their choices provide additional insight onto why they made particular choices. These comments are interesting because they serve to shed light on the reasons for which a given VERV reranking is not necessarily preferred for a given scenario across the board.

The most interesting factor at play is that a specific VERV reranking may sacrifice user perceptions of the topical relevance of the videos in the results list to the query. This is illustrated by the query "black hawck down" in Figure 5.4. When judging the reranking of `here` under the informative scenario, one participant made this comment, *"List A [`orig`] contains one concise informative comment. List C [`here`] appears to have an informative comment, but the video isn't actually Black Hawk Down"* (Figure 5.4), revealing the importance of the video, as well as the snippet, being obviously topically relevant to the query.

In general, the topical relevance of the query to the results is complex for queries closely associated with entertaining content, such as movies and music. We mention this issue explicitly since most of the queries in our query set are of this type, due to the nature of content on YouTube. Topical relevance at the level of the video appeared to impact the preference of users for one list over another. The comments revealed that sometimes one reranking was preferred because it contained performances by the original artist rather than a cover version. However, topical relevance at the deep link level was less clearly evident or important to the user study participants, revealed by the re-

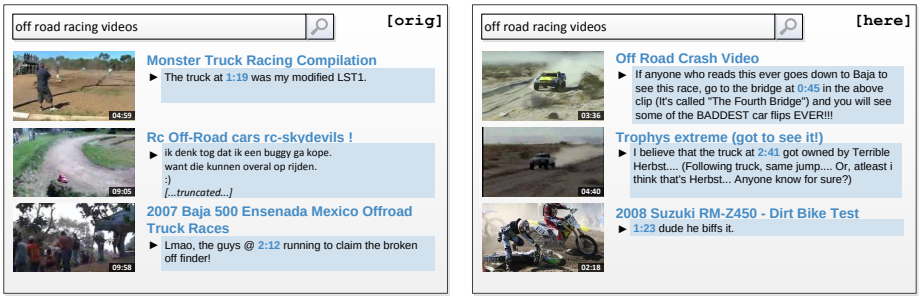


Figure 5.5: Comparison between the original results list and the list ranked with respect to the `here` VERV class for the query "off road racing videos". Participants in the user study found the `here` list to be the most informative list of results.

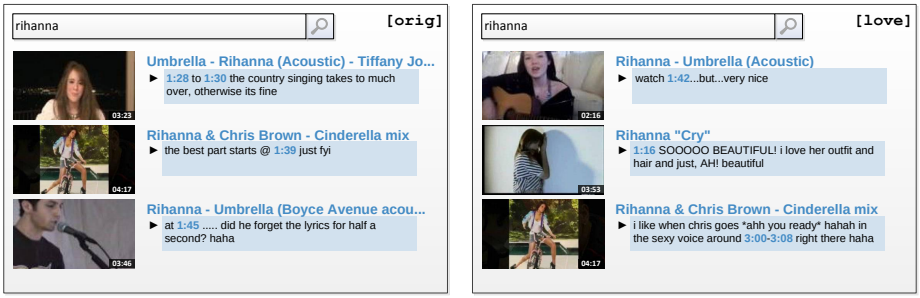


Figure 5.6: Comparison between the original results list and the list ranked with respect to the `love` VERV class for the query "rihanna". Here, the reranked list shows deep-link comments displaying appreciation as opposed to criticism.

marks, (*"Honestly, none of them seem too informative"*), and (*"Funny videos [the query] is not a subject that you can be easily informed about"*).

We also mention a set of practical considerations. Comments containing profanity turned off some participants and, consequently, were at times completely ignored. Non-English comments slipped through as filtering was done at the video level (Section 5.4.1). However, interestingly, some participants were still able to make some sense out of non-English comments as evidenced by the following excerpt: *"Even though one of the videos wasnt in english the emoticons allowed me to infer that it was really funny"*.

Next we turn to consider the queries for which the impact of VERV reranking did not achieve statistical significance. We find that in a number of cases the study participants provide explanations for their choices that indicate that their preferences are indeed impacted by VERV. One case, is the comment refering to emoticons just mentioned. Other participants reported looking for "laughing comments", and used these to conclude that the result was funny or entertaining. When they were asked to look for a ranking that contained informative results, we found that participants generally looked for the most descriptive comments. However, we also found personal preference plays a role here. As such, we encountered explanations like *"List C was mostly (more or less) useless 'lol' comments, which gives no real indication as to what's under them."*, which show that some

people prefer descriptive comments even when they were asked to look for entertaining results. On the other hand, some participants explicitly ignored very descriptive comments in this scenario, reasoning that *"they tell what is happening which kind of gives away the punchline"*. Personal preferences also played a role when it came to songs. Some participants made their final choice for a particular ranking because they thought the list contained a better selection of songs.

We observed evidence that the difference between video-level relevance and time-point level relevance is subtle, and is difficult to distinguish for some participants. For example, one participant reasoned: *"List A includes an interview which would likely to be informative"*. In a different query and scenario, one participant mentioned: *"List A may entertain more than other two as it has a live concert video"*. Note, however, that the videos in the experiment were selected based on the VERV relevance of the retrieved comments, not the relevance of the video as a whole, and thus it may happen that, e.g., a very informative interview gets included in a `love` list because of a funny moment.

Still, a clear example of VERV reranking that shows promise is shown in Figure 5.5. Here, participants clearly preferred the `here` list in the `INF` scenario. They preferred this list for reasons pertaining to both the type of videos (*"List C includes a dirty bike test which could be useful."*) and the informative comments (*"A clue [...] is found in the first comment because a name of bridge and where to find it is given. Similarly, the second comment is informative because it gives the name of a truck owner at a specific time."*). Another example is shown in Figure 5.6. While no statistically significant preference was observed (Table 5.12), here we can see that the `love` list contains deep-link comments that display appreciation for the moments within the video rather than criticism. This is also reflected within the comments from the participants, e.g., *"It seemed like the comments were favorable for all of the videos in this set, as opposed to the other sets that had one or two videos with negative comments."*

Finally, while participants were not explicitly asked during the study whether they found the VERV reranking useful, mainly to avoid default 'yes' answers, we did find some participants commenting on its usefulness. One participant commented on YouTube comments in general: *"I personally tend to read comments whenever possible to inform myself if a video is relevant or enjoyable to watch."* We also found supporting evidence for findings by Potthast et al. [76] as discussed in Section 5.3: *"The opinions expressed in the comments don't always provide information on the content of the videos. However they prove to be slightly more useful when it comes to choosing videos for entertainment value."* Besides using comments to assess whether a video as a whole is relevant, several participants commented on how deep links themselves are useful and how these deep links enabled non-linear access to interesting segments: *"The comments seem like they give me some useful points to watch"*, and *"the time-points would allow me to jump right into the action"*.

5.7. DISCUSSION

In the experiments just presented, we investigated the usefulness of deep links for improving video search results and how they could open up a new perspective on the relevance of a video by focusing on individual video segments rather than the video as a whole. Under this perspective, deep-link comments offer a way of providing non-

linear video access via their time-codes, which can match alternate dimensions of user needs in addition to the topical relevance of the whole video. We developed a Viewer Expressive Reaction Variety (VERV) typology that covers six different types of deep-link comments (*RQ1*, Section 5.4) that capture how users deep-link to noteworthy moments in a video. The six different VERV classes were intuitive to human annotators as they were able to come to a reasonable agreement (*RQ2*, Section 5.4.4). The most surprising conflict was that some annotators find "funniness" not a personal opinion, but a neutral statement instead. Using annotations obtained from the human annotators, we were able to train machine classifiers that outperform a dominant-class baseline using unigram text features (*RQ3*, Section 5.5). Using the trained classifiers to rerank videos with deep-link comments, we set up an initial crowdsourcing user study to compare our reranking method to a popularity baseline that preserved the original ranking and selected the most liked deep-link comments. The study showed that people indeed notice a difference (*RQ4*, Section 5.6.2), and that they report the reranking to be useful (*RQ5*, Section 5.6.2). Even when a difference is not statistically significant for a single query, manual inspection of individual cases shows that some participants in the study are still able to motivate why one particular reranking is preferred using convincing arguments. In several cases, participants explicitly noted in their general feedback how deep links would help them to directly access useful and interesting segments, which shows that the use of deep links for enabling non-linear video access looks promising.

As already mentioned in the introduction, the proof-of-concept presented here is not the only potential application of our work to enable non-linear video access. In addition to serving as the basis to create a results list of jump-in points, deep links can also be used for other applications that allow users to interact with video on the basis of time codes. Here, we discuss the possibility of using deep links to offer a rich form of navigating through a video that the user has selected for playback. For example, deep links of various VERV types could be marked on the video's timeline, allowing the user to quickly jump to a different part of the video. We carried out an additional experiment with the goal of shedding light on how our approach would perform in such an application. Our experiment was designed to approximate the proportion of deep links per video that would be correctly classified if a user was presented with a timeline of a video that was created from deep-link comments belonging to a certain Viewer Expression Reaction variety. For this purpose, we needed a classification score for every point in our dataset, and not just the test data. In order to obtain the classification scores, we carried out leave-one-out cross-validation. We restricted our investigation to the 428 videos in our dataset associated with three labeled deep-link comments. The result was that for 74% of the videos at least two out of three deep-link comments were correctly classified.

For future research on the use of deep links, we see three important directions. First, future research should focus on further improving automatic classification of deep-link comments into specific VERV classes. For instance, when large amounts of resources are available to annotate data, a larger training set can be used. Second, future research should investigate modalities beyond text, exploring how deep-link comments relate to the video content could help to further improve classification. For example, one can think of leveraging specific audio cues, such as the presence of laughter near the linked time-point. Third, our user study also revealed that it is worthwhile to conduct addi-

tional evaluation studies to investigate the right balance between the VERV relevance of the deep link and the topical relevance of the video.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme under grant agreement № 287704 (CUbRIK). We thank everyone who participated in our experimental studies by carrying out the task on Amazon Mechanical Turk.

6

COLLECTING REALISTIC VIEWING BEHAVIOR FROM THE CROWD FOR NON-LINEAR VIDEO ACCESS

In this penultimate chapter, we investigate the creation of enriched video representations using collective playback*, an aggregation of the viewing patterns of individual viewers. We build on the assumption that viewers' play and skip behavior reflects parts of videos that are most memorable and worth re-watching. Such playback information is collected in large quantities by commercial online video platforms, but is not generally available to researchers outside of the companies that run those platforms. We present a methodology that addresses this challenge, and provide insights by studying data collected with this methodology. The novelty of the approach is that it allows us to collect a larger amount of realistic playback behavior than previously possible outside of commercial settings. The methodology is applied in the form of a crowdsourcing campaign that is designed to elicit realistic playback behavior. Our investigation of the study results provides three insights on collective playback behavior: first, it converges, second, it shows promise of being useful for non-linear video access and, third, it contributes added value to non-linear video access above and beyond existing content analysis approaches.*

This chapter is currently being reviewed as **Raynor Vliegendhart, Martha Larson, and Alan Hanjalic. Collecting realistic viewing behavior from the crowd for non-linear video access. Under review** [100].

6.1. INTRODUCTION

In order to enjoy a video, or to absorb its content, viewers can take one of two approaches. They can watch the video linearly from end to end, or they can watch the video *non-linearly*, which means they skip through the video to access only specific parts. Linear video access is well supported by the standard online video players in widespread use today. In comparison, support for *non-linear video access* is relatively underdeveloped. Today's conventional video players are limited to providing jump-in functionality and a simple time-slider control. They lack indication of what can be found where in a video. Users who do not wish to watch a video end-to-end must scrub in order to find the parts of the video they would like to watch. More recently, some video players, such as YouTube's [115], offer a thumbnail preview strip. Although this strip is helpful, effectively, its functionality is no more sophisticated than the fast-forward feature found on old VCRs.

This paper studies the usefulness of *collective video playback behavior* to support non-linear video access. Specifically, we derive video representations from collective playback behavior by aggregating information about which parts of videos are most frequently watched and re-watched by viewers. Prior research on this topic has certain limitations, such as the use of short videos, non-realistic viewing conditions or limited sample sizes. In this work, we consider a representation *useful* for non-linear video access if users feel it helps them to easily find good starting points in the video or portions of the video that they find enjoyable and interesting. We choose to focus on one particular form of non-linear access, namely, providing users with heat maps as in Figure 6.1. Our techniques for aggregating playback behavior can be anticipated to apply to other forms of content consumption, such as highlight summaries, which are covered in the related work, but not in our experimental investigation.

A key, novel aspect of this paper is that we go beyond useful aggregation of collective playback behavior, to also tackle the issue of how researchers can gather such playback behavior to begin with. Currently, in order to investigate realistic playback behavior at a meaningful scale, researchers must have access to playback data from commercial platforms. Here, we present a methodology for using crowdsourcing to collect playback behavior that is realistic in nature, and non-trivial in volume. We will release the software and collected data through GitHub.[1] The innovation of our methodology is simple, yet effective: instead of first defining a set of videos, and then asking study participants to watch them, we first define a set of participants, and study video content that they are interested in watching independently of our experiment. Our goal in presenting this methodology is to enable researchers outside of industry or without access to commercial data to also be able to develop techniques using playback data.

6.1.1. MOTIVATION AND SIGNIFICANCE

The motivation driving this paper is the wish to revitalize research into non-linear video access. In particular, we would like to make substantial amounts of realistic viewer playback behavior available outside of industry. To this end, we turn to a generally accessible group of internet users, namely the workers on commercial crowdsourcing platforms.

[1] https://github.com/mmc-tudelft

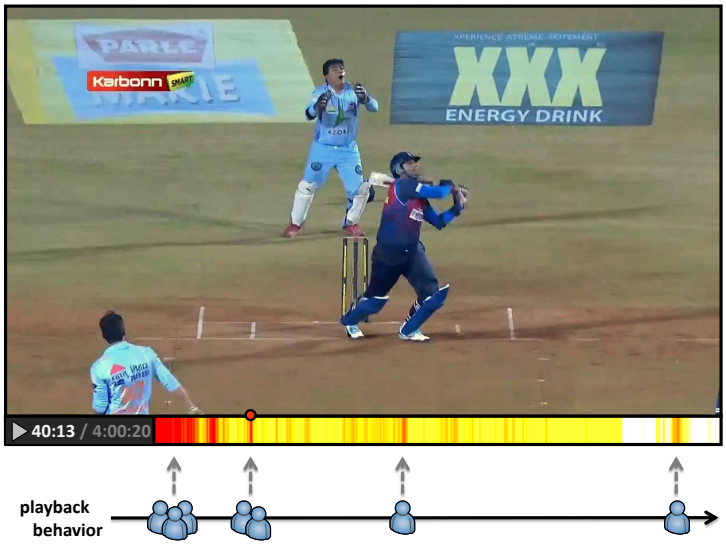


Figure 6.1: Collective viewing behavior reveals the overall interest patterns of the viewing community. This signal can be visualized as a heat-map seekbar, as seen here. The heat-map allows the viewer to grasp the video's structure at a glance and makes it possible to directly jump to potentially the best parts of a video.

In this section, we present evidence of the importance of non-linear video access for viewers of on-line video. Prior to beginning this work, we carried out an open-ended survey as a task on Amazon Mechanical Turk (`www.mturk.com`). Amazon Mechanical Turk (AMT) is a crowdsourcing platform that allows *requesters* to post small Human Intelligence Tasks (HITs) for *workers* to carry out in return for a small monetary reward per completed task. In our preliminary survey, we asked 100 people to describe the ways in which they watch video content. Specifically, we asked whether they watch content from beginning to end or whether they engage in one or multiple types of *non-linear viewing behavior* instead, e.g., skipping the intro, seeking interesting parts or refinding a particular segment. Other questions in the task ensured that we were targeting people who viewed online content, and that they were taking the survey seriously.

We manually reviewed how people described their viewing behavior, and found that the descriptions could be easily sorted into four categories (number in parentheses is number of responses in that category): linear (68), non-linear (29), near-linear (25), and other (2). If the responder described skipping only the intro and/or outro of a video, we labeled it as *near-linear* viewing behavior. We conclude that non-linear video access is important since one quarter of our respondees report using it.

The survey results revealed a further insight. Respondees reported using non-linear viewing behavior to fulfill a variety of information needs, including finding a single song in a long performance, and looking for a controversial act of a coach during a game. Responses covered a variety of content such as how-to videos, interviews and sports.

In this paper, we pursue one aspect of this insight further. Clearly, most information needs mentioned by the study participants as driving their viewing behavior are

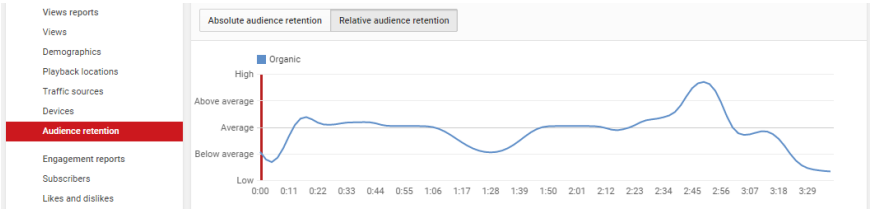


Figure 6.2: Platforms such as YouTube provide uploaders insight into which parts of their videos retain the viewers' attention.

not highly personal information needs. Rather, the interest of the viewer for a particular part of the video (for example, a song, or a controversial moment) can be assumed to be shared by other viewers. We do not deny the existence of cases in which the viewer is looking for something personal (e.g., the moment where my cousin, who attended the game, is visible in the stands). Rather, we point out the importance of these relatively *non-personalized* cases, as revealed by our preliminary survey. In short, these examples show that viewers would benefit from a generic form of non-linear access, such as could be provided by using the playback signal, to determine what, in general, viewers of a video find interesting.

Our desire to dig into collective viewing behavior is further motivated by the surprisingly limited amount of existing research on aggregating viewing behavior to support video navigation. To our knowledge previous work has been restricted to small studies, which have been few in number. These studies will be discussed in Section 6.2.2.

We attribute the shortage of work in this area, to a substantial degree, to the difficulty in obtaining adequate amounts of user interaction data outside of industry, already mentioned. We note that playback behavior is made available to individual uploaders. For example, YouTube makes graphs of video playback available, cf. Figure 6.2, but these can only be accessed by channel owners, i.e., the people who uploaded the video. Playback behavior cannot be collected for the large number of on-line videos needed for research.

6.1.2. CONTRIBUTIONS AND STRUCTURE

This paper makes contributions to the study of viewer playback behavior for non-linear video access in three areas:

- First, we present a methodology designed to collect meaningful amounts of playback behavior that is more *realistic* than what has until now been used in previous work.
- Second, we provide insight into the usefulness of collective viewing behavior with crowdsourcing experiments addressing the three research questions below.
- Third, we provide the resources needed to reproduce our results, including the text of the crowdsourcing tasks[2], the software used to collect the playback behavior[3]

[2] `https://mmc-tudelft.github.io/mturk-playbackbehavior`
[3] `https://github.com/mmc-tudelft/commonhit`

and a list of the videos that we use. We will also release the raw user playback data that we collected in a properly anonymized form.

Our insight into the usefulness of collective viewing behavior for non-linear video access takes the form of the answers to three sets of research questions on the usefulness of video playback behavior. These questions start by establishing a foundation and build towards an understanding of the added value of playback behavior in practice.

RQ1 *(Convergence)* Does collective viewing behavior actually converge to a pattern? Over time are all time points in a video watched equally? Does convergence require a prohibitively large number of viewers (Section 6.5.1)?

RQ2 *(Usefulness)* Do viewers find the information in collective viewing behavior useful (Section 6.5.2)?

RQ3 *(Added value)* Does collective viewing behavior go beyond the information that would be provided by concept detection (Section 6.5.3)? Does it duplicate the information provided by multimedia content analysis (Section 6.5.4)?

The paper is structured as follows. In Section 6.2, we cover related work, and position this paper with respect to it. Section 6.3 describes the methodology that we propose for collecting realistic viewer playback behavior outside of an industry setting. Section 6.4 describes the three main crowdsourcing tasks that we have carried out and defines the main concepts and notation. The results of these tasks, which answer our three research questions, are presented in Section 6.5. We conclude with discussion and outlook in Section 6.6.

6.2. RELATED WORK

In this section we provide a background on video enrichment, viewer signals, and viewer interest. We cover previous work on collective playback for non-linear access, and highlight its limitations, which are addressed by this paper.

6.2.1. VIDEO ENRICHMENT

Non-linear access relies on video enrichment in order to direct users to specific parts of a video. Content analysis is a classic approach to enriching video. Key examples are visual concept detection [90], e.g., finding video frames that depict a car or a cow, and computing affect over time [37], e.g., finding parts of the video reflecting excitement. Content analysis has also been used to directly offer viewers more interesting methods of browsing videos. Schoeffmann et al. [84] survey a large collection of user interfaces for browsing videos, many of them employing some form of content analysis. Some examples include the use of colored seek bars to represent the dominant color in the visual channel and seek bars depicting amount and the direction of motion. With *RQ3*, this paper investigates how collective playback complements information provided by content analysis.

6.2.2. VIEWER SIGNALS

Research on enriching video with explicit user feedback was driven by the idea that what users find interesting in video content is related to not only what the content depicts, i.e., *perception*, but also how the audience reacts to the content, i.e., *reception*. The importance of audience reaction has led to a large body of work on viewer signals. Here, we discuss work on both explicit and implicit viewer signals.

Explicit viewer signals have been used to enrich video. Twitter has provided enrichment for, e.g., political debates [17], NFL football matches [96], and soap operas [33]. In [52], an interface was proposed that offers non-linear access by using visual content analysis to align tweets mentioning popular concepts to the moments those concepts appear in the video.

Implicit viewer signals have also been researched extensively. Important insights have been achieved with electroencephalography (EEG) and eye gaze tracking, e.g., [91]. The Interest Meter [75] records the viewer's body language during full linear playback of a video. We also mention [65], which looked at the use of sensors resembling those available in smart phones to record audience response, as expressed by e.g., movement, during a live performance. Viewer response research must necessarily make a trade-off between the invasiveness of the sensors recording response, and the types of insights that can be achieved. A minimally invasive approach is that adopted by [119], who take the mouse movements of users watching videos online to reflect interest. A similar assumption is made by work, such as ours, that connects the user playback signal to interest.

We point out that the idea of leveraging the collective playback signal for non-linear access falls under the larger paradigm of using community activity to understand content, discussed by [85]. Applications include ranking [117], authoring [8], and categorization [113]. Here, we put our focus on issues related to our application of non-linear video access.

6.2.3. VIEWER INTEREST

Formally, viewer interest is the internal state of a person viewing a video. Soleymani [91] defines interest as, "an affective state that drives users' attention, and in combination with users' intent, it constructs users' preferences and shapes their behavior on multimedia delivery platforms". Interestingness research faces formidable challenges. First, a person's internal state is affected by a range of stimuli that include aspects of the content of the video, but also aspects of the context, including recent events and the viewing condition. It is difficult to control these stimuli to isolate the ones directly responsible for the state. Second, people are themselves unaware of their own internal states and their self reports may contradict physiological measurements. Because of these challenges, much research on video interestingness is directed at generating enrichment information that is *useful* in an application setting. For example, [16] focuses on selecting frames to pique viewers' interest in content on a Video On Demand website. An application can be highly successful using a best educated guess of the stimuli and also benefits by balancing conscious and unconscious interests.

In this paper, we take exactly such a pragmatic, application-driven approach to interestingness. Our methodology is designed to carefully set up the preconditions in

which video viewers would experience and express genuine interest. We are motivated by strong statements of the importance of studying realistic viewing behavior, cf. [85]. In our study, we prompt viewers to predict what other viewers would find memorable, and reflect on what they themselves enjoy.

Our methodology sets us apart from other work, which also takes a pragmatic approach to interestingness, but fails to consider the realism of the viewing conditions. For example [119], mentioned above, collects ground truth using AMT, but chooses videos of no more than 5 minutes in length that do not require domain-specific knowledge. Workers are asked to click thumbs up or thumbs down when something interesting or uninteresting is being shown. Effectively, they are forced to find something interesting in a very short video. Our approach directly addresses the limitations of this method.

6.2.4. PLAYBACK SIGNAL FOR NON-LINEAR ACCESS

Next we turn to looking at previous work that uses a collective playback signal in a way that supports non-linear access. We mention the key limitations of existing studies: limited number of contributors to the collective signal, limited length of the videos, and low attention to realism. Note that we do not cover work such as [24], which exploits the viewing history of a single user, rather than collective playback.

Early work used playback signals alongside of audio and slide transitions to create presentation summaries [38]. The tests involved four presentations of about an hour viewed by around 40 users each. Also noteworthy are the degree of interest (DOI) functions were presented by [71] for use with American Football. The functions use a mix of game specific features (e.g., touchdowns, field goals) and/or previously recorded viewer activity (i.e., replay a play, skip or change angle.) The test involved one game and 11 users. We point out that there is a small but important difference between summaries and the type of non-linear access under study here. Successful summaries require a complete set of highlights. In contrast, we focus here on the usefulness of the jump-in points that we can offer to users and not their completeness.

VideoSkip [30] makes use of user interactions with a web player to detect video-events. A playback signal was collected from 23 users for 3 videos of around 10 minutes each. This work explicitly avoided long videos since they would be tiresome for users. It adopts the question-based approach of [117], namely, driving the playback behavior by a set of questions given to the user. Further research on this data of [30] is reported in [4] and [47]. Another approach that uses questions to drive browsing behavior is [95]. This approach used the playback signals to generate video previews by training a hidden Markov model. Our work also can be considered to drive viewing with a task. However, we choose a general task closely that is linked with a natural viewer browsing goal, namely, finding memorable moments.

An early proposal for a navigation systems leveraging other user's interactions was made by [66], but was unfortunately not evaluated. The work that is closest to our own introduces the VCR (View Count Record) system [3], which aggregates segment-level view count statistics over multiple users. The video's time-line is visualized by a series of keyframe thumbnails representing the segments. The size of the thumbnail reflects the viewer count. However, the system is evaluated with the playback signal of only 6 users for 5 videos of 3-5 minutes in length.

It is striking that in nearly every study viewers are subjected to forced interest. In other words, the study design is not sensitive to the fact that the playback signal was collected from people who possibly were not truly interested in the content of the video. An exception is [93], which discusses realism and recruits a large number of users (103) who are interested in the video content. Unfortunately, only one, six-minute video is used. Follow-up work in [5] suffered the same constraint.

In sum, we see in the related work, that videos are short and that users are either required to watch them end to end or receive questions to drive their browsing behavior. Little attention is devoted to whether the users would actually have watched the video if it were not for the experiment, and playback behavior is not necessarily a result of genuine interest.

6.3. METHODOLOGY

This section presents our methodology for collecting realistic feedback behavior. Our methodology addresses the shortcomings of existing work in terms of limitations on the amount of playback data that it is feasible to collect, and limitations on the naturalness of the viewing behavior of the users that generated the playback data. We follow two design concepts, *Viewer-First Dataset Design* and *Experience-Embedding Task Design*, discussed here in turn.

6.3.1. VIEWER-FIRST DATASET DESIGN

We design our dataset, as already mentioned above, by first identifying our viewer population, and then allowing the viewer population to dictate the video dataset that we use for our investigation. This procedure is the inverse of the standard procedure that is applied outside of industry. Conventionally, researchers first choose the video dataset, and then go in search of users in order to view the video content to provide feedback. Our motivation for choosing the inverse procedure is the observation that genuine enjoyment of the video content will lead to the most natural interaction behavior. We want to avoid forced interest and maximize the chance that the viewer would have watched the video independently of our research. If we can minimize the feeling of the viewers that they are obliged to watch something that falls outside of their usual viewing interests, then, we reason, we can maximize the realism of the viewer playback behavior.

In order to reach a large viewer population, we turn to a crowdsourcing platform (AMT). We start from an observation made during our initial open-ended survey, which was described in Section 6.1.1. In the survey answers, we found that about 25% of the respondees mentions viewing sports video.

A second observation that we build on is that a large number workers on AMT are based in India, corresponding to previous estimates that about 20% of the worker population on AMT is based in India [44]. In India, the game of cricket is popular throughout all segments of the population, and constitutes a billion dollar industry [1]. Our previous research experience has included game design, which also successfully leveraged the power of cricket for engagement in the Indian context [55].

These considerations led us to choose the Celebrity Cricket League (CCL) as our source of video content. This content source matches the natural interests of the viewer

base to which we have access via AMT. Additionally, it fulfills three necessary research requirements, which we discuss now in turn.

Suitability As revealed in our preliminary study, non-linear access is used for certain types of video. Short videos, we assume, are more likely to be watched linearly, from end to end. Non-linear access should instead focus on longer videos. Cricket matches are known for being long. The average Celebrity Cricket League match lasts four hours, making the chance high that viewers prefer to watch or re-watch by skipping from moment to moment rather than from start to finish.

Availability As a practical necessity for the collection of viewer feedback behavior, and for reproducibility of our experiments, the videos must be publicly available and permit embedding. CCL matches fulfill these availability requirements.

Generalizability The immense popularity of sports makes sports an attractive choice for viewer-first dataset-design. However, sports games have a particular structure, dictated by the rules of the game. Ideally, our viewer-first dataset choice should be simultaneously both popular with our viewer database, and also restricted as little as possible to a single genre. From this perspective, Celebrity Cricket League is a particular good choice. The non-professional Celebrity Cricket League (CCL) exists in India alongside professional cricket. It involves eight teams of film actors from Indian regional film industries [109]. The CCL has been referred to as *cricketainment* since it combines elements of cricket and entertainment: It showcases interviews with stars, who are both players and ambassadors, and is appreciated because of the emotion that they express about the game [97]. These characteristics of CCL increases the chance that the findings of this paper, which are based on CCL data, will go, in the future, beyond sports events and transfer to other forms of video, most specifically, reality TV.

In sum, by putting the viewer first, we have arrived at the dataset for use in studying viewer playback behavior. We identified a viewer population large enough to allow us to collect playback behavior from a sizable number of users, and chose a dataset that fits their interests, and fulfills requirements of *Suitability*, *Availability*, and *Generalizability*. The final dataset used for the experiments reported on in this paper consists of full matches from the fifth Celebrity Cricket League (CCL) from 2015, which are available on YouTube. The fifth league ran from 10 January to 1 February 2015, and the crowdsourcing experiments described in this paper were run shortly thereafter. Details on the videos are available in Table 6.1.

In order to validate our viewer-first dataset decision, we ran a pilot study in which we offered a pilot version of our task to both US workers and Indian workers. The pilot task, which ran for a week, attracted more workers from India than from the US (22 vs. 3) and workers from India spent more time on average on the whole task (23 minutes vs. 4 minutes). We interpret these results to mean that the Indian workers engage more strongly and naturally with the content, and that our viewer-first dataset design indeed allows us to collect feedback behavior on AMT that is more natural than what was previously considered possible.

Table 6.1: Cricket video dataset

	Team 1	Team 2	Pool	Duration	YouTube ID
1.	MH	VM	B	4:53:25	yrtXAFahKgM
2.	KB	BD	A	4:11:15	kA1OAzwUdrU
3.	CR	KS	B	3:53:29	6CvnO7I2mO4
4.	TW	BT	A	3:44:41	oGey3G-ML-w
5.	MH	KS	B	4:34:20	NZh7EjTb6wU
6.	BD	TW	A	3:36:55	z8U_bakITDw
7.	CR	VM	B	3:22:41	oA_og6KmCRY
8.	KB	BT	A	4:16:34	VH0xsWpqkdU
9.	TW	KB	A	4:17:17	i4L3HEV6D3E
10.	KS	VM	B	4:14:38	r7zN9HjKzAs
11.	MH	CR	B	3:32:02	8WyjDMljVeg
12.	BD	BT	A	4:00:21	xoWYjBOqMyQ
13.	CR	KB	Semi	4:10:50	W02IHehf77o
14.	TW	MH	Semi	4:12:45	BvxQ5_srVLk
15.	TW	CR	Final	4:18:43	vDoj_nsrn88
			Total:	61:19:56	

Teams: Bhojpuri Dabanggs (BD), Bengal Tigers (BT), Chennai Rhinos (CR), Karnataka Bulldozers (KB), Kerala Strikers (KS), Mumbai Heroes (MH), Telugu Warriors (TW), Veer Marathi (VM).

6.3.2. EXPERIENCE-EMBEDDING TASK DESIGN

Now that we have chosen the dataset, we turn to the question of how to design an AMT task (i.e., a HIT) that will allow us to maintain the naturalness of viewer behavior to the largest extent possible. To this end, we would like to minimize the impact of the fact that on AMT we are actually paying viewers to watch our video data set. In this section, we discuss the issues, and how we address them.

Although some workers use AMT to kill time or have fun [43], we assume that a major motivation of workers is to earn money. We would like to shield our viewers from the temptation to speed through viewing in order to earn money as quickly as possible. In order to accomplish this shielding, we design our HITs to embed the experience of participating in a contest inside the experience of carrying out a task on the AMT platform. The contest is designed in such a way that the chances of success in the contest are improved the more time the worker spends watching a video. Since the motivations introduced by the contest cancel out motivations to engage in other behavior (i.e., speed through the HIT), we assume that there is a high chance that users will simply default to natural viewing behavior.

The contest consists of two parts, the `Select` Task and the `Vote` task, which are described here in turn. In the `Select` Task, the contest is introduced. Workers are presented with the video of a Celebrity Cricket League match and are asked to watch it in a natural way and select three memorable moments, which are described as '...moments that other people would want to rewatch'. Workers are informed that it is okay to seek, skip, and watch specific segments. Workers win in the contest if they select moments that receive many votes from other workers. Winning workers receive an AMT bonus.

In addition to collecting memorable moments from the workers, playback behavior of the workers is also recorded during the `Select` Task for further study. The specific formulation of the `Select` Task is created to encourage workers to adopt the type of non-linear access behavior that viewers would use to re-view a lengthy segment of video content. We assume that a major motivation of re-viewing recorded matches is to dis-

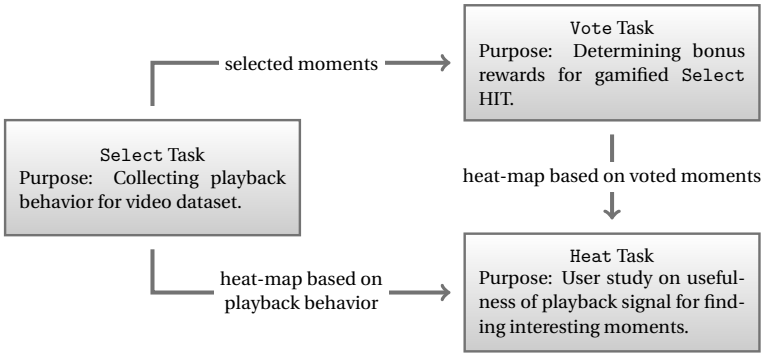


Figure 6.3: Overview of the experimental setup showing the relationship between the three different crowdsourcing tasks.

cuss them with colleagues, friends, and family. For this reason, providing a bonus for memorable moments that are popular with other workers emulates a natural motivation for watching match recordings in a real-world setting.

In the `Vote` Task, workers are asked to view a set of selected moments from the `Select` Task and to vote on the three moments that they find the most memorable. We do not use any playback behavior collected during the `Vote` Task. Rather, we use the task to determine winners and to collect explicit confirmation of the usefulness of specific video moments. Figure 6.3, which depicts all HITs used in our study, illustrates the connection of the `Select` Task with the `Vote` Task. Details of the execution of these tasks and also of the `Heat` Task, which assesses heat maps created with aggregated playback behavior, are provided in the next section.[4]

6.4. EXPERIMENTAL STUDY

In this section, we apply the methodology introduced in Section 6.3 to carry out studies on the CCL dataset described in Section 6.3.1 by gathering information with three HITs, summarized by Figure 6.3 and introduced in Section 6.3.2. Technically, the HITs all incorporate a player capable of collecting playback behavior, and also presenting heat maps such as the one in Figure 6.1. The player is based loosely on LikeLines [99], our open source system for collecting playback behavior (e.g., play, pause, scrub), and has been adapted for use in AMT. We cover each HIT in turn, detailing how the feedback collected in one HIT flows into the next HIT in the chain, and contributes to our analysis.

6.4.1. COLLECTING PLAYBACK BEHAVIOR

The `Select` Task, as explained above in Section 6.3.2, asked workers to watch a CCL match video, but also to take notes in order to select three memorable moments at the end of the task. At the start of the task, workers were presented with instructions and a general pre-video survey, in which they were asked whether they have watched this video before and if they were supporting a particular team. The task was divided into

[4] Recall that all HITs are available at `https://mmc-tudelft.github.io/mturk-playbackbehavior/`.

well-delineated steps that had to be completed before the worker could continue, i.e., instructions, pre-video survey, watching the video, selecting moments, and epilogue. The software we used to implement the task templates and to capture the worker's playback behavior during video viewing is available on GitHub.[5] Details of the published `Select` HIT are as follows. The HIT was run for all 15 matches in the dataset. For each cricket match, 53 workers could submit three memorable moments. The reward for each assignment was set to US$0.50 and the HIT ran in March 2015.

On the basis of this playback behavior, we create a *collective playback signal*, using the following approach. Let $p_i(n)$ denote a playback signal describing how often a single person in a viewing session $i \in S_v$ has viewed the nth second of video $v \in V$. Let $|v|$ denote the length of video v in seconds. We create $\tilde{p}_i(n)$, a smoothed version of $p_i(n)$, obtained using a median filter with a window of 11 under the assumption that any segment of the video played for 5 seconds or less is an artifact from scrubbing or seeking. The individual playback signal of a single viewing session $\hat{p}_i(n)$ is calculated by normalizing $\tilde{p}_i(n)$ such that the signal's range lies within $[0,1]$:

$$\hat{p}_i(n) = \frac{\tilde{p}_i(n)}{\max\limits_{0 \le x < |v|} \tilde{p}_i(x)}. \tag{6.1}$$

We then define the collective playback signal to be the sum of all individual playback signals in S_v, the set of all viewing sessions for video v:

$$h_{S_v}(n) = \sum_{i \in S_v} \hat{p}_i(n) \tag{6.2}$$

Next, in order to reduce the impact of incidental differences, we create a smoothed version of Equation 6.2, denoted as $\tilde{h}_{S_v}(n)$. We again use a median filter, but this time we use a filter window of 5. This window size was chosen on the basis of data collected during our pilot task.

We next define a second version of the collective playback signal, the *consensus-based collective playback signal* that takes into account some form of consensus. First, we introduce $m_{v,k}(n)$ to denote a discrete signal that is equal to 1 when at least k viewers have watched the nth second of a video v:

$$m_{v,k}(n) = \begin{cases} 1 & \text{if } \left|\{i \mid i \in S_v \wedge p_i(n) \ge 1\}\right| \ge k, \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

We can then use Equation 6.3 as a mask to suppress parts of the collective signal watched by less than k viewers. We use $k = 3$ to obtain:

$$c_{S_v}(n) = h_{S_v}(n) \cdot m_{v,3}(n) \tag{6.4}$$

Again, as with Equation 6.2 above, we use $\tilde{c}_{S_v}(n)$ to denote a smoothed version of Equation 6.4 using the same median filter with a window of 5.

[5] `https://github.com/mmc-tudelft/commonhit`

6.4.2. COLLECTING MOMENT JUDGMENTS

Next we move to the `Vote` HIT, which, as discussed above was used to decide which of the memorable moments that viewers submitted in the `Select` HIT would be considered winning moments. Note that the 'memorable moments' are the moments explicitly selected by the workers.

For each video, we collected all moments submitted by workers in the `Select` HIT. The moments were randomly placed in brackets of six moments. For each bracket, workers had to pick their three moments of choice and explain which of the three would be their most favorite. After the HIT is finished, the top three moments that have received the most votes in a bracket are declared as winning moments for that bracket. Prize money for that bracket is then evenly be split across the workers that submitted these winning moments. Just like the `Select` HIT, this HIT contained the same short pre-video survey.

Details of the published `Vote` HIT are as follows. The HIT was run for 349 brackets (each of 6 moments) in total for the 15 cricket matches in the dataset. For each bracket, 3 workers could cast three votes. The prize money for each bracket was US$0.75, which was equally distributed over the winning moments. The reward for each assignment was set to US$0.17 and the HIT ran in March 2015.

6.4.3. USER STUDY ON THE USEFULNESS OF COLLECTIVE BEHAVIOR

In the final HIT, the `Heat` HIT, we carried out a user study on the usefulness of the collected viewing behavior signal observed in the `Select` HIT. For this HIT, we visualize the collective playback signal as a heat-map for the purpose of navigation (cf. Fig 6.1). We compare it to two baselines, 'Voted moments', explicitly chosen memorable moments that emerged as 'winning moments' in the `Vote` HIT, just described in 6.4.2 and 'Random moments'.

The `Heat` HIT is structured as follows. The task starts with an overview and instructions. An example of a heat-map seek bar is shown, together with a color map explaining the scale of less interesting to more interesting.

Then, the worker carries out the same pre-video survey as in the other two HITs. Next the worker is presented with a video with the heat-map seek bar to be tested. The worker is instructed to watch a cricket match video. The HIT is framed such that workers must imagine they missed a cricket match from the CCL league and have limited time to watch a recording of it. Clicks on the heat-map seek bar are recorded and are used in the next and final step of the HIT.

Finally, the worker is asked for an explicit evaluation of the heat map. The video with the heat-map seek bar are presented again and the worker is asked three questions. For the first question, ($Q_{relevance}$), the worker is shown five of the time-codes that s/he had previously jumped to by clicking on the heat-map seek bar. The time-codes are selected by random sampling of all the worker's previous clicks. For each time-code, the worker is asked to specify whether it was a good starting point.

As the second question, (Q_{agree}), the worker is asked to judge (on a 5-point Likert scale) the following three statements about the heat map.

S1) "This specific heat-map made watching the video fun (and not frustrating);"

S2) "This heat-map seek bar for this video allows me to find good starting points in the video;"

S3) "This heat-map seek bar helped me to find the interesting moments in the match more quickly."

As the third question, (Q_{free}), workers are asked to give general impressions using a few sentences and optionally provide questions or suggestions.

Details of the published `Heat` HIT are as follows. The HIT was run for each of the 4 conditions for each of the 15 cricket matches in the dataset. For each condition and video, 30 workers could participate in the evaluation, resulting in a total of 1,800 assignments. The reward for each assignment was set to US$0.20 and the HIT ran in April 2015.

Next we describe in detail the four different conditions (types of heat maps) that are tested in `Heat` HIT. Each condition uses a different type of heat map each using a different source signal for the heat-map seek bar. Two of them are based on the collective playback signals, i.e., smoothed versions of Equation 6.2 and Equation 6.4.

Since these signals are not yet normalized to fall between $[0,1]$ and since observations from the pilot `Select` HIT showed that these signals tend to contain extremely high peaks at the beginning of the video, we adjusted the signals for display. Specifically, we clipped the signal to retain all values within two standard deviations of the mean before and then applied normalization, as follows. Let $\bar{h}_{S_v}$ denote the mean and let $\sigma_{h_{S_v}}$ denote the standard deviation of $\tilde{h}_{S_v}(n)$. Then

$$\hat{h}_{S_v}(n) = \frac{\min(h_{S_v}(n), \bar{h}_{S_v} + 2\sigma_{h_{S_v}})}{(\bar{h}_{S_v} + 2\sigma_{h_{S_v}})} \tag{6.5}$$

denotes the smooth normalized *collective playback signal* of Equation 6.2. Using the same notation for $\tilde{c}_{S_v}(n)$, the smooth normalized *consensus-based collective playback signal* of Equation 6.4 is then defined as follows:

$$\hat{c}_{S_v}(n) = \frac{\min(c_{S_v}(n), \bar{c}_{S_v} + 2\sigma_{c_{S_v}})}{(\bar{c}_{S_v} + 2\sigma_{c_{S_v}})} \tag{6.6}$$

The 'Voted moments' and the 'Random moments' baselines are defined as follows. The 'Voted moments' baseline uses results from the `Vote` HIT. Here, for a given video v, we select all moments that received at least three votes. Let M_v denote this set of time-points (in seconds). On average, M_v contained 10 time-points (ranging from 8 to 16 points). We place a Gaussian kernel at each point, since we have no knowledge whether the kernel should favor the moment before or after the time-point. The kernel's σ was set to 30 to make the moments visibly discernable in the resulting heat-map seek bar. In order to prevent nearby time-points amplifying each other, we use max rather than addition as the combining operator. Then, given a set of points $P = M_v$, the resulting signal for the third condition is defined to be:

$$f_P(n) = \max\left\{ e^{-\frac{(n-x)^2}{2\sigma^2}} \mid x \in P \right\}, \quad \text{where } \sigma = 30 \tag{6.7}$$

The 'Random moments' baseline is formed in the same way as the 'Voted moments' baseline, except that the time-points are instead randomly chosen. For each video v, we draw 10 points from the uniform distribution $\mathscr{U}(0,|v|)$ under the assumption that events occur at a fixed rate. We choose 10, since M_v contains on average 10 time-points. We designate the randomly drawn point set as R_v. Then, we set P=R_v, and calculate f_{R_v} by again applying Equation 6.7.

6.5. RESULTS

This section presents the results achieved by applying the methodology (Section 6.3) to implement crowdsourcing experiments (Section 6.4) that, taken together, enable insights into the usefulness of collective playback behavior. We analyze the data collected from the three AMT HITs (refer back to Figure 6.3), in which a total of 272 workers participated (141 in the `Select` HIT, 82 in the `Vote` HIT, 142 in the `Heat` HIT). We address each of our three sets of research questions (*RQ1*, *RQ2*, and *RQ3*), introduced in Section 6.1.2.

6.5.1. EMERGENCE OF TRENDS IN THE PLAYBACK SIGNAL

First, we investigate *Convergence*: Whether and how quickly trends emerge in the collective playback signal (*RQ1*). Answering this question provides insight into how many users must contribute to the collective playback signal before it can be meaningfully used in an application.

To answer these questions, we analyze data from the `Select` HIT. In total, the `Select` HIT collected 727 viewing sessions for the 15 videos, or, on average, 48.5 sessions per video. In these 727 sessions, workers watched in total 140 hours of footage, or 11.5 minutes on average. As mentioned before, a total of 141 workers participated in the HIT. They watched in total 9,488 segments, which means an average of 13 segments per viewing session. Colored area of Figure 6.7 plots the collective playback signal for each of the videos.

We use the data that we collected with the `Select` HIT to simulate a large number of possible ways in which the collective playback signal could accumulate over time by randomly ordering the sessions $i \in S_v$ that we collected for a video v. Note that we can safely apply any reordering to the sessions recorded in the `Select` HIT, since each session was collected independently of all others. Let $o_j : S_v \rightarrow \mathbb{N}$ denote a function representing a particular ordering j that assigns to a session $i \in S_v$ a sequence number $k \in \{1,2,\ldots,|S_v|\} \subset \mathbb{N}$. We can then modify Equation 6.5 as follows to represent the collective signal under a given ordering o_j after t viewing sessions have been observed:

$$\hat{h}_{S_v;o_j;t}(n) = \hat{h}_{\{i \in S_v \mid o_j(i) < t\}}(n) \tag{6.8}$$

We can compare this signal at a particular time t_1 to the signal at a later time $t_2 > t_1$ to see how much the signal has changed after more viewing sessions have been observed for a particular ordering. Note that for $t = |S_v|$, Equation 6.8 is equal to the full collective playback signal $\hat{h}_{S_v}$.

We study the evolution of the collective playback signal over time by looking at the pattern with which never-before-watched seconds of video are added to the collective playback signal by each additional user playback session. Specifically, we look at the

number of zero bins in the signal that become non-zero between two time points, computed as follows:

$$\Delta z(h_1, h_2) = |\{n \in \mathbb{N} \mid h_1(n) \neq h_2(n) \wedge h_1(n)h_2(n) = 0\}| \tag{6.9}$$

In our analysis, we combine Equations 6.8–6.9 and compare the collective signal at two adjacent points in time:

$$conv_{S_v;o_j}(t) = \Delta z(\hat{h}_{S_v;o_j;t-1}, \hat{h}_{S_v;o_j;t}) \tag{6.10}$$

Figure 6.4 shows the value of the function defined in Equation 6.10 over time averaged over 500 random orderings of viewing sessions, o_j, for each of the 15 videos using interaction data recorded in the `Select` HIT. We illustrate the standard deviation for single video using error bars. Figure 6.5 illustrates the standard deviation for each point in time for a typical video (the other videos are comparable).

Figures 6.4 and 6.5 allow us to make two observations. First, we observe the existence of a characteristic convergence pattern. In the top plot of Figure 6.4, the convergence curve begins to flatten at around $t = 10$ sessions. By $t = 50$ sessions it is very flat, but still well above the 0 line. This pattern contrasts with what we would expect if the playback behavior did not converge: in such a case the line would fall more slowly, but also more steadily towards zero, the point at which all timepoints in the video have been watched.

One possible explanation for the observed convergence pattern is the patterns of interest among our viewer population. Recall from Section 6.1.1, that we are not aiming to satisfy *personalized* requirements for non-linear access, but rather focus on relatively *non-personalized* cases. However, the nature of sports competition means that viewer interest cannot be considered homogeneous. Instead we can expect a pattern: some viewers will support one team and some will support the other.

Figure 6.6 is a version of the top plot in Figure 6.4 which includes only sessions of the viewers who supported the same team supported by the majority of the viewers. The fact that the plot in Figure 6.6 ends after about 30 sessions reflects the fact that the majority is about 60% of the total viewers. The critical point to notice in Figure 6.6 is that it has the same shape as the top plot in Figure 6.4. This fact suggests that the convergence that we observe is a useful convergence leading to a collective playback signal that is universally applicable to viewers, rather than being specific to the supporters of one team.

Second, we can see that videos have two different types of convergence behavior. Up until now we have only considered the top plot in Figure 6.4, which shows the patterns for videos 1–11 and 15. Examining the top three lines of the bottom plot of Figure 6.4, we see that videos 12–14 exhibit a different pattern. These curves differ in their mean value of the curves and also in their smoothness. Upon first consideration, these videos seem to contradict our conclusion that the collective playback signal converges. However, closer consideration reveals that for these videos a hand full of viewers watched very long stretches of the video consecutively. Apparently they were caught up in the game, and were not watching exclusively for the purpose of selecting memorable moments. The bottom three lines of the bottom plot of Figure 6.4 show the convergence pattern when sessions with long continuous playback segments have been removed. We see that the plots now fit the same pattern as observed with videos 1-11 and 15 in the top plot of Figure 6.4.

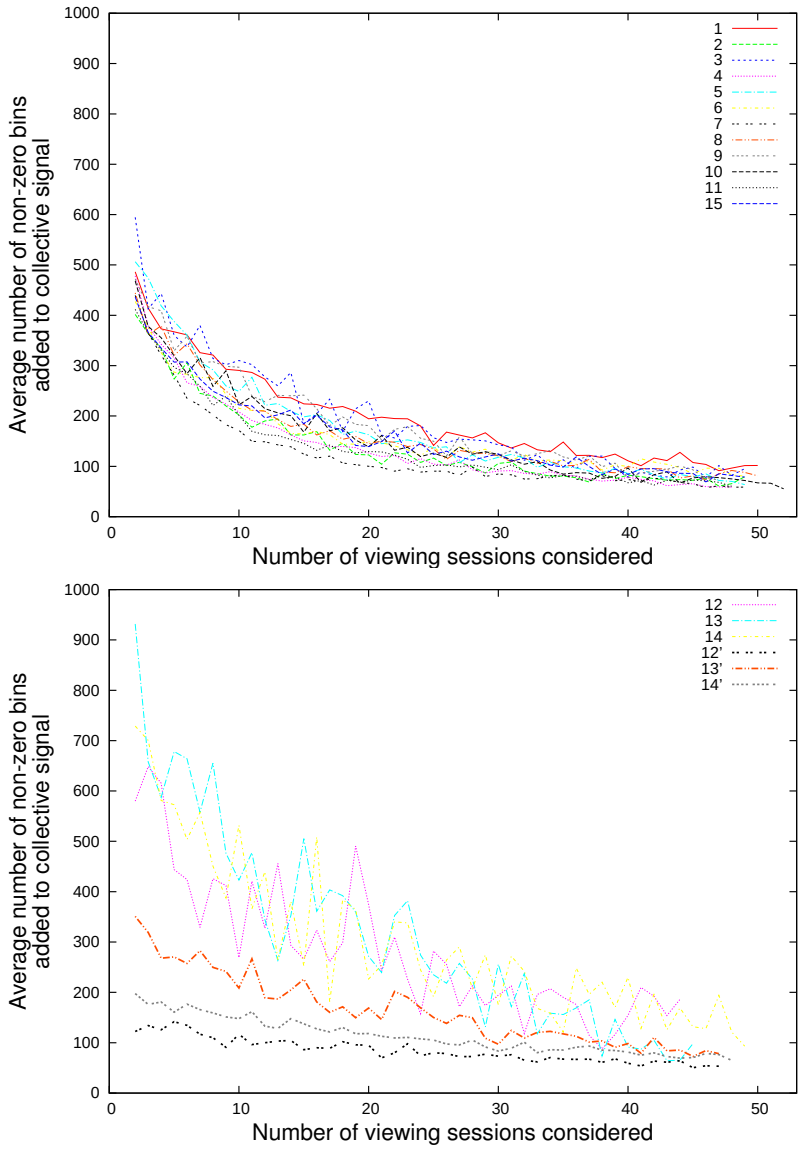


Figure 6.4: Convergence of the overall collective playback signal for each of the 15 videos as the number of user sessions considered grows larger. For reasons of presentation, the videos are divided between the top and bottom graph. The top graph contains videos with typical convergence patterns and the bottom graph contains videos with atypical convergence patterns (top 3 lines, 12–14) caused by sessions with long continuous playback segments. Averaged over 500 random orderings of sessions. Curves labeled n' represent a subset of sessions in which sessions with long continuous playback segments have been removed.

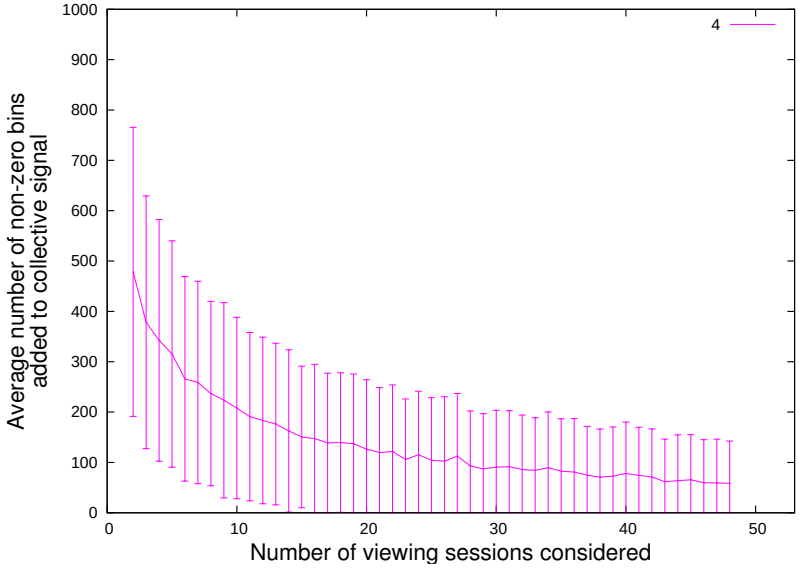


Figure 6.5: Further detail for a typical video (Video 4). Convergence behavior of a the collective playback signal as the number of user sessions grows larger; averaged over 500 random orderings of sessions (as in Figure 6.4). Standard deviation is depicted using errorbars.

We finish this section by looking ahead to Figure 6.7 to point out how these two observations are also evident there. In this figure, the collective playback signal is plotted as a colored histogram that stretches over the timeline of the video. We see that for most videos, there is a lot of playback activity at the beginning of the video, but then islands of interest emerge. We see that the ca. 50 viewers who have interacted with the video do not spread their interest evenly, but instead engage in playback behavior that leads to peaks of interest. This observation is particularly interesting given the fact that each user is watching the video independently, and does not see the playback signal from previous viewers. The videos for which we observe that the playback curve is non-zero over long stretches of the timeline are exactly videos 12–14, exactly those videos which, as we pointed out above, were watched by a handful of viewers whose behavior was non-linear. In, sum, our observations provide the following answer to *RQ1*: Viewing behavior converges to a pattern, which may be hidden unless viewing sessions in which the viewer engaged in linear playback behavior are removed. Convergence occurs after a relatively low number of viewers, implying that a limited number of viewers can already create a playback signal potentially useful in practice. In the next section, we dive into the question of usefulness in more detail.

6.5.2. USEFULNESS OF PLAYBACK SIGNAL FOR NAVIGATING IN VIDEOS

Here, we investigate *Usefulness*: whether users find any value in the collective playback signal (*RQ2*). We analyze the results of the `Heat` HIT, i.e., the user study of the heat-map seek bars described in Section 6.4.3.

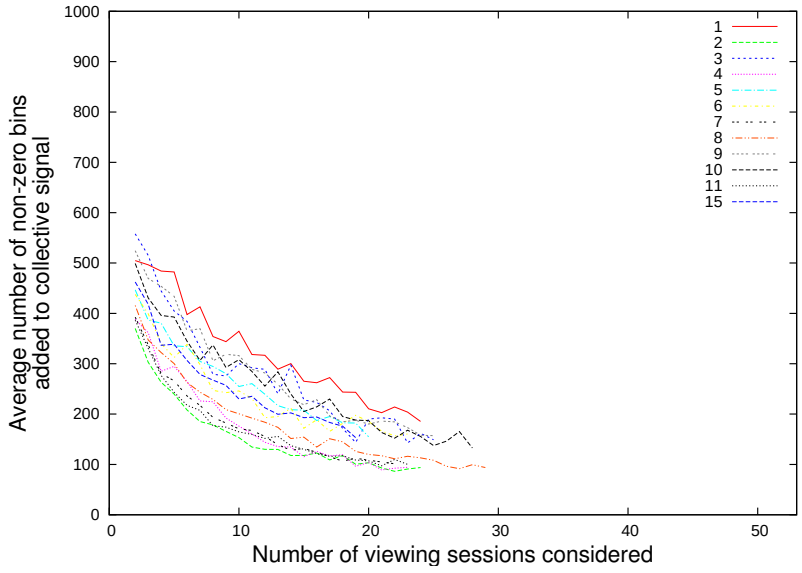


Figure 6.6: Convergence of overall collective playback signals based on a homogeneous subset of viewers. For each video, viewers were divided into groups based on team preference and only viewing sessions from the largest group was considered.

The first question ($Q_{relevance}$) asked workers to judge jump-in points that they have clicked. Table 6.2 reports the precision, defined as the percentage of jump-in points judged to be good. The most interesting insight to be drawn from these results is that the collective playback signal, Equation 6.5, achieves an acceptance rate quite close to that of the 'Voted moments'. This finding is significant, since the 'Voted moments' is based on a time-consuming process of collecting explicit feedback from users, and the collective playback signal requires only recording implicit behavior.

Further, we see from Table 6.2 that all conditions outperform the 'Random moments' baseline. However, only the collective playback signal, Equation 6.5, and the 'Voted moments' outperform it significantly. On the basis of this observation, we conclude that filtering the collective playback signal with consensus does not lead to a more useful signal. Also interesting is the fact is that the random precision is relatively high, and the 'Voted moments', which are explicitly handpicked, demonstrates a significant, but not particularly dramatic contrast with random moments (the difference in precision is only about 0.06).

Next, we turn to the answers that we collected to the three statements of (Q_{agree}) introduced in Section 6.4.3. Table 6.3 gives the mean opinion scores We can see that workers gave highly positive answers. S1 asked whether the heat-map seek bar was fun and not frustrating. Here, we see that the 'Random moments' did not provide more frustration than the other conditions. S2 asked about good starting points and S3 about interesting moments. For these answers, a difference can be observed between the 'Random moments' and the other conditions, but it is not statistically significant. Taken together

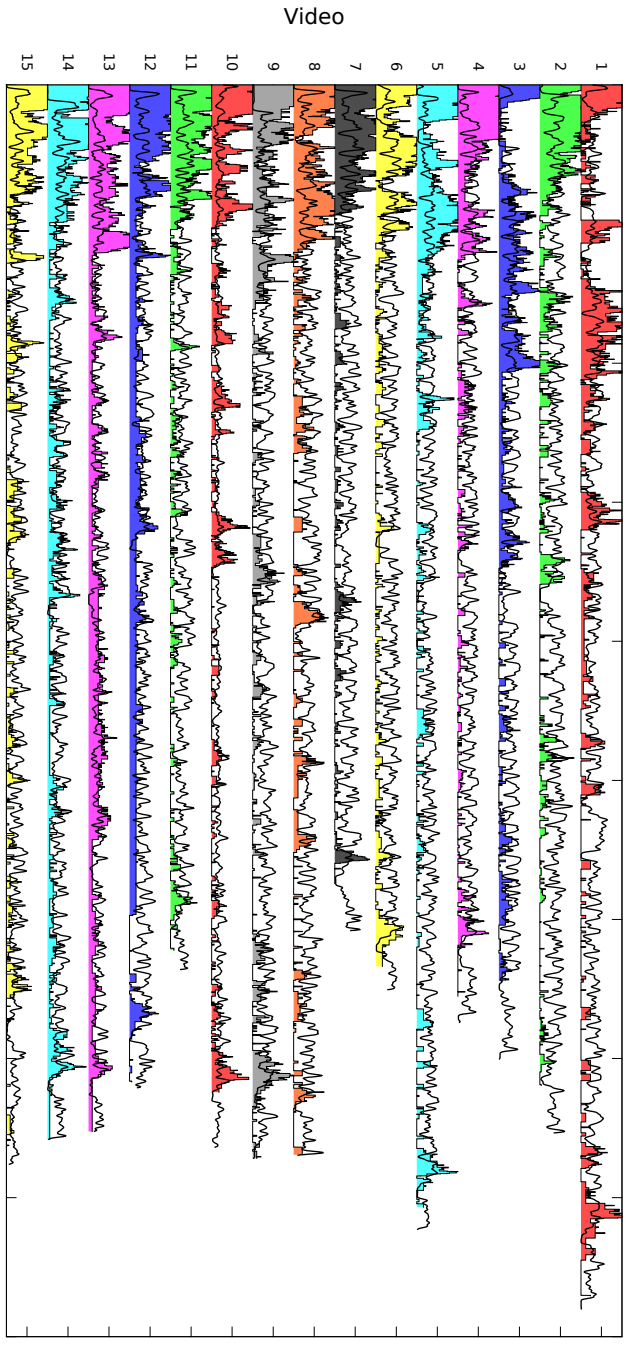


Figure 6.7: Collective playback signal observed in the `Select` HIT (colored area) compared to arousal curve in black for each of the 15 videos.

Table 6.2: Average fraction of relevant heat-map clicks ($Q_{relevance}$)

	Condition	Average precision	Mean worktime	Student's t-test (vs. condition 4)
1.	Collective playback signal, Equation 6.5	0.6170	469s	$p = 0.0323$ *
2.	Consensus-based collective playback signal, Equation 6.6	0.6053	418s	$p = 0.1001$
3.	Voted moments, Equation 6.7, $P = M_v$	0.6294	451s	$p = 0.0070$ **
4.	Random moments, Equation 6.7, $P = R_v$	0.5698	500s	—

(*) $p < \alpha = 0.05$ (**) $p < \alpha = 0.01$

Table 6.3: User study opinion score and completion time statistics (Q_{agree})

	Condition	Mean opinion score S1	S2	S3
1.	Collective playback signal, Equation 6.5	4.01	4.10	4.13
2.	Consensus-based collective playback signal, Equation 6.6	4.01	4.08	4.11
3.	Voted moments, Equation 6.7, $P = M_v$	4.05	4.12	4.16
4.	Random moments, Equation 6.7, $P = R_v$	4.01	4.01	4.06

the answers to $Q_{relevance}$ and Q_{agree} point to the conclusion that users appreciate a heat map in general, that an "incorrect" heat map will not frustrate them. The ability of collective playback signal and hand-selected moments to improve a heat map is competitive.

Next we turn to analyze the free-text answers provided to Q_{free}. Most relevant to our question of the usefulness of the collective playback signal were comments that addressed the sparsity of jump-in points in the 'Voted moment' condition. One worker commented, *"There were many interesting moments and the seek bar didn't cover all."*, Another did not find the sparsity a problem, *"Now this seek bar did not have many selections. But all that was there was enough to cover the proceedings."* We obtain two interesting insights from these comments. First, an advantage of the collective playback signal over explicitly chosen moments is that not only can it be collected effortlessly, there is also no difference in effort between providing a relatively sparse and a dense heat map. Second, these comments suggest that people explore parts of the video beyond what the heat-map seekbar points to. This behavior could explain why they are relatively tolerant of an 'incorrect' heat map, as represented by the 'Random moments' condition.

Many other comments provided tips on usability. These are less relevant for our research question, which is about the usefulness of the collective playback signal, and not about the particular design and implementation of the seek bar, but we include them for completeness. The learning curve appears to be shallow. One worker commented, *"After attempting a few of these hits, I understood the working of heat map and find it very interesting and useful tool too."* Some workers remarked that it is sometimes hard to

Table 6.4: Top 25 most frequent unigrams in workers' written motivations for choosing memorable moments in the CCL matches. Cricket-related concepts are indicated in bold.

1.	good	6.	**toss**	11.	**runs**	16.	excellent	21.	**fielder**
2.	**wicket**	7.	**catch**	12.	**six**	17.	**bowling**	22.	warriors
3.	**ball**	8.	**boundary**	13.	**match**	18.	**hit**	23.	team
4.	**shot**	9.	**four**	14.	won	19.	nice	24.	telugu
5.	first	10.	**batsman**	15.	chennai	20.	**bowler**	25.	great

click precisely on a certain part of the heat-map seek bar due to resolution, and one also offered a suggestion, e.g., *"Have a zoomed version of heat map near the mouse pointer (like spotlight search on iOS). I thought I was clicking on red, but happened to click on the yellow area beside it. Zoomed heat map where the cursor is, will prevent this."* Other suggestions included adding a thumbnail preview to the seekbar similar to the functionality already provided by the YouTube player and including more semantic information in the seekbar, which has actually previously been explored in the literature on collaborative tagging [89]. Finally, the remark of one worker, *"If buffering to directly go to the selected part was a bit faster then one could enjoy it all. Slow loading makes a viewer wait for some seconds."*, reveals that existing video platforms are not suitable for non-linear access, especially if you were to consider common bandwidth configurations found in a crowdsourcing setting [18]. However, it also tells us that approaches such as [10], which seek to reduce latency by pre-fetching could be driven by collective playback behavior.

6.5.3. TYPES OF INTEREST POINTS WHICH PEOPLE SEEK OUT

Now, we turn to *Added value*, our third research question *RQ3*, and look first at whether the collective playback signal duplicates information that would be provided by multimedia content analysis. As mentioned above, one of the user-study participants suggested more semantic information in the seekbar. Specifically, that worker commented, *"Have different colors in the heat bar to identify which moments are boundaries, wickets etc. So if I want to see only moments related to wickets, I can use the respective color."* It is conceivable that automatic content analysis would be able to identify cricket concepts such as boundaries and wickets with a high level of confidence. Here, we are interested in exploring whether such automatic analysis would make the collective playback signal unnecessary.

To this end, we turn back to the information gathered during the `Select` HIT. Recall that workers had to pick three memorable moments and justify their selections with a short explanation. In Table 6.4, we present the most frequently occurring single words (unigrams) in the workers' explanations. It can be seen that the motivation of the workers is heavily grounded in the cricket domain: 15 out of the top 25 most frequently used unigrams are words directly related to cricket (shown in bold). The remaining unigrams are mainly evaluative adjectives (e.g., "good" and "excellent") or are proper nouns referring to CCL teams.

An important insight from Table 6.4 is that not every cricket event in the whole game is memorable. Instead, users qualify moments with evaluative terms such as "excellent" and "great", we see that workers mainly nominated events that exceeded that their ex-

pectations in some way. We carried out an analysis of the ten most frequently used cricket terms in the explanations we collected. Specifically, we looked at cases in which one of these cricket terms was preceded by another word. We found that in 30% of the cases, the word the preceded the cricket term was an evaluative adjective. Most notable are "shot" and "catch", for which the percentages are 66% and 71%, respectively. Terms like "wicket", "ball" and "runs", are less likely to be preceded by evaluative terms. Instead, they are more often qualified by adjectives providing an objective qualification, e.g., "wide ball" or "first wicket". We conclude that although providing users with a clickable timeline of cricket events, such as could be provided by concept detection, would support them in non-linear access, it does not provide complete coverage of their interests. Conventional concept detection is not able to detect the quality of an event, which is clearly a part of why users find certain moments memorable or worthwhile.

Further analysis of the explanations that we collected, reveals that users find moments memorable not because of an event that occurs, but because of an expected event that *fails* to occur, e.g., *"A dropped catch. One of those exciting moments of cricket that takes us to the edge of our seats."* Occasionally, controversial moments were pointed out, e.g., one worker said: *"I liked 30:01 moment. The ball hit the edge of the bat and landed in the hands of wicketkeeper. The field umpires couldn't conclude the decision. The third umpire made the decision of declaring that the batsman got out"*. Previously, users interests in mistakes and bloopers has been observed by studying comments containing deeplinks to particular time points in YouTube videos [103].

Finally, we note that the explanations for many of the moments were not specifically related to cricket. These moments included shots of people and the location, e.g., the audience or celebrities close-ups, beautiful scenery, e.g., *"the camera shows moon in the dark night"*, or side-events not directly related to the match, e.g., dance performances prior to the start of the match and interviews. These explanations did not seem to be constrained to a particular semantic domain. Detecting such moment with a concept detector would require predefining a list of relevant concepts. Given the breadth of these comments, defining a complete list would require a human curator to watch the individual matches.

In sum, our analysis of viewers explanations has led to insight on *RQ3*. We have seen that visual concept detection may support non-linear video access, but cannot fully cover the full range of moments found interesting by users. In particular, evaluative judgments, mistakes, and unexpected topic material are all required techniques going above and beyond concept detection.

6.5.4. CORRESPONDENCE BETWEEN PLAYBACK BEHAVIOR AND AROUSAL

In the past, multimedia research has investigated the value of an arousal curve estimated directly on video content, e.g., [37]. In sports events, the arousal curve captures the reaction of the audience. As such, we continue addressing *RQ3* with an investigation whether or not this curve, which can be pre-computed before the video has been viewed, can substitute for the collective playback signal. We have computed the arousal curve for all 15 videos based on the method described in [37]. The curve is depicted as a black line in Figure 6.7.

We first investigated the hypothesis that users stop watching and decide to skip when the arousal curve starts falling off. To test this hypothesis, we looked at parts of the collective playback signal at which the value of the signal dropped to zero, i.e., points at which viewers stopped watching the video and skipped to a different part. For each of these points in the video, we checked the last 10 seconds of the arousal curve signal and computed the slope in this window. We found that for each video, the mean of the slopes of the arousal curves at skip points is nearly zero. We cannot conclude that arousal drops can help us predict points at which people stop watching.

Next, we turned to investigate whether there is an underlying correlation too subtle to be evident in Figure 6.7. For this purpose, we use $a_v(n)$ to denote a function describing the value of the arousal curve for video v at the nth second. We compare the full collective playback signal of each video (Equation 6.5) with its corresponding arousal curve using both Pearson's correlation coefficient and normalized mutual information measures. This analysis revealed no evidence of a notable correlation. For all but two videos, $|\rho(a_v, \hat{h}_{S_v})| < 0.18$, indicating the relationship between the two types of signals is negligible. Only for video 10 and 15 it was the case that the Pearson's coefficients [72] were 0.24 and 0.21, respectively, indicating a weak relationship. Using normalized mutual information as defined in [107], we found that for all v, $0.5 < N_{\mathrm{MI}}(a_v, \hat{h}_{S_v}) < 0.52$, which indicates the two signal types share little information.

Putting all the insights discussed above together, we can now answer *RQ3* by stating that the evidence points to collective playback behavior being a source of information useful for non-linear access that goes above and beyond the information that can be gained from other forms of video enrichment, specifically visual concept detection, and affective analysis.

6.6. CONCLUSIONS AND OUTLOOK

This paper has introduced a methodology for collecting a large amount of realistic playback using a commercial crowdsourcing platform. Our goal was to open the study of collective playback behavior to researchers who do not have access to playback data collected by large commerical video platforms. We applied our methodology to develop a set of HITs, which were run on AMT and carried out by 272 workers. The data collected in this way was analyzed in order to yield insights on the use of the collective playback signal for supporting non-linear video access. We state these insights here in the form of three conclusions. First, if the collective playback signal is to support non-linear access, it is important to know how many viewers are needed for convergence to a useful pattern (*RQ1* on *Convergence* answered in Section 6.5.1). Our study shows that after only about 10 viewing sessions, a collective playback signal starts to take shape.

Second, it is important to know if users find the collective playback signal useful (*RQ2* on *Usefulness* answered in Section 6.5.2). Here, we focused on the collective playback signal visualized in the form of a heat map, as in Figure 6.1. Our study revealed that adding extra effort to hand pick particular moments does not lead to a heat map that viewers find more useful than the one created from the collective playback signal. Further, viewers seem to be relatively tolerant to a sub-optimal heat map, suggesting that video platforms can experiment with using the collective feedback signal as a heat map for supporting non-linear video access without undue worry that users will be put off by

their perceived quality of the entry points. Third, it is important to know if the collective playback signal has the potential to provide information about and beyond what can already be automatically inferred using content analysis techniques (*RQ3* on *Added value* answered in Sections 6.5.3 and 6.5.4). We found that both visual concept detection and affective content analysis have potential to support non-linear video access, but that the collective playback signal contains information that is important for viewers that they cannot replace.

We would like to conclude with a set of research directions for future work. First, in this paper we have analyzed explicit viewer feedback in the form of explanations for their choice of memorable moments, and in the form of votes on memorable moments. In the future, it would be interesting to take the comparison between explicit feedback and the implicit feedback of the collective playback signal one step further. In particular, the comparison of user satisfaction with heat maps created by expert curators, and heat maps created by collective feedback could be interesting. Further, we are interested in how feedback impacts the development of collective behavior: if the developing heat map is displayed to users, how will this change the patterns with which feedback is collect? Similarly, do we notice a difference if the heat map is initialized by time moments selected by expert curators.

Second, one of the reasons that we choose the Celebrity Cricket League is because it goes beyond being a sports game, and also qualifies as reality TV. In the future, we are interested in expanding our studies to other forms of realty TV, and of long-play video in general. Our methodology requires finding viewers that we can connect with first, and only then choosing the content. In order to broaden the range of possible viewer interests, future work should widen its scope for recruiting study participants to online fora.

Finally, much multimedia retrieval and access research assumes that the path to improving systems involves collecting more data from more users, and creating highly personalized systems. The insights of this paper point in the opposite direction. We find that the collective playback signal is a universal signal. The collective supporters of a particular cricket team and the collective signal of all viewers of all viewers of a cricket match demonstrated the same convergence behavior. Personalization may have much less of an impact than is otherwise assumed, and a useful universal signal can possibly be obtained by recording the playback behavior of only a few users. Moving forward, it is important to investigate the trade-offs between personalization, large-scale user behavior tracking, and user satisfaction in more depth. The results of this paper suggests that we should not assume that can know a priori what the optimal trade-off is.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme under grant agreement № 287704 (CUbRIK). We thank everyone who participated in our experimental studies by carrying out the task on Amazon Mechanical Turk.

IV

Outlook

7

CONCLUSIONS AND OUTLOOK

In the preceding chapters up to this point we have seen a variety of approaches to incorporating crowd perspectives into multimedia retrieval systems. From consulting crowds on crowdsourcing platforms to inform design decisions to directly leveraging collective intelligence through activity found on social side-channels and in the actual multimedia systems themselves, each approach captured the spirit of open-endedness in one way or another.

In this final chapter, we look back on the work presented in the preceding chapters and reflect on the central question of this thesis (Section 7.1). We also use our experience gained in many of the crowdsourcing campaigns carried throughout this thesis to provide practical pointers that could greatly improve future work (Section 7.2). Finally, by tying the idea of multiple perspectives into a single concept of interpretive crowdsourcing*, we open up three new potential research directions that could be pursued (Section 7.3).*

7.1. DISCUSSION

The objective of this thesis was to investigate the research question of how we can incorporate the perspectives of the crowd into multimedia retrieval systems. In order to answer this question, the preceding chapters started with a simple example of involving the crowd and slowly built up to applying crowd-informed designs to retrieval systems by leveraging different forms of collective intelligence in various ways. In this section, we reflect on the work presented in the preceding chapters so far.

The approach of starting small and slowly building up is also reflected in the thesis' structure. Chapter 2 of Part I presented a simple case of mining information from social networks (*social computing* in Figure 1.4) as a method for ranking anchors in broadcast videos. The underlying assumption of this approach was tested using a classical case of crowdsourced relevance judgment.

Moving on to Part II, Chapters 3A to 4B focused on developing methodologies for *crowdsourcing* studies in order to use crowdsourcing beyond simple labeling and evaluation. Chapter 3A laid down the foundation of our framing methodology for designing crowdsourcing tasks that help people to picture particular scenarios that are potentially outside of their typical daily life. This methodology was then first used for evaluating a new browsing feature of a multimedia retrieval system in Chapter 3B. In a sense, the methodology developed here enabled new and interesting crowdsourcing tasks not typically seen before on crowdsourcing platforms with a broader potential for multimedia applications.

Continuing the second part of the thesis, Chapter 4A focused on studying elicitation techniques for obtaining useful responses from the crowd that could be used for informing the design of new features in retrieval systems. The final chapter of Part II, Chapter 4B, built on the preceding chapter's results and investigated the feasibility of conducting user tests for fully realized and already deployed retrieval system features through the use of virtual machines in a crowdsourcing setting. In sum, these two chapters showcase one way of eliciting crowd perspectives and integrating them into a retrieval system.

Finally, the foundations laid down in the previous chapters were then used in Part III of the thesis to advance non-linear video access. Chapter 5 combined *crowdsourcing* and *social computing* and directly applied methodologies of Part II in researching new relevance dimensions at a video time-code level. The resulting crowd-informed typology for categorizing user comments with time-codes was then evaluated through a carefully framed crowdsourcing study in which hypothetical search scenarios were presented to the participants. In Chapter 6 a methodology for capturing realistic viewing behavior in a crowdsourcing setting was developed, with which it would become possible to study the utility of this viewing behavior signal as means for supporting non-linear video access.

Looking back on the work presented in this thesis, we see the success of our framing methodology reflected in users' appreciation of the multimedia information retrieval technology that was built on enrichment collected through said methodology. Especially Sections 5.6.3 and 6.5.2 showed that users reacted positively to some of the new ideas presented in this thesis. The framing methodology shows potential and is feasible. The new methodologies introduced within this thesis for incorporating crowd perspectives into multimedia systems open up new opportunities for further research. As the next two

sections will detail, we will provide a few practical pointers for future work (Section 7.2) and look at how we could increase the potential of using the crowd for multimedia information retrieval research (Section 7.3).

7.2. PRACTICAL FUTURE WORK

Based on the discussion presented in this chapter and the experience gained through all crowdsourcing campaigns carried out in this thesis, we would like to make the following recommendations for future work on *interpretive crowdsourcing*—a term discussed in more detail in the next section. We believe that following up on these recommendations would bring fundamental improvements to incorporating crowd perspectices into multimedia retrieval systems.

1. Reduction of manual post-processing workload

The requirement of being open-minded when processing responses from crowdsourcing workers makes it hard to automate the processing step. As a result, the lack of a definition of what is right means there are no clear rules that can be easily translated into a single script that checks and aggregates the results from a crowdsourcing campaign. In this thesis, a considerable amount of time was spent on manually checking whether collected answers were sincere and interpreting free text. Most of the time, the work was done by a single individual researcher. Hence, research with a large dependency on crowdsourcing tends to be less agile and could be more efficient if the burden of manual post-processing were to be reduced to a minimum. Future research should focus on improving this efficiency, while maintaining the qualities of interpretive crowdsourcing. One of the foreseeable challenges includes devising workflows that allow for distributing the workload of manually processing results among multiple researchers from a research group and thus in a sense parallelizing the human computation. The additional benefit is that more than one set of eyes could check the collected responses. Once this challence has been overcome, further research could look into crowdsourcing the post-processing step, giving research teams access to a larger pool of people over which the work could be distributed. Inspiration for this research could be drawn from [53], in which the authors created a dataset that was constructed and verified entirely through crowdsourcing.

2. Design patterns in interpretive crowdsourcing tasks

The crowdsourcing tasks used within this thesis' research were carefully designed, but setting up the crowdsourcing campaigns also took a considerable amount of work. These setup costs could be attributed to one of the following factors. First, some tasks were novel. As a result, each of these tasks basically had to be designed from scratch and then tested through iteration and refinement. Second, some tasks were similar to certain previous tasks, yet also different. In this particular case, designing these tasks feels like reinventing the wheel at times, or when the task is designed by editing a copy of a previous task, it feels tedious and error-prone. While it is true that through repetition one gains insight and experience, making the process of designing crowdsourcing tasks easier, it is also true that this experience and insight are not tangible artifacts that could

be easily shared with others. This thesis presented methodologies for designing interpretive crowdsourcing tasks, but these are like general guidelines to adhere to and do not help directly in implementing a task. To alleviate this problem, future work should focus on discovering patterns in crowdsourcing task design. Like design patterns in software [29], patterns for crowdsourcing tasks should be reusable artifacts with four essential elements: a pattern *name*, a *problem* description, a *solution* and *consequences* of applying the pattern. These design patterns would be the materialization of experience and insight researchers have accumulated through their crowdsourcing experiments. The challenge in pursuing this research is obtaining a sizable number of task designs to extract patterns from. The actual task designs are not always included in research papers, but are instead described briefly rather than in detail, most likely due to constraining factors such as time and page limits. In a sense, task design is often secondary to the main results of a paper. To overcome this challenge, one ideal situation and one practical approach can be considered. The ideal situation would be that, from now on, publishing venues would require full details on crowdsourcing task design to be included for each publication. A practical approach would be to crawl crowdsourcing platforms for tasks.

3. Crowdsourced card-sorting

Our final recommendation on future work is a concrete specialization of our first recommendation on reducing individual workload during post-processing crowdsourcing results. We recommend that the topic of card-sorting in a crowdsourcing context should be researched in more detail. Card-sorting is a technique used extensively throughout this thesis as a method for aggregating free-text responses from the crowd in order to extract a set of abstract categories that is more manageable than dealing with individual answers. This abstraction procedure often requires manual effort of sifting through the responses. It is a time-consuming task often carried out by only a limited number of people and hence involves a limited number of perspectives. Research on crowdsourcing the card-sorting process could offer a solution that tackles both the time-consuming aspect and the aspect of limited number of perspectives. For this research, existing responses from past crowdsourcing campaigns that have been card-sorted by researchers before could be reused. The initial challenge of this research would be the challenge of designing an intuitive interface for the crowdsourcing task and studying whether carrying out card-sorting on a crowdsourcing platform is feasible in the first place. Past this challenge, further research should study the stability of the card-sorting process. Possible factors that could influence the outcome of a card-sorting process could be how the tasks are framed and the number of items a worker has to sort.

7.3. CONCEPTUAL OUTLOOK

By tracing similarities throughout the thesis, we can tie the different ideas involving multiple perspectives into a single concept. In this thesis, crowdsourcing tasks that are *interpretive* are a recurring pattern. Interpretive crowdsourcing tasks are related to the concept of imaginative load. We consider a task to be interpretive when it involves a frame and it lacks a definition of what is right. The presence of a frame implies an imag-

inative load. Persons carrying out such a task are aligning their mental models to fit the frame as best as possible according to their judgment (as depicted back in Chapter 1 in Figure 1.7) and thereby adjusting their perspective on the task's subject matter. By not providing a definition of what is right, the frame and any formulated questions are still open to some interpretation. The lack of a clear-cut, black-or-white criterion is related to the notion of open-endedness. It allows for multiple interpretations, or world views so to say, and prevents people's mental models to be squished into a single point, i.e., a single permitted world perspective.

We argue that interpretive crowdsourcing should be studied in its own right, independently of other forms of crowdsourcing without an interpretive aspect. Interpretive crowdsourcing, i.e., forms of crowdsourcing with interpretive tasks, is an interesting, but not yet a well-studied, research subject that could well aid in meeting the needs of various groups of people in a world of variety of multimedia beyond the limited potential of one-size-fits all approaches. This section presents a selection of work that uses interpretive crowdsourcing related to three conceptual directions in which future work should move. The work discussed here is chosen because it contrasts or complements with what has been presented in this thesis and reveals questions yet to be answered.

1. Universal domain

The first work chosen carries part of the concept of interpretive crowdsourcing, but is void of any particular framing. In work by Krishna et al. [53], the authors construct a dataset that could serve as a visual genome through dense image annotations obtained from the crowd. The goal of this visual genome is to be the makeup of how we humans understand our visual world through the detection of objects, descriptions of these objects and their interactions, similarly how the human genome describes the genetic makeup of humans. For the construction of the dataset, there is no strong notion of what is considered right or wrong. Annotations collected from the crowd are unconstrained free-text descriptions, but each annotation must be judged by three other crowdsourcing workers for correctness before it gets accepted by the system. Besides this crowd-enforced correctness check, the system is open to multiple interpretations as it allows multiple annotations for a single image, object or relationship. In that view, the work by Krishna et al. is an example of interpretive crowdsourcing. Unfortunately, their work does not discuss how the crowdsourcing tasks have been framed, if at all. However, considering the goal of the authors is to advance the field on how humans understand the visual world, one question that arises is the following: To what extent could interpretive crowdsourcing yield generic or domain independent solutions? Is it possible to gather universal input from the crowd or do you always need a domain?

2. Relatable framing

The second work chosen is selected for its particular choice of framing. Similar to the work just discussed, Agrawl et al. introduced the task of free-form and open-ended visual question answering (VQA) [2]. In the task of VQA, a system takes an image and a question formulated in a natural sentence as input and must produce a natural-language answer as a response. This kind of task is applicable to scenarios that visually-impaired users

could encounter when they need information about some visual scene. To support this task, the authors constructed a VQA dataset through crowdsourcing, in which both questions and answers for a collection of images were collected from the crowd. There were no restrictions imposed on the questions and answers that the crowd could submit, but at the same time, the authors were also interested in collecting interesting, diverse and well-posed questions. In order to achieve this goal, the authors framed their crowdsourcing task as follows. Workers had to image a very smart robot and to think of a question about a given image which they would think the robot most likely would not be able to answer. By not having a notion of what is considered right and by presenting workers of a frame, the crowdsourcing task used in constructing the VQA dataset matches our definition of interpretive crowdsourcing. The particular choice of framing used in this work opens up an important question for further research on the topic of framing relatability. Robots are currently rather far removed from the daily life experience of many and one might wonder how well people could relate to the posed scenario. How far from current reality could we go before framing fails to capture a common understanding?

3. Closing the loop

The third and final work discussed here is selected for its perfect fit to the notion of interpretive crowdsourcing, yet having an untapped potential. In the work by Liem et al., the authors focus on unearthing the connotative layer of music [58]. The authors collect cinematic scene descriptions from the crowd for a collection of music fragments. They then show that people are able to make the reverse association from a given scene description back to the original music fragment, an insight that could lead to new forms and scenarios of multimedia retrieval. The crowdsourcing task used in the experiments made use of elaborate framing, asking participants in the experiment to imagine themselves being a great film director working on their latest magnum opus. The film director asked the best composer imaginable to arrange the film's score and this composer understands the director's vision perfectly. The participants in the crowdsourcing study are given a music fragment that was composed by the composer, which fits a particular scene in the movie perfectly. Based on just this music fragment and how the task was framed, the participants then have to describe in various details this particular scene using unrestricted, free-form text. This description of the crowdsourcing task nicely fits our notion of what constitutes interpretive crowdsourcing, namely the presence of a frame and the absence of a definition of right and wrong. The work in [58], however, does not follow up with an implementation of a multimedia information retrieval system in which the collected data was put to use, missing the opportunity to test the full potential of the crowd-gathered perspectives on music fragments. This opens up a question for future research: Does closing the loop give us new insights on how to effectively incorporate perspectives from the crowd into multimedia retrieval systems?

These three sets of questions posed in this section can serve as a starting point for future research on interpretive crowdsourcing tasks. Even if none of these questions are followed up on, it is important to take the following idea to heart for any new multimedia retrieval research:

The systems and the crowdsourcing tasks presented in this thesis were not bound by limits set by an individual researcher with a particular world view, but instead fueled by views of many.

BIBLIOGRAPHY

[1] Why Indians love cricket. *The Economist, 4 February*, 2014.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[3] Abir Al-Hajri, Matthew Fong, Gregor Miller, and Sidney Fels. Fast forward with your VCR: Visualizing single-video viewing statistics for navigation and sharing. In *Proceedings of Graphics Interface 2014*, GI '14, pages 123–128, 2014.

[4] Markos Avlonitis and Konstantinos Chorianopoulos. Video pulses: User-based modeling of interesting video segments. *Advances in Multimedia*, 2014:2:2–2:2, January 2014.

[5] Markos Avlonitis, Ioannis Karydis, and Spyros Sioutas. Early prediction in collective intelligence on video users' activity. *Information Sciences*, 298:315–329, 2015.

[6] Arslan Basharat, Yun Zhai, and Mubarak Shah. Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding*, 110(3):360–377, June 2008.

[7] Steven Bird, Edward Loper, and Ewan Klein. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[8] Dick C. A. Bulterman, Pablo Cesar, and Rodrigo Laiola Guimarães. Socially-aware multimedia authoring: Past, present, and future. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(1s):35:1–35:23, October 2013.

[9] Peter Caputi and Prasuna Reddy. A comparison of triadic and dyadic methods of personal construct elicitation. *Journal of Constructivist Psychology*, 12(3):253–264, 1999.

[10] Axel Carlier, Vincent Charvillat, and Wei Tsang Ooi. A video timeline with bookmarks and prefetch state for faster video browsing. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 967–970, New York, NY, USA, 2015. ACM.

[11] Sergiu Chelaru, Claudia Orellana-Rodriguez, and Ismail Sengor Altingovde. How useful is social feedback for learning to rank YouTube videos? *World Wide Web*, pages 1–29, 2013.

[12] Mauro Cherubini, Rodrigo de Oliveira, and Nuria Oliver. Understanding near-duplicate videos: a user-centric approach. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 35–44, New York, 2009. ACM.

[13] Konstantinos Chorianopoulos, Ioannis Leftheriotis, and Chryssoula Gkonela. SocialSkip: pragmatic understanding within web video. In *Proceedings of the 9th international interactive conference on Interactive television*, EuroITV '11, pages 25–28, New York, NY, USA, 2011. ACM.

[14] Bram Cohen. Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, pages 68–72, 2003.

[15] Paolo Costa, Vincent Gramoli, Márk Jelasity, Gian Paolo Jesi, Erwan Le Merrer, Alberto Montresor, and Leonardo Querzoni. Exploring the interdisciplinary connections of gossip-based systems. *ACM SIGOPS Operating Systems Review*, 41(5):51–60, 2007.

[16] Claire-Helène Demarty, Mats Sjöberg, Gabriel Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do, and Hanli Wang. Predicting Interestingness of Visual Content. In *Visual Content Indexing and Retrieval with Psycho-Visual Models*. 2017.

[17] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.

[18] Eelco Dolstra, Raynor Vliegendhart, and Johan Pouwelse. Crowdsourcing GUI tests. In *IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST)*, pages 332–341, March 2013.

[19] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie F. Cranor. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2399–2402, 2010.

[20] David Easley and Jon Kleinberg. The structure of the Web. In *Networks, crowds, and markets: Reasoning about a highly connected world*, chapter 13, pages 375–395. Cambridge University Press, 2010.

[21] Carsten Eickhoff and Arjen P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[22] Maria Eskevich, Robin Aly, Roeland Ordelman, David N. Racca, Shu Chen, and Gareth J. F. Jones. SAVA at MediaEval 2015: Search and anchoring in video archives. In *MediaEval*, Wurzen, Germany, September 2015.

[23] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[24] Matthew Fong, Abir Al Hajri, Gregor Miller, and Sidney Fels. Casual authoring using a video navigation history. In *Proceedings of the 2014 Graphics Interface Conference*, GI '14, pages 109–114, Toronto, Ont., Canada, Canada, 2014. Canadian Information Processing Society.

[25] Jay W. Forrester. Counterintuitive behavior of social systems. *Theory and Decision*, 2(2):109–140, 1971.

[26] Fay Fransella, Richard Bell, and Don Bannister. *A Manual for Repertory Grid Technique*. Wiley, 2nd edition, 2003.

[27] Gerald Friedland and Ramesh Jain. Multimedia: A definition. In *Multimedia Computing*, chapter 2, pages 6–14. Cambridge University Press, 2014.

[28] Petra Galuscáková and Pavel Pecina. CUNI at MediaEval 2015 search and anchoring in video archives: Anchoring via information retrieval. In *CEUR Workshop Proceedings, no. 1436, 2015. MediaEval 2015 Workshop, Wurzen, Germany, 15-15 September, 2015*. CEUR-WS, 2015.

[29] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[30] Chrysoula Gkonela and Konstantinos Chorianopoulos. VideoSkip: event detection in social web videos with an implicit user heuristic. *Multimedia Tools and Applications*, pages 1–14, 2012.

[31] Olaf Görlitz, Sergej Sizov, and Steffen Staab. PINTS: Peer-to-peer infrastructure for tagging systems. In *International Conference on Peer-to-Peer Systems*. USENIX, 2008.

[32] Catherine Grady and Matthew Lease. Crowdsourcing document relevance assessment with Mechanical Turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's MTurk*, pages 172–179, 2010.

[33] Seyong Ha, Dongwhan Kim, and Joonhwan Lee. Crowdsourcing as a method for indexing digital media. In *CHI '13 Extended Abstracts*

on Human Factors in Computing Systems, CHI EA '13, pages 931–936, New York, NY, USA, 2013. ACM.

[34] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[35] Alan Hanjalic, Christoph Kofler, and Martha Larson. Intent and its discontents: The user at the wheel of the online video search engine. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1239–1248, New York, NY, USA, 2012. ACM.

[36] Alan Hanjalic, Reginald L. Lagendijk, and Jan Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):580–588, June 1999.

[37] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[38] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 489–498, New York, NY, USA, 1999. ACM.

[39] Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, 1st edition, September 2009.

[40] Djoerd Hiemstra and Wessel Kraaij. Evaluation of multimedia retrieval systems. In Henk M. Blanken, Henk Ernst Blok, Ling Feng, and Arjen P. de Vries, editors, *Multimedia Retrieval*, pages 347–366. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[41] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[42] Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. Ranking comments on the social web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, volume 4 of *CSE '09*, pages 90–97, Washington, DC, USA, 2009. IEEE Computer Society.

[43] Panagiotis G Ipeirotis. Demographics of Mechanical Turk. In *CeDER-10-01*, CeDER Working Papers, 2010.

[44] Panagiotis G Ipeirotis. Demographics of Mechanical Turk: Now live! (April 2015 edition). `http://www.behind-the-enemy-lines.com/2015/04/demographics-of-mechanical-turk-now.html`, April 2015.

[45] Adam Kapelner and Dana Chandler. Preventing satisficing in online surveys: A "Kapcha" to ensure higher quality data. In *CrowdConf ACM Proceedings*, 2010.

[46] Rianne Kaptein, Djoerd Hiemstra, and Jaap Kamps. How different are language models and word clouds? In *ECIR 2010*, volume 5993 of *LNCS*, pages 556–568, Berlin, March 2010.

[47] Ioannis Karydis, Markos Avlonitis, Konstantinos Chorianopoulos, and Spyros Sioutas. Identifying important segments in videos: A collective intelligence approach. *International Journal on Artificial Intelligence Tools*, 23(02), 2014.

[48] George A. Kelly. *The Psychology of Personal Constructs, volume one: Theory and personality*. Norton, New York, 1955.

[49] Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 572–580, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[50] Christoph Kofler, Martha Larson, and Alan Hanjalic. Intent-aware video search result optimization. *IEEE Transactions on Multimedia*, 16(5):1421–1433, 2014.

[51] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, February 2009.

[52] Svetlana Kordumova, Christoph Kofler, Dennis C. Koelma, Bouke Huurnink, Bauke Freiburg, Joris Kleinveld, Manuel van Rijn, Marco van Deursen, Martha Larson, and Cees G. M. Snoek. SocialZap: Catch-up on interesting television fragments discovered from social media. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 538:538–538:540, New York, NY, USA, 2014. ACM.

[53] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[54] Rodrigo Laiola Guimarães, Pablo Cesar, and Dick C. A. Bulterman. Let me comment on your video: Supporting personalized end-user comments within third-party online videos. In *Proceedings of the 18th Brazilian symposium on Multimedia and the web*, pages 253–260. ACM, 2012.

[55] Martha Larson, Nitendra Rajput, Abhigyan Singh, and Saurabh Srivastava. I want to be Sachin Tendulkar!: A spoken English cricket game for rural students. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1353–1364, 2013.

[56] Hao Li, Meng Wang, and Xian-Sheng Hua. MSRA-MM 2.0: A large-scale web multimedia dataset. In *Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on*, pages 164–169, December 2009.

[57] Stan J. Liebowitz and Alejandro Zentner. Clash of the titans: Does Internet use reduce television viewing? *Review of Economics and Statistics*, 94(1):234–245, 2012.

[58] Cynthia C. S. Liem, Martha Larson, and Alan Hanjalic. When music makes a scene. *International Journal of Multimedia Information Retrieval*, 2(1):15–30, 2013.

[59] Babak Loni, Maria Menendez, Mihai Georgescu, Luca Galli, Claudio Massari, Ismail Sengor Altingovde, Davide Martinenghi, Mark Melenhorst, Raynor Vliegendhart, and Martha Larson. Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 72–77. ACM, 2013.

[60] Eng Keong Lua, Jon Crowcroft, Marcelo Pias, Ravi Sharma, and Steven Lim. A survey and comparison of peer-to-peer overlay network schemes. *Communications Surveys & Tutorials, IEEE*, 7(2), 2005.

[61] Marco Lui and Timothy Baldwin. `langid.py`: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Demo Session, Jeju, Republic of Korea*, 2012.

[62] Amy Madden, Ian Ruthven, and David McMenemy. A classification scheme for content analyses of YouTube video comments. *Journal of Documentation*, 69(5):693–714, 2013.

[63] Thomas W. Malone, Robert Laubacher, and Chrysanthos Dellarocas. Harnessing crowds: Mapping the genome of collective intelligence. Technical Report CCI Working Paper 2009-001, Massachusetts Institute of Technology, 2009.

[64] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, 2008.

[65] Claudio Martella, Ekin Gedik, Laura Cabrera-Quiros, Gwenn Englebienne, and Hayley Hung. How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances. In

Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, pages 201–210, 2015.

[66] Robert Mertens, Rosta Farzan, and Peter Brusilovsky. Social navigation in web lectures. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, HYPERTEXT '06, pages 41–44, New York, NY, USA, 2006. ACM.

[67] Sean A. Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '10, pages 1457–1466, 2010.

[68] Nico Nico Douga. `http://nicovideo.jp`.

[69] Nielsen. Screen wars: The battle for eye space in a tv-everywhere world. `http://nielsen.com/us/en/insights/reports/2015/screen-wars-the-battle-for-eye-space-in-a-tv-everywhere-world.html`, 2015.

[70] Donald A. Norman. Some observations on mental models. In Dedre Gentner and Albert L. Stevens, editors, *Mental Models*, chapter 7, pages 7–14. Lawrence Erlbaum Associates, Inc., 1983.

[71] Dan R. Olsen and Brandon Moon. Video summarization based on user interaction. In *Proceedings of the 9th International Interactive Conference on Interactive Television*, EuroITV '11, pages 115–122, New York, NY, USA, 2011. ACM.

[72] Anthony J. Onwuegbuzie, Larry Daniel, and Nancy L. Leech. Pearson product-moment correlation coefficient. In Neil J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 751–756. SAGE Publications, Inc., Thousand Oaks, CA, USA, 2007.

[73] Roeland J.F. Ordelman, Maria Eskevich, Robin Aly, Benoit Huet, and Gareth Jones. Defining and evaluating video hyperlinking for navigating multimedia archives. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 727–732, New York, NY, USA, 2015. ACM.

[74] Manoj Parameswaran and Andrew B. Whinston. Social computing: An overview. *Communications of the Association for Information Systems*, 19(1):37, 2007.

[75] Wei-Ting Peng, Wei-Ta Chu, Chia-Han Chang, Chien-Nan Chou, Wei-Jia Huang, Wen-Yan Chang, and Yi-Ping Hung. Editing by viewing: Automatic home video summarization by viewing behavior analysis. *IEEE Transactions on Multimedia*, 13(3):539–550, June 2011.

[76] Martin Potthast, Benno Stein, Fabian Loose, and Steffen Becker. Information retrieval in the commentsphere. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):68:1–68:21, September 2012.

[77] Johan A. Pouwelse, Pawel Garbacki, Jun Wang, Arno Bakker, Jie Yang, Alexandru Iosup, Dick H. J. Epema, Marcel Reinders, Maarten R. van Steen, and Henk J. Sips. Tribler: A social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, 20(2):127–138, 2008.

[78] Johan A. Pouwelse, Jie Yang, Michel Meulpolder, Dick H.J. Epema, and Henk J. Sips. Buddycast: an operational peer-to-peer epidemic protocol stack. In G.J.M. Smit, D.H.J. Epema, and M.S. Lew, editors, *ASCI 2008*, 2008.

[79] Alexander J. Quinn and Benjamin B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[80] Justus J. Randolph. Free-marginal multirater kappa (multirater κ_{free}): An alternative to Fleiss' fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*. ERIC, 2005.

[81] Michael Riegler, Mathias Lux, Vincent Charvillat, Axel Carlier, Raynor Vliegendhart, and Martha Larson. Videojot: A multifunctional video annotation tool. In *Proceedings of International Conference on Multimedia Retrieval*, page 534. ACM, 2014.

[82] Gordon Rugg and Peter McGeorge. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93, 1997.

[83] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 595–604. ACM, 2011.

[84] Klaus Schoeffmann, Frank Hopfgartner, Oge Marques, Laszlo Boeszoermenyi, and Joemon M. Jose. Video browsing interfaces and applications: a review. *Journal of Photonics for Energy*, pages 018004–018004–35, 2010.

[85] David A. Shamma, Ryan Shaw, Peter L. Shafton, and Yiming Liu. Watch what I watch: Using community activity to understand content. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pages 275–284, New York, NY, USA, 2007. ACM.

[86] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)*, 8(3):17:1–17:39, July 2014.

[87] Anca-Roxana Simon, Guillaume Gravier, and Pascale Sébillot. IRISA at MediaEval 2015 search and anchoring in video archives task. In *CEUR Workshop Proceedings, no. 1436, 2015. MediaEval 2015 Workshop, Wurzen, Germany, 15-15 September, 2015*. CEUR-WS, 2015.

[88] James Sinclair and Michael Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, 2008.

[89] Ewine Smits and Alan Hanjalic. A system concept for socially enriched access to soccer video collections. *IEEE Multimedia*, 17:26–35, 2010.

[90] Cees G. M. Snoek and Marcel Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, April 2009.

[91] Mohammad Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 919–922, 2015.

[92] Mohammad Soleymani and Martha Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 4–8, 2010.

[93] Antonia Spiridonidou, Ioannis Karydis, and Markos Avlonitis. Mimicking real users' interactions on web videos through a controlled experiment. In *Engineering Applications of Neural Networks*, pages 60–69. Springer, 2013.

[94] Kathryn T. Stolee and Sebastian Elbaum. Exploring the use of crowdsourcing to support empirical studies in software engineering. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '10, pages 35:1–35:4, 2010.

[95] Tanveer Syeda-Mahmood and Dulce Ponceleon. Learning video browsing behavior and its application in the generation of video previews. In *Proceedings of the Ninth ACM International Conference on Multimedia*, MULTIMEDIA '01, pages 119–128, New York, NY, USA, 2001. ACM.

[96] Anthony Tang and Sebastian Boring. #epicplay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1569–1572, New York, NY, USA, 2012. ACM.

[97] Bina Trivedi. CCL: Superstar cricketainment. *Times of India, 20 January 2012*, 2012.

[98] Raynor Vliegendhart, Eelco Dolstra, and Johan Pouwelse. Crowdsourced user interface testing for multimedia applications. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 21–22. ACM, 2012.

[99] Raynor Vliegendhart, Martha Larson, and Alan Hanjalic. LikeLines: collecting timecode-level feedback for web videos through user interactions. In *Proceedings of the 20th ACM international conference on Multimedia*, MM '12, pages 1271–1272, New York, NY, USA, 2012. ACM.

[100] Raynor Vliegendhart, Martha Larson, and Alan Hanjalic. Collecting realistic viewing behavior from the crowd for non-linear video access. *IEEE Transactions on Multimedia*, under review.

[101] Raynor Vliegendhart, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse. Investigating factors influencing crowdsourcing tasks with high imaginative load. In *WSDM'11 Workshop on Crowdsourcing for Search and Data Mining*, February 2011.

[102] Raynor Vliegendhart, Martha Larson, Christoph Kofler, and Johan Pouwelse. A peer's-eye view: network term clouds in a peer-to-peer system. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1909–1912. ACM, 2011.

[103] Raynor Vliegendhart, Martha Larson, Babak Loni, and Alan Hanjalic. Exploiting the deep-link commentsphere to support non-linear video access. *IEEE Transactions on Multimedia*, 17(8):1372–1384, 2015.

[104] Raynor Vliegendhart, Martha Larson, and Johan Pouwelse. Discovering user perceptions of semantic similarity in near-duplicate multimedia files. In *Proceedings of the 1st International Workshop on Crowdsourcing Web Search*, pages 54–58. CEUR-WS.org, April 2012.

[105] Raynor Vliegendhart, Cynthia C.S. Liem, and Martha Larson. Exploring microblog activity for the prediction of hyperlink anchors in television broadcasts. In *CEUR Workshop Proceedings, no. 1436, 2015. MediaEval 2015 Workshop, Wurzen, Germany, 15-15 September, 2015*. CEUR-WS, 2015.

[106] Raynor Vliegendhart, Babak Loni, Martha Larson, and Alan Hanjalic. How do we deep-link?: Leveraging user-contributed time-links for non-linear video access. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 517–520, New York, NY, USA, 2013. ACM.

[107] Nicholas Vretos, Vassilios Solachidis, and Ioannis Pitas. A mutual information based face clustering algorithm for movie content analysis. *Image and Vision Computing*, 29(10):693–705, 2011.

[108] Shoko Wakamiya, Daisuke Kitayama, and Kazutoshi Sumiya. Scene extraction system for video clips using attached comment interval and pointing region. *Multimedia Tools and Applications*, 54(1):7–25, 2011.

[109] Wikipedia. Celebrity Cricket League — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Celebrity+Cricket+League&oldid=779755718`, 2017. [Online; accessed 14-May-2017].

[110] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2005.

[111] Xiao Wu, Alexander G. Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th international conference on Multimedia*, MM '07, pages 218–227, New York, 2007. ACM.

[112] Meng Yang and Gary Marchionini. Exploring users' video relevance criteria—a pilot study. *Proceedings of the American Society for Information Science and Technology*, 41(1):229–238, 2004.

[113] Jude Yew, David A. Shamma, and Elizabeth F. Churchill. Knowing funny: Genre perception and categorization in social video sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 297–306, New York, NY, USA, 2011. ACM.

[114] YouTube — Deep Links. `http://support.google.com/youtube/bin/answer.py?hl=en&answer=116618`.

[115] YouTube — Looking ahead in the YouTube player. `http://youtube-global.blogspot.com/2012/03/looking-ahead-in-youtube-player.html`.

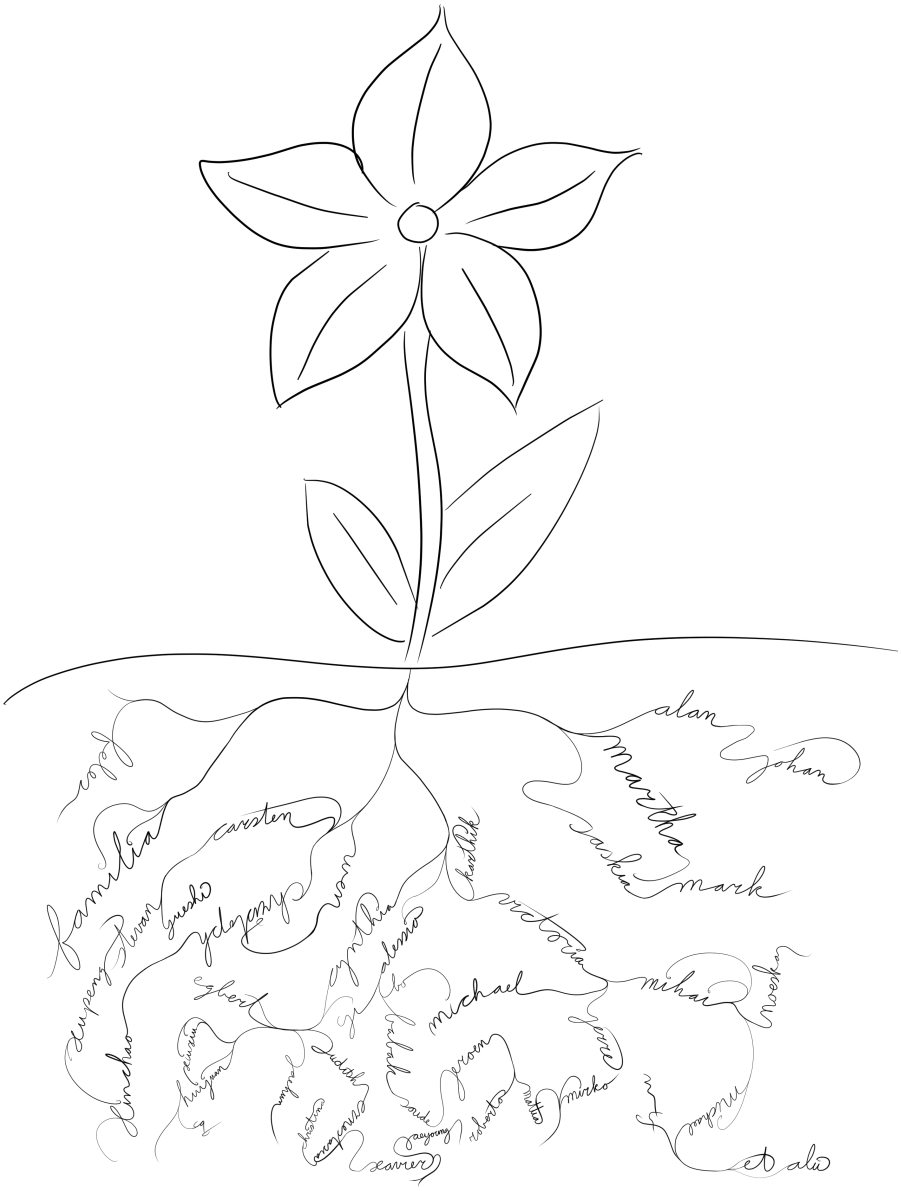
[116] YouTube — Statistics. `http://web.archive.org/web/20150407044201/http://www.youtube.com/yt/press/statistics.html`, 2015.

[117] Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang. Video summarization based on user log enhanced link analysis. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 382–391, New York, NY, USA, 2003. ACM.

[118] Niels Zeilemaker, Mihai Capotă, Arno Bakker, and Johan Pouwelse. Tribler: P2P media search and sharing. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 739–742. ACM, 2011.

[119] Gloria Zen, Paloma de Juan, Yale Song, and Alejandro Jaimes. Mouse activity as an indicator of interestingness in video. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 47–54, 2016.

Acknowledgements



CURRICULUM VITÆ

Raynor Vliegendhart was born on July 14, 1985 in Waddinxveen, The Netherlands. From 1997 until 2003, he received secondary education (gymnasium level) at De Goudse Waarden in Gouda.

After completing his secondary education, Raynor entered university. He obtained both his Computer Science BSc degree (2003–2007) and MSc degree (2007–2010) from the Delft University of Technology. During his studies, he was a part-time teaching assistant for several lab courses.

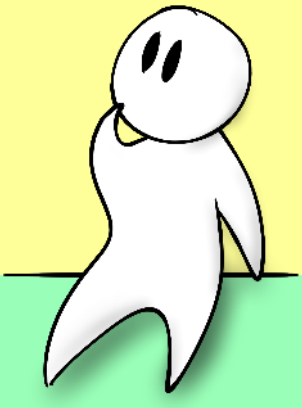
Right after obtaining his master's degree, he joined the Delft Multimedia Information Retrieval Lab, which later became the Multimedia Computing group, as a scientific programmer. In this group at the Delft University of Technology, he carried out field trials for the EU FP7 project PetaMedia from 2010 to 2011.

From November 2011 onwards, he pursued a PhD degree at the Delft University of Technology under supervision of prof. dr. Alan Hanjalic and prof. dr. Martha Larson. During his PhD studies he received the ASCI Best Poster Award, 2nd prize, for his poster presentation at ICT.Open 2012, the Dutch conference for ICT research, and a distinctive mention from the organizers of the Search and Anchoring in Video Archives (SAVA) task at MediaEval 2015 for "Bringing with Twitter the most innovative approach to the SAVA task".

In March 2016, Raynor started to work at Ingenieursbureau Geodelta, at which he was responsible for maintaining photogrammetric software and carrying out related research and development (R&D) activities. In June 2017, he also became an employee at The Netherlands' Cadastre, Land Registry and Mapping Agency, at which he continued to work on photogrammetric software in the context of aerial photography.

The twenty-first century has brought plentiful computational power and bandwidth to the masses and has opened up access to multimedia recording devices for everyone. With these developments, a shift in the landscape of multimedia took place: from traditional one-to-many programming (the paradigm of traditional television) to many-to-many creation of diverse content. Nowadays, everyone can become a content creator and connect with new audiences, which has resulted in an explosion of diverse and available multimedia content. In tandem with this change, user needs have evolved as well. Yet, existing multimedia retrieval systems have been struggling to keep up with what users are looking for.

In this thesis, we argue that a multi-perspective approach is desired in order to cater to a diverse range of user needs. In order to know which perspectives should be taken, we turn to the crowd as a source of information on which perspectives would be actually helpful for serving users of multimedia retrieval systems. The central question underlying the research presented in this thesis is: How can we incorporate these perspectives of the crowd into multimedia retrieval systems?



ISBN 978-94-6299-787-5