METHODS FOR MANAGING, VALIDATING AND RETRIEVING EVENT–RELATED INFORMATION IN EVOLVING CONTEXTS

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

DOKTOR DER NATURWISSENSCHAFTEN

Dr. rer. nat.

genehmigte Dissertation
von

M. Sc. Andrea Ceroni

geboren am 10. Juli 1988, in Genua, Italien

Hannover, Deutschland, 2018

Referent: Prof. Dr. techn. Wolfgang Nejdl
Korreferent: Prof. Dr. Francesco G.B. De Natale
Tag der Promotion: 05.06.2018

ABSTRACT

Events have always been fundamental building blocks of individual lives as well as of the whole world. Nowadays, thanks to the several technological advances achieved within the digital age, the processes of capturing, describing and spreading events have never been so simple and intuitive. This results in an ubiquitous presence of event-related information, which is digitally embedded in any form of media. Both the pervasiveness of such information as well as the benefits of its exploitation for many purposes have fostered decades of research effort to detect and summarize it. However, several issues emerge at subsequent stages and shall be addressed to support the proper exploitation and consumption of event-related information. The work presented within this thesis is indeed committed to this goal.

The aforementioned ubiquity of events makes them exhibit different characteristics and appear in a diverse range of scenarios. Therefore, we categorize events according to three main aspects that come into play when considering the management and usage of event-related information over time, once it has been created. These are the *degree of privacy*, as events can be of public domain or rather pertain to a more personal sphere, the *type of description*, which is the form (e.g. textual or visual) in which events are described, and the *time of usage*, namely the temporal horizon over which event-related information is expected to be accessed and used. The problems addressed in this thesis regard different combinations of such aspects, each one subject to specific issues to be dealt with.

Concerning the private sphere, we aim at properly *managing* large amounts of photographs taken during personal events, so that they can be easily revisited and enjoyed in the future. The common habit of dumping every single picture, encouraged by the availability of cheap storage devices, poses serious threats to their future revisiting and calls for more selective strategies to identify the most important pictures from an entire collection, thus making the future reminiscence of the related events more enjoyable and less tedious. In fact, going through the whole stored photo collections can be such a cumbersome procedure to discourage from doing it at all. We present a selection method that learns to identify the photos that the collection owner would like to keep from a whole collection for future reminiscence, outperforming approaches based on clustering and on the concept of coverage.

Then, moving towards more public settings, we consider the problem of *validating* the occurrence of events of public domain in the real world based on the information contained in textual document collections. In scenarios where events are detected from large amounts of natural language text by automatic procedures, which might introduce false positive detections, being able to retain true events while discarding the false ones becomes fundamental for a proper exploitation of the detected event-related information for any subsequent purpose. We therefore validate the verity of events by checking whether they are reported within a set of documents, which serve as ground truth, reaching substantial agreement with human evaluators. Moreover, when performing event validation as a post-processing step of event detection, we observed an increase of precision within the set of detected events.

Finally, we make a temporal jump and consider a scenario where descriptive information of public events (e.g. news articles) are read after few decades. Since the original context of an event, needed for its proper comprehension, might have been forgotten or never known at all after such a relatively long time, we aim at *retrieving* contextualizing information to support the understanding of old events in presence of wide temporal and contextual gaps. We investigate methods to formulate queries from event descriptions as seeds for retrieving topically and temporally relevant information from a context source, particularly aiming at high recall. Targeting recall as query performance criterion makes the set of retrieved results a favorable starting point for pursuing additional objectives at subsequent stages.

Keywords: *personal photo selection, event validation, recall-based query formulation*

ZUSAMMENFASSUNG

Schon immer waren Ereignisse wesentliche Bausteine sowohl des indiviudellen Lebens, als auch der ganzen Welt. Dank diverser technologischer Fortschritte im digitalen Zeitalter sind die Prozesse des Erfassens, Beschreibens und Verbreitens von Ereignissen simpler und intuitiver als jemals zuvor. Dies führt zu einer allgegenwärtigen Präsenz ereignisbezogener Informationen eingebettet in unseren digitalen Medien. Sowohl die Verbreitung als auch die Vorteile ihrer Nutzung für viele verschiedene Zwecke haben Jahrzehnte wissenschaftlicher Bemühungen zur Entdeckung und Zusammenfassung dieser Informationen gefördert. Dabei entstehen unterschiedliche Probleme, mit denen wir uns befassen, um die geeignete Nutzung und Verarbeitung von ereignisbezogenen Informationen zu unterstützen. Die hier präsentierte Arbeit widmet sich diesem Ziel.

Die erwähnte Allgegenwärtigkeit von Ereignissen führt zu verschiedenen Ereignisformen mit unterschiedlichen Charateristiken. Daher kategorisieren wir Ereignisse nach drei Hauptaspekten, die beim Einbeziehen des Managements und der Verwendung von ereignisbezogenen Informationen über die Zeit eine Rolle spielen. Diese Aspekte sind der *Privatheitsgrad*, der angibt, ob ein Ereignis tendenziell mehr öffentlich oder mehr in einen privaten Kreis eingeordnet wird, der *Beschreibungstyp*, der die Form, in der das Ereginis repräsentiert wird, angibt sowie die *Nutzungszeit*, der zeitliche Rahmen, in dem voraussichtlich auf ereignisbezogene Informationen zugegriffen wird. Die Probleme, die in dieser Arbeit behandelt werden, beziehen sich auf verschiedene Kombinationen dieser Aspekte.

Hinsichtlich der Privatsphäre beschäftigen wir uns mit der geeigneten Verwaltung großer Fotomengen, die im Rahmen persönlicher Ereignisse aufgenommen wurden, damit Nutzer diese in der Zukunft einfacher wiederaufrufen und sich an ihnen erfreuen können. Die weitverbreitete Angewohnheit, jedes Bild dauerhaft abzuspeichern, die insbesondere auf preisgünstige Speichergeräte zurückzuführen ist, stellt eine ernstzunehmende Bedrohung für die Wiedernutzung dar und verlangt nach selektiveren Strategien zur Identifizierung der wichtigsten Bilder einer Sammlung, die das Wiederansehen alter Bilder angenehmer gestalten. Wir stellen eine Auswahlmethode vor, welche lernt, jene Fotos zu identifizieren, die der Besitzer der Sammlung dauerhaft behalten möchte, wobei wir existierende cluster- und umfangsbasierte Ansätze in der Leistung übertreffen.

Im Hinblick auf öffentlichere Einstellungen beschäftigen wir uns mit dem Problem der Validierung des Auftretens von Ereignissen der öffentlichen Domäne in der realen Welt basierend auf Informationen, die in Sammlungen von Textdokumenten enthalten sind. In Szenarien, in denen automatische Verfahren Ereignisse in großen Textmengen von natürlicher Sprache entdecken, kann es zu false positive-Entdeckungen kommen. In diesen Fällen ist das Behalten echter Ereignisse und das Verwerfen falscher Ereignisse fundamental für eine geeignete Nutzung der erkannten ereignisbezogenen Informationen in nachfolgenden Anwendungen. Hierzu validieren wir die Wahrheit von Ereignissen, indem wir prüfen, ob diese Ereignisse in einer Menge von Dokumenten erwähnt werden. Diese Dokumente fungieren dabei als Grundwahrheit und erreichen deutliche Zustimmung von menschlichen Bewertern.

Schließlich machen wir einen zeitlichen Sprung und beschäftigen uns mit einem Szenario, in dem beschreibende Informationen eines öffentlichen Ereignisses (z.B. Nachrichtenartikel) nach einigen Jahrzehnten gelesen werden. Da der ursprüngliche Kontext eines Ereignisses, der für das Verständnis benötigt wird, nach relativ langer Zeit vergessen sein kann, streben wir danach, Kontextinformationen zu extrahieren, um das Verstehen vergangener Ereignisse zu unterstützen. Wir untersuchen Methoden, um Anfragen aus Ereignisbeschreibungen als Seeds zur Extrahierung von thematisch und zeitlich relevanten Informationen aus einer Kontextquelle zu formulieren. Dabei streben wir hohe Recall-Werte an. Recall als Performance-Kriterium macht die Menge der extrahierten Ergebnisse zu einem günstigen Startpunkt für das Verfolgen weiterer Ziele in nachfolgenden Schritten.

Schlagwörter: *Persönliche Fotoauswahl, Ereignisvalidierung, Recall-basierte Anfragenformulierung*

FOREWORD

The work presented in this thesis has been disseminated through various peer-reviewed publications, as follows.

Chapter 2 is built upon the work published in:

- Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. *To Keep or not to Keep: An Expectation-oriented Photo Selection Method for Personal Photo Collections.* In Proceedings of the 5th ACM International Conference on Multimedia Retrieval, ICMR'15, pages 187–194, 2015. [CSN+15]

- Andrea Ceroni, Vassilios Solachidis, Mingxin Fu, Nattiya Kanhabua, Olga Papadopoulou, Claudia Niederée, and Vasileios Mezaris. *Investigating Human Behaviors in Selecting Personal Photos to Preserve Memories.* In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops '15, pages 1–6, 2015. [CSF+15]

- Mingxin Fu, Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. *Learning Personalized Expectation-oriented Photo Selection Models for Personal Photo Collections.* In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops '15, pages 1–6, 2015. [FCS+15]

- Andrea Ceroni. (2018) *Personal Photo Management and Preservation.* In: Mezaris V., Niederée C., Logie R. (eds) Personal Multimedia Preservation – Remembering or Forgetting Images and Videos. Springer, Berlin, Heidelberg. [Cer18]

In Chapter 3 we describe our research presented in:

- Andrea Ceroni and Marco Fisichella. *Towards an Entity-based Automatic Event Validation.* In Proceedings of the 36th European Conference on Information Retrieval, ECIR'14, pages 605–611, 2014. [CF14]

- Andrea Ceroni, Ujwal Gadiraju, and Marco Fisichella. *Improving Event Detection by Automatically Assessing Validity of Event Occurrence in Text.* In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM'15, pages 1815–1818, 2015. [CGF15]

- Andrea Ceroni, Ujwal Gadiraju, Jan Matschke, Simon Wingert, and Marco Fisichella. *Where the Event Lies: Predicting Event Occurrence in Textual Documents.* In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR'16, pages 1157–1160, 2016. [CGM+16]

- Andrea Ceroni, Ujwal Gadiraju, and Marco Fisichella. *JustEvents: A Crowdsourced Corpus for Event Validation with Strict Temporal Constraints.* In Proceedings of the 39th European Conference on Information Retrieval, ECIR'17, pages 484–492, 2017. [CGF17]

In Chapter 4 we describe contributions included in:

- Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini, and Marco Fisichella. *Information Evolution in Wikipedia.* In Proceedings of the 10th International Symposium on Open Collaboration, OpenSym'14, pages 1–10, 2014. [CGG+14]

- Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. *Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-contextualization.* In Proceedings of the 8th ACM International Conference on Web Search and Data Mining, WSDM'15, pages 339–348, 2015. [TCKN15a]

During the course of my doctoral studies I have also published other papers related to information retrieval and content analysis, which are not discussed within this thesis due to its space limitations and scope. The complete list of such publications is reported hereafter.

- Tuan A. Tran, Andrea Ceroni, Mihai Georgescu, Kaweh Djafari Naini, and Marco Fisichella. *Wikipevent: Leveraging Wikipedia Edit History for Event Detection.* In Proceedings of the 15th International Conference on Web Information Systems Engineering, WISE'14, pages 90–108, 2014. [TCG+14]

- Andrea Ceroni, Nam Khanh Tran, Nattiya Kanhabua, and Claudia Niederée. *Bridging Temporal Context Gaps Using Time-aware Re-contextualization.* In Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'14, pages 1127–1130, 2014. [CTKN14]

- Marco Fisichella, Andrea Ceroni, Fan Deng, and Wolfgang Nejdl. *Predicting Pair Similarities for Near-duplicate Detection in High Dimensional Spaces.* In Proceedings of the 25th International Conference on Database and Expert Systems Applications, DEXA'14, pages 59–73, 2014. [FCDN14]

- Ujwal Gadiraju, Kaweh Djafari Naini, Andrea Ceroni, Mihai Georgescu, Dang Duc Pham, and Marco Fisichella. *Wikipevent: Temporal Event Data for the Semantic Web.* In Proceedings of the 13th International Semantic Web Conference (Posters & Demonstrations Track), ISWC'14, pages 125–128, 2014. [GNC+14]

- Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. *Time-travel Translator: Automatically Contextualizing News Articles.* In Proceedings of the 24th International Conference on World Wide Web (Companion Volume), WWW'15, pages 247–250, 2015. [TCKN15b]

- Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, and Vasileios Mezaris. *Expo: An Expectation-oriented System for Selecting Important Photos from Personal Collections.* In Proceedings of the 7th ACM International Conference on Multimedia Retrieval, ICMR'17, pages 187–194, 2017. [CSN+17]

- Andrea Ceroni, Chenyang Ma, and Ralph Ewerth. *Mining Exoticism from Visual Content with Fusion-based Deep Neural Networks.* In Proceedings of the 8th ACM International Conference on Multimedia Retrieval, ICMR'18, pages 37–45, 2018. [CME18]

Contents

List of Figures

List of Tables

1 Introduction

Events are of paramount importance as they shape our lives and trigger the evolution of today's world. The advances brought by the digital age have amplified their echo, dramatically simplifying the creation of descriptive event-related information in different forms of media. No matter whether such events are milestones of a person's life, complex political affairs, or simply trivial situations of the daily life, all of them can be at least partially captured by digital objects such as text, pictures, videos. Such event descriptions are shared and spread through the Web in different forms, for instance, as news articles by news broadcasters, posts on Social Media and online communities, collaboratively curated content, or more informal discussions in blogs. This last dissemination phase, however, might not happen in case of more personal and affective events, which might rather remain private or shared only within a restricted circle of relatives and close friends. In any case, today event-related information is pervasive in the digital world and it is essential for purposes like studying the past, understanding the present, anticipating medical, economical, political situations, and, moving to a more personal sphere, keeping memories of our lives vivid.

The ubiquity and potential of event-related information have motivated decades of research on event detection, extraction, and summarization, which nowadays encompasses hundreds of approaches tailored to different kinds of data and domains [AK15, All02, CG16b, HFK$^+$16, TMM$^+$16]. However, less interest has been devoted to what happens once events have been captured, which is indeed the primary subject of this thesis. While initiatives for organizing events and their descriptions do exist, for both the public (e.g. the GDELT project [LS13], Wikipedia's Current Events[1]) and the more personal ones (e.g. personal photo management software by Apple, Google, Microsoft), still several issues arise and shall be addressed for the sake of a proper exploitation and consumption of this event-related information over time. The above mentioned event ubiquity becomes a mixed blessing in this regards. While being one major reason of interest in events, it also scatters them over a diverse spectrum of scenarios, each one with different characteristics and issues to be dealt with.

[1] https://en.wikipedia.org/wiki/Portal:Current_events

The content of this thesis reflects this fact, as it embraces events that differ from each other under two main aspects. One is the degree of privacy, where events can either be of public domain or belong to the personal life of individuals; the other is the way in which events are represented, namely either with textual or visual information. In any of the considered scenarios, as we will see, the temporal dimension has a remarkable role. In this regard, we are particularly interested in ensuring the long-term access to event-related information, an issue that has not received the deserved attention yet.

In this thesis we address three problems related to the management and exploitation of event-related information. The first one consists in *validating* the verity of public events, by finding evidences within sources of information to affirm whether they truly happened or not. This is especially important when events are detected from large amounts of text in natural language by automatic methods, as these could introduce false positive detections. Recognizing and discarding false events would benefit any subsequent stage that makes use of the collected event-related information. Second, we assume events to be properly stored along with their descriptive information and question how they could be understood in the future, e.g. after a few decades, when the context in which they happened has evolved and, consequently, the original one might have been forgotten. Given such lack of contextual information about old events, which might hinder their comprehension, our goal is *retrieving* information to bridge these contextual gaps and support the understanding of old events. Third, moving towards the private sphere, we aim at properly *managing* increasing amounts of photographs depicting personal events, so that they can remain accessible and enjoyable over time. The common habit of dumping every single picture, possibly due to the difficulty of anticipating future information needs and the effort required for revisiting and pruning potentially large photo collections, seriously threatens the reminiscence of event memories at later points in time.

In the rest of this introductory chapter we will explain the reasons that brought us to focus on the three above mentioned problems and we will highlight the main research contributions that we have achieved in each of them.

1.1 Event Dimensions

This thesis is all about events and, therefore, it is important to clarify how they are represented for our purposes. The first fact to point out is that there is not a common consensus on how events should be defined, due to the wide range of domains and applications they appear in. A well known and very general definition is the one given by the Topic Detection and Tracking (TDT) project [APL98], which looks on events as "something that happens at some specific time and place". This has been extensively used as reference point within the Information Retrieval community. Other works extend such definition either to conform to specific media, e.g. News [XWL+15] or Social Media [XLY+16], or to introduce additional notions, e.g. interestingness

[AS12] or significance [MMJ13]. Differently, the Natural Language Processing (NLP) community embraces more fine-grained representations of events, which, besides slight differences, are regarded as propositions containing subjects, objects and predicates, possibly with further temporal and spatial information. We refer the reader to the survey by Sprugnoli et al. [ST17], which retraces the event definitions within the past 25 years of NLP research. Finally, also the multimedia domain attributes many different meanings to events (see [TMM$^+$16] for an overview). These range from simple notions, e.g. "changes of states" in multimedia streams or sets of "actions" performed by one or more agents, to more high-level ones for describing news, social events, as well as more personal ones [LTM03]. For the sake of completeness we also point out that, in parallel to the variety of event definitions mentioned above, a wide range of event models have been designed over the years (e.g. SEM [vHMS$^+$11], LODE [STH09]). We refer the reader to [ST17, TMM$^+$16] for an overview on this matter.

We adopt a simple notion of event, close to the one given by the TDT project, so that it can hold for all the different scenarios considered in this thesis. It is based on the fact that the temporal dimension is one of the few aspects that recur across all the event definitions. In more detail, we regard an event as "a set of participants related together within a given time period" (similarly to works on event detection, e.g. [DSJY11]). We will augment this core definition with additional descriptive information, e.g. type of the participants, textual or visual event descriptions, locations, depending on what is available in the particular scenario at hand. On top of this simple definition of event, we introduce 3 hyper-dimensions based on which events can be further categorized, namely their *degree of privacy*, *type of description*, and *time of usage*. We highlight these dimensions as they define the space that we explore in this thesis: all the problems that we tackle exhibit a different combination of them. The dimensions that we consider are described hereafter, while their combinations that we picked as scenarios to be investigated are presented in Section 1.2.

1.1.1 Degree of Privacy

Events continuously occur, everywhere, but they are characterized by different degrees of privacy and visibility. Some of them, for instance political elections, sport matches, natural disasters, concerts, are publicly known and their related information can be easily accessed either via digital or physical media. Others, such as vacations, work or free time, belong to the personal life of individuals and their descriptions, whatever their form is, might be rather kept private among who participated in the event. Events shared on Social Media are a kind of middle ground, as they can belong to the personal lives of inner circles of people but still they can be exposed to a wider audience. In this thesis we address problems related to either the private or public domains, nevertheless we will also refer to the social one in the following preliminary discussions for the sake of comparison.

There are some particular differences between the public and private spheres that should be taken into account when handling event-related information. The first one lies in how an event is perceived. In case of a public event, each individual might feel more or less involved and interested in it due to various reasons, e.g. the topic of the event or the geographic locations affected by it. Nevertheless, most of the people are external observers with their own point of view on what happened. Things change when moving towards more private situations (e.g. vacations, home life, ceremonies), no matter whether they are shared on Social Media or not, where there is a clearer separation between who attended the event and any other external individual, who might not even know the event participants at all. This becomes crucial when apprising data regarding a personal event, e.g. for the sake of identifying and preserving those most valuable to the owner, as the importance he or she attributes to any information item might be influenced by the presence of personal attachment, background knowledge, and memories, which are not known to anyone outside the context of the event. For instance, one might be emotionally attached to a picture because it recalls memories of things happened when it was taken, even though it has bad quality or it does not depict anything particularly appealing.

Another point of distinction concerns the kind of available information related to events. The public ones are often reported in form of news articles, which are structured by means of standard elements (title, headline, lead paragraph, body) that carry different meanings. News articles can also be given categories and topics they belong to, and they can contain external references to other articles or Web pages. Event descriptions within online encyclopedias, such as Wikipedia, possesses an extensive set of metadata as well as references to either internal articles or external resources that can be exploited. Furthermore, information concerning social interactions (e.g. friendships, followings, liking, popularity) becomes available when events are shared on Social Media in form of textual or visual content. This whole amount of information is simply not available when personal events are kept private within forms of local storage. Or more precisely, it is up to the willingness of the owners whether to invest effort in organizing and enriching their event-related data with additional information. In case of pictures taken during an event, for instance, one could assign a descriptive title to each of them, annotate the people appearing in them, an so forth. Unfortunately, little time seems to be dedicated to activities like these [KSRW06, RW03]. The lack of such extra knowledge represents a further challenge, as automatic approaches for processing personal events have to infer any information from the rough data.

1.1.2 Type of Description

There are several ways of capturing and documenting events, the most common ones being writing textual descriptions, shooting pictures or videos, recording sounds. They offer complementary perspectives of the same event and also pose very different challenges regarding their automatic processing and mining. In this thesis we handle two types of event descriptions, namely text and images. The choice of considering textual data is nearly self-explanatory, since it is the most traditional form of spreading information in both the physical and digital world. Pictures are also common means of communication, which have become especially pervasive in the last decades thanks to the diffusion of digital devices, such as cameras and smartphones, along with software to manage and share them [MW13]. Images are intuitive, easy to grasp, and informative: as a popular idiom says, “a picture is worth a thousand words”. While videos are largely used to document events as well, we chose to focus on images for two reasons: first, the latter seem to be used more frequently in personal settings [KSFS05, Dug13], possibly because their creation, visualization and revisiting take less time; second, many of the approaches developed for processing images might be reused when handling videos.

While this dimension is in principle independent from the previous one, as events with any possible combination of privacy degree and description type can be observed, there are still a couple of them that are more common than the others. In the public sphere, events in the Web are nowadays reported in a multimedia fashion, where traditional textual descriptions are accompanied by pictures and videos. However, we believe legit to regard text still as the primary form in which public events are described, for different reasons. First, the textual descriptions of events reported in news articles (or Wikipedia articles) are self-contained and can stand on their own, with accompanying pictures or videos being a useful yet not indispensable supplement. The opposite does not usually hold, as the visual content within a news article can be hardly interpreted in the proper way without the full textual event description or, at least, a summarizing caption. Exceptions are the “TV-like” videos with a narrating voice providing facts on the reported event, but such audio can be regarded as a summarized oral version of the article’s content. Second, textual event descriptions are available in several and diverse information sources: not only news articles, but also encyclopedias (e.g. Wikipedia), Blogs, Knowledge Bases (e.g. DBpedia [ABK+07], YAGO [HSBW12]).

In contrast, in the private sphere, the spread of relatively cheap digital devices fostered the process of taking pictures and videos of both memorable events and everyday life. This made them become the uncontested form in which personal events are captured, with a special preference for images with respect to videos, at the expenses of textual descriptions such as personal diaries. Text is clearly used when planning events, e.g. by emails or (instant) messaging, but not really to describe what happened during them. This does not hold for more social events, whose related information shared on Social Media encompasses both images and short texts. Also Blogs might contain descriptions of events attended by the author, and to some extent they can be regarded as a social equivalent of the above mentioned personal diaries. Nevertheless, when considering Social Media and Blogs, we already moved out of the strictly private sphere as the events described there become visible to a broader audience.

1.1.3 Time of Usage

Time is certainly fundamental for events and this is indeed one of the few aspects of consensus among the existing event definitions. As a consequence, the temporal dimension is spotlighted and explicitly taken into account within most of (if not all) the approaches handling event-related information, no matter what their particular goals are. This thesis is no exception as we do encompass time in our basic definition of an event discussed at the beginning of this section. However, we go beyond that by adding an additional outer dimension representing the time interval between when an event happened and when it is accessed and used. The width of such temporal gap changes based on the objectives of the usage scenario, and it also influences the kind of issues to be taken into account. For instance, long intervals in the order of decades might introduce evolutionary and forgetting aspects, which might not be significant or even observable in short-term usage.

One example of short-term usage of information, not restricted to events, is the process of sharing data either on Social Media or over more private means (e.g. emails, messaging, shared folders). In these situations there is an intention of letting other people know that something is happening or has recently happened, without necessarily caring about whether and when the shared data might be accessed again in the future. The mobile application Snapchat even poses ephemerality as one of its fundamental and distinctive principles, since it makes the shared data accessible only within a short time window (order of seconds). Moving to more public settings, news articles are commonly read to satisfy short-term information needs and to be up-to-date on current events and topics. Web pages of newspapers and media broadcasting are updated multiple times within a single day, making the same news article visible for a relatively short period (days). While such Web sites might make archives of the previously published news available, they are not explicitly brought to the attention of the readers anymore.

Let us now discuss examples where event-related information is subject to long-term usages. Remaining within the public sphere, news articles do not only satisfy short-term information needs as mentioned before, but they can also be accessed in the future by professionals conducting retrospective studies, e.g. journalists, historians, librarians or scholars. Also event descriptions contained in online encyclopedias, e.g. the Current Events portal of Wikipedia, are written to keep a documentation of the most significant events over time, which can be inspected at future time points. Such temporal gaps can harm the comprehension of past events, either reported in news articles or in event repositories, as it might require the awareness of old contextual information from the time when the events happened. Long-term usage is even more recurrent for personal events, in all those cases where they are captured (most commonly by pictures) for the sake of a future reminiscence instead of an

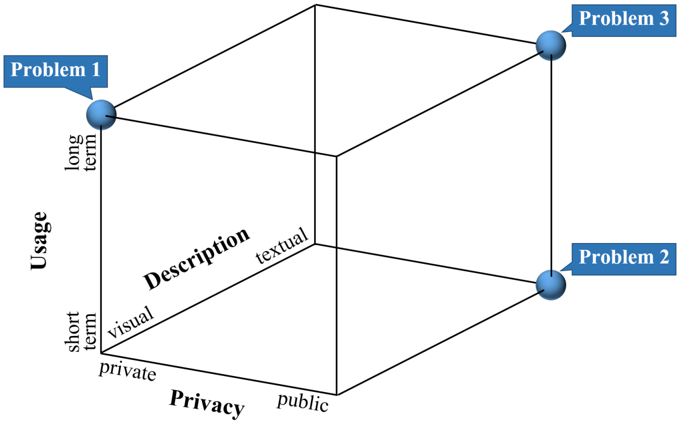


Figure 1.1 Main problems covered in the thesis, positioned with respect to three event hyper–dimensions.

immediate and evanescent sharing. Such need of preservation regards not only life's milestones but also more mundane moments of everyday life. One point that should be taken into account is that moving from immediate sharing to long-term preservation can alter the set of information items relevant for those purposes, i.e. one might want to share something (e.g. a picture) with friends without necessarily willing to keep it for decades, and vice versa. Furthermore, the common habit of merely keeping every single piece of data from an event, encouraged by the diffusion of relatively cheap devices to capture and store them [RW03, Mee16], might impose additional challenges regarding the management and access of such amounts of personal data accumulated over time.

1.2 Scope of the Thesis

We now state the three main problems covered in this thesis, which are related to the management, validation, and retrieval of event-related information, respectively. We also discuss how they are positioned with respect to the three event hyper-dimensions introduced before. Visually, this is illustrated in Figure 1.1.

Problem 1 *How to identify, within large photo collections depicting personal events, those pictures that the collection owners would like to keep for future reminiscence?*

The amount of digital pictures taken to document memorable moments as well as the daily life has dramatically increased in the last decade, thanks to the joint diffusion of digital photography and cheap storage devices. People can easily take hundreds of pictures during one single event, such as a vacation or a business trip. Specifically for this scenario, we regard each personal event as a photo collection, which provides a visual description of it and satisfies our basic event definition by encompassing both temporal information (as metadata) and participants (people appearing in the pictures). While the possibility of storing literally every bit of data is perceived as a blessing at first, as it allows to circumvent the onerous sorting and pruning procedures, it turns into a curse when the stored photo collections are revisited in the future. In fact, going through even a single collection of hundreds of photos to enjoy the most salient ones can be such a demanding procedure to discourage from doing it at all.

This suggests to be more selective in the first place and identify which photos the collection owner would like to keep for future reminiscence of the related event. This does not imply that the other photos should be deleted immediately, but rather that the selected ones would receive a particular spotlight as concise subset of important photos that is more enjoyable and easier to revisit than the whole collection. Also, to cope with possible storage breakdowns and format changes over years while reducing the money to be invested for it, additional preservation procedures and services could be dedicated only to the set of selected photos. However, choosing what would be best to have in future and potentially different contexts is not trivial even for humans, who would just tend to keep everything. This is exacerbated by the personal nature of this scenario, where memories and affective attachment could influence the perceived importance of each photo as previously discussed in Section 1.1.1.

In Chapter 2 we investigate this problem and propose an approach for selecting important photos from personal collections with the purposes of a future preservation and revisiting. Referring to Figure 1.1, Problem 1 falls into the space defined by the three previously discussed hyper-dimensions as follows. Events in this regard belong to the *private* life of individuals, involving a relatively small circle of people and not being necessarily meant to be shared with a broader audience, and are described in terms of *visual* data (photographs). This particular combination has been already discussed in Section 1.1.2. Finally, the event-related photo collections are subject to a *long-term* usage, because they are expected to be accessed in the future and to provide reminiscence of past personal events.

Problem 2 *How to validate the occurrence of events in the real world from the information contained in textual document collections?*

Events shape the world continuously and pervasively, therefore they are reported and discussed in many forms of media, such as news articles, blogs, social media. Such event-related information is valuable for various purposes: supporting the work of professionals such as journalist, librarians, historians; satisfying the curiosity and interest of general readers; serving as input for other IR tasks like indexing, ranking, document enriching; signaling emergency situations, e.g. natural disasters or civil disorders, and triggering timely reactions of the responsible parties. For these reasons, there has been a large interest and research effort in detecting real-world events from textual sources.

The problem that we address in this scenario does not concern how events are detected from text, but rather how to validate their verity and occurrence. In fact, errors introduced in the detection phase might jeopardize the proper exploitation of the extracted event-related information for any of the aforementioned purposes. While one arguable remedy would consist in enhancing event detection in the first place, our intuition is that making the detected events undergo a further validation step, which exploits information that was unavailable as input to event detection, could eventually improve the quality of the detected events. Furthermore, event validation can be used by end users to verify the verity of manually specified events and, in case they did happen, to find information that describes and corroborates their occurrence.

Verifying whether an event has actually happened is a cumbersome process, as it requires the inspection of a potentially large and cluttered source of trustworthy information (generally different from the data where events have been detected), which serves as a ground truth for seeking evidences of the given event. Besides checking for the mere appearance of the event participants, automatic methods should also assess whether they participated together within the same event as well as whether they did so during the time when the event has been declared to happen. In Chapter 3 we define the problem of event validation and propose methods to solve it automatically.

Let us locate Problem 2 into the 3-dimensional space shown in Figure 1.1. The events that we consider are of *public* domain and span different topics and degrees of newsworthiness. Applying event validation in more private settings would encounter many issues such as, for instance, the need of personal event descriptions to be used as a ground truth. Instead, in case of publicly known events, the validation can draw information from content published online. Both events and the ground truth are represented in *textual* form as it is the prevalent means for referring to public events (see Section 1.1.2), although the problem of data verification in the multimedia domain exist as well (see e.g. [BPK+14, BPAK17]). While the actual time of usage depends on the purpose for which the validated events are used, we categorize this problem under the *short-term* usage for different reasons. First, since events are mostly reported and discussed while they are happening and shortly afterwards, validating an event at time points far from its expected occurrence would not gain from the presence of much more related documents in the ground truth. Second, many of the envisioned applications mentioned before do exploit the validated events in a short-term, for instance acquiring signals of emergency situations and planning reactions accordingly, or enriching news articles with related event descriptions to better satisfy the information needs of the readers. Third, also for those scenarios involving long-term usage such as maintaining event repositories for retrospective inspections, the core formulation of event validation does not explicitly address any effect that might arise due to long-term dynamics, which are expected to be handled by the given application using the validated events.

Problem 3 *How to retrieve concise information related to old events, as a basis for their proper understanding in presence of wide temporal and contextual gaps?*

Properly understanding an event always requires a certain amount of background knowledge, for instance regarding its main topic and participants. This fact is exacerbated in presence of old events, as explained hereafter. Let us suppose that an event, either reported in form of news article or described within encyclopedic event repositories (e.g. Wikipedia's Current Events), is read after a relatively long time period, at least few decades. Readers might encounter difficulties in properly understanding such an old event description as it assumes background knowledge from the time when it has been written. Such information was considered as part of the generally known context at that time, and therefore it was not explicitly stated. However, this original context has evolved over time and eventually has turned into the current one, potentially very different in case of large time gaps, with the result that the original context might have been forgotten or never known at all.

The temporal and contextual gaps emerging in this scenario call for the provision of additional information, which should be temporally related to the time of the event and also relevant to its content, to support the proper understanding of the event itself. The retrieval of such event-related information is affected by several issues from its very first stages, such as understanding what parts of the event description should be used as seed for querying an available context source, and how to combine them together in actual queries. It is also desirable to retrieve candidate contextual information with high recall, so that the result set can serve as versatile starting point for subsequent re-ranking phases based on additional objectives. In fact, these would benefit from handling a bigger amount of temporally and topically relevant results in the first place. This problem will be investigated in Chapter 4.

Referring to the event categorization based on our three hyper-dimensions, the large amount of historical events contained in news archives and encyclopedias as well as the need for a collective awareness of the past made us focus on *public* events. However, the process of retrieving contextualizing information for old events is also relevant to private settings and could be indeed tackled in a similar way as long as a personal context source is available. This is unfortunately not as common as for the public domain, where many knowledge bases and encyclopedias are maintained over time and can be leveraged as sources of context. As already discussed in Section 1.1.2, we consider *textual* representations as they are prevalent for public events, especially for the old ones, but the inclusion of visual data for representing either input events or the context source could be possible as well. Finally, Problem 3 involves a *long-term* usage of events because they are expected to be accessed and understood at future points in time, and the retrieved event-related information should ensure their proper understanding in presence of potentially big temporal and contextual gaps.

1.3 Contributions of the Thesis

In this thesis we tackle the problems presented before and achieve three major contributions, which are summarized hereafter.

- Regarding the long-term management of personal photo collections characterizing Problem 1, we present a selective approach that aims at identifying those photos that are perceived as most important by the collection owner for the purposes of long-term preservation and revisiting. Having such a concise set of highly important pictures makes the future revisiting more enjoyable and less tedious. In order to achieve this, we consider a multifaceted notion of image importance driven by user expectations, which represents what photos users perceive as important and would select. User expectations have been acquired during a user study, asking participants to provide their own photo collections and to select those most important to them for preservation and revisiting. The presented approach exploits supervised Machine Learning and an extensive set of visual and semantic information, extracted solely from the image content without expecting any external or manually provided data, to estimate such notion of long-term image importance and to select subsets of important pictures based on it. Since a wide part of state of the art methods is driven by the concept of coverage, we also investigate how to combine our expectation-oriented selection with more explicit modelings of coverage.

- Concerning Problem 2, we develop methods for validating the verity of real-world events based on their occurrences in a corpus of textual documents, which is used as ground truth. First, for a given event to be validated, we query the ground truth to retrieve a set of candidate documents that match the input event to some extent. Second, for each retrieved document, we check whether the event participants truly take part to the same event in the document and strictly within the timespan of the event. Our best performing approach carries out this task using supervised Machine Learning and exhibits a substantial level of agreement when compared with human evaluators. Thanks to the generality of the problem definition and the lack of assumptions regarding the handled data, our approaches stay suitable to different kinds of event representations and document corpora. We also investigate the insertion of event validation as a post-processing step of event detection, which results in an increase of precision without affecting recall, i.e. wrongly detected events are discarded while retaining the true ones. Finally, we describe and release a benchmark for event validation to foster future researching on this task.

- In order to cope with wide temporal and contextual gaps threatening the understanding of old events, a central matter of Problem 3, we present methods to formulate queries from event descriptions and retrieve contextualizing information from a given document corpus. The most effective query formulation method learns to combine parts of a given event description into queries by means of recall-based Query Performance Prediction. The features utilized within this model take into account both content-wise relevance and temporal proximity between retrieved results and input events, as these are necessary characteristics that the results should exhibit for being good contextualization candidates. Furthermore, our approach explicitly optimizes recall by considering it as query performance criterion, which consequently makes queries that produce high recall more likely to be formulated. Maximizing the recall within the set of retrieved results makes our query formulation more versatile and suitable for pursuing additional objectives at different subsequent computations (e.g. re-ranking), since any of them would have the advantage of handling a bigger quantity of topically and temporally relevant results in the first place.

2 Managing Memories from Personal Events

Thanks to the spread of digital photography and available devices, taking pictures has become effortless and tolerated nearly everywhere. This makes people easily ending up with hundreds or thousands of photos, for example, when returning from a holiday trip or taking part in ceremonies, concerts, and other events. Furthermore, photos are also taken of more mundane motives, such as food and aspects of everyday life, further increasing the number of photos to be dealt with. The decreased prices of storage devices make dumping the whole set of images common and affordable. However, this practice frequently makes the stored collections a kind of "dark archives", which are rarely accessed and enjoyed again in the future. The big size of the collections makes revisiting them time demanding.

This suggests to identify, with the support of automated methods, the sets of most important photos within the whole collections and to invest some effort into keeping them accessible. Evaluating the importance of pictures to their owner is a complex process, which is often driven by personal attachment, memories behind the content and personal tastes that are difficult to capture automatically. Therefore, to better understand the selection process for photo preservation and revisiting, the first part of this chapter presents a user study on a photo selection task where participants selected subsets of most important pictures from their own collections.

In the second part of this chapter, we present methods to automatically select important photos from personal collections, in light of the insights emerged from the user study. We model a notion of image importance driven by user expectations, which represents what photos users perceive as important and would have selected. We present an expectation-oriented method for photo selection, where information at both image- and collection-level is considered to predict the importance of photos.

2.1 Introduction

Photos are excellent means for keeping and refreshing memories, they can illustrate situations we have gone through and serve as memory cues [ME00, ME07] to bring back reminiscences of experiences, events, and people from our past [RW03]. In the recent years we have been witnessing a huge increase in the production of pictures, mostly due to the wide spread of digital devices such as cameras, smartphones, tablets. People easily take hundreds or even thousands of pictures during relatively short and memorable events, e.g. vacations, ceremonies, concerts, or depicting more mundane aspects of everyday life [MHE14], like shopping, eating, working, free time. These numbers can amount to Terabytes of data over years. Considering only the pictures that are shared on social media like Flickr, Facebook, Instagram and Snapchat, a study conducted in 2013 [MW13] reported that around 500 million images (most of them from personal photo collections) were uploaded to the Internet every day, and this amount has nearly doubled in each of the subsequent years [Mee16].

This scenario points out the significance of properly dealing with such increasing volumes of pictures. Due to decreasing storage prices [Mee16] and offers of cloud storage services, e.g. by Microsoft or Google, it is not a problem to store personal photos somewhere. As a matter of fact, directly dumping photo collections spending little or even no time in activities like pruning, editing, sorting, or naming has become a popular procedure [KSRW06, RW03]. This comes at a price: storage devices tend to become a kind of "dark archives" [LD04] of photo collections, which means that the stored pictures, although still available, are rarely accessed and revisited again in the future. The big size of the stored collections makes going through them such a tedious activity to prevent the viewers from accessing them at all. As an additional challenge, there is the risk of losing photos by a random form of "digital forgetting" [KNS13]: over decades storage devices break down, and formats and storage media become obsolete, making random parts of photo collections inaccessible (Digital Obsolescence [SVDHG11]). One example is how difficult it would be today to access photos stored years ago in .mos format in a floppy disk.

Both the threats of personal dark archives and digital forgetting raise the following question: how can photos be kept enjoyable and serve their original purpose as memory cues, where large photo collections tend to get dumped on hard disks and other types of storage? We propose a transition from dumped contents to more selective personal digital memories, supported by automatic methods for information value assessment, to support long-term personal data management. Regarding photo collections, this means identifying the most important pictures from an entire collection and investing some effort to keep them accessible and enjoyable on the long run. Having a reduced sub set of important photos would make the revisiting easier and more pleasant for the user. However, understanding the importance of pictures to their owner for preservation and revisiting purposes is a complex process due to the presence of hidden factors, which are hard to model and capture automatically. These can be, for instance, memories, context, relationships to whom is in the picture, or simply personal tastes [KSFS05, WSS14].

Therefore, the first part of this chapter summarizes a user study for a photo selection task where participants were asked to provide their personal photo collections and to select the subsets of photos that they would want to preserve and revisit again in the future. The study involved 35 participants, each one contributing at least one personal collection containing some hundreds of pictures. The goal of this part is to better understand the human selection process for photo preservation and revisiting, identifying insights, patterns, and challenges that can shape the development of automatic selection approaches. Moreover, the gathered data will be employed for the development and evaluation of automatic selection methods. The user study was complemented by a survey, which we asked the participants to fill after completing the photo selection task.

In the second part of this chapter, we present methods to automatically select important photos from personal collections, inspired by the insights emerged from the user study. Many approaches to photo selection for summarization are centered around the concept of coverage, aiming at creating summaries that resemble the original collection as much as possible. However, we believe that the complex decision making behind the selection of photos from personal collections, characterized by personal attachment due to memories, might reduce the importance of coverage. Therefore, we model a notion of image importance driven by user expectations, which represents what photos users perceive as important and would have selected. We present an expectation-oriented method for photo selection, where information at both image- and collection-level (incorporating a relaxed notion of coverage) is considered to predict the importance of photos. We also investigate the role of coverage further by combining it with the expectation-oriented selection in different ways, showing that coverage plays only a secondary role in this task.

This chapter contains four major contributions: (i) we conduct a user study on the photo selection process for the purposes of long-term preservation and revisiting; (ii) we present a low-investment method, driven by user expectations, for selecting important photos from personal photo collections for that task; (iii) we study the role of coverage in a systematic way by combining our expectation-oriented selection with an explicit modeling of coverage in different ways, showing that comparable results to our method can be achieved only when coverage is not considered as a primary selection aspect; (iv) we explore further directions for improvements, namely user personalization and the inclusion of additional features into the selection model.

The rest of the chapter is structured as follows. In Section 2.2, we outline related works and current approaches to photo selection. Section 2.3 describes the user study while the selection methods are presented and compared in Section 2.4. Sections 2.5 and 2.6 contain further investigations on personalization and inclusion of additional features in the model, respectively. Finally, in Section 2.7 we summarize and conclude the chapter.

2.2 Related Work

The discussion of the works related to personal photo selection and preservation is organized in three parts. The first one is on qualitative user studies, the second part reviews automatic approaches for photo selection and summarization tasks, the last one gives an overview on the image processing techniques employed in this chapter.

2.2.1 User Studies and Surveys

A considerable research effort has been dedicated to investigate issues related to photo management and preservation from a Human-Computer Interaction perspective [Cop14, KSRW06, WSS14, WNL14, WNR+15]. Kirk et al. [KSRW06] introduced the notion of *photowork* as the set of activities performed with digital pictures after capturing them and before any end usage like sharing or revisiting. One of their findings is that people spend little time in activities like reviewing, pruning, editing, sorting, as these are cumbersome and time consuming procedures. This fact clearly supports the topic and goal of this chapter. In the context of preservation of public images, [Cop14] reports a qualitative study assessing their value for representing social history. The evaluators were asked to rate five images from Flickr based on their worthiness for long-term preservation. This study is limited in (i) not considering personal photos and (ii) the small data considered. Interestingly, the participants expressed a clear inclination to preserve all the pictures irrespective of their actual value. The authors hypothesized two possible reasons for this, namely the difficulty of anticipating a future information need and the effort required for organizing and pruning increasing amounts of data. In any case, they recognized this as a problem and pointed to the need of methodologies for information appraisal and selection.

Wolters et al. [WNL14] investigated which photos from an event people tend to delete over time. In this study, the participants took pictures during a common event and then they were asked for deletion decisions at different points in time. While this work is certainly related to our study, which drew inspiration from it especially regarding the formulation of the survey, there are nevertheless some differences. Despite preservation ("keep") and "delete" decisions are related, we explicitly asked our evaluators to make selection decisions for the purposes of preservation and revisiting of images, rather than for deletion. Moreover, in our study the users were asked to make joint selection decisions (i.e. select a sub-collection) instead of making decisions for each individual picture in isolation. This is potentially a key difference, since selecting one photo might affect the decisions for other similar photos. Finally, instead of taking pictures of a common event explicitly for the study, we work with personal real-world collections belonging to diversified events. A subsequent work by Wolters et al. [WNR+15] presented a large-scale survey of 72 young people and students, with the goal of supporting the design of personal and mobile preservation systems. The main message of the study is coherent with what emerged from our survey: a large part of the participants acknowledge the importance of preserving pictures for future generations. Interestingly however, only a small fraction of them carries out practices to support photo management and preservation. In another user study [WSS14], participants wearing eye tracking devices were asked to select subsets of pictures from two collections depicting two social events. This work focuses more on the selection process than on preservation matters. The survey on the aspects driving the selection process shares with our experiment both similarities (e.g. most of the highly rated aspects were subjective) and differences (e.g. quality was highly rated there).

2.2.2 Photo Selection and Summarization

Automated photo selection has already been studied in various other contexts, such as, photo summarization [LLT03, SBS14, SMJ11, TIWB14, WLS+16], identification of appealing photos based on quality and aesthetics [LLC10, YHBO10], selection of representative photos [CL08, WSS14], and the creation of photo books from social media content [RSB10]. We consider the task of selecting important photos from personal collections (e.g. for revisiting or preservation), which meet user expectations.

The work of Wang et al. [WLS+16] is very related to ours, as their model of image importance does not explicitly include coverage and diversity. They introduce the notion of *event–specific image importance*, meaning that the importance of images for selection purposes depends on the category of the event they belong to. The intuition is that, within a collection depicting a certain type of event, the set of images commonly perceived as important can be identified based on the event type. There are, however, substantial differences regarding the task definition and the ground truth. First, the ground truth was not gathered considering photo selection, since ratings were assigned by the evaluators to each image in isolation without explicitly deciding what subset of the collection should be kept. Second, individuals different than the collection owner rated the importance of images, potentially ignoring any personal attachment due to memories or hidden context. The influence of the event type on image importance remains an interesting aspect to be explored within our context of personal photo selection. In this regard, employing works on event recognition [BGVG13, DDNDN14, ACBDN17] and "event saliency" [RBDN15, ACDN18] would provide useful information on the event type and on where event-related information lies within the images. Image importance has been also considered in [LLC10, YHBO10], but based on quality and aesthetic criteria. Instead, we explicitly consider selections preferences and expectations of users both for training and as evaluation criterion. Walber et al. [WSS14] also consider human judgments to evaluate selections, but the users have to wear eye trackers when using the system to make automatic selections since gaze information is used as features in the model.

Different photo selection and summarization works consider coverage by identifying clusters of images based on time and visual content [CL08, LLT03, RSB10]. Differently, our approach does not impose such a strict notion of coverage but rather considers clusters and other global information together with image-level information, learning their different impact in a single model. The works in [MBK16, SBS14, SMJ11, TIWB14] are closer to ours, as they consider coverage in a relaxed way as part of a multi-goal optimization, but they still consider coverage as a key component. Moreover, [SBS14, SMJ11] do not consider user assessments in their evaluation and make partial use of manually created text to associate semantic descriptors to images, while our method does not require any manual input, once the models for both feature extraction and importance estimation have been learned. Image collection summarization is performed in [MBK16, TIWB14] by applying structured prediction methods for learning weighted mixtures of submodular functions. The attention is drawn to two aspects that good summaries should exhibit, namely fidelity (coverage) and diversity, which are represented as a set of non-negative submodular functions and combined together in a single weighted submodular scoring function. There are two main differences with respect to the work presented in this chapter. First, their goal is purely summarization, aiming at optimizing coverage and diversity of output summaries, without considering whether they contain the most valuable pictures. This is strengthened by the utilization of the recall-based V-ROUGE metric (a criterion for summary evaluation inspired by the ROUGE metric [Lin04], used for document summarization) within the loss function. Second, the way the ground truth was collected is heavily oriented towards coverage: the evaluators, not the owners of the collections, were told to produce reference summaries that summarize the original collections *in the best possible way*, and those exhibiting low coverage were discarded. Conversely, we asked the collection owners to select the most important pictures according to their memories and perceptions, without any mention to coverage or diversity.

Besides [SBS14, SMJ11], other works rely on external knowledge to accomplish the summarization task [CG16a, SM16, YDYX14]. Camargo et al. [CG16a] combine textual and visual contents of a collection in the same latent semantic space, using Convex Non-Negative Matrix Factorization, to generate multimodal summaries. Domain-specific ontologies are required as input in [SM16]. They provide knowledge about the concepts in a domain and are used to derive a set of ontology-based features for measuring the semantic similarity between images. Finally, [YDYX14] jointly leverages image content and associated tags and encodes the selection of images in two vectors, for the visual and textual domain respectively, whose non-zero elements represent the images to be included in the summary. The optimization process makes use of a similarity-inducing regularizer imposed on the two vectors to encourage the summary images to be representative in both visual and textual domains.

Summarizing, our approach is different from all the previous works under at least one of the following aspects: (i) our notion of image importance is based on selection decisions made by people on their own photo collections; (ii) we do not estimate image importance using single indicators (e.g., quality, presence of faces, representativeness of the cluster a photo belongs to), but we rather combine these aspects and learn their impact; (iii) we use selection decision made by the collection owners themselves as ground truth for evaluation; (iv) we do not rely on any kind of photo tagging or descriptive annotation provided manually.

2.2.3 Image Processing

We finally provide a brief survey on existing approaches related to the image processing techniques employed in this chapter.

Concept Detection. The purpose of concept detection is to analyze the visual content of an image and automatically assign semantic concept labels to describe it. Many methods for concept detection follow a processing pipeline that involves (i) selecting specific locations on the image grid where features should be computed and extract at these locations local descriptors such as SIFT, SURF, and others [Low04, vdSGS10], (ii) building a global image representation from the local features using BoW, VLAD, Fisher vectors [AZ13], (iii) using such representations of ground-truth-annotated training corpora to train concept detectors based on Machine Learning (e.g. SVMs). Recently, Deep Convolutional Neural Networks (DCNNs) have been employed to learn feature representations directly from the raw image pixels. The different layers within a DCNN are able to learn image representations at different levels of abstraction and in an unsupervised manner, thus dramatically alleviating the effort of feature engineering. Different network architectures have been proposed in the last few years, e.g. AlexNet [KSH12], GoogLeNet [SLJ+15], VGG ConvNet [SZ14], which can be used both as standalone classifiers for concept detection and as feature generators for other models (e.g. SVMs, Logistic Regression, Random Trees).

Clustering. Several general-purpose clustering methods (e.g. k–means [Llo82] or hierarchical clustering [Joh67]) can be used in combination with visual features (e.g. HSV histograms) or concept vectors, as done in [PM14]. Since time is the dominant dimension for event clustering, several time-based clustering methods have been presented in the literature [CFGW05, DDN14]. They can incorporate other data dimensions, e.g. geolocation and visual information (either low level similarity or high level concept scores). Geolocation information can be used either to merge two subsequent sub-events if they are spatially close, or to split a cluster into more new sub-events if the locations of the images are spatially diverse. Visual content can also serve as complementary information if geolocation information is unavailable. In case of multiple collections created with different devices, the images may refer to different temporal offsets and a synchronization is therefore required before being able to cluster them. Recent advances in this regard have been presented in [SAC+17], which synchronizes multiple collections based on the visual similarity of their pictures.

Near–duplicate Detection. Near–duplicate pictures depict slightly different variations of the same scene. Many approaches have been proposed for near-duplicate detection, such as employing multi-resolution histograms [WLC07] or aggregating local descriptors into global representations [JPD+12]. More recently, Wang et al. [WSL+14] proposed Deep Learning techniques to learn a similarity metric that can be used for near–duplicate detection.

Face Detection and Clustering. One of the most popular and successful approaches to face detection is the Haar-like-feature-based detector introduced in [VJ04].

It has been widely extended over the years by means of color–based segmentation combining HSV and RGB color spaces [RB13], skin face color segmentation [RA12], rotated features [LM02], locally assembled binary features [YSCG08], and other approaches. Face Clustering is often performed by computing similarity matrices based on different facial representations, e.g. SIFT features [ANP07], gradient and pixel intensity [HYL+03], contextual and co-occurrence information [ZTL+06], local binary patterns and clothing colors [CWX+07], and then applying standard clustering algorithms (often spectral or hierarchical) on such matrices. Deep Learning has also brought advances in both Face Detection [SCWT14, PVZ+15, YLLT15] and Face Clustering [SKP15, OWJ17].

Quality and Aesthetics. The visual quality of an image can be quantified by looking at the presence of artifacts such as blur, noise, and low contrast. The available techniques for image quality assessment can be divided into full-reference, reduced-reference, and no-reference image quality assessment, depending on the amount of information provided regarding the undistorted image [WWS+06]. A variety of no–reference techniques have been proposed for detecting quality degradations such as image blur, e.g. in [MM14]. Image Aesthetics is a more complex and subjective measure, which tells how much an image is perceived as appealing and pleasant. Some of the existing approaches assess aesthetics purely based on low–level features, such as HSV distributions, textures, color distributions [TLZ+04, WBT10, DJLW06], while others also include more complex composition rules like object positioning, simplicity, balance, rule of thirds [DOB11, YHBO10, MM15]

2.3 User Study

As a preliminary step towards the development of automatic methods, we describe a user study conducted on a photo selection task, whose objective is the gathering of insights, challenges, and behaviors exhibited by humans when selecting personal photos for preservation and revisiting purposes. Using their own photo collections depicting personal events, participants were asked to select a subset of pictures that they would like to stay accessible and enjoyable in the future. Such data, i.e. the whole collections along with the selections done by the users, have been used for the training and evaluation of the selection methods described in Sections 2.4, 2.5 and 2.6. Upon completion of the task, the participants were also asked to fill a survey about it, which is reported in Section 2.3.2.

2.3.1 Task Setup

The setup of the performed photo selection task involves the gathering of both participants and their photo collections, instructions on how the task should be accomplished, and, of course, the development of a software application to perform the selection in a comfortable way.

Participants. The experiment involved 35 users (28.6% females and 71.4% males) with 15 different nationalities: 25.7% of the participants came from Greece, 17.1% from Germany, 11.4% from Italy, 11.4% from China, 5.7% from Vietnam, and the rest from Ethiopia, Turkey, Kosovo, Iran, UK, Thailand, Sweden, Brazil, Albania, and Georgia. Regarding their ages, 60.0% of the participants are between 20 and 30 years, 25.7% between 30 and 40, 11.4% between 40 and 50, 2.9% between 50 and 60.

Photo Collections. Previous works mostly consider either public photo collections, for instance available on social media like Facebook and Flickr [CG16a, RSB10, SM16, WLS+16], or pictures from a shared event in which all the evaluators took part [WSS14]. One difficulty we see with using public collections of photos from different people, even if they attended the same event, is that according to the different experiences of the individuals in the event they might also have a different level of appreciation for the same photo, thus influencing their decisions. In contrast, we use personal photo collections. For instance, these can be photos from business trips, vacations, ceremonies, or other personal events the evaluator participated in. This means that each collection is not just a bunch of pictures, which might exhibit different degrees of quality and aesthetics, but there are experiences, sub-events, and memories that might influence the selection behavior. We decided to focus on such personal collections because we wanted to observe the personal photo selection decisions in a setting that is as realistic as possible. In total, 39 collections were used in the experiment (four users evaluated two collections), resulting in 8,528 images. The size of collections ranges between 100 and 625 images, with an average size of 219 and a standard deviation of 128.7. These collection sizes also emphasize the need for automated selection support, since manually browsing for photo selection becomes time-consuming. We asked participants for further information about their collections, such as, the main topic of the collection, whether they were previously pruned (e.g. by discarding low quality images), and when the photos were taken. Overall, 51% of the collections represent vacations, 30% business trips, and 19% other events like music festivals and graduation ceremonies. In addition, 23% of the collections were already pruned before the evaluation. The time when the collections were taken spans from 2007 to 2014 (64% in 2013-2014, 17% in 2011-2012, the rest in 2007-2010).

Task Definition. Since our task of selecting photos for preservation is not an everyday task for the users, it was important to find a good metaphor for supporting the task. After discussing a number of options with cognitive experts, we decided to use the metaphor of a "magic digital vault", which incorporates the ideas of protection, durability, and a sort of advanced technologies to keep things accessible in the long-term. Therefore, the task consisted in selecting a subset of valuable photos to be put in the magic digital vault, which would protect the images against loss and would ensure that they remain readable and accessible over the next decades.

User Interface. To perform the photo selection task, we developed a desktop application, which enabled the participants to import their own photo collections and to select pictures in a comfortable way. It is depicted in Figure 2.1, where the

Photos already selected for preservation by the user for this photo collection

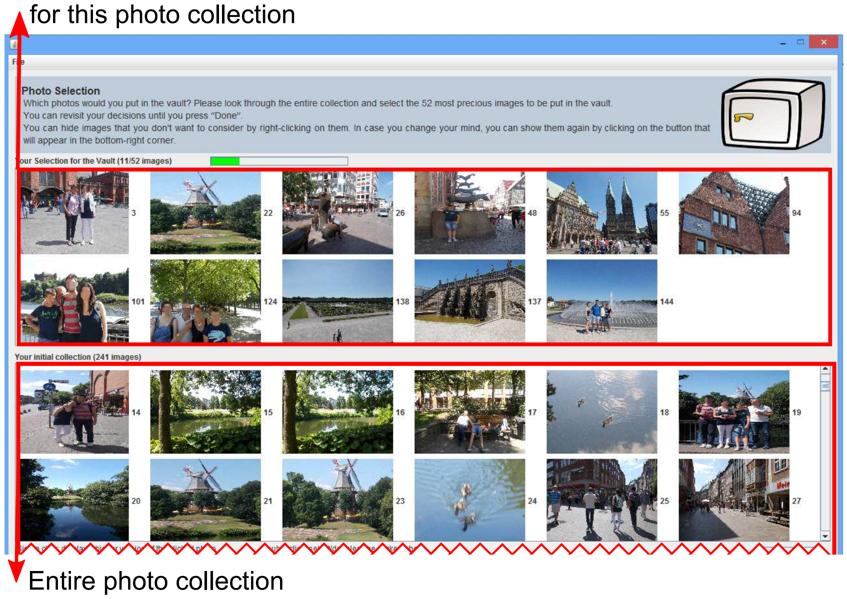


Photo Selection

Which photos would you put in the vault? Please look through the entire collection and select the 52 most precious images to be put in the vault.

You can revisit your decisions until you press "Done".

You can hide images that you don't want to consider by right-clicking on them. In case you change your mind, you can show them again by clicking on the button that will appear in the bottom-right corner.

File
our Selection for the Vault (11/52 images)
3
22
26
48
55
94
101
124
138
137
144
Your initial collection (241 images)
14
15
16
17
18
19
20
21
23
24
25
27

Entire photo collection

Figure 2.1 GUI for browsing collections and selecting the photos to preserve.

images contained in the imported collection are displayed in the bottom panel, while the ones selected are shown in the top panel. Faces appearing in Figure 2.1 have been blurred for the sake of privacy (only for inclusion in this thesis). The photos are selected and deselected by double-clicking on them, and they can be enlarged to inspect them better and appreciate their quality, although no explicit reference to the quality aspect was made in our instructions to the users. The images in the collection were shown in the same order in which they were taken because this makes the browsing, remembering, and selection easier for the users [GGMPW02]. We checked that keeping the original order did not introduce any significant bias in the selection in favor of the early photos in the collection. This could have been a risk, since users might lose attention or even complete the task without parsing the entire collection.

User Evaluation Methodology. Before starting the evaluation, the users were personally introduced to the photo selection task as well as to the application that they were asked to use. Further remarks and clarifications about both the task and the usage of the application were given, where needed. However, no guidelines were given about the criteria to use for selection, in order not to influence the selection process. After the users imported their collections, the application asked them to select 20% of photos from them for preservation and revisiting purposes. This selection percentage (20%) has been empirically identified as a reasonable amount of representative photos, after a discussion with a subset of users before the study. We also checked the adequacy of this chosen amount with the users in the survey by asking them whether they would have selected more photos if they could: 45% of them answered yes, the rest no. This balance means that 20% was a meaningful threshold, neither being to low (the majority of the users would have answered yes in this case) or too high.

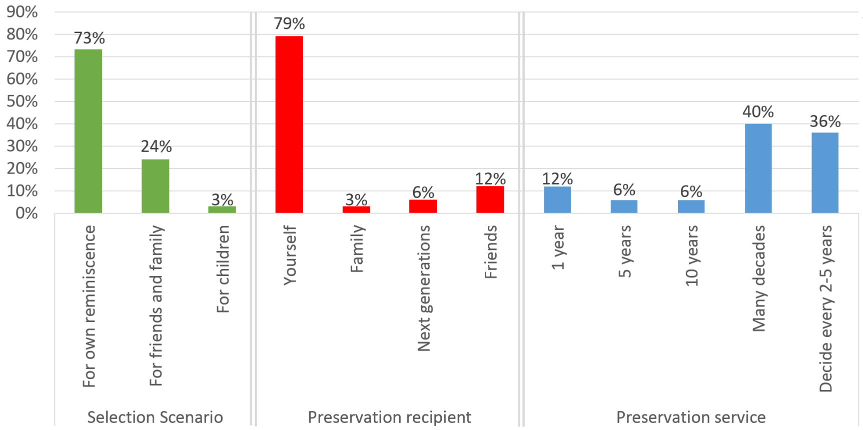


Figure 2.2 Survey results with respect to preservation scenario, preservation target group, and preservation as a service.

2.3.2 Survey and Discussion

After the photo selection step, the users filled a survey organized into two groups of questions: one refers to the scenario of photo selection for personal preservation, while the second one looks into the criteria that were considered during the selection.

Regarding the first group of questions, the users were asked to provide information about (1) which scenario they had in mind when selecting the images; (2) for whom they are preserving the images; (3) whether they would be ready to pay, and for how many years, if preservation was a paid service. The answers to each question were posed as multiple choices and are reported in Figure 2.2. The answers to questions (1) and (2) reveal that the process of long-term preservation is centered around the owner of the photos: more than 70% of the evaluators said that they thought about own future reminiscence when they selected the photos, and almost 80% indicated themselves as a main consumer of the preservation outcome. Looking at the preservation as a valuable service to be paid (question (3)), the evaluators were mostly split into two groups: either being ready to pay for many decades (39%) or needing flexibility to make new preservation decisions every 2-5 years (36%). In both cases, these answers highlight a clear need for preservation of personal photo collections.

In the second group of questions, we suggested different photo selection criteria and asked the users to rate how much each criterion was considered during the selection. The suggested criteria, which are in line with the insights on “keep” and “delete” decisions in [WNL14], were rated via star ratings on a scale between 1 and 5 (5 stars mean very important, 1 means not important at all). The criteria along with statistics about their ratings are reported as box plots in Figure 2.3. Note, that medians are represented as horizontal bold bars, while sample mean is indicated with a bold cross. For the sake of clarity, we grouped the criteria into three classes: (1) *content-based criteria* refer to objective and subjective measures for individual

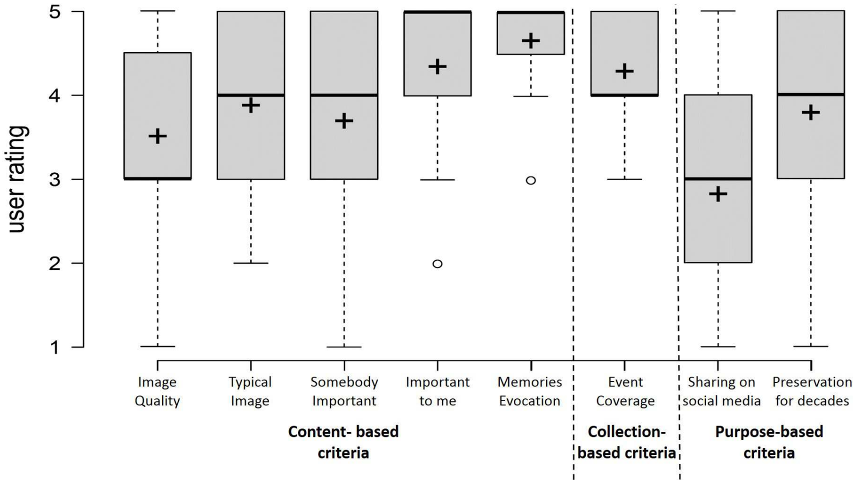


Figure 2.3 Boxplots of the different selection criteria.

images such as image quality, image typicality (i.e. how suitable it is for serving as an iconic summary of the event), the presence of important people in images, whether images are generally important, and the evocation of memories, (2) *collection-based criteria* - here represented by coverage of events - consider an image in the context of its collection, and (3) *purpose-based criteria*, indicating the importance of different selection goals (in our case, sharing and preservation).

An important finding of this evaluation is that the objective quality of photos is rated as the second least important selection criterion, after the sharing intent. This shows that quality and aesthetics, although being important and used for "general purpose" photo selection [LLC10], are not considered very important in case of selecting photos for preservation. In contrast, criteria more related to reminiscence, such as event coverage, typical image, and "the picture evokes memories" are all rated high, with highest ratings for memory evocation. The remaining two criteria "picture is important to me" and picture "shows somebody important" refer to the personal relationship to the picture and are also both rated high. These results anticipate that the task of predicting images to be selected for long-term preservation is likely to be difficult, since many of the criteria that are rated high, e.g. memory evocation, personal importance and "typical image", are difficult to assess for a machine, because they contain a high level of subjectivity. Another complicating fact is that there is no single dominant selection criterion, but a combination of highly rated criteria. In these ratings, there are differences with respect to those given to the partially overlapping set of criteria reported in [WSS14], where photos of shared events were used and the selection was not directly related to preservation and reminiscence. In that work, higher ratings are given to criteria such as quality, whereas event coverage and importance of depicted persons are rated relatively low (although with high variance). Interestingly, photos that capture memories are also rated high in this case.

2.3.3 Image Clustering and Human Selections

We see two main uncertainties in applying existing summarization methods for the task of photo selection for preservation. First, they are developed for other purposes, e.g. identifying sub-sets of photos that provide comprehensive summaries of the initial collections [RSB10, SMJ11, TIWB14]. Second, they often do not compare the performances of their output with selections done by users, or, when they do, they consider judgments based on more objective criteria such as aesthetics [LLC10, YHBO10]. Since a wide part of the state of the art methods for photo selection and summarization considers clustering and/or coverage for generating selections and summaries (as discussed in Section 2.2.2), we clustered photos by applying the event-based clustering technique described in Section 2.4.2 and compared the clustering results with the human selections. This analysis is corroborated by the fact that the *event coverage* criterion, representable through clustering, has been identified as important during our study (Section 2.3.2).

In our opinion, one of the main risks of applying clustering to emulate human selections for long-term preservation is that not all the clusters might be important for the users. There might be pictures from a sub-event that the user either simply does not like or considers less important than others. We supported this hypothesis by counting the number of human-selected images in each cluster identified in our collections. As to be expected, only for a few clusters (7.3%) all images of the cluster were selected. However, for a considerable part of the clusters (43%) no images were selected at all. Given these statistics, the selection done by any pure coverage–based method that picks an equal number of images from each cluster will contain at least 43% of images that would not have been selected by the user. Another statistics worth to be mentioned refers to the possibility for cluster–based selections of picking centroids as representative images. From our collections, it resulted that only the 26% of the centroids was actually selected by the users. This reveals that information about how much an image is representative of a wider group is only one of the aspects considered by the users when selecting photos.

Finally, making the assumption that bigger clusters might be more important for the users (as indicated by the users' choice to take more photos that capture that part of the event), we consider the size of the clusters with respect to the number of user-selected images that they contain. Figure 2.4 shows the correlation between relative size of clusters (x axis) and the percentage of selected images in them (y axis). It is possible to observe that the selections done by the users result in many clusters with few selected images in each, which is coherent with the notion of coverage. However, what is more interesting is that the size of the cluster seems to be only marginally correlated with the importance of the cluster (i.e. the number of selected images it contains). This is potentially another limitation for all those methods that select an amount of images from each cluster proportionally to its size.

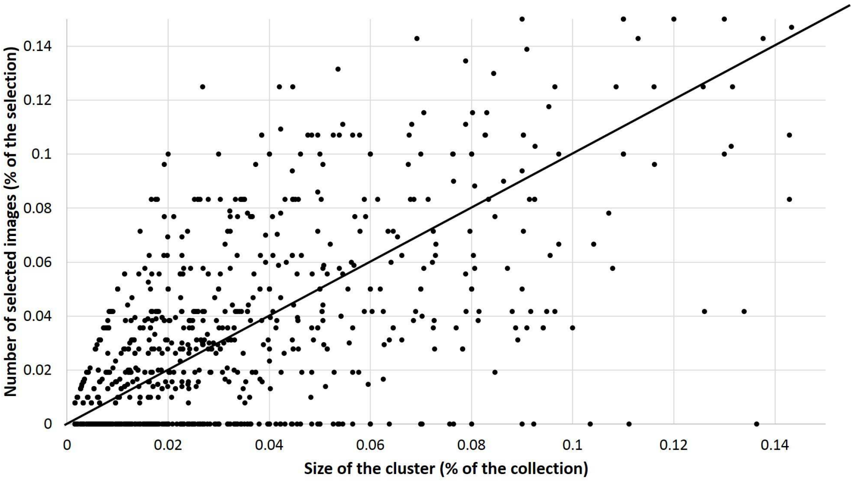


Figure 2.4 Amount of selected images in clusters (with respect to the size of selection) versus relative size of clusters.

2.4 Photo Selection

We present an automatic method to identify, within big personal collections, those photos that are most important to the user, in order to invest more effort in keeping them accessible and enjoyable in the future. The availability of such a method alleviates the problems of *digital forgetting* and *dark archives*, discussed in Section 2.1, which affect the archival of images and their access, respectively. From one side, preservation effort could be invested only on those photos that are worth to be preserved for the owner. From the other side, having a reduced sub set of important photos would make the revisiting and enjoying easier and pleasant for the user. Moreover, to foster adoption, our approach has to keep the level of user investment low: we do not rely on any additional user investment such as photo annotation with text [RSB10, SBS14, SMJ11] or eye tracking information [WSS14], because we believe it is exactly the reluctance of further investment that lets large photo collections unattended on our hard disks. To alleviate errors in automatically created selections and to accommodate user preferences, our approach can be regarded as a semi-automatized procedure, where users can interact with it and modify the suggested selections.

When developing methods for semi-automatic photo selection, it is important to consider human expectations and practices. Photo selection is a complex and partially subjective process, where the selection decision taken for a given image both affects the decisions for other photos and depends on the ones already selected. For this reason, many state-of-the-art methods for photo selection and summarization are driven by the aspect of coverage, which means attempting to create summaries that resemble the content of the original collection as much as possible. Some of them perform a two-step process of first clustering the photo collection (for reflecting subevents in the collection) and subsequently picking the most representative photos from the clusters [LLT03, RSB10]. Others [SBS14, SMJ11, TIWB14] consider coverage as part of a multi-goal optimization, along with the concepts of quality and diversity within the summary. While coverage surely plays an important role for many photo selection tasks (see e.g. [WSS14]), we believe that the complex decision making behind the selection of photos from personal collections, characterized by subjectivity and personal attachment possibly due to memories, might reduce the importance of coverage. For instance, considering photos taken during a trip, the user might want to discard the ones depicting boring or joyless moments.

Therefore, we model a multifaceted notion of image importance driven by user expectations, which represents what photos users perceive as important and would have selected. User expectations have been acquired during the study described in Section 2.3, where participants have been asked to provide their own photo collections and to select those most important to them for preservation and revisiting. We present an expectation-oriented method for photo selection, where information at both image- and collection-level is considered to predict the importance of photos (Section 2.4.3). This information consists of: (a) concept detection, to capture the semantic content of images beyond aesthetic and quality indicators; (b) face detection, reflecting the importance of the presence of people within pictures; (c) near-duplicate detection, to take the redundancy of many pictures of the same scene as a signal of importance, and to eliminate very similar images; (d) quality assessment, since good quality photos might be preferred in case of comparable photos. This is complemented by (e) temporal event clustering and, more generally, collection-level information, to reflect the role of coverage in photo selection. This information is combined via Machine Learning to predict image importance. The selections performed by the users from their own collections are used as labels to train the selection model, so that the predicted importance of photos represents what the user would have selected. For the sake of comparison, in Section 2.4.4 we investigate how the expectation-oriented selection can be combined with more explicit ways of modeling coverage, showing that coverage plays only a secondary role in this task.

Before delving into the details of the selection method, a general consideration on the comparison between the features considered in the model and the user study presented in Section 2.3 has to be done. The aspects that resulted to be important from the user study, e.g. evocation of positive memories, image typicality, personal importance of images, are highly subjective and not directly recognizable by a machine, especially when only relying on the image content without any other contextual information. Given these challenges and constraints imposed by the task, our attempt to address the insights emerged from the study is threefold: (i) we model event coverage, which resulted to be an important aspect in the user study, through clustering and the Hybrid Selection methods described in Section 2.4.4; (ii) we employ concept detection to model more semantic and abstract aspects; (iii) we also include image quality, although perceived as not very important within the user study, for the sake of comparison with the other features.

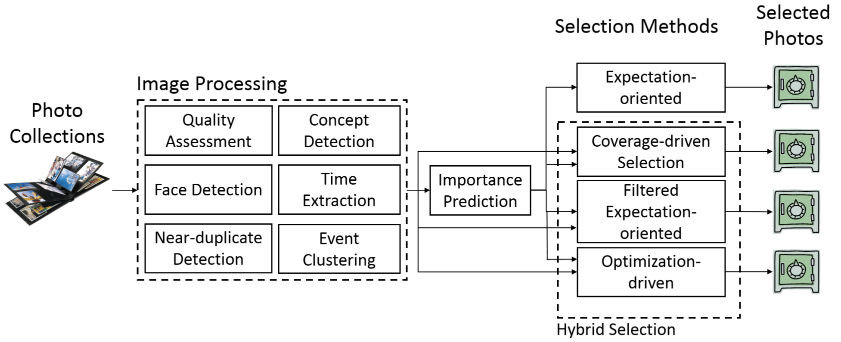


Figure 2.5 Approach overview of our automatic photo selection.

2.4.1 Overview

Definition 1 *Let a photo collection P be a set of N photos, where $P = \{p_1, p_2, \ldots, p_N\}$. The photo selection problem is to select a subset S of size θ ($S \subset P$ and $|S| = \theta$), which is as close as possible to the subset S^* that the user would select as the photos most important to her, i.e. S meets user expectations.*

We represent each photo collection as a set $C = \{P, CL, ND\}$, where P is the set of original photos, and CL and ND are sets of image clusters and near-duplicate photos identified in the collection, respectively. A cluster $cl \in CL$ contains a set of photos P_{cl} grouped together with respect to a defined notion of similarity, whereas a near-duplicate set $nd \in ND$ is a set of highly similar photos P_{nd}. Each photo $p \in P$ is modeled as a set of features $p = \{\mathbf{q}, \mathbf{c}, F, t\}$, where $\mathbf{q} \in \mathbb{R}^{n_q}$ is the quality vector of the photo, $\mathbf{c} \in \mathbb{R}^{n_c}$ is the concept vector of the photo, F is the set of faces f appearing in the photo, t is its timestamp. Each face $f = \{f_l, f_s\}$ is described by its location f_l and relative size f_s in the photo. For each photo p, we will estimate the importance value I using the extracted features.

Figure 2.5 gives an overview of our approaches to photo selection. Given a photo collection, we first extract information from the images by applying different image processing techniques described in Section 2.4.2. Our main approach is named Expectation-oriented selection (Section 2.4.3), which learns to generate selections by taking into account user selections from personal collections as training data. Furthermore, we present three different Hybrid Selection methods (Coverage-driven, Filtered Expectation-oriented, Optimization-driven), with the goal of investigating whether our method can be improved by combining it with methods that explicitly consider coverage. The Hybrid Selection methods will be discussed in detail in Section 2.4.4.

2.4.2 Image Processing

The image processing techniques described hereafter have been developed within the scope of the FP7 ICT project ForgetIT[1].

[1] https://www.forgetit-project.eu

Concept Detection. Concept Detection involves analyzing the visual content of an image and automatically assigning concept labels to it. It allows to identify abstract concepts like *joy*, *cheering*, *entertainment*, as well as more concrete ones like *crowd*, *girl*, *stadium*. The methodology that we followed has been introduced in [MPP+15]. We trained 346 concept detectors, each one represented by a Support Vector Machine (SVM), for the 346 concepts defined as part of the TRECVID 2013 benchmarking activity [OAM+13]. We used SIFT, SURF, and ORB local descriptors and their color variants [MPP+15] for visual feature extraction. Then, PCA was applied on each descriptor for reducing their dimensionality to 80 and VLAD encoding [AZ13] was applied to the local descriptors of reduced-dimensionality so as to calculate the final image representation. As training corpus, the TRECVID 2013 dataset comprising 800 hours of video was used. Despite this being a video dataset, the use of only static (key frame based) features in our approach makes the learned concept detectors applicable both to videos and images.

Near-Duplicate Detection. The presence of near–duplicate images, generated by photographers shooting the same scene multiple times, can be evidence of the importance of the scene. We used the near-duplicate detection method described in [APM14], which detects at most 500 keypoints using the SIFT detector (Harris-Laplace) and extracts their corresponding descriptors using the SIFT extractor. It then forms a vocabulary by applying k-means on SIFT features using 192 cluster centers, and then encodes the image using VLAD encoding. The images with very similar VLAD vectors are efficiently retrieved using a KD forest index (using 5 KD–trees). Finally, Geometric Coding [ZLLT11] is used to check geometric consistency of image's keypoints and to accept or reject the hypothesis that they are near-duplicates.

Image Quality Assessment. We followed the procedure proposed by Mavridaki et al. [MM14] to compute four quality measures, namely blur, contrast, darkness, and noise, along with their aggregation by weighted pooling using Minkowski metric. Blur degradation is estimated by initially computing the Fourier transform of the entire image and of each of its 9 (3×3) equal sized blocks. The frequency histograms of the 10 image/image blocks are calculated and then used as input to an SVM that provides the probability of an image being blurred. Contrast is measured via four indicators: Root Mean Square contrast, normalized (luminance extrema independent) Michelson contrast, RGB histogram, and chromatic purity, which are used as input to an SVM regression model. Regarding darkness, the image is partitioned into 9 equal blocks as for blur detection, and the luminance (Y channel of YCbCr) and color intensity (V channel of HSV) histograms are calculated for each block. They are again fed input to an SVM regression model which estimates the probability for an image to be dark. Noise is estimated using the Blind Image Quality Index's (BIQI) technique [MB10], which employs Natural Scene Statistics in the frequency domain and a supervised learning approach in order to quantify the image quality.

Temporal Event Clustering. In a photo collection captured by a single camera, time is a dominant dimension to reveal the sub-events that are represented in the collection. Thus, we apply time-based clustering in order to cluster collections into sub-events. Additional data dimensions that can be used after the time-based clustering (in order to either split or merge the formed clusters) are geolocation and visual information. We preliminarily experimented also with pure visual clustering and a combination of temporal and visual clustering in cascade, but both exhibited worse performances than pure temporal clustering: the former often grouped images that are far from each other in time, while the latter tended to produce too fragmented clusters. Also, while nowadays geolocation information is more frequently attached to photos as metadata, we did not assume it to be available at the time of our study (in fact, only 1 collection in our dataset contained GPS metadata).

Our method follows the approach presented in [CFGW05]. Images are sorted according to the time information extracted from them, and a similarity matrix of the ordered images is constructed using the time information and a sensitivity parameter. Then, for different values of the sensitivity parameter, the novelty scores and then their first derivatives are calculated. First derivatives which are greater than a threshold based on the maximum peak (at least 0.5 times greater than the maximum peak) are selected. The procedure results is a set of boundary lists (one list per sensitivity parameter). Finally, the confidence measure for each boundary list is calculated and the list that corresponds to the largest confidence measure is selected.

Face Detection. We introduce an approach that combines several face detectors to maximize the number of detected faces. We apply the approach in [VJ04] using four pre-trained Haar Cascades detectors (frontalface_alt, frontalface_alt_tree, frontalface_alt2, frontalface_default) publicly provided by OpenCV[2]. Adopting the union of different face detectors allows to decrease the missed detection error but, on the other hand, the percentage of false positives can significantly increase. In order to reduce the amount of false positive detections, we consider the detected faces as potential facial regions and, in every region, (i) we try to detect other facial characteristics (eyes, nose, and mouth) and (ii) we calculate the percentage of skin-like pixels. If facial characteristics are detected and are located in a valid location (e.g. eyes are centered and on the upper half for facial region) and the skin-like pixels percentage is above a threshold (set to 0.3 after experiments), then the detected region is classified as face. Furthermore, a facial region is accepted as a face if it has been detected by all face detectors (regardless of the existence of facial features or the ratio of skin-color pixels). This last case is effective in dark images where facial characteristics can not be detected and face color is too dark to be considered as skin-like.

2.4.3 Expectation-oriented Selection

The photo selection model presented in this section aims at meeting human expectations when selecting photos that are most important to the user from a collection, for revisiting or preservation purposes. We believe that selecting photos that

[2] https://github.com/Itseez/opencv/tree/master/data/haarcascades

are important to a user from personal collections is a different task than generating comprehensive summaries: the set of images important to the user might not be a proportioned subsample of the original collection. For this reason, we do not impose a strict notion of coverage but rather consider clusters and other global information as a set of features, along with photo-level features, learning their different impact in a single selection model by means of supervised Machine Learning.

A key characteristic of our features is that they do not require any manual annotation (e.g., tags, textual descriptions, file names) or external knowledge, differently from other works [RSB10, SM16, SBS14, SMJ11] that make partial use of manually created text associated to photos. This means that users do not have to invest time and effort in preparing the photos before feeding them into our system.

Features

Four groups of features have been designed to be used in the photo selection task, based on the information extracted from images as presented in Section 2.4.2.

Quality-based features. They are the 5 quality measures described before: blur, contrast, darkness, noise, and their fused value. They are all numeric features between 0 and 1, where 0 represents the worst quality and 1 the best. The assumption behind using this information is that users might tend to select good quality photos, although their impact seems to be less important when selecting from personal pictures as emerged from previous work [WSS14] and from our user study (Section 2.3). Nevertheless, quality could play a role in case of near-duplicate images with different quality: the one with best quality might be picked in these cases.

Face-based features. The presence and position of faces in a picture might be an indicator of importance and might influence the selection. Some people might prefer images with persons instead of landscape images, some people might like group photos more than single portraits. We capture this by considering, for each photo, the number of faces within it as well as their positions and relative sizes. Each photo is divided in nine quadrants and the number of faces and their size in each quadrant are computed, resulting in 19 features: two for number and size of faces in each quadrant, plus an aggregated one representing the total number of faces in the photo.

Concept-based features. High-level and semantic information has been thoroughly investigated in the past years within the scope of digital summarization (e.g. [SBS14, SMJ11]). The semantic content of pictures, which we model in terms of concepts appearing in them, is expected to be a better indicator than low-level image features, because it is closer to what a picture encapsulates. We associate to each photo a vector of 346 elements, one for each concept, where the i-th value represents the probability for the i-th concept to appear in the photo.

Collection-based features. All the previously mentioned features are extracted from images in isolation. However, when users have to identify a subset of important photos, instead of just making decisions for each photo separately, the characteris-

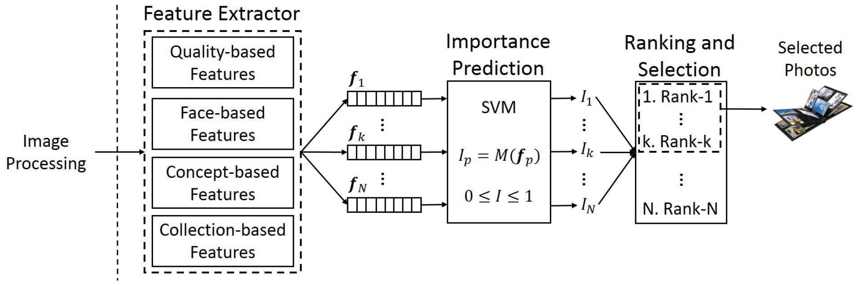


Figure 2.6 Workflow of the importance prediction and photo selection.

tics of the collection a photo belongs to might influence the overall selection of the subset. For the same reasons, but moving to a finer granularity, it might be worth considering information about the cluster a photo belongs to. For each photo, we consider collection-based features to describe the collection and, if any, the cluster and near-duplicate set the photo belongs to. Regarding the whole collection, we consider its size, the number of clusters and near-duplicate sets in the collection, the number of not near-duplicate images, the size of the clusters (avg, std, min, max) in the collection, the size of near-duplicate sets (avg, std, min, max) in the collection, the quality of the collection (avg, std), the number of faces in the collection (avg, std, max, min). Regarding clusters, we first perform event–based image clustering for each collection. Then, given the cluster a given image belongs to, we compute its size, its quality (avg, std, max, min), and the number of faces within it (avg). Finally, since the redundancy introduced by shooting many pictures of the same scene can be evidence of its importance for the user, who replicates a scene to ensure its availability and quality, we also extract features regarding whether the given image has near-duplicates or not, as well as how many they are.

Importance Prediction and Ranking

Given a set of photos p_i, their vectors $\mathbf{f}_{p_i}$ containing the features presented above, and their selection labels l_{p_i} (i.e. *selected* or *not selected*) available for training, an SVM is trained to predict the selection probabilities of new unseen photos, i.e. their importance. In the last two decades SVMs have been popular and effective models for a wide variety of image classification tasks [Wan05, MG14, CHV99], and they resulted to perform best on our task as well when compared to other classifiers, such as MLPs, Logistic Regression, and Random Forests. Also with the advent of Deep Learning, which has advanced the state-of-the-art for image classification (see e.g. [KSH12, SLJ+15]), using SVMs to learn associations between image representations learned by Convolutional Neural Networks and task-dependent labels still exhibited competitive performances [Tan13, RASC14].

Figure 2.6 shows how the importance prediction and ranking of photos is performed for new unseen collections. Feature vectors $\mathbf{f}_p$ are constructed from the images and the importance of each unseen photo p is computed as:

$$I_p = M\left(\mathbf{f}_p\right) \tag{2.1}$$

which is the probability of the photo to be selected by the user. Once the importance of each photo in a collection is predicted, the photos are ranked based on this value and the top-k are finally selected. The parameter k represents the desired size of the selection and its choice will be discussed during our evaluation (Section 2.4.5).

2.4.4 Hybrid Selection

Given the wide exploitation of the concept of coverage in many approaches, we want to better understand its role in personal photo selection, in order to see if and in which way our previously described method can be improved by combining it with explicit representations of coverage. The notion of coverage resulted to be highly important from our user study (Section 2.3) as well, which is another motivation for further investigating its potential contributions and limitations. It is interesting to note that, despite the participants declared coverage as highly important, the selections that they made in the study exhibited a poor degree of coverage.

We investigate three ways of combining our importance prediction model with coverage-oriented approaches, which we call *hybrid selection* methods. Although kept into account within the expectation-oriented selection via the *collection-based* features (Section 2.4.3), the concept of coverage is more prominent here.

Coverage–driven Selection

The coverage-driven selection is based on the widely used two-step process of first clustering and subsequently picking photos from the clusters. First, for a given collection C, a set of clusters CL_C is computed as described in Section 2.4.2 and the importance $I(p)$ of each photo $p \in P_C$ is computed according to our importance prediction model (Equation 2.1). Given the clusters CL_C, we use the importance $I(p)$ for each photo $p \in P_C$ to pick an equal number of top-ranked photos from each cluster in order to produce the selection S of required size k.

Cluster Visiting. When picking photos from each cluster, there are different possible ways of iterating over them until the requested size of the selection is reached. After experimenting a number of alternatives, we identified a round-robin strategy with a greedy selection at each round as the best performing one. The pseudo-code is listed in Algorithm 1. Given an initial set of candidate clusters CL_{cand}, the greedy strategy in each step selects the cluster cl^* containing the photo p^* with the highest importance, according to the prediction model M. The photo p^* is added to the selection S and removed from its cluster cl^*. The cluster cl^* is then removed from the set of candidate clusters for this iteration, and the greedy strategy is repeated until the candidate set is empty. Once it is, all the not empty clusters are considered

Algorithm 1: Coverage–driven Selection (Greedy)

```
Input: clusters $CL$, selection size $k$, prediction model $M$
Output: selection $S$
Set $S = \emptyset$
while $|S| < k$ do
    Set $CL_{cand} = CL$
    while $|CL_{cand}| > 0$ do
        $\{cl^*, p^*\}$ = get_most_important_cluster $(CL_{cand}, M)$
        $S = S \cup \{p^*\}$
        $P_{cl^*} = P_{cl^*} - \{p^*\}$
        $CL_{cand} = CL_{cand} - \{cl^*\}$
        if $|cl^*| = 0$ then
            $CL = CL - \{cl^*\}$
        end
    end
end
return $S$
```

available again and a new iteration of the cluster visiting starts. This procedure continues until the requested selection size k is reached. We also experimented with a regression model to predict the number of photos to select from each cluster, but it did not lead to satisfactory results.

Cluster Filtering. The intuition behind cluster filtering is that not all the clusters identified in a collection are equally important to the user. For instance, considering photos taken during a trip, there might be pictures depicting exciting moments along with other more boring situations, which the user might want to discard. We tackle this issue by proposing a cluster filtering method to automatically predict the clusters that are not important for the user, in order to ignore them when picking photos from each cluster. We train a classifier (SVM) to detect and filter out clusters which are not important to the user. First, each cluster is described with the following features: size, quality vector (avg, std), average concept vector, number of faces (avg, std, min, max), number of near-duplicate sets and near-duplicate photos in it, near-duplicate sets size (avg, std, min, max), photo time (avg, std, min, max), photo importance (avg, std, min, max). The label associated to a cluster is *good* if it contains at least one selected photo, *bad* otherwise. Given a training set made of clusters c_i, their corresponding feature vectors $\mathbf{f}_{c_i}$, and their classes l_{c_i}, an SVM is trained and the learned model N is used to predict the class $L = N(\mathbf{f}_{c_{new}})$ of new unseen clusters c_{new}. Details on the training process are reported in Section 2.4.5.

Given the clusters CL_C in a collection and a classifier trained on a different portion of the dataset, applying cluster filtering removes from CL_C all those clusters that are classified as *bad* by the classifier. The iteration and picking phase are then performed only with the remaining *good* clusters.

Filtered Expectation-oriented Selection

The coverage-driven selection is characterized by two steps: first clusters are identified and handled by possibly filtering and sorting them, and then photos in each cluster are ranked based on their predicted importance. Differently, within the *filtered expectation-oriented selection*, we give priority to importance prediction. The photos in a collection are first ranked based on the predicted importance and then cluster filtering is applied. The result is a ranked list of photos, where those belonging to clusters classified as *bad* have been removed. Note that the second phase of this paradigm, which contains cluster filtering in our case, can incorporate any other computation that exploits cluster information. The way photos are selected after applying cluster filtering is the same as the one described in Section 2.4.3: the selection S of size k is created by choosing the top–k photos in the list.

Optimization–driven Selection

Besides applying clustering, another way of explicitly incorporating coverage in photo selection is to consider it as part of a multi-goal optimization problem. This has been done in [SMJ11] to generate representative summaries from personal photo collections, with the objective of having concise sub-collections that resemble the original one as much as possible. In more detail, in this work *quality*, *coverage*, and *diversity* are jointly optimized and the optimal summary S^* of a requested size k is defined as:

$$S^* = \underset{S \subset P_C}{\arg\max}\, F\left(Qual\left(S\right), Div\left(S\right), Cov\left(S, P_C\right)\right) \tag{2.2}$$

where $Qual\left(S\right)$ determines the interestingness of the summary S and it aggregates the *interest* values of the individual photos within S, $Div\left(S\right)$ is an aggregated measure of the diversity of the summary measured as $Div\left(S\right) = \min_{p_i, p_j \in S, i \neq j} Dist\left(p_i, p_j\right)$, and $Cov\left(S, P_C\right)$ denotes the number of photos in the original collection C that are represented by the photos in the summary S with respect to a concept space.

We incorporate our expectation-oriented selection within this framework, creating the *optimization–driven selection*, by computing the $Qual\left(\cdot\right)$ function in Equation 2.2 based on the importance prediction model (Equation 2.1), that is:

$$Qual\left(S\right) = \sum_{p \in S} M\left(p\right) \tag{2.3}$$

Since part of the concepts in [SMJ11] are discrete categorical attributes, associated to photos using textual information and external knowledge bases not available in our task, we binarized the elements of our automatically detected concept vector (which includes the probability that a given concept appears in the photo) by using a threshold τ such that $c_i = 1\ if\ c_i > \tau$, and $c_i = 0$ otherwise. The threshold has been empirically identified as $\tau = 0.4$ as the value that led to the most meaningful binary results. The rest of the calculation of the $Div(\cdot)$ and $Cov(\cdot)$ functions in Equation 2.2 is performed as in the original work. In more detail, the *distance* between two photos, used to measure the *diversity* within a summary, is computed based on exif features, time, and concept vectors (as in the original work, however we use the automatically extracted concepts), while the *coverage* of a summary is calculated based on the number of pictures in the original collection that are represented by the photos in the summary in a concept space (considering binarized concepts vectors when needed).

Regarding the solution of Equation 2.2, which is an NP–Hard problem, we experimented with the different approaches presented in [SMJ11] and the best performing one consisted in combining quality, diversity, and coverage in a linear way:

$$S^* = \underset{S \subset P_C}{\arg\max} \left[\alpha \cdot Qual\,(S) + \beta \cdot Div\,(S) + \gamma \cdot Cov\,(S, P_C)\right] \tag{2.4}$$

and performing a greedy optimization, which has proved performance guarantees (please refer to [SMJ11] for further details). We will discuss the values used for the α, β, γ parameters in the experimental analysis.

2.4.5 Experiments

Experimental Setup

Dataset For our experiments we use personal photo collections with importance judgments given by the owners of the collections as dataset. These can be photos from business trips, vacations, ceremonies, or other personal events a person participated in. We decided to focus on personal collections because we wanted to observe the selection decisions in a setting that is as realistic as possible. This gives us a ground truth for assessing user expectations. For each collection, we only considered the selections done by its owner because we wanted our ground truth to include also the effects of subjective aspects (e.g. memories, personal attachment, context) on the selection process and evaluate different automatic approaches based on it. This would have been unlikely to happen if we had let other users evaluate collections that represent events they did not take part in.

Given the unavailability of such a dataset of real-word personal collections, with selections done by the owners based on their perceived importance, we considered the data collected during the user study previously described in Section 2.3. As a short reminder, participants were asked to provide their photo collections and to select the 20% that they perceive as the most important for revisiting or preservation purposes. The selection percentage (20%) was empirically identified as a reasonable amount of representative photos, after discussing this matter with a subset of participants before the study. In order to make the evaluation results more statistically significant, we expanded the original dataset by repeating the same evaluation with other participants and photo collections. Such extended dataset consists in 18,147 photos organized in

91 collections and belonging to 42 users. The collection sizes range between 100 and 625 photos, with an average of 199.4 (SD = 101). The use of crowdsourcing would have allowed to scale up the size of our dataset, but we decided not to employ it for two reasons. First, engaging the annotators and detecting malicious behaviors is not straightforward due to the complexity of the task (difficult to decompose in atomic units), its potentially long duration, and the unavailability of gold standard to control the selections made by the workers from their own collections. Second, we presumed that the workers could have been reluctant to share personal data with strangers.

Near-duplicates have been detected and filtered by considering the centroid of each set as representative photo, as done in [CL08]. For sets containing two photos, the one with better quality is chosen as representative. Similarly to [SEL00], each representative is marked as selected if at least one photo in its set has been marked as selected, and marked as not selected otherwise.

Evaluation Metrics Since the overall goal of our work is emulating the human behaviors in selecting the subsets of photos from a personal collection, we compare the automatic selections generated by our methods with the ones done by the users.

The selection methods presented in this chapter generate a selection S of size k from the original collection, where k can assume different values. We evaluate the different methods considering the precision P@k of the selection S of size k that they produce, computed as the ratio between number of photos in S that were originally selected by the user and the size of S. Since the collections in our dataset have high size variability (from 100 to 625 photos), absolute values of k, although traditionally used in Information Retrieval tasks, would result in selecting very different relative portions of the collections depending on their sizes. This makes the impact of the selection different among collections. We, therefore, decided to express k as a percentage of the collection size, instead of an absolute value. In particular, we compute the precision for $k = 5\%, 10\%, 15\%, 20\%$, which are indicated as P@5%, P@10%, P@15%, P@20%, respectively. We concentrate the discussion on P@20%, because our ground truth was gathered by asking users to select the 20% of their collections. We will also give comments about the recall of the selections generated by the different methods.

The collections in our dataset were split by 10–fold cross validation (used for training and evaluating the classifiers) and all the values reported in this section are averaged over the test sets of each split. Statistically significant improvements were assessed with a two-tailed paired t-test and marked as $^{\blacktriangle}$ and $^{\vartriangle}$ (with $p < 0.01$ and $p < 0.05$, respectively) in the tables. If not stated otherwise, the significance outcomes refer to the comparisons with both the baselines described in the following section.

Feature Selection Information Gain Feature Selection was performed to reduce the dimensionality of the classifiers trained on different subsets of features. We retained the top–200 features when training with the whole set, while the top–160 were kept when training only using concept features. These values were identified via 10-fold cross validation (grid search [50; 300] with steps of 10). We did not apply any feature selection to the other smaller feature sets.

Implementation and Parameter Settings The SVMs used in this chapter have Gaussian Kernels and their hyper-parameters were tuned by grid search and 10-fold cross validation at collection level to $C = 1.0$, $\gamma = 1.0$ for the expectation-based model and to $C = 3.0$, $\gamma = 1.4$ for the cluster filtering. The classifiers were trained using the Support Vector Machine implementation of LibSVM[3], while OpenCV[4] and OpenIMAJ[5] were used as a support for feature extraction.

Baselines

We compare our method with two baselines, one based on clustering and one representing the optimization framework presented in [SMJ11].

Clustering. Similarly to what was described at the beginning of Section 2.4.4, for a given collection C, a set of clusters CL_C is computed. The selection is built by iterating the clusters, temporally sorted, in a round–robin fashion and picking at each round the most important photo from the current cluster (until the requested selection size is reached). Instead of using our expectation-based model, the importance of each photo $p \in P_C$ is modeled as

$$I(p) = \alpha \cdot \left\|\mathbf{q}_p\right\| + (1-\alpha) \cdot \dim(F_p) \tag{2.5}$$

which is a weighted sum of the quality vector of the photo and the number of faces in it. This notion of image importance covers different works in the literature, e.g. [LLT03, RSB10]. We experimented with different values of the parameter α, identifying the best value as $\alpha = 0.3$, which gives more importance to the number of faces. We report the performances obtained with this parameter value in our evaluation.

Summary Optimization. We implemented the approach presented in [SMJ11] as another baseline, where summaries are generated by optimizing *quality*, *coverage*, and *diversity* as in Equation 2.2. It differs from the hybrid method described in Section 2.4.4 in how photo importance is modeled, as here the expectation-based model is not considered. Instead, the *quality* of summaries is computed by summing the *interest* of photos in it, defined as a measure dependent on photo quality and presence of portraits, groups, and panoramas. We computed the interest of photos as in the original work, using the concepts *face*, *3 or more people*, and *landscape* available in our concept set to represent portraits, groups, and panoramas respectively. Also *diversity* and *coverage* of summaries are computed coherently with their original computation, as already described in 2.4.4. Giving equal weights to the α, β, γ parameters gave us the best results, thus we will report the performances for only this setup in the following evaluation, denoting it *SummOpt*.

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [4] http://opencv.org/ [5] http://openimaj.org/

2.4.6 Results

The discussion of the results is organized as follows. First, we show the performances of our expectation-oriented selection with respect to the baselines. Second, we present the results of the hybrid selection methods and we compare them both with the baselines and with the expectation-oriented selection. Third, we make a general comparison of the methods based on recall performances.

Expectation-oriented Selection

This section presents the evaluation of our expectation-oriented selection with respect to the two baselines defined in Section 2.4.5. Different importance prediction models have been trained by using the subsets of the features described in Section 2.4.3, so that the impact of different groups of features on the precision can be analyzed.

The results for different selection sizes (k) are listed in Table 2.1. The two baselines exhibit comparable performances, with *SummOpt* performing slightly better for all considered values of k (5%, 10%, 15%, 20%). Regarding our model, *quality* features are the ones that perform weakest individually, which has already been observed for other photo selection tasks [WSS14]. This corroborates the idea that low quality photos might be kept anyway because they contain and recall memories and events important to the user. *Faces* features alone already show better performances than the baselines: the presence, number, and position of people in photos, largely used as one selection criterion in other works, is indeed a meaningful indicator of importance. The performance achieved when only using *concepts* features (top-160 as explained in Section 2.4.5) is better than the ones of *quality* and *faces*: they are able to capture the semantic content of the photos, going beyond their superficial aesthetic and quality. Examples of concepts with a high importance in the model are *person*, *joy*, *entertainment*, and *crowd*. The model trained with the combination of all the aforementioned features, denoted *photo-level* because the features are extracted from each picture in isolation, slightly improves the performance of using concept features alone. This indicates that leveraging quality and faces features in addition to semantic measures, such as concepts, can ameliorate the overall performance.

If we include global features for each photo representing information about the collection, the cluster, and the near–duplicate set the photo belongs to, we get a comprehensive set of features, which we call *all* (top–200 features). The precision of the selection for this global model further increases for every selection size: this suggests that decisions for single photos are not taken in isolation but they are also driven by considering general characteristics of the collection the photo belongs to: e.g. number of photos, clusters, average quality of photos in the collection and in the same cluster, how many duplicates for the photo there are. This is a point of distinction with respect to state-of-the-art methods (represented by the two baselines), because our selection approach does not strictly handle collection-level information by imposing clustering

Table 2.1 Precision of the expectation-oriented selection, distinguishing different sets of features.

	P@5%	**P@10%**	**P@15%**	**P@20%**
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Expectation-oriented Selection				
quality	0.3431	0.3261	0.3204	0.3168
faces	0.4506▲	0.3968▲	0.3836△	0.3747△
concepts	0.5464▲	0.4599▲	0.4257▲	0.4117▲
photo-level	0.5482▲	0.4760▲	0.4434▲	0.4266▲
all (Expo)	0.7124▲	0.5500▲	0.4895▲	0.4652▲

(*Clustering*) or optimizing measures like coverage and diversity along with photo importance only based on quality and presence of people (*SummOpt*). It rather takes this global information in consideration in a flexible way through a set of features, whose impact to the selection is learned from user selections and expectations. The expectation-oriented model using all the available features (named *Expo* in the rest of the evaluation) leads to a relative improvement of 38.5% and 33.75% over *Clustering* and *SummOpt* respectively, considering P@20%, and even higher improvements when considering smaller values of k (90.4% and 84.6% for P@5%).

The P@20% metric is of primary importance because we asked users to select exactly 20% of their collections during the data acquisition. However, another point of discussion is the trend of precision performances over different values of k: all the models reach higher precision values for smaller selection sizes. This can be due to the presence of a limited number of selected photos that are relatively easy to identify for the methods, which give them highest selection probability.

Hybrid Selection

This section discusses the precision of the hybrid selection methods (Section 2.4.4) with respect to the baselines, along with a comparative analysis between them. The results are listed in Table 2.2, where they have been split based on the three different classes of hybrid selection described in Section 2.4.4. For coverage-driven selection, we report results of different combinations: *basic* refers to the coverage–driven selection which only uses our importance prediction model defined in Section 2.4.3 as photo importance measure, picking photos in a round-robin fashion from clusters temporally ordered; the term *filtered* means the use of cluster filtering, while the presence of the term *greedy* indicates the use of the greedy visiting strategy. The filtered expectation-oriented selection is denoted *F-Expo*.

Table 2.2 Precision of the hybrid selection methods.

	P@5%	**P@10%**	**P@15%**	**P@20%**
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Coverage-driven Selection				
basic	0.4732$^{\blacktriangle}$	0.4113$^{\blacktriangle}$	0.3902$^{\vartriangle}$	0.3809$^{\vartriangle}$
filtered	0.5351$^{\blacktriangle}$	0.4617$^{\blacktriangle}$	0.4325$^{\blacktriangle}$	0.4170$^{\blacktriangle}$
filtered+greedy	0.6271$^{\blacktriangle}$	0.4835$^{\blacktriangle}$	0.4391$^{\blacktriangle}$	0.4262$^{\blacktriangle}$
F-Expo	0.7065$^{\blacktriangle}$	0.5502$^{\blacktriangle}$	0.4863$^{\blacktriangle}$	0.4600$^{\blacktriangle}$
SummOpt++	0.7115$^{\blacktriangle}$	0.5533$^{\blacktriangle}$	0.4937$^{\blacktriangle}$	0.4708$^{\blacktriangle}$
Expo	0.7124$^{\blacktriangle}$	0.5500$^{\blacktriangle}$	0.4895$^{\blacktriangle}$	0.4652$^{\blacktriangle}$
Filtering with Oracle				
greedy+oracle	0.6499$^{\blacktriangle}$	0.5107$^{\blacktriangle}$	0.4665$^{\blacktriangle}$	0.4484$^{\blacktriangle}$
F-Expo+oracle	0.7150$^{\blacktriangle}$	0.5606$^{\blacktriangle}$	0.4982$^{\blacktriangle}$	0.4753$^{\blacktriangle}$

For the optimization-driven method, we experimented the different optimization methods described in [SMJ11] after introducing our importance prediction model in place of the original importance measure used in that work ($Qual\,(\cdot)$). We found out that the best performing method was still the greedy optimization of a linear cost functional combining importance, diversity, and coverage (Equation 2.4) but with a parameter combination that gives more importance to the quality of the photos (0.6 *Qual*, 0.3 *Cov*, 0.1 *Div*). We consider the results of this setup in the following evaluation. This difference in weights with respect to the *SummOpt* baseline already anticipates that our expectation-based measure of importance has a bigger impact in the performances than the native quality measure defined in [SMJ11]. The method will be referred to as *SummOpt++*.

The results in Table 2.2 show that all hybrid methods outperform the baselines, with statistical significance, showing that the inclusion of the importance prediction model to assess photo importance has a strong impact compared to the baselines methods, which model photo importance with simple functions of quality and people occurrence. Similarly to the performances of the expectation-oriented models, both the absolute precision values and the improvements with respect to the baselines increase for decreasing k. Concerning the *coverage-driven selection*, the results in Table 2.2 also show that cluster filtering increments the precision of the *basic* approach of an amount between 9.48% (P@20%) and 13.1% (P@10%). The greedy visiting strategy leads to improvements as well. Statistical significance was observed in the improvements introduced by *filtered* and *filtered+greedy*.

Comparing the results of the different hybrid selection methods, *F-Expo* and *SummOpt++* achieve better precision performances than the coverage-driven methods, and a t-test confirms that these improvements are statistically significant. This shows that the measure of photo importance modeled by our importance prediction has a bigger impact in the precision of the selection than coverage, and those methods that strictly model it through clustering (*coverage-driven selection*) get a smaller benefit when incorporating the expectation-oriented model. On the other side, methods that either give priority to expectations (*F-Expo*) or consider expectations, coverage, and global information in a flexible way via optimization (*SummOpt++*) can better exploit the expectation-oriented model.

Expectation vs. Hybrid Analysis

In this section we make a comparative analysis between the expectation-oriented selection model exploiting all the available features (*Expo*), and the hybrid selection models. Considering Table 2.2, we can observe that the performances of *Expo* are better or comparable with the ones of the hybrid-selection models. In particular, the improvements of *Expo* with respect to the *coverage-driven* methods are statistically significant. The only improvements over *Expo* (which anyway are not statistically significant) are obtained when considering methods that prioritize expectations (*F-Expo*) or possess a relaxed consideration of coverage and global information in general (*SummOpt++*). These results further support our assumption that, for the photo selection task involving personal data, a strong consideration of coverage overstresses this aspect as a selection criterion. Instead, the users might not follow a strict idea of coverage when making selections, generating selections that are not as proportioned samples of the original collections as purely coverage-based methods would suggest. Only for the methods with a more flexible consideration of coverage the performances are similar to the pure expectation-oriented method.

Cluster filtering is an attempt to eliminate clusters uninteresting to the user, and in order to further alleviate this aspect we conducted experiments considering only important clusters, i.e. those ones containing at least one selected photo. This is done by assuming to have a perfect classifier, i.e. an *oracle*, to filter out not important clusters and to focus the hybrid selection strategies only on the important ones. Although getting improvements compared to *filtered+greedy* and *F-Expo*, the performances when using such oracle, reported in the bottom part of Table 2.2, did not lead to consistent and statistically significant improvements with respect to *Expo*. *Greedy+oracle* does not beat *Expo*, while *F-Expo+oracle* only introduces a limited and not statistically significant improvement. These results show that the aspect that mostly drives user selections and expectations is the personal perception of importance, although this can produce unbalanced selections which are not representative of the original collection. Another issue related to clustering, which is not addressed here, is the decision of how many photos to pick from each of the important ones.

Table 2.3 Recall of different selection methods.

	R@20%	R@30%	R@50%	R@75%
Clustering	0.3358	0.4555	0.7000	0.9231
SummOpt	0.3478	0.4354	0.6884	0.9253
Expo	$0.4652^{\blacktriangle}$	$0.5310^{\blacktriangle}$	$0.7356^{\vartriangle}$	0.9310
filtered+greedy	$0.4262^{\blacktriangle}$	$0.5129^{\vartriangle}$	0.7232	0.9231
Filtered Expo	$0.4600^{\blacktriangle}$	$0.5361^{\blacktriangle}$	$0.7433^{\vartriangle}$	0.9275
SummOpt++	$0.4708^{\blacktriangle}$	$0.5408^{\blacktriangle}$	$0.7405^{\vartriangle}$	0.9315

Recall-based Analysis

Finally, we make a comparative analysis of the different method based on recall. The motivation of considering recall is that a user might accept to increase the size of the automatically created selection in order to include more important photos than the ones included when remaining strict to the ideal size of 20% (considered during the user study). In Table 2.3 we show the recalls of the best performing methods from each selection class, computed for different selection sizes. Note that R@20% always coincides with P@20%, since users were asked to select 20% of their collections. The results are coherent with the analysis already done for the precision: both the expectation-based model and the hybrid-selection methods outperform the baselines, and the former is overall better then or comparable to the latter class. Only methods that prioritize user selections (*Filtered Expectation-based*) or consider expectations, coverage, and global information in a flexible way via optimization (*optimization-driven selection*) can reach slightly higher recall values than the one of the expectation-oriented model. In the future, this consideration could be the starting point for a photo selection method that maximizes recall, or at least considers it in the learning model along with precision-based criteria.

2.5 Personalization

Although the expectation-oriented selection method presented in Section 2.4.3 has been proved to be more effective in meeting user expectations than approaches based on coverage, it applies the learned selection model for any user and collection. Nevertheless, the photo selection process (especially for personal data) can be highly subjective and the factors that drive the selection can vary from individual to individual [ODOO10, SEL00, YHBO10]. General selection models, although capable of representing common selection patterns (e.g., photos depicting people might be usually appreciated), might be improved by considering the preferences of each single user separately and derive personalized models for them. Some users might be particularly interested on photos depicting many people, while others might prefer pictures with landscapes or buildings. Besides variations in the set of appreciated concepts, also selection aspects that are ignored by some people might become more important for others. It is therefore worth investigating how personalized selection models that adapt to the preferences of different users could be developed.

Huang et al. [HXW$^+$11] proposed an image retrieval system for personalized portraits ranking based on four kinds of features and two user interfaces to acquire personalized feature weights from the user for ranking. A similar approach has been presented in [YHBO10], where the ranking is not restricted to portraits. In both works, the user preferences are explicitly expressed through the interfaces instead of implicitly learning them from the data, which is indeed what we aim at. The work in [YBO14] is closer to ours, as it learns to rank photos from a dataset of public rankings and realizes personalization by exploiting examples of personal rankings for re-ranking (i.e. personal and public rankings are considered together when re-ranking). Besides the different learning algorithm employed, this work differs from ours because the ranking is done only based on aesthetic features. Relevance feedback has been used to iteratively refine search results based on human feedback [RH00, ZZSH10]. These approaches are not directly applicable to our scenario, where there is not any explicit query to be considered as reference when refining the result set. Also collaborative filtering techniques have been used (e.g. in [UK05]) to accumulate records of user feedback and exploiting such relations among images to help future users.

We aim at refining personalized models starting from the general one learned via the approach described in Section 2.4.3, denoted *general model* hereafter. Starting from it, selection decisions done by a given user on new collections are acquired and the selection model is updated according to them. Feeding the revisions of the user for automatically generated selection back into the selection model can, on the long run, bridge the gap between the general selection model and the user preferences. Moreover, in order to tackle the problem of having limited initial data to train the model (cold-start scenario), we experiment whether the exploitation of data from other users can boost the adaptation of the model to a given user when a limited amount of personal training data is available.

2.5.1 Approach

We adopt an incremental learning strategy to achieve personalization, re-training the model each time new data (selection decisions) is provided by the user. The annotated photo collections available to train the general model are first pre-processed through image processing techniques and features are extracted from them, in the same way described in Section 2.4.3. For each new collection provided by the user, a first selection is made by the trained general model and the selected photos are displayed to the user, who gives feedback revising the automatically generated selection. The training dataset is then expanded by adding the feedback data and the general model is retrained with the updated training dataset. Iterating this process, it is expected that the gap between user expectations and model's selections gets lower, due to the adaptation of the model towards the selection preferences of the user.

Incremental Learning. A recurrent matter in machine learning is continuously managing new data, so that the existing model can be updated to accommodate new information and to adapt to it. Two common approaches for updating the model to new incoming data are *online learning* [BEWB05], where the model is updated only considering the new data, and *incremental learning* [CP00], where the model update considers the old training data along with the incoming data. We consider the latter strategy and re-train the model each time new data (i.e. selection decisions) was provided by the user because, in our scenario, the updated model has to be aware of the entire data available, not just of the most recent one. Although efficient and effective incremental versions of off-line learning algorithms exist (e.g., [CP00]), we perform the model update by including the new data in the training set and re-train the model from scratch. We implemented such more straightforward but functionally equivalent approach because our scenario does not impose strict time constraints for the model update, thus making the efficiency benefit of incremental versions of secondary importance. The time taken by a user to produce a new collection (e.g. after a trip or vacation) can be considered sufficient to re-train the model with the whole available data. Should the temporal constraints of the envisioned scenario become stricter, the incremental version of the employed algorithm could be plugged in without changing the functionalities of the whole application.

Model Update. Personalized photo selection models, one for each user, are built by re-training the model every time that a new collection is imported and the automatic selection done by the current model is revised by the user. The procedure of the model update is the following. The annotated photo collections available to train the general model are first subject to image processing and feature extraction, in the same way as described before in Sections 2.4.2 and 2.4.3. In Section 2.5.2 we will experiment with different ways of creating the initial training set. For each new collection provided by the user, a first selection is made by the trained general model as described in Section 2.4.3 and the selected pictures are displayed to the user (along with the discarded ones), who gives feedback revising the automatically generated selection. The training dataset is then expanded by adding the feedback data and the general model is retrained with the updated training dataset. Iterating this process, it is expected that the gap between user expectations and model's selections gets lower, due to the adaptation of the model towards the selection preferences of the user. This workflow represents the envisioned behavior once the whole system has been finalized and released to the end user. However, in order to easily repeat evaluations when testing the model, we collected the data from each user once for all, i.e. users evaluated all the collections from scratch without revising any automatically generated selection. Although we are aware that the selections done by the user starting from an automatically generated selection might differ from those done when selecting photos from scratch, repeating the evaluation multiple times when designing the system would have been unfeasible for the users. Moreover, acquiring evaluations done from scratch is unbiased from the initial selection proposed automatically.

Cold–Start Problem. Usually, the adaptation of a system within the initial rounds of user interactions is affected by the so called *cold-start problem*: there is not enough (or even not at all) training data to let the model adapt to the user. This holds in our scenario as well, where the selection model might not make proper predictions due to the lack of annotated collections in the initial training set. We consider two ways of building the initial training set. One consists in using one annotated collection of the given user as initial training set. The other is based on using annotated collections from other users to train the initial selection model, hopefully boosting the adaptation of the model to a given user when a limited amount of personal training data is available. The latter approach is based on the assumption that, despite the subjectivity of the task, common selection patterns exist and could be captured through a sample of selections done by other users.

2.5.2 Experiments

Experimental Setup

Dataset. We used the same dataset described in Section 2.4.5 for our experiments. In order to assess personalization performances, we consider users who contributed at least 5 collections as test users. Among the overall 91 photo collections, there are 11 users who provided more than 5 collections (10 users contributed 5 collections, 1 user contributed 6 collections) which result in 56 collections totally. According to this, our dataset is split into two parts: one part contains 35 collections from 31 users, whereby each user provided at most 2 collections, which is called *general dataset*; another part contains 56 collections from 11 users, whereby each user provided at least 5 collections, which is called *personalized dataset*.

Evaluation Metrics. The evaluation metrics are the same as the ones reported in Section 2.4.5. In particular, we compute the precision for $n = 20\%$, which is indicated as P@20%, coherently with our user study where participants were asked to select the 20% most important photos from their collections. In order to assess the adaptation of our personalized model to users, we apply the personalization process described at the beginning of Section 2.5.1 to the collections of each user separately and average the P@20% among the test collections available at each iteration k, where k denotes the number of collections that are used for training the personalized model.

Parameter Settings. The classifier employed for importance prediction, built using the Support Vector Machine implementation of LibSVM[6], has Gaussian Kernels and has been trained via 10–fold cross validation on the training set. Note that the training set is expanded at each iteration (i.e. each time a new annotated collection of the user is provided), and the training via 10–fold cross validation is repeated each time. The open parameters were tuned via grid search and updated at each iteration. The ones identified for the *general dataset* were $C = 1.5$, $\gamma = 0.25$.

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Training and Test Sets

We evaluate the model update over different rounds of adaptation. The *personalized dataset* is split by users where each user owns 5 or more collections. At each iteration k, for each user with N collections, k collections are added to the initial training set to learn the personalized model of the user, and $N-k$ collections are used for testing. The ways in which the original training set is built are described later in this section.

We experiment all the values of k ($k = 0, 1, 2, 3, 4$), and for each of them we repeat the split and evaluation 5 times so that all the collections could be selected the same times as training collections. Note that the iteration $k = 0$ corresponds to the situation when the selection model is trained only on the initial training set. The selection strategy to select training collections is the following. When $k = 1$, we ensure that each collection of the user that we are considering is selected once as initial training data and the remaining four collections are treated as test data, then we average the performances. When $k = 2$, we pick two collections at each time from 5 collections, with the constraint that each collection could only be selected twice in all 5 repetitions (to be fair to all collections). We then average the performance achieved at each time. The cases when $k = 3$ and $k = 4$ can be done in the same manner. Finally, we average the performances over users for the same value of k.

The ways in which we build training sets are described hereafter. The model update and the split in train and test set described before are the same in each case.
Stand-alone. The initial model is trained with one random collection of the user, and the model update is incrementally done considering the remaining collections (starting from iteration $k = 1$, since the training set would be empty at $k = 0$).
Collaborative. The initial training set at $k = 0$ is formed by all the collections within the *general dataset*. This case represents the situation where, in absence of large amount of annotated personal data for training, annotated collections of other users are used to alleviate the cold-start problem.
User-agnostic. Similarly to the *collaborative* case, the *general dataset* is used as initial training set. However, at each iteration k, instead of including k collections of the user that we are considering, we add k randomly selected collections from the other test users. This case is motivated by the assumption that, if one collection, which is not from the user that we are considering, is included in the training set at each iteration, then the adaptation performances should be smaller than including collections that are from the user that we are considering. This would highlight the importance of incorporating selection information of the user in the training set when making selections for new collections of the same user.

Results

As a motivation to the need of personalization in photo selection, we trained a not personalized selection model on the *general dataset* and we tested its performances (P@20%) on the *personalized dataset*. From the results, we observed a large amount of

Table 2.4 P@20%, standard deviation, and performance gain of the personalized models at each iteration.

	k = 0		**k = 1**		**k = 2**		**k = 3**		**k = 4**	
	P@20%	**Δ**	**P@20%**	**Δ**	**P@20%**	**Δ**	**P@20%**	**Δ**	**P@20%**	**Δ**
Stand-alone	-	-	0.353 ± 0.060	-	0.374 ± 0.068	+5.9%	0.383 ± 0.067	+2.4%	0.402 ± 0.069	+5.0%
Collaborative	0.427 ± 0.057	-	0.430 ± 0.054	+0.7%	0.432 ± 0.055	+0.5%	0.437 ± 0.050	+1.2%	0.444 ± 0.061	+1.6%
User-agnostic	0.427 ± 0.057	-	0.427 ± 0.052	+0.0%	0.428 ± 0.055	+0.2%	0.429 ± 0.053	+0.2%	0.426 ± 0.055	-0.7%

variability in performances over the different collections, with precision values ranging between 0.190 and 0.722. The same pattern was observed when grouping collections by test users, although the differences in performances were less prominent. This suggests that the overall performances of the system could be improved by learning selection models personalized to each single user.

The results of our personalization procedure, considering the three different ways of constructing the training set described before, are shown in Table 2.4. Along with the precision when selecting the 20% of the original collection (P@20%) and its standard deviation over the test users, we also explicitly report the relative gain (Δ) obtained between two consecutive iterations. For instance, the Δ for $k = 3$ represents the relative gain in P@20% with respect to the one achieved for $k = 2$. It is possible to observe that the precision of both *stand-alone* and *collaborative* increases at each iteration, i.e. with the increase of the number of user's collections considered for training the model. This shows that having a selection model partially aware of the user preferences (by exploiting a certain amount of the selection behavior in the training phase) can improve the precision of new unseen collections of the same user. The precision of *collaborative* is higher than the one of *stand-alone*, especially at the first iterations, showing that the selection data from other users can alleviate the cold-start problem. The gain Δ of *stand-alone* at each iteration is higher than the one of *collaborative*, because the initial model is weaker (due to the limited training set) and the inclusion of new training collections has a higher impact on the learning. We also observe relatively high values of standard deviation in the table. Comparing *user-agnostic* and *collaborative*, the former exhibits an almost null gain in performances over iterations (it is even negative for $k = 4$), while the latter leads to a higher and increasing performance gain iteration after iteration. This shows that the increase of performance at each iteration is due to the inclusion of a new collection of the same user in the training set and not simply because the training set is expanded at each iteration, since in this case the gain of *user-agnostic* should have been higher as well.

As a conclusion, this evaluation led to promising results, showing that (i) including new annotated collections for the same user when training the model can benefit the selections on new unseen collections of the same user, and (ii) exploiting annotated collections from other users as initial training data can boost the system performances in cold-start scenarios. It is important to clarify that the standard deviation observed in these experiments was relatively high. This can be due to a mixture of aspects, such as a limited size of test set (both in terms of users and iterations) and intrinsic changes of difficulty among collections of the same user. For this reason, although a promising user adaptation emerged from this study, the inclusion of an extended amount of users, collections, and iterations would help make the results more evident and statistically significant.

2.6 Exploiting Additional Information

Besides personalization, another complementary way to improve the whole selection process is to extend the general, not personalized model with additional information, to come up with a more comprehensive description of the image's content. The goal is finding further information to model those selection patterns that are still hidden and have not been considered before. This enriched selection model would be also a better starting point for personalization itself. The newly extracted information is translated into different sets of features, which are added to the ones already available. In the rest of this section we describe the different extracted features and show the results that we achieved when including them in the learning process.

2.6.1 Feature Description

The kinds of information described hereafter are, to some extent, orthogonal to each other and together can give a more comprehensive description of the image's content. This information consists of low–level visual information, aesthetics, concept detection based on Deep Learning, emotional concepts, and face clustering.

Low-level Visual Information

We extract a set of 45 low–level visual features that have been previously employed for other image classification tasks (e.g. in [TLZ+04, WBT10]). The motivation of the employment of this kind of features is that the low-level visual content of an image represents basic visual signals (e.g. colors, textures, lines) that might unconsciously capture the attention and interest of the observer (in our case, the collection owner).

HSV Statistics. We represented pictures in the HSV color space and computed avg, std, min, max for each dimension.

Pleasure, Arousal, Dominance. A psychological study [VM94] showed that particular linear combinations of Saturation and Brightness fairly correlate with the sentiments of *pleasure*, *arousal*, and *dominance*. We compute such quantities as those sentiments might play a role in the selection process: $0.69V + 0.22S$ for pleasure, $-0.31V + 0.60S$ for arousal, and $0.76V + 0.32S$ for dominance.

Colors. We extract two kinds of color-related information, namely *colorfulness* and *color names*, to measure the variety and distribution of colors within an image. The former is measured by computing the Earth Mover's Distance (EMD) between the histogram of an image and the histogram having a uniform color distribution (one for each R,G,B channel). For the latter, we used the algorithm presented in [vdWSV07] to classify pixels into one of the 11 basic colors (black, blue, brown, green, gray, orange, pink, purple, red, white, yellow) and we counted the percentage of pixels for each distinct color.

Tamura Textures. We compute two Tamura features [TMY78] representing textural aspects such as *coarseness* and *directionality*. *Coarseness* is related to texture scales and reflects the size of the largest texture within an image. For each image point (x, y), we take the averages over neighborhoods of sizes $2^k \times 2^k$ (with $0 \leq k \leq 5$):

$$A_k\left(x, y\right) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} f\left(i, j\right) / 2^{2k}$$

where $f\left(i, j\right)$ is the gray–level at (x, y). Then, for each point (x, y) and for both horizontal and vertical dimensions, we consider the averages computed at not overlapping neighborhoods on opposite sides of (x, y) and compute the difference between them, i.e. $E_k(x, y) = |A_k(x+2^{k-1}, y) - A_k(x-2^{k-1}, y)|$ along the horizontal direction. Finally, for each point, we take the best neighborhood size $S_{best}(x, y) = 2^{k_{best}}$ which maximizes E in either direction and we compute the coarseness value for the whole image by averaging $S_{best}(x, y)$ over all the points. Regarding content *directionality*, it is computed in two steps. First, for each point, we take horizontal (Δ_h) and vertical (Δ_v) derivatives and we compute the local edge direction $\theta = tan^{-1}\frac{\Delta_v}{\Delta_h} + \frac{\pi}{2}$. Second, a histogram is constructed by quantization of the θ values in 16 bins and the overall directionality value is computed by summing the second moments around each peak (peak sharpness) in the histogram, as done in [TMY78].

GLCM Textures. A Gray–Level Co–Occurrence Matrix (GLCM) [Har79] represents how often different combinations of pixel intensities (gray levels) occur in an image and it is a well known means to extract textural features. Given the normalized GLCM $P(i, j)$ of an image, we compute *correlation*, *energy*, and *homogeneity* [HR04]:

$$\begin{aligned} correlation &= \sum_i \sum_j P\left(i, j\right) \frac{\left(i - \mu\right)\left(j - \mu\right)}{\sigma^2} \\ energy &= \sum_i \sum_j P^2\left(i, j\right) \\ homogeneity &= \sum_i \sum_j \frac{P\left(i, j\right)}{1 + \left|i - j\right|} \end{aligned}$$

where μ and σ are mean and variance of P, respectively.

Dynamics. Studies (e.g. [Itt73]) have suggested that the presence and slope of lines in pictures can trigger different emotions. For instance, horizontal lines are associated with calmness, while slant lines indicates dynamism. Therefore, we identify lines using the Hough transform and count the number and length statistics of *static* lines (horizontal and vertical) and *slant* lines (a line was classified as static if its angular coefficient was within $[-15°; 15°]$ or $[75°; 105°]$).

Skin. The amount of skin in an image is a signal of people appearance. We consider the color spectrum suggested in [LSK09] that represents the color of skin in the YCbCr color space and count the percentage of pixels belonging to it.

Image Aesthetics

Similarly to low-level information but moving to an higher level of abstraction, also image aesthetics might raise interest in the observer. It reflects how an image is well posed, attractive and pleasant to an observer, for instance considering how colors, shapes, and objects are arranged together. We took inspiration from previous works in computational aesthetics [YHBO10, MM15] to extract the following 6 features.

Rule of Thirds. It consists of splitting the image content in vertical and horizontal thirds and placing the main subjects at their intersections (power points). First, the main subjects are identified by segmenting the image, assigning a saliency score to each pixel according to [AHES09], and computing the saliency of each segment by averaging the scores of the pixels belonging to it. Second, the rule of thirds is measured by aggregating, over all the segments, their sizes A_i, saliencies S_i, and distances D_i to the closest power point:

$$\frac{1}{\sum_i A_i S_i} \sum_i A_i S_i e^{-\frac{D_i^2}{2\sigma}}$$

with $2\sigma = 0.34$ used for distance normalization. Intuitively, main subjects close to power points will make the feature value higher.

Simplicity. We compute two values to represent *simplicity*. For the first, we build the Region of Interest (ROI) map based on saliency (computed as for rule of thirds), we binarize it according to

$$ROI(i,j) = \begin{cases} 1, & \text{if} \quad S(i,j) > \alpha S_{max}, \quad \alpha = 0.67 \\ 0, & \text{otherwise} \end{cases}$$

and we sum the areas A_i of all the bounding boxes of the not overlapping saliency regions identified in the map $\sum_i \frac{A_i}{img\_size}$ as simplicity value. The second value, regarding simplicity as the "attention distraction of the objects from the background" [LT08], is calculated by separating subject and background regions and using the color distribution of the background to evaluate simplicity [YHBO10]. A histogram h of 4096 bins is generated from the RGB channels of the background (16 levels per channel) and the simplicity value is measured as $\frac{\|S\|}{4096}$, where $S = \{x | h(x) \geq 0.1 \cdot h_{max}\}$.

Contrast. We compute two measures of contrast, defined as the degree of diversity among the components of an image [YHBO10]. The first one is the Weber Contrast, which assesses it based on the diversity of intensity $I(x,y)$ within the image:

$$\frac{1}{X \cdot Y} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \frac{I(x,y) - \hat{I}}{\hat{I}}$$

where $\hat{I}$ is the average intensity. For color contrast, we segment images and use the CIEDE2000 color difference equation [SWD05], which estimates color disparity between each pair of segments based on their average colors and sums them together:

$$\sum_{i} \sum_{j=i+1} (1 - d(i,j)) \frac{c(i,j)}{A_i A_j}$$

with $d(i,j)$ and $c(i,j)$ being the distance and color dissimilarity between two segments i, j, respectively, and A_i, A_j their sizes.

Intensity Balance. Content balance can transmit equilibrium and calmness to whom is watching the picture. We assess balance in terms of pixel intensity, computing the Chi-Square difference between two intensity histograms, one for the left-hand and one for the right-hand part of the image.

Concept Detection with Deep Learning

Although Concept Detection has been already included in the selection model, the recent advances of Deep Learning for image classification tasks (see e.g. [GLO+16] for an overview) have induced us to experiment with it as well. To this aim, we use the GoogLeNet network [SLJ+15] pre-trained for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). GoogLeNet is a very deep convolutional neural network, made of 22 layers, where *inception modules* are stacked on top of each other and perform in parallel 1 × 1, 3 × 3, 5 × 5 convolutions, and max-pooling before concatenating the output of such filters into a single output vector becoming the input for the next module. The activations are subject to average-pooling. The approach that we used to extract concepts has been presented in [MMT+15] and produces two sets of concepts. The first one is obtained by using GoogLeNet as a standalone classifier, where the direct output of the network (corresponding to the final class label predictions for the 1000 ISLVRC categories) constitutes the returned set of concept scores. The second is made of the same 346 concepts already considered before in Section 2.4.3, but the input features used to train the concept detectors (SVMs) are different. GoogLeNet has been applied on the keyframes of the TRECVID 2013 dataset (800 hours of video) and the 1000 output values have been used as features for each of the 346 SVMs, which were trained separately for each target concept.

Emotional Concepts

The concepts considered so far in the selection model have almost always a neutral meaning and interpretation. Concepts like animal, building, beach do not directly suggest any particular positive or negative sentiment. However, being personal pictures possibly characterized by memories and personal affections, sentiments could indeed matter in our scenario. In order to introduce emotional and sentimental aspects in the selection model, we applied the concept detectors available in SentiBank [BCJC13] to model the presence of 1200 Adjective Noun Pairs (ANP) in images. The output value of each detector represents the probability for the related ANP to appear in the photo. That approach can be summarized as follows. The 24 emotions defined in Plutchik's Wheel of Emotions [Plu80] were used to retrieve images and videos from Flickr and YouTube. From the tags of the retrieved images and videos, nouns (concepts) and adjectives carrying sentiments are combined together to form ANPs. So, by such construction, ANPs are formed by a noun, which represents a neutral concept, and an adjective, which instead associates a particular emotion to the concept. For instance, for the same neutral concept *eyes*, the concept set contains ANPs like *gorgeous eyes*, *sad eyes*, and *tired eyes*, each of them carrying a different emotional impact for the same concept. Finally, individual detectors (Linear SVMs) for each ANP are trained using Flickr images tagged with a specific ANP, and the 1200 most accurate ones have been retained and publicly released. In this set there are 286 distinct "neutral" concepts (e.g. airport, beach, office) combined with 181 distinct adjectives (e.g. pretty, ancient, scary) to give the formers emotional connotations.

Face Clustering

Face detection is a way of assessing people appearance in photos and has been already considered in the selection model. However, it does not reveal any clue about the relationships between the owners and the people appearing in their collections. Obtaining this kind of information often requires a certain investment of the user in tagging and annotating, as well as the awareness of social relationships, which are both not assumed to be available in the considered scenario. We therefore adopt a simpler and unsupervised way of estimating the "role" of a given face within the collection, which is based on how much that person is popular in it (in terms of occurrence frequency). A person related to who took the photos, e.g. a friend, husband, wife, colleague, will probably occur many times in the collection. On the contrary, random people appearing by chance, e.g. in outdoor crowded environments, will have a low occurrence frequency.

The face clustering technique presented in [PMS+16], which is a variant of [SKP15], has been applied on the faces detected as in Section 2.4.2. We summarize the method hereafter, and we refer the reader to [PMS+16] for an extensive description. A DCNN (4 convolutional layers each one followed by a pooling layer) is employed in a Siamese architecture [BGL+93], where the network is replicated twice so that it can handle pairs of faces to learn their difference along with their individual representations. Each branch of the network produces a 10-dimensional vector as feature representation for the input face, and the network is trained based on the Euclidean distance of the representations in each face pair. After the initial training on a face dataset, the network weights are fine-tuned by re-training the network imposing constraints based on the hair and costume regions adjacent to the faces in order to increase the clustering accuracy. Finally, the learned face representations are then used as input to the clustering method (modified version of the Rank Order Distance based clustering presented in [ZWS11]).

Table 2.5 Precision of the expectation-oriented selection enriched with additional features.

	P@5%	**P@10%**	**P@15%**	**P@20%**
Expo				
quality	0.3431	0.3261	0.3204	0.3168
faces	0.4506	0.3968	0.3836	0.3747
concepts	0.5464	0.4599	0.4257	0.4117
all	**0.7124**	**0.5500**	**0.4895**	**0.4652**
Expo++				
low level	0.4399	0.3913	0.3729	0.3697
aesthetics	0.4406	0.3923	0.3732	0.3639
face popularity	0.4692	0.4101	0.3977	0.3945
concepts (DCNN)	0.5694	0.4945	0.4553	0.4436
concepts (SentiBank)	0.6124	0.5172	0.4674	0.4502
all	**0.7426**$^{\triangle}$	**0.6155**$^{\blacktriangle}$	**0.5330**$^{\blacktriangle}$	**0.5121**$^{\blacktriangle}$

Each detected face cluster represents one distinct person and contains all the occurrences (faces) in the images within a collection. We leveraged this information to derive features about the popularity of faces and then to have aggregated measures of the popularity of an image. First, for each face, we compute its popularity as the size of the face cluster it belongs to (normalized by the total number of faces in the collection). Second, for each image, we consider the popularity values of all the faces contained in it and compute statistics (avg, std, min, max) about them.

2.6.2 Results

Finally, we report the performances of the selection model when using the different previously described sets of features within the learning process. The experimental setup is the same one used for the evaluation of the original selection model (Section 2.4.5). For feature selection (Information Gain), we used the top–200 features when training with the whole extended set, while the top–160 were kept when training only using concept information (also for emotional concepts). No feature selection was applied to the other smaller feature sets. The hyper-parameters of the SVM trained with the expanded set of features are $C = 2.0$ and $\gamma = 0.5$, while those of the Siamese network for face clustering were chosen according to [ZWS11]. Caffe[7] was used to work with the DCNNs for concept detection and face clustering.

The results are listed in Table 2.5, distinguishing over different subsets of features. The values referring to the additional sets of features are under the name *Expo++* and we also report the results of the previous feature sets (Section 2.4.6) for the sake

[7] http://caffe.berkeleyvision.org/

of comparison. The *Expo++* model incorporating *all* the additional features (top–200) outperforms the previous *Expo* model for all the selection sizes k, with relative improvements ranging from 11.9% (P@10%) to 4.2% (P@5%). These improvements have been proved to be statistically significant. This shows that expanding the selection model with a more variegate set of features does help improve the selection precision.

We now discuss the performances of the different subsets of features in isolation. Regarding the semantic information, both *concepts (DCNN)* and *concepts (SentiBank)* improve the precision of the *concepts* features: the presence of both more precise concept detectors thanks to the employment of Deep Learning (*concepts (DCNN)*) and a set of concepts carrying sentiments and emotions (*concepts (SentiBank)*) helps in the selection task. It is important to make a distinction between the precision in the concept detection task and the one regarding our photo selection task. When using the image representations learned by GoogLeNet as input features to the concept detectors (SVMs), we observed a relative improvement in mean average precision of over 40% with respect to the previous concept detectors relying on hand-crafted features (we do not report these results here because the advances of concept detection per se is not the main topic of this work). However, the improvements that we get when applying these more accurate detectors on our photo selection task are more limited. We believe that this is due indeed to the difference between the two tasks: concept detection is fairly objective, while photo selection (especially with personal data) can be characterized by the interconnection of hidden factors such as memories, emotions, personal attachment and tastes. Therefore, no matter how accurate the concept detection can become, its contributions to the task of personal photo selection would saturate at some point.

The inclusion of face clusters information to assess face popularity also exhibited a slight improvement over the *faces* features alone, although popularity features were expected to have a stronger impact. This might be due to the way we estimated people popularity, which is done based on face clusters in an unsupervised manner. The introduction of more complex person recognition techniques, for instance to know who is actually appearing in the pictures and what the relationship with the collection owner is, might probably bring more substantial improvements. However, this would also require an additional investment of the user in tagging and annotating, as well as the awareness of social relationships that might not be available in case of personal data (i.e. out from social networks). Finally, both *low level* and *aesthetics* features resulted to be more useful than the mere *quality* features extracted via quality assessment, but still their performances are lower than the ones of the other features sets (especially the ones related to concepts). This is a further confirmation that, for the task of photo selection from personal collections, the semantic and emotional aspects are dominant with respect to those related to surface visual content and aesthetics.

2.7 Conclusion

In this chapter, we considered the problem of keeping personal photo collections enjoyable over time. Given the explosion in the production of digital pictures within the recent years and the common practice of merely dumping such data on cheap storage devices or using storage services, the stored photo collections are rarely accessed and revisited afterward. To some extent, their content tends to be forgotten because the big collection size makes their revisiting a fatiguing process. As a remedy, we proposed a selective approach to long-term data management that aims at identifying what is most important to the user and investing in the longevity and enrichment of this content, in order to make the future revisiting more enjoyable and less tedious. The development of such automated method was preceded by a user study, to lay the foundations of the task and better understand its challenges.

The user study was centered around a photo selection task where 35 participants contributed their own photo collections and selected from them those most important to them, namely the pictures that they would like to preserve for future revisiting. One important outcome was that many hidden and subjective criteria (memory evocation, personal importance, image typicality) were rated high, anticipating the difficulty of automatizing the selection task. Moreover, the more objective criterion of image quality was rated as less important. Another aspect emerged as important was coverage, which means that the set of selected pictures should fairly represent the content of the original collection. Although this was stated by the participants, their selections exhibited a poor degree of coverage.

Afterward, we presented a method for photo selection exploiting an extensive set of image- and collection-level features, to estimate the long-term photo importance based on user expectations. The evidence of user expectations has been derived from the personal data provided during the user study and has been used to train the selection model. The goal of this method is supporting users in selecting the most important photos for creating an enjoyable sub-collection of a personal collection for preservation and revisiting purposes. Since a wide part of state of the art methods is driven by the concept of coverage, which resulted to be highly rated in our user study as well, we also investigated how to combine the expectation-oriented selection with more explicit modelings of coverage. Experiments with real world photo collections showed that (i) our method outperforms such state of the art works when considering human selections as evaluation criterion and (ii) comparable results to our method can be achieved only when coverage is not considered as a primary selection aspect.

Finally, we investigated the personalization of the general selection model to each user. Promising adaptation patterns emerged from the analysis, especially on the utility of incorporating user feedback into the selection model as well as on the exploitation of training data belonging to other users in cold-start scenarios, while a relatively high standard deviation indicated the need of further data to make such adaptation effects more prominent.

3 Validating Event-related Information in Text

Verifying whether an event has truly happened is an onerous and time consuming task, which comprises the manual inspection of a trustworthy source of textual information to check whether the given event is reported in it. Such manual examination is not a scalable approach and it becomes infeasible in case the events to be scrutinized are continuously gathered at high rates, e.g. by automatic methods. In this chapter, we introduce *event validation* as the task of determining whether a given event occurs in a given document or corpus, along with methods to perform it automatically. The employment of automatic event validation can boost the precision of the input event set, discarding false events and preserving the true ones. Also, it can find documents to corroborate and explain the occurrence of those events.

We propose two novel methods to address event validation. The first one validates events by estimating the temporal relationships among their participants within documents via a set of hand–crafted validation rules. The second one learns to identify event occurrences by means of supervised Machine Learning. Our experiments showed that the latter method is more accurate and achieved a *substantial* level of agreement (according to [LK77]) when compared with validations done by humans. Moreover, when applied as a post-processing step of event detection, it increased the precision within the set of detected events while preserving recall.

The ground truth used for evaluating the methods is made of real-world events and documents and has been subject to crowdsourcing. Events and documents are coupled in pairs, whose validity has been judged by human evaluators based on whether the document in the pair contains evidence of the given event. In contrast to the notion of relevance considered in available datasets for event detection, validity judgments in our work strictly consider whether a document reports an event within its timespan as well as the number of event participants reported in the document. These requirements make the generation of manual validity judgments an onerous procedure.

3.1 Introduction

Events are the crucial building blocks of all forms of news media. Since a large amount of online space is consumed in describing and discussing events, they are embedded within news articles, forums, blogs and different online social media. Events and their descriptions can provide concise summaries of news articles, making the search for information more enjoyable for general readers and more effective for professionals like historians and journalists. Also, event-related information can be a valuable known input for tasks like document enriching, indexing, ranking, and summarization [SG12, NPD17, BHDR11, MB16]. This has produced a considerable interest and research effort in the detection of real-world events from natural language text, a task that has been historically denoted as *topic detection and tracking* [APL98] and is out of the scope of this chapter. In fact, we do not focus on how events are detected or extracted from text but we are rather interested in assessing their verity and occurrence within a given document or corpus. We refer to this problem as *event validation*. Manually assessing whether an event occurs in a document collection is a cumbersome task and becomes infeasible in scenarios where events are continuously and automatically detected from news streams on a large-scale.

In this chapter, we propose to automate event validation by learning to assess the occurrence of events in a given unannotated corpus. We model an event as a set of participants related within a given time period. This is in line with both the basic definition of *event* given by the Topic Detection and Tracking (TDT) project ("something that happens at some specific time and place") [APL98] and with more recent works on event detection, e.g. [DSJY11, MMJ13, FBJ15]. For instance, the event $\{($*Novak Djokovic*, *Roger Federer*, *US Open*$)$, *30/08/2015 to 15/09/2015*$\}$ represents the participation of two tennis players in the 2015 US Open. Given a set of events, an effective event validation method would allow to (i) reduce the number of false events with respect to a given corpus, and (ii) find documents to corroborate the occurrence of events and enrich available knowledge bases (e.g. [HSBW12]) with such event descriptions. We particularly tackle the former application, introducing event validation as a post processing step of event detection to boost the precision within the detected set of events. This has to be accomplished without affecting recall, i.e. preserving accurately detected events.

Being able to handle true events and ignoring the false ones is of crucial importance to ensure the quality of any application that uses event-related information for any other purpose. Of course the first and most intuitive remedy to such problem consists in making event detection algorithms themselves more and more accurate. However, we claim that event validation could improve the quality of a given set of detected events because it can exploit information, such as participants and timespan of the detected events, which is clearly not available as input to event detection (it is actually what it produces as output).

Since the verity of occurrence of an event depends on the considered documents in a corpus, we distinguish two levels of event validation: *document-level validation* validates the occurrence of an event in a given document, while *corpus-level validation* considers the occurrence within the entire corpus. This distinction is compliant with the two aforementioned use cases: a user or application would benefit by either handling a cleaner event set (corpus-level) or having more available information related to the event (document-level).

We develop and compare two novel methods to validate events. In the first one the occurrence of events in documents is assessed by considering the presence of dates within the event timestamps, estimating the regions of text associated to these dates, and returning the percentage of event keywords present in these regions. This approach is based on hand-crafted validation rules. The second one extracts features from event-document pairs and, based on them and validity labels for each pair, it employs supervised Machine Learning to learn a model for judging whether an event occurs in a given document. Given the lack of assumptions on the nature of events and documents, our model for event validation can be applied to a wide range of events and related corpora.

We also present a novel dataset for training and evaluating the performances of event validation methods, consisting of 250 real-world events and 6,457 candidate documents used as a base for assessing event occurrence. Events and documents are coupled in pairs and are associated with validity judgments, indicating the percentage of participants acting together within the event timespan in the document. These validity judgments were assigned by human evaluators via crowdsourcing. The dataset has been publicly released to foster advances and comparisons in event validation and other related fields.

The contributions of the work presented in this chapter are summarized hereafter: (i) the definition of the *automatic event validation* task, whose model is applicable to a wide range of scenarios due to the few assumptions regarding the handled data; (ii) a novel and effective approach to the task based on supervised Machine Learning, which exhibits a substantial level of agreement when compared with human evaluators; (iii) the introduction of automatic event validation in cascade to event detection, resulting in a significant boost of precision with a low decrease in recall within the set of detected events; (iv) the release of a benchmark for event validation, where the occurrence of events in documents is assessed by taking both relations among event participants as well as their temporal validity into account.

The rest of the chapter is structured as follows. In Section 3.2, we outline existing works and benchmarks related to event validation. Section 3.3 contains the definition of the event validation task. In Section 3.4 we describe a benchmark for event validation (and possibly also usable for event detection) while our automatic approaches are presented and compared in Section 3.5. Finally, in Section 3.6 we summarize and conclude the chapter.

3.2 Related Work

In this section we discuss existing work on topics related to event validation, such as event detection, fact validation, fake news and rumors detection. Since one of the contributions of this chapter is a novel dataset explicitly created for event validation, we also describe available datasets for event detection and show that they do not meet the requirements of event validation.

3.2.1 Event Detection and Validation

The automatic detection and tracking of events from natural language content has been widely studied in the past decades. Most classical approaches come from the Information Retrieval community and include exploiting generative models to cluster documents based on semantics and timestamps [APL98, LWLM05], as well as discovering events by studying distributions of their entities over documents and time [DSJY11, FYYL05, HCL07]. Differently, the majority of the works within the NLP community have been focusing on the extraction of events as more fine-grained relations among entities, time, places. Such approaches encompass, among others, dependency parsing [MSM11], transductive inference [HZM$^+$11], event coreference resolution [BH10, CN16], structured prediction [LJH13], and convolutional neural networks [NG15, CXL$^+$15].

We do not further delve into the review of the event detection literature because this topic, while being evidently related, does not constitute the main interest of this chapter. We claim that event detection is different from event validation, since the output of event detection methods (i.e., a set of *events*) constitute the known input for event validation methods. Events are used as known information also in other works. For instance, in [HFG$^+$12] temporal references in documents are attached to input events, while [SGG$^+$16] presents a method for answering event queries made of imprecise geographical or temporal information (i.e. at a coarse granularity). In our work we focus on the validation of events in a corpus, independent of how the input event listings are acquired. Although the task of event validation with respect to a corpus can be posed as being analogous to an *Information Retrieval* task, we find that checking the mere appearance of event participants in text (e.g. via keyword matching) is insufficient to validate the occurrence of events in documents while establishing mutual relationships and temporal conformation. Available system implementations do not tackle event validation, but they rather consider other applications like event extraction [KW14], tracking [VdVBM15], retrieval of event-related information [MRRG$^+$13], visualization [MTY$^+$14], and retrospective exploration [MB15].

3.2.2 Fact Validation

Fact validation, belonging to the broader task of textual entailment recognition, has drawn research effort in the last decade [DGM06, MMS+14] and is probably the topic most related to event validation. It consists in identifying whether a piece of text is entailed from a given set of sentences relevant to it, i.e. a human reading the latter set would deduce that the former text is true. The majority of the approaches to fact validation [MMS+14] exhibit two main differences with respect to our problem definition and approaches: first, input facts can be general statements and claims, which do not necessarily carry event-related information; second, the temporal dimension is either ignored or not considered in a strict way. Araki et al. [AC14] performed historical fact validation by casting the problem as *Passage Retrieval*. Differing from our problem definition and approaches, they assess event validity in terms of the textual similarity (ignoring temporal information) between facts and passages of fixed length, thus assuming the evidence of facts in documents to be restricted to the pieces of adjacent content. Moreover, their work is focused and evaluated only on historical facts, while we aim at validating events of different granularities and topics. A more recent approach has been presented by Samadi et al. [STVB16], where Probabilistic Soft Logic (PSL) is used to jointly evaluate claims and estimate the trustworthiness of the sources (Web documents) used as ground truth for validation. The estimation of source credibility is an interesting matter, which our methods do not address. While the overall evaluation workflow is similar to ours, the main differences mentioned before still hold in this case: input claims (e.g. "chicken is healthy") are not events and, therefore, the temporal dimension is not considered within the validation.

Validating facts is also regarded as an important task for maintaining knowledge bases. The work presented in [LGMN12] validates RDF triples by looking for evidence of their validity within textual documents in the Web, eventually associating a confidence score to an input fact and collecting a set of excerpts mentioning it in natural language as proof of its validity. Besides the input being represented as RDF facts, the most crucial difference with respect to our approaches is that the the temporal dimension is not considered in this approach. Its extension [GEL+15] does include temporal information as part of the validation. However, the time granularity is broad (years) and temporal validity is assessed in cascade by further processing the proof excerpts collected by [LGMN12], while in our methods participants occurrence and temporal validity are considered jointly. Finally, [Leb17] presents a query language based on first-order logic to represent facts and claims along with algorithms for query answering to check their validity within a given ontology (ground truth). This work assumes a formal description of both input facts and ground truth based on first-order logic, while our methods work on rough text in natural language.

3.2.3 Fake News and Rumor Detection

Fake news and rumors are forms of false information. The former are "news articles that are intentionally and verifiably false" [AG17], while the latter are commonly referred to as "unverified and instrumentally relevant information statements in circulation" [DB07]. They are created and spread for reasons like political propaganda, sarcasm, satire, or merely to generate confusion [RCCC16, Wan17, AG17]. While the detection of fake news and rumors can be considered as similar to the one of facts and claims discussed before in Section 3.2.2, as they are all forms of misinformation, there is one substantial difference coming from the digital environments, such as social networks and blogs, where fake news and rumors are created and spread. The content in such platforms can be generated by any user with almost no moderation, with the result of being highly opinionated and potentially untrustworthy [CMP11]. Most of the approaches to fact validation, instead, rely on a trustworthy ground truth as a basis for the validation. Of course, the presence of fake information in such ground truths could harm the results of fact and event validation. Although we do not assess source trustworthiness in our approaches, we did not observe the presence of rumors in our dataset due to the exclusion of content coming from social media, which is regarded as one of the biggest sources of misinformation [Wat05, CMP11]. Also, there is a rich network of user relationships and interactions that is usually not present in the context of fact validation. Another difference particularly with respect to the definition of event validation is that our notion of validity is strictly temporal, i.e. we are interested in determining whether events occurred strictly within the claimed timespan, while the goal of rumor detection is primarily the assessment of their verity regardless of when they happened.

We now provide a brief overview of existing works on fake news and rumor detection. One group of approaches aim at determining the veracity of a given textual excerpt only based on its linguistic characteristics such as, among others, distributions of parts of speech [ZBNT04], syntax [FBC12], story coherence via rhetorical structure theory [RL15], satirical cues [RCCC16]. However, the majority of the works go beyond the exploitation of linguistics alone. Castillo et al. [CMP11] studied information credibility on Twitter using a set of features extracted from message content, users profiles, topics, message propagation, which has been expanded by many following works (e.g. by [QRRM11, YLYY12, WYZ15]). Finally, other approaches investigate the dynamics of rumor propagation given the available social media data. These are based on temporal properties of tweet volume [KCJ+13], variations over time of the rumor descriptions either based on hand-crafted featured [MGW+15] or learned via Recurrent Neural Networks [MGM+16], and learning the structure of rumor propagation [MGW17].

3.2.4 Datasets

Available corpora for event detection like the TREC[1] and TDT[2] datasets are related to ours as they consist of annotated event-document pairs. The TREC 2014 Temporal Summarization Track used a set of 15 events along with related documents belonging to news streams. In the TREC Web and Ad-hoc Retrieval Tracks input queries (topics) can be regarded as high level events, and relevance judgments for documents are provided. The TREC Novelty Track dataset contains 25 event and 25 opinion topics, each one including 25 relevant documents and irrelevant ones in addition. The TDT5 corpus consists of 15 news "sources" in different languages, where documents are annotated with topic relevance judgments considering 250 topics (events), 126 for the English portion. This corpus is sparsely labeled: among over 250,000 English news articles only around 4,500 are annotated with topic relevance.

The main difference with respect to our dataset resides in the meaning of topics and relevance judgments, since these corpora were designed for ad-hoc retrieval. Input topics can be regarded as high-level named events (e.g. "Costa Concordia disaster") and relevance judgments are given based on the classical query-document relevance in IR. Differently, the events in our dataset are characterized by a set of participants related together within a given time period, and the validity judgments for each event-document pair represent the degree of occurrence of the event within the document, measured with the number of participants that conformed together in the document, strictly within the event time span. Acquiring annotations by inspecting text and checking for mutual relationships and temporal constraints is a more onerous process than producing classical relevance judgments. Moreover, our dataset also contains false events (due to either unrelated participants or wrong time span) and partially true events, where one or more intruders are present along with the true event participants. This allows to test event validation methods not only in case of true and clean events, but also for false or ambiguous ones. The amount of manually annotated data in our dataset is comparable to the one in the above mentioned corpora.

McMinn et al. [MMJ13] presented a corpus with more than 500 events and related tweets as a ground truth for event detection on Twitter, with an event model close to ours. Besides the different nature of documents (tweets instead of news articles), this work considers a more general notion of relevance, which does not count for either the number of event participants or the temporal validity.

Other works focus on collecting large sets of events, like YAGO2 [HSBW12], DBpedia[3], GDelt[4] and Wikipedia Current Events[5], without particularly focusing on relations between events and supporting documents. Kuzey et al. [KVW14] present methods for populating knowledge bases by extracting and organizing named events from news corpora. The ground truth used to evaluate the grouping of documents into events consists of around 100 named events and 1600 articles in Wikinews and news

[1] http://trec.nist.gov [2] http://www.itl.nist.gov/iad/mig/tests/tdt [3] http://wiki.dbpedia.org
[4] http://gdeltproject.org/ [5] https://en.wikipedia.org/wiki/Portal:Current_events

sources referenced in Wikipedia articles. Moreover, such ground truth is built without reporting mutual conformation of event participants with temporal constraints.

3.3 Overview

In this section we introduce the concepts and notations that will be used in the rest of the chapter, namely our data model (Section 3.3.1) and the Event Validation task (Section 3.3.2). We also elaborate on how event validation can be exploited to increase the precision of event detection algorithms (Section 3.3.3).

3.3.1 Data Model

Event. An *event* e is a tuple $e := (K_e, t_e^0, t_e^f)$, where $K_e = \{k_e^1, \ldots, k_e^{n_e}\}$ is a set of keywords representing the participants of the event, t_e^0 and t_e^f indicate the timespan within which the event occurred. Although this is in line with event definitions used in previous works [DSJY11, HCL07, MMJ13, FBJ15], other representations like named events, temporal subject-object predicates [HSBW12], or sentences in natural language can be adapted to our model by aligning the keyword sets accordingly.

Document. A document d is defined by its textual content that is subject to scrutiny in order to assess the validity of an event.

Evidence. An event e is said to have evidence of its occurrence in a document d with a threshold $\tau, 0 < \tau \leq 1$, iff at least $\tau\%$ of the keywords K_e participate together in an event reported in d, strictly within the timespan $\{t_e^0, t_e^f\}$ of the event.

Pair. Given an event e and a document d, we logically couple them within an *(event, document)* pair $p := (e, d)$.

3.3.2 Event Validation

We now present a general model for event validation, which we use as reference for our automatic approaches (Sections 3.5.1 and 3.5.2). We distinguish between two levels of event validation: document-level and corpus-level. An (e, d) pair may be judged as *invalid* if the document d contains no evidence of the event e, although the event might be true with respect to other documents in the corpus. Conversely however, a pair that is judged as being *valid* may sufficiently indicate the validity of an event at the corpus-level. In both our automatic approaches we design the event validation at corpus-level by combining validation decisions made at the document-level. It is important to clarify that we perform event validation with respect to either one document (document-level) or a collection of documents (corpus-level). That is, we do not aim at stating whether an event is true or false in general, but whether it is valid with respect to its occurrence in a document or a collection.

Document-level Validation. Given a pair p and an evidence threshold τ, we define the *document–level validation* as a function

$$\gamma_d = V_d\left(p, \tau\right) \tag{3.1}$$

which determines the evidence γ_d of e in d with threshold τ. The codomain of $V_d\left(\cdot\right)$ can vary based on the application requirements. In case a binary classification between valid and invalid pairs based on the threshold is required, then $\gamma_d \in \{valid; invalid\}$; otherwise, one can measure evidence as the percentage of conforming event keywords without considering any threshold, and $\gamma_d \in [0, 1]$.

Corpus-level Validation. Given an event e, a set of documents D, and an evidence threshold τ, we define the *corpus–level validation* as a function

$$\gamma_c = V_c\left(e, D, \tau\right) \tag{3.2}$$

which determines the evidence γ_c of e in the corpus D. Just like at the document-level, the codomain of $V_c\left(\cdot\right)$ depends on the application requirements. The function depends on which corpus is used as reference to validate the occurrence of the given event. While an intuitive way of performing corpus-level validation consists in combining validation decisions taken for each document, which is indeed the procedure implemented by our automatic methods, corpus-level validation could be in principle process a whole set of documents without explicitly resorting to the document-level validation.

3.3.3 Precision Boosting

We introduce automatic event validation as a post-processing step of event detection to boost its precision. Given the set of events originally detected, only the ones judged as valid by our validation at corpus-level are retained in the set. In addition to advances introduced within the event detection itself, we believe that automatic event validation can boost the overall performances by processing the detected events in a dedicated way, exploiting information that is unavailable as input to event detection (i.e. participants and timespan of events). Event detection usually identifies events in a document collection based on a vague input, represented by a wide set of entities [DSJY11] or terms with high occurrence frequency [FYYL05, HCL07]. In contrast, we do not intend event validation to discover new events, since entity relationships and their durations are already suggested by event detection in form of candidate events. We instead focus on verifying their occurrence in documents.

We call this effect *precision boosting*, because the goal is to increase the precision within the set of detected events by discarding invalid events. However, this has to be accomplished without affecting recall, i.e. the valid events originally detected by the event detection phase should not be discarded during the validation. The benefits of this approach will be shown in Section 3.5.3.

3.4 A Benchmark for Event Validation

Our publicly available dataset[6] complies with the data model and problem definition given in Section 3.3: events and documents are coupled into pairs, whose validity is assessed based on if and how many event participants conform together in the document within the event timespan. Such validity judgments have been assigned by human evaluators via crowdsourcing.

Available datasets, for example those released within the context of TREC and TDT, are built mostly for ad-hoc retrieval, and therefore their high-level topics and corresponding relevance judgments are not suitable for event validation. Moreover, the validity judgments available in our dataset strictly take the number of event participants related together in the document and their conformation to the event timespan into account. Reading a document to assess (i) whether it contains a given event, (ii) how many event participants are mentioned and related together, and (iii) whether such relationships can be associated within the timespan of the event, is more complex than assessing the overall relevance of a document to an event/topic. The different judgment criteria and the effort to produce them make our ground truth more valuable. Event detection methods relying on the same event model as ours (e.g. [DSJY11, HCL07, FBJ15]) can use our dataset as ground truth. Moreover, given the high focus on the temporal aspect and conformation during the event timespan, our dataset can be relevant to any event-related research or analysis involving the temporal dimension in a strict manner.

To the best of our knowledge, this dataset is the first publicly available corpus for event validation, where the occurrence of events in documents is assessed by taking both relations among event participants as well as their temporal validity into account. In the rest of this section we describe how events, documents, and their validity judgments have been gathered.

3.4.1 Events

Each event is made of (i) a set of participants, and (ii) a start and end date, indicating the timespan within which the event occurred. This follows event definitions used in previous works [DSJY11, HCL07, MMJ13]. We applied the algorithm introduced by Tran et al. [TCG+14] to detect events, working on the Wikipedia Edit History of more than 1.8 million Wikipedia pages representing persons, locations, artifacts, and groups. Titles of Wikipedia pages are considered as event participants. The considered time period spans from 18^{th} January 2011 to 7^{th} February 2011. We chose this period because it covers newsworthy events, such as the Arab Spring, the Academy Awards Nominations, the Australian Open, the Super Bowl. The minimum granularity of event duration, a parameter of the applied algorithm, has been set to one week.

[6] http://github.com/xander7/JustEvents

Table 3.1 Overall statistics of the event set.

Events	250
Distinct participants	456
Participants per event	2.94 ± 1.4

Table 3.2 Overall statistics of the document set.

Documents	6,457
Avg document length (char)	5,428
Documents per event	25.8 ± 7.4

In total, we detected 250 events, whose main characteristics are listed in Table 3.1. The distribution of the events over different categories, along with examples, is reported in Table 3.3. These categories were assigned manually based on the inspection of each of the 250 events. The considerable fraction of events related to sport (35%) is due to the actual occurrence of popular and newsworthy events within the considered time period, such as the Australian Open tennis tournament, the Super Bowl, and the Freestyle World Ski Championships. Moreover, complex events lasting many days or even weeks (such as the Australian Open) can trigger the detection of different sub-events within them.

Since events have been detected automatically, the event set also contains false events (due to either unrelated participants or wrong time span) as well as partially true events, where one or more intruders are present along with the true event participants. These events have been retained in the set and were also subject to manual evaluation since event validation has to deal with not only true and clean events, but also with false or ambiguous ones. Our comprehensive dataset thus supports evaluation of event validation methods for all potential cases, and contains a corresponding ground truth for them. This aspect will be further discussed in Section 3.4.3.

3.4.2 Documents

Documents in our dataset consist of Web pages that have been subject to scrutiny in order to assess the validity of events. We chose the Web as a source for documents due to its easy accessibility and wide event coverage. For each event, queries have been constructed by concatenating the name of event participants along with the months and year covered by the dataset (one distinct query for January 2011 and another for February 2011). We used the Bing Search API to perform queries and to retrieve the *top-20* Web pages for each query. Plain text has been extracted by using BoilerPipe[7], while Stanford CoreNLP[8] has been used for POS tagging, named entity recognition, and temporal expression extraction. After removing duplicates and discarding both non-crawlable Web pages and those with no content extractable by BoilerPipe, we have 6,457 documents corresponding to the 250 events. Titles and URLs of documents are provided along with plain texts. Some overall characteristics of the document set are summarized in Table 3.2.

[7] http://code.google.com/p/boilerpipe/

[8] https://stanfordnlp.github.io/CoreNLP/

Table 3.3 Distribution of events over different categories, with corresponding examples.

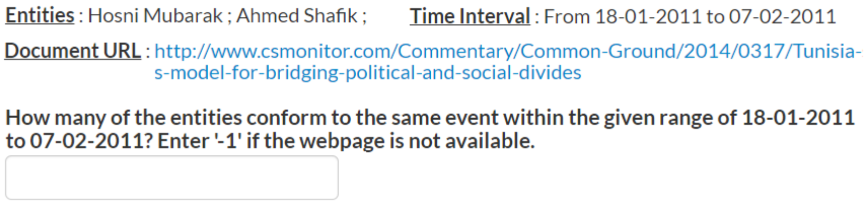
Category	%	Examples
Cinema	13%	{James Franco, Academy Award for Best Picture, 127 Hours, 83rd Academy Awards}, 25/01/2011 - 31/01/2011
Music	7%	{Jessie J, Price Tag, Who You Are}, 25/01/2011 - 31/01/2011
Nature and Disasters	3%	{Rio de Janeiro, Floods, Mudslides}, 18/01/2011 - 24/01/2011
Sport	35%	{Kim Clijsters, Li Na, Caroline Wozniacki, Australian Open, Svetlana Kuznetsova}, 25/01/2011 - 31/01/2011
Politics	14%	{Gamal Nasser, Ahmed Shafik, Smartphone, Cairo, April 6 Youth Movement, Gamal Mubarak, National Democratic Party}, 18/01/2011 - 07/02/2011
Science and Economics	4%	{World Economic Forum, Rosneft}, 25/01/2011 - 31/01/2011
TV and Entertainment	16%	{John Cena, Booker T, Royal Rumble}, 25/01/2011 - 31/01/2011
Other	8%	{Andy Gray, Loose Women, Richard Keys}, 25/01/2011 - 31/01/2011

Although the latest content of a Web page can be retrieved at any time via its URL, it might be different from the one considered at the time of the evaluation and available in our dataset. Therefore, validity judgments have to be related with the stored content of Web pages, not with the available content according to their latest versions. Moreover, due to the extraction of plain text via BoilerPipe, the stored content might slightly differ from the one visible in the Web pages and considered at the time of the evaluation.

3.4.3 Validity Judgments

To manually evaluate the validity of the 6,457 (event, document) pairs in the dataset, we decompose the task of assessing whether or not a document contains evidence of the occurrence of an event into atomic units and deploy them on CrowdFlower[9], a premier crowdsourcing platform. For each pair, workers were presented with the event (participants and timespan) and the document URL. The event timespan, specified by a start and end day, was strictly considered during the tasks. The workers were then asked to report the number of event participants conforming to the same event in the document and within the event timespan (see Figure 3.1 for an example).

[9] http://www.crowdflower.com/



Entities : Hosni Mubarak ; Ahmed Shafik ; **Time Interval** : From 18-01-2011 to 07-02-2011

Document URL : http://www.csmonitor.com/Commentary/Common-Ground/2014/0317/Tunisia-s-model-for-bridging-political-and-social-divides

How many of the entities conform to the same event within the given range of 18-01-2011 to 07-02-2011? Enter '-1' if the webpage is not available.

Figure 3.1 Sample question from the crowdsourced microtask.

Workers also had the possibility to specify whether the Web page was not available and whether the temporal bearings of the document were unclear. We followed task design guidelines to engage the workers [MS13] and employed gold standard questions to detect untrustworthy workers as suggested by previous works [EdV13]. We offered monetary rewards on successful task completion by paying 20 USD cents for each set of 10 pairs.

For each pair, we gathered at least 5 independent validity judgments resulting in over 32,285 responses in total. Based on these, we identified the most frequent judgment given by workers for the same pair as the aggregated validity judgment for each pair (in case of a tie, we considered the judgment that is closest to the average of all judgments for the same pair). These aggregated values give a more robust and intuitive indication of pair validity, coping with user disagreement and outliers. The independent judgments are made available in the dataset, for the remainder of the description we will refer to the aggregated judgments. Both the independent and the aggregated judgments can be utilized further depending on the application requirements. For instance, binary validity labels for pairs (i.e. *valid* or *invalid*) are derived for the evaluation of our approaches (Section 3.5) depending on whether the real-valued aggregated judgment exceeded a given validity threshold or not. This allowed to pose event validation as a classification problem. Among all the evaluated pairs, 6,336 (98.1%) have a proper aggregated judgment indicating how many event participants conform to the same event in the document and within the event timespan. For the other pairs, the Web page was either not available during the evaluation (110 pairs, 1.7%) or contained an unclear temporal setup (11 pairs, 0.2%). To show how pairs and events distribute over aggregated judgments, we present three cumulative frequency distributions (CFD) in Figure 3.2. *Pairs (All Events)* represents the CFD of all the 6,336 pairs that received proper aggregated judgments. *Pairs (Positive Events)* is a CFD only considering pairs related to events that had at least one associated pair with the aggregated judgment greater than '0' within the entire dataset. *Events* is the CFD of events with respect to the maximum judgment over all their pairs.

Figure 3.2 shows a relatively low amount of pairs with validity greater than '0', despite the retrieved documents matching the event queries to an extent. As mentioned in Section 3.2.4, due to the fact that the events were generated by an automatic

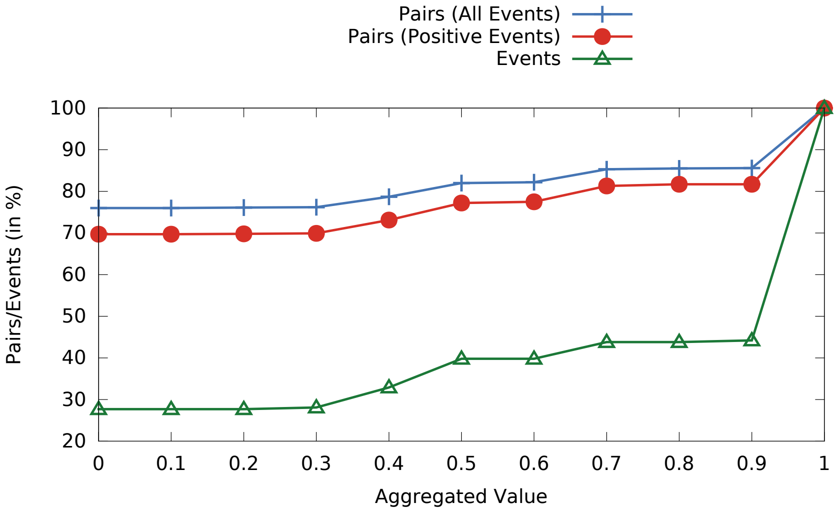


Figure 3.2 Distribution of pairs and events over validity judgments.

method, the event set also contains false events, which introduce only pairs with judgment equal to '0'. If false events are ignored (*Pairs (Positive Events)* CFD), the amount of pairs with judgment greater than '0' increases. Nevertheless, such increase is limited due to: (i) keyword matching considered to retrieve candidate documents is insufficient to ascertain document-level validity (as proved in [CGF15]), (ii) the mutual conformation of participants has to satisfy (narrow) temporal constraints, and (iii) even true events might not occur in all the retrieved documents. Differently, when considering events and the maximum judgment that they received over their pairs (*Events* CFD), less than 30% of the events are completely false, i.e. those having all pairs with aggregated judgments equal to '0', while more than half of them have all the participants truly conforming together within the same time period (at least one associated pair with judgment equal to '1'). The remaining events are judged as having intermediate verity, i.e. only a subset of the participants conform together.

Note that the absence of valid pairs corresponding to an event does not mean that the event is false in general. Some of these events may really be false events in the global sense (i.e., such an event did not occur), while some others may be true events in the global sense (i.e., such events actually did occur) but the considered documents do not provide supportive evidence of their occurrence.

3.4.4 Scope and Limitations

We finally elaborate on the scope and limitations of the presented dataset. Event detection methods relying on our event model (e.g. [DSJY11, HCL07, FBJ15]) could also use our dataset as ground truth. Since the validity judgments provided by crowd workers were bound by strict temporal constraints that were laid down in accor-

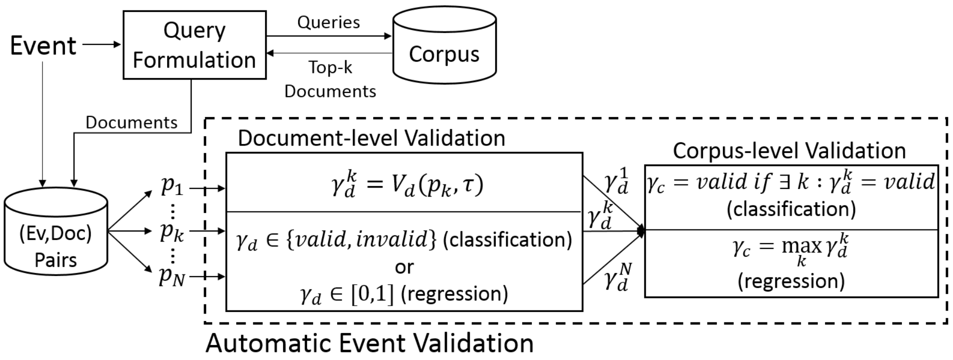

Event
Query Formulation
Queries
Top-k Documents
Corpus
Documents
(Ev,Doc) Pairs
p_1
p_k
p_N
Document-level Validation
$\gamma_d^k = V_d(p_k, \tau)$
$\gamma_d \in \{valid, invalid\}$ (classification)
or
$\gamma_d \in [0,1]$ (regression)
γ_d^1
γ_d^k
γ_d^N
Corpus-level Validation
$\gamma_c = valid\ if\ \exists\, k : \gamma_d^k = valid$
(classification)
$\gamma_c = \max_k \gamma_d^k$
(regression)
Automatic Event Validation

Figure 3.3 Overview of the automatic event validation.

dance to the event definition, this dataset does not fit to scenarios where atemporal event validation is required. However, given the high focus on the temporal aspect and conformation during the event timespan, our dataset can be relevant to any event-related research involving time in a strict manner. The time period of events considered within this dataset is narrow, therefore it may be unsuitable for event validation purposes that contain events spanning a larger granularity of time. However, it is noteworthy that the complex and elaborate manner of task decomposition and consequent acquisition of human judgments makes the dataset a rich source of event validation for events with similar timespans. The quality and quantity of human judgments in our dataset makes it a valuable resource and follows the order of magnitude that is typical in works related to event detection and validation (Section 3.2.4).

3.5 Automatic Event Validation

In this section we describe and evaluate our two automatic approaches to event validation. Before delving into their descriptions, we give an overview of how they are meant to operate as a black box. The high-level workflow is depicted in Figure 3.3.

For each input event, the query formulation generates one or more queries by taking into account the participants and timespan of the event. Then, such queries are performed to retrieve candidate documents from a given corpus, used as ground truth. Pairs are created by coupling the event with each retrieved document and they are fed into the automatic event validation. It is made of document-level validation followed by a corpus-level validation, according to the problem definition given in Section 3.3.2. We assume the data model to be the one defined in Section 3.3.1 and input (event, document) pairs to be collected as in Section 3.4.2, but the whole workflow holds for other choices of data models, query formulations, and corpora.

The two methods described in this section provide different implementations of the document-level validation, while the one at corpus-level is the same for both and consists in aggregating validation decisions done for each pair. Note that, due to computational constraints, the corpus-level validation is “approximated” by processing a subset of candidate documents instead of the whole corpus. However, if the data size and requirements of a given application scenario allow so, nothing prevents from retrieving and subjecting to validation the entire corpus.

3.5.1 Rule-based Event Validation

The first and simpler approach validates the occurrence of events in documents by checking a set of manually defined conditions. Given an event e and a document d, it first looks for temporal expressions in d within the timespan of e and then it counts the percentage of event participants mentioned together after each valid temporal expression and before an invalid one. This is the region of text assumed to be associated to each valid temporal expression. The maximum percentage found within the document is the evidence of e in d.

Document Processing. Terms and temporal expressions are extracted from each document d and they are used as basis to determine whether an event e occurs in d. Each term is a tuple $w := (v_w, p_w)$ where v_w is a non stop word and p_w is its position within the text. Each temporal expression consists of a tuple $t := (v_t, p_t)$, where v_t is the temporal expression value and p_t is its position within the text. Thus, for each candidate document d, a set of terms $W_d = \{w_1, \ldots, w_{n_W}\}$ and a set of temporal expressions $T_d = \{t_1, \ldots, t_{n_T}\}$ is extracted.

Document-level Validation. Given a (e, d) pair, the validation procedure described in Algorithm 2 computes the evidence $\gamma_d \in [0; 1]$ of the event e in the document d. It takes as input a document (the terms W_d and temporal expressions T_d extracted from it) and an event e (its participants and timespan). For every temporal expression $t \in T_d$ included in the event timespan $[t_e^0; t_e^f]$, the boundaries of the document region associated to t are computed by considering $p_t^{start} = p_t$ as starting position and $p_t^{end} = p_{\tilde{t}}$ as ending position, where $\tilde{t}$ is the first temporal expression after t which is not included in $[t_e^0; t_e^f]$. Matching of event keywords K_e are then checked within the set W_t of terms whose position is included in $[p_t^{start}; p_t^{end}]$. In order to handle the cases where the event keywords have multiple words, we say that a keyword k_e is contained within W_t if at least one word of k_e matches a term in W_t. The number k_t of event keywords in K_e having at least one matching in W_t is computed for every t, and the maximum value k_{max} is used to compute the document-level evidence $\gamma_d = \frac{k_{max}}{|K_e|}$. If a binary classification between valid and invalid pairs is required, then the real-valued γ_d can be binarized by imposing an evidence threshold τ as previously mentioned in Section 3.3.2. In this case, $\gamma_d \in \{valid; invalid\}$.

Corpus-level Validation. The validation of a given event e with respect to a set of candidate documents D is done by applying the document-level validation procedure described above for each document $d \in D$. Finally, the corpus-level validation

Algorithm 2: Rule-based Validation

```
Input  : document $d = (W_d, T_d)$, event $e = (K_e, t_e^0, t_e^f)$
Output: event occurrence $\gamma_e$
Set $k_{max} = 0$
for each $t \in T_d$ do
    if $t \subset [t_e^0; t_e^f]$ then
        Set $k_t = 0, p_t^{start} = 0, p_t^{end} = 0$
        $[p_t^{start}, p_t^{end}]$ = get_region_boundaries($t$)
        $W_t$ = get_terms_in_region($p_t^{start}$, $p_t^{end}$, $W_d$)
        for each $k \in K_e$ do
            if $k \subset W_t$ then
                $k_t = k_t + 1$
            end
        end
        $k_{max} = \max(k_{max}, k_t)$
    end
end
return $\gamma_d = \frac{k_{max}}{|K_e|}$
```

function is $V_c(e, D) = \max_{d \in D} \gamma_d$, while in case of binary classification it returns *valid* iff $\exists d \in D : \gamma_d = valid$ for a given evidence threshold τ. This means that the best (e, d) pair evaluated at document-level drives the decisions at corpus-level. We chose this policy, as opposed to others such as averaging the validity values computed at document-level, because we believe that the evidence of event occurrence in a single document is sufficient to indicate event validity in many scenarios.

3.5.2 Learning-based Event Validation

The validation method presented in Section 3.5.1 estimates the occurrence of events in documents based on a hand-crafted validation procedure. Differently, the approach described in this section learns a validation model by combining via supervised Machine Learning a set of features extracted from events and documents.

Features

For each (event, document) pair, we extract features from the event and document in isolation, as well as by considering them jointly. Features are extracted from plain text, without requiring any information about structure, title, urls, publication date, or markup, making our approach potentially applicable to any kind of corpora. We use the term *statistics* to indicate average, standard deviation, minimum, and maximum values of a given feature.

Event Features. Event features describe the event without coupling it with any document, under the hypothesis that the characteristics of events (e.g. number and type of keywords) might serve as *prior* indicators of the event validity. They consist of the number of event keywords, their length statistics, as well as the percentage of keywords representing people, locations, organizations, and artifacts.

Document Features. Document features are extracted from each document, independent of the event to be validated. We believe that characteristics of the content of documents can provide indications about the likelihood that the document contains events. For each document, we compute the percentage of words representing different parts of speech (nouns, verbs, adjectives) and named entity types (people, locations, organizations, artifacts), both when considering and ignoring duplicates. For each of the pos and named entity types mentioned above, we also compute statistics about their positions and mutual distances. Since events have a temporal lifespan that form a key basis for evidence, we derive features from time information in each document, such as number of temporal expressions, statistics of time (in days), statistics of temporal expressions positions and mutual distances in the document. We also consider the length of the document, number of sentences, length statistics of words, and length statistics of sentences.

Pair Features. Pair features are extracted from (e, d) pairs to give information about the extent to which a document contains evidence of the event. First, we compute the percentage of event keywords that have at least one match in the document, statistics about the number of matches of event keywords, and the percentage of event keywords that fully appear in the document (for multi-term keywords). Since distance between the matched event keywords in the document might affect their actual relationship (e.g. words that are far apart might have a higher probability of being unrelated), we compute statistics regarding the positions and mutual distances of matching keywords. The presence of event keywords in a document is a necessary but not sufficient condition for the event to have evidence in the document. One possible situation is that they might refer to a different timespan than that of the event. Therefore, we derive features considering the document temporal expressions within the event timespan (called matching dates hereafter). We compute the percentage of matching dates as well as statistics about positions and mutual distances of both matching and not matching dates (the proximity of not matching dates to matching dates might be a hindrance to assessing the validity of the timespan). We also consider the position of the first date and whether it belongs to the event timespan as features, supposing that the first date reported might have more impact than other dates in the text (e.g. as an implicit publication date). Finally, in order to consider the joint presence of matching keywords and dates, we compute features representing distances between them. For each matching date and each distinct keyword, the matching word nearest to the date is considered, and statistics about their distances to the date are computed. These distance-based features allow to estimate the likelihood of event keywords being mutually related, while referring to a matching date.

Event Validation

Document-level Validation. Once pairs have been described in terms of the features, a Support Vector Machine (SVM) is trained to predict the validity of new unseen pairs. The meaning of the pair labels l_p depends on which of the scenarios discussed in Section 3.3.2 is considered: in case of binary classification based on a validity threshold, then $l_p \in \{valid; invalid\}$; in case of regression of the percentage of event keywords conforming to the same event in the document and within the event timespan, $l_p \in [0, 1]$.

Given a set of pairs p_i, their corresponding feature vectors $\boldsymbol{f}_{p_i}$, and their validity labels l_{p_i}, an SVM is trained and the learned model V_d is used to predict the validity of unseen pairs p_{new}:

$$\gamma_d = V_d\left(\boldsymbol{f}_{p_{new}}\right) \tag{3.3}$$

The model was trained via 10-fold cross validation over the pairs.

Corpus-level Validation. We perform corpus-level validation by combining document-level decisions, in the same way of the previous approach in Section 3.5.1. Given an event e, a set of candidate documents D, a validity threshold τ, and a document-level validation function $V_d\left(\cdot\right)$, our corpus-level validation function $V_c\left(e, D, \tau\right)$ is designed as follows. In case of classification, it returns *valid* iff $\exists d \in D : V_d(e, d, \tau) = valid$, while in case of regression $V_c\left(e, D, \tau\right) = \max_{d \in D} V_d(e, d, \tau)$. We chose to let the pair having highest validity shape the corpus-level validation because the evidence of event occurrence in a single document may be a sufficient signal of event validity in many cases.

3.5.3 Experiments and Results

In this section we present the experimental setup of our evaluation and the performances of the two approaches described in Sections 3.5.1 and 3.5.2.

Experimental Setting

Dataset. We used the dataset described in Section 3.4 as basis for our evaluation. As a short reminder, it is made of a set of 250 real-world events (participants and timespan) and candidate documents (Web pages) associated with each one of them. This information results in a total of 6,457 (event, document) pairs, which are subject to validity assessment through a crowdsourcing evaluation. For each pair, at least 5 independent workers had to report the number of event participants involved in the same event mentioned within the document and strictly within the timespan of the event. The way such independent judgments have been aggregated for each pair is described in Section 3.4.3. From these real-valued pair labels l_p, we derived binary validity labels $l_{p,\tau}$ for pairs (i.e. *valid* or *invalid*) by applying a validity threshold τ as discussed in Section 3.3.2, such that $l_{p,\tau} = valid$ iff $l_p > \tau$ and $l_{p,\tau} = false$

otherwise. At the corpus-level, the same binarization is applied based on the pair with highest label among all the pairs for the same event (as described in Sections 3.5.1 and 3.5.2 about the corpus-level validation). These binary labels are used when considering event validation as a classification task (Section 3.3.2) and allow to have a more intuitive notion of document-level and corpus-level validity. In our experiments, we consider three different values for τ: 0.5, 0.65, and 1.0 (depending on the fraction of the event keywords required to conform to the event within the given timespan). The distributions of the percentages of *valid* pairs and events for different thresholds τ can be inferred from Figure 3.2 in Section 3.4.3.

Evaluation Metrics. We center the evaluation around the Cohen's Kappa between validity labels and the output of our methods for automatic validation, both at document and corpus-level. The Cohen's Kappa (named K hereafter) is especially useful in case of unbalanced datasets, as it determines the level of agreement between two judges by considering the probability that they agree by chance. We also report the accuracy (ACC) of the methods for sake of completeness. We compute ACC and K for each validity threshold previously discussed. For instance, K with $\tau = 0.5$ represents the Cohen's Kappa observed after binarizing the validity labels with a threshold of 0.5. Since the original validity labels are real-valued, we also report the Pearson Correlation Coefficient (r) when predicting them via regression.

Parameter Settings. The classifiers employed to predict validity of pairs, built using the SVM implementation of LibSVM[10], have Gaussian Kernels and were trained via 10–fold cross validation. The values reported in the rest of this section are averaged over the test sets of each split. The parameters were tuned to $C = 6.2, \gamma = 1.0$ in case of classification, and to $\nu = 0.5, \gamma = 0.3$ for regression.

Baseline. The baseline that we consider, named *Keyword Matching* (KM), validates pairs by counting the percentage of event keywords present in documents. It is therefore based on the mere appearance of event participants in text. In case of multi-term keywords, the matching of one term is considered a match for the entire keyword. The thresholds imposed on validity labels are also applied to this baseline.

Results

We now present the results of our event validation methods, both at document and corpus level. We also discuss the effects of applying event validation as post-processing phase of event detection.

Document-level. The results of event validation at the document-level, considering different validity thresholds ($\tau = 0.5, 0.65, 1.0$) and performance metrics (accuracy ACC, Cohen's Kappa K, Pearson Correlation Coefficient r), are reported in Table 3.4. The rule-based validation (Section 3.5.1) and learning-based validation (Section 3.5.2) are abbreviated as *RB* and *LB*, respectively. The learning-based method has been evaluated distinguishing two different models, namely *LB-pairs* and *LB-all*: the

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 3.4 Performances of the automatic validation at the document-level.

	τ = 0.5		τ = 0.65		τ = 1.0		-
	ACC	**K**	**ACC**	**K**	**ACC**	**K**	**r**
Baseline							
KM	0.575	0.091	0.575	0.090	0.668	0.191	0.314
Presented Methods							
RB	0.867	0.401	0.868	0.400	0.870	0.286	0.540
LB-pairs	0.910	0.659	0.909	0.655	0.916	0.605	0.713
LB-all	0.925	0.728	0.923	0.719	0.926	0.680	0.758

former only considers pair features in the learning, while the latter exploits all the available features described in Section 3.5.2. These names will be used within the rest of this section. Also, if not mentioned otherwise, we will refer to Cohen's Kappa during the discussion.

Both our approaches outperform the baseline under all the criteria. The *LB* models are the best performing ones overall, showing that combining information from events and documents via Machine Learning is more effective than (i) considering mere keyword matching (*KM*), or (ii) designing validation rules that cater for relationships between event keywords and temporal conformity to the event timespan (*RB*). In particular, according to the Cohen's Kappa ranges defined in [LK77], the values achieved by the *LB* models show a *substantial* level of agreement. The improvement of *LB-all* over *RB* is of 81% for $\tau = 0.5$ and of 138% for $\tau = 1$. *LB-pairs* alone sufficiently outperforms both *KM* and *RB*. However, we obtain higher agreement when exploiting information from events and documents as well (*LB-all* model).

Considering different validity thresholds, we observe a decreasing trend of performances for increasing validity thresholds (except *KM*, characterized by low and noisy results). We allude this to the fact that the methods face hindrance when validating those pairs where all the event keywords appear in a document, but only a subset of them actually conform to the same event within the given timespan. Let us assume a pair where the event has 4 keywords and all of them appear in the document, but only 3 are related together within the event timespan. In this case, the binary labels resolved by setting the validity threshold to 0.5 and 1.0 would be *valid* and *invalid*, respectively. If a method evaluates this pair as *valid*, driven by the fact that all the keywords are present in the document, then this would be counted as a mistake only for threshold 1.0. This matter persists although the model is trained with the proper binarized label (i.e. the training is aware of it since labels are changed according to the threshold). Also the presence of multi-term keywords partially formed by terms with broad usage and meaning (e.g. the keyword *Arab World*) might increase the presence of false positives when considering higher thresholds.

Corpus-level. Table 3.5 shows the results regarding the validation at corpus-level. The methods and performance metrics are the same ones considered before.

Table 3.5 Performances of the automatic validation at the corpus-level.

	$\tau = 0.5$		$\tau = 0.65$		$\tau = 1.0$		-
	ACC	**K**	**ACC**	**K**	**ACC**	**K**	**r**
Baseline							
KM	0.663	0.022	0.663	0.022	0.743	0.178	0.410
Presented Methods							
RB	0.823	0.653	0.827	0.647	0.824	0.601	0.701
LB-pairs	0.904	0.799	0.908	0.802	0.904	0.806	0.862
LB-all	0.876	0.740	0.896	0.782	0.901	0.797	0.845

		predicted				
		p	**n**		**p**	**n**
actual	**p′**	12.8%	5.2%	**p′**	7.9%	6.5%
	n′	2.3%	79.7%	**n′**	1.9%	83.7%
		(a)			(b)	

Figure 3.4 Confusion matrices (document-level) for (a) *LB-all* with $\tau = 0.5$ and (b) *LB-pairs* with $\tau = 1.0$.

Again, our methods outperform the *KM* baseline under all the criteria, with *LB-all* achieving the highest performances. Its improvements over *RB* range from 13.4% ($\tau = 0.5$) to 32.6% ($\tau = 1.0$). Comparing the results of the document-level and corpus-level evaluation (Tables 3.4 and 3.5), it can be observed that the absolute values in the latter are relatively higher. This is because at the corpus-level a single valid (e, d) pair is sufficient to validate the event, despite the presence of errors in the other pairs corresponding to the same event. In other words, at corpus-level any valid event evaluated as true will not introduce any mistake as long as at least one pair for that event is judged as true, although there might be errors even for all the other pairs. Correct validations at corpus-level might be also due to mistakes at document-level. Let us consider an event for which two pairs A and B are available, where A is true and B is false. If the document-level evaluation validates A as false and B as true, then there would be two errors at document-level but no errors at corpus-level.

Another point of difference between the results of document and corpus level validation is that the values of K at corpus-level increase for increasing values of validity threshold τ, while they decrease at document-level. In order to explain this trend, in Figure 3.4 we show the confusion matrices at document-level obtained for the limit cases *LB-all* with $\tau = 0.5$ (case "a") and *LB-pairs* with $\tau = 1.0$ (case "b"). We have mentioned before that increasing the validity threshold at document-level introduces difficulties in correctly evaluating valid pairs. The confusion matrices show that, to reduce classification errors, the learning process adapts by increasing

Table 3.6 Effects of applying event validation after event detection.

	$\tau = 0.5$			$\tau = 0.65$			$\tau = 1.0$		
	P_{det}	P_{val}	R_{val}	P_{det}	P_{val}	R_{val}	P_{det}	P_{val}	R_{val}
Baseline									
KM	0.632	0.636	0.993	0.627	0.631	0.995	0.604	0.628	1.000
Presented Methods									
RB	0.632	0.842	0.773	0.627	0.845	0.771	0.604	0.863	0.618
LB-all	0.632	0.894	0.900	0.627	0.913	0.913	0.604	0.919	0.899

the percentage of negative detections (either true or false) from 84.9% in (a) to 90.2% in (b). On the contrary, when the validity threshold is low ($\tau = 0.5$), the learning process is more confident in classifying valid pairs and the resulting output rate of positive detections is higher, from 9.8% in (b) to 15.1% in (a), leading to an increase of false positive pairs as well. The increase of such errors in (a) generates more false positive events at corpus-level, because one single false positive pair is sufficient to declare an invalid event as valid. On the contrary, models with reduced rate of positive predictions at document-level, like in (b), ensure a lower number of false positives at corpus-level despite being less accurate at document-level.

Precision Boosting. The effects on the performances of event detection after applying event validation as post-processing phase are reported in Table 3.6 for different values of τ. P_{det} refers to the precision values (one for each τ) within the set of detected events, which are comparable with the performances reported in [DSJY11]. The precision and recall (with respect to the true events discovered through event detection) in the event set after applying event validation are indicated as P_{val} and R_{val}, respectively. Our method boosts the precision achieved by event detection up to 0.894 for $\tau = 0.5$ and 0.919 for $\tau = 1.0$. Moreover, the high values of recall (0.899 or higher) indicate that most of the true events present in the original set are preserved by the validation. Our method beats *CF* both in precision and especially in recall, showing a better capability of not discarding true events. Regarding *KM*, the low increase in precision and the high recall show that performing validation via keyword matching has no effect on the performances of event detection: most of the events are judged as valid and retained, since the appearance of their keywords in documents is a sufficient condition of event validity.

3.6 Conclusion

In this chapter, we introduced *event validation* as the problem of validating the verity of real-world events by means of their occurrence in a corpus of textual documents. We distinguished between two levels, namely *document–level validation*, looking for event occurrence within a single document, and *corpus–level validation*, which assesses the verity of events based on an entire corpus of documents. Given the effort required to manually inspect a potentially large set of documents looking for evidences of a given event, we proposed two automatic methods to accomplish such validation. The first and simpler one validates events via a hand-crafted validation policy, which looks for portions of text referring to the timespan of the event to be validated and then returns the highest percentage of event participants mentioned together in such excerpts. The second one learns to assess the occurrence of events in text based on features extracted from event-document pairs. Experimental results showed the superiority of machine learning, which could discern the verity of events with a *substantial* agreement with respect to human evaluators. We also introduced automatic event validation as a post-processing step of event detection to boost the precision within the set of detected events while preserving recall.

Moreover, we presented a crowdsourced corpus for evaluating the performances of event validation methods with respect to a given document corpus. The dataset comprises of 250 events and 6,457 corresponding documents. The event set spans several categories ranging from cinema, music, TV and entertainment to sport, politics, nature and disasters. Each (event, document) pair is associated with multiple independent validity judgments, representing the number of event participants conforming together to the same event in the document and within the event timespan. These judgments have been acquired through a crowdsourcing process where the manual task of event validation has been decomposed to make it fit for the easy consumption of workers. To the best of our knowledge, this is the first dataset of its kind that is made publicly available: a corpus for event validation, where event occurrence in documents is a strict function of participants and their conformation within a specified timespan.

4 Collecting and Retrieving Event-Related Information

Keeping events accessible and reusable poses a number of challenges that go beyond their detection. In this chapter we discuss two of them, one coming from the ubiquity of events and the other concerning their understanding over time.

Storing events and their descriptive information is the very first step for their reuse after they have been identified, either automatically or manually, and validated. One fact that should be considered at this stage is that events can exhibit huge differences among them, under aspects like topic, detail of description, duration, complexity, due to their pervasiveness in any part of the physical and digital world. This can result in different sources of events that give special attention to particular combinations of event characteristics. However, to be applicable and satisfy information needs in different contexts, event collections should be as comprehensive as possible and accommodate events with diverse characteristics. We present a comparative study involving three event sources, showing their differences and how they can be complementary to each other. This analysis motivates the importance of identifying and merging diverse event sources, to come up with a more comprehensive event set.

In the second part of the chapter we bring up the temporal dimension and investigate how events, properly stored along with their descriptive information, can be understood over time. When storing an event in a repository for future reuse, important information might not be included in its description as being part of the generally known context at that time. However, the evolution of such context and the partial forgetting of the original one might hinder the understanding of the event only based on the information stored initially. While this might not happen for relatively short time periods, e.g. few years, it would probably become evident after few decades. To overcome this problem, we aim at formulating queries from the description of an old event to retrieve information for its proper understanding with high recall, which can be further used for re-ranking purposes. Our method formulates queries by means of a recall-based query performance prediction and outperforms other baseline methods.

4.1 Introduction

Chapter 3 has been devoted to ensuring the verity of events, which can be collected either automatically or manually. That was an important step to discard false events and let any subsequent phase handle as accurate event-related information as possible. In this chapter we move forward and discuss two other phases concerning the use of events after their detection. One is perhaps the very next step after event validation, namely the storing and indexing of events coming from multiple sources. The second phase is temporally farther away and consists in properly understanding events that have been collected and stored in the past.

Since events are ubiquitous and valuable for a wide range of end users, such as professionals, students, or simply people being interested in a certain topic, properly collecting and organizing event descriptions is a significant step to support their search and consumption over time. An important observation is that events can be very diverse from each other. They can of course pertain to different topics, but there are other aspects that might vary even within the same topic. Some events can have a worldwide impact and reach a broader news coverage (newsworthiness) than others, they can involve different numbers of participants related to each other in multiple possible ways (complexity), or yet their timespans can range from hours to weeks. Also the very same event can be described by different viewpoints or granularities. Distinct event collections, e.g. those gathered by news agencies, by online collaborative communities, or by automatic methods, might be sources of events possessing only a specific subset of those characteristics. For instance, news agencies might give more attention to events like the Arab Spring or the US presidential elections, which have a direct or indirect effect on many countries, while other less newsworthy events like a movie premier or the release of a smartphone might go viral in Social Media.

Given these premises, in the first part of this chapter we compare three ways of collecting events, which we call event sources, with the objective of highlighting how their different characteristics can complement each other. The three event sources included in the study are Wikipedia *Current Events* portal[1], the *YAGO2* knowledge base [HSBW12], and an event detection algorithm that we have previously developed [TCG+14]. While we are aware of other initiatives aiming at collecting and structuring events (see Section 4.2), we chose them as a minimal but still very diverse set of ways to collect events. We describe the events from each source under the same model, which is simple and accordant with other existing models [STH09, vHMS+11], and we compare them in terms of their participants and event complexity focusing on the topic of politics. Although our study is limited in the number of event sources, topic, and comparative metrics, we believe it shows the potential of merging complementary event sources, to come up with a more comprehensive and not redundant repository that encompasses events with different topic, newsworthiness, complexity, and time duration.

[1] http://en.wikipedia.org/wiki/Portal:Current_events

Let us now assume that events are stored along with their descriptions, possibly coming from different and complementary event sources as suggested previously. In the second part of the chapter we study the problem of properly understanding events after a relatively big amount of time, such as few decades. The event descriptions collected when the corresponding events happened, although being exhaustive at that time, might not be enough in presence of such big temporal and contextual gaps. This because, while time passes, the context in which a certain event happened is likely to evolve and, at some point, the original one would be forgotten or at least fade over time. Such lack of original contextual information, which was generally known at the time of the event and therefore not stored, might hinder the comprehension of that event in the future. Note that the need for dealing with content from the past is not restricted to expert users, such as journalists, historians or scholars. In fact, due to the growing age of the Web, general users increasingly have to handle old content that assumes knowledge of the context at the corresponding creation time for its interpretation.

In order to ease the understanding and interpretation of old events, we generate queries from their descriptions and perform them into an available context source to retrieve candidate contextualizing information with high recall. Such retrieved information should be relevant to the content of the event and temporally close to its date. We focus on recall so that the result sets retrieved by our methods can be used for re-ranking based on precision and any additional side objective, with the benefit of having retrieved more relevant results in the first place, even though they might not be ranked as high as desired. For example, candidate contextualizing results retrieved by our methods have been re-ranked in [TCKN15a] to push up those results containing complementary information that is not already present in the event description. Our best query formulation method is based on query performance prediction exploiting a set of novel temporal features and optimizing recall.

Summarizing, the contribution of this chapter is twofold: (i) we elaborate on the convenience of gathering events from diverse and complementary sources to come up with a more comprehensive and not redundant event repository, along with a comparative analysis among three different event sources to support our argument; (ii) we compare different query formulation methods to retrieve candidate contextualizing information for understanding old events, with the most efficient one based on a recall-oriented query performance prediction using a set of novel features for adaptivity to the difficulty of the events to be understood.

The rest of the chapter is structured as follows. In Section 4.2, we discuss existing works on both collecting events and retrieving context information. Section 4.3 contains the comparative analysis on three event sources and their complementarity, while the query formulation methods for retrieving contextualizing event-related information are presented and compared in Section 4.4. Finally, in Section 4.5 we summarize and conclude the chapter.

4.2 Related Work

The review of the related work for this chapter is organized in two parts: one on existing approaches and initiatives for collecting and organizing events, the other on methods for formulating queries and retrieving contextualizing information for news articles.

4.2.1 Event Collections

The analysis reported in Section 4.3 consists in a comparison between three means of collecting events (called event sources), namely the YAGO2 knowledge base, the Wikipedia *Current Events* portal, and an event detection method developed by Tran et al. [TCG+14]. We chose them to have a sample of very different ways of collecting/detecting events in our analysis while limiting their number. However, given the usefulness of collecting and organizing events to facilitate their access over time, many other approaches and projects exist with this goal. Here we give a brief survey of the most popular ones that have not been included in our analysis.

Perhaps the most remarkable effort in collecting events and related news articles is being carried by the GDELT project [LS13], which periodically gathers and makes available a set of events from all around the world. Every day, it monitors news broadcasters on the Web in more than 100 languages and collects events reported by them along with a rich list of descriptive information. Each event has a category (300 in total), and over 60 fields like time, geolocation, participants (people, organizations), topics, sources, number of mentions (over all sources), tone (from extremely negative to extremely positive), potential impact on a country. Event categories are organized in a hierarchy, which allows to estimate the specificity of events. The whole dataset comprises over 250 million events from 1979 up to the present date. While already existing at the time of our study, the popularity and usage of GDELT significantly increased after it. This is the reason why we did not include it within our analysis, but it would be definitely interesting to compare the events collected by GDELT with those reported in more informal sources such as blogs and social media.

McMinn et al. [MMJ13] presented a methodology for the creation of an event corpus using state-of-the-art event detection approaches as well as Wikipedia Current Events, employing the basic event model of the TDT project (something that happens at some specific time and place). The resulting corpus is made of more than 500 events along with related tweets and is meant to be used as a ground truth for event detection on Twitter.

An approach for populating knowledge bases with named events extracting from news articles have been introduced in [KVW14]. Input news articles are crawled by using the external links of all Wikipedia pages, thus encompassing a variegate set of online newspapers and other news broadcasters. Then, events are “distilled” from the input documents by assigning them semantic event types (using Wikipedia categories and WordNet event classes) and hierarchically partitioning them into groups, each one of them representing an event. Storylines are also built by chaining related events in chronological order. The identified events along with their semantic annotations (same as those attached to input documents) and descriptive metadata (e.g. time, locations, entities, etc.) are eventually inserted into the given knowledge base. Overall, the generated corpus is made of 25,000 events and 300,000 news articles.

Finally, there exists datasets in the Linked Open Data (LOD) cloud that contain events, such as DBpedia [ABK+07], which is similar to YAGO2 as both contain named events extracted from Wikipedia infoboxes, and EventMedia [KT16], which gathers events and their descriptions from four public event repositories (Last.fm, Eventful, Upcoming and Laynrd5) and two social media (Flickr and Twitter).

4.2.2 Retrieving Contextualizing Event-related Information

Entity Linking, which consists in detecting entity mentions in text and linking them to entity descriptions like Wikipedia pages or representations in other knowledge bases (e.g. YAGO, DBPedia, Freebase), can be regarded as a way of adding context to a piece of text. Early works, like [MC07, MW08], disambiguate terms individually by exploiting Wikipedia and its structure of links and categories. More sophisticated approaches [HSZ11, HYB+11] perform entity linking by taking into account how the disambiguation decision of one term affect the disambiguation of other terms. We claim that the additional information provided by entity linking is not suitable for our task, where the reader is likely to experience large temporal and contextual gaps when reading about an old event. The information contained in entity descriptions such as Wikipedia pages usually covers a wide topical and temporal range, being neither focused on the topic and participants of the event nor temporally close to its date. The whole Wikipedia pages are also too long to be inspected by the reader in a short time and without being distracted, which is the reason why we consider paragraphs as size of contextualizing information pieces to be retrieved. Finally, the need for additional information might regard other terms not categorized as entities (as already noted in [CTKN14]).

Given these challenges, the problem of retrieving additional and contextualizing information for documents has received increasing interest in the last decade. Štajner et al. [vTP+13] formulate the task of selecting the set of most *interesting* tweets related to a given news article as an optimization problem, whose objective function takes into account diversity, popularity, user authority, and opinions of the tweets within a candidate set. Social media utterances that discuss a given news article are discovered in [TdRW11]. Given a news article, different queries are formulated based on its internal structure, term selection strategies, and utterances explicitly containing links to the article itself. The ranked lists of results retrieved by each query are then merged through data–fusion techniques. Gao et al. [GLD12] tackle the problem of generating a representative but not redundant summary of a given event by means of a topic modeling approach that jointly exploits information from news articles and Twitter. Complementarity between tweets and sentences within news articles is measured as a combination of their similarity and difference. All these methods ignore the temporal dimension, which is instead crucial in our problem. Also, our query formulation methods could be added in the first place to ensure higher recall in the initial sets of documents handled by them.

The automatic formulation of queries from an event description is one of the contributions of this chapter and an important step towards retrieving contextualizing information to better understand the input event. Formulating queries from a document, or a piece of text in general, has been actively studied since the beginning of the last decade [HCMB03]. In case no further information (e.g. annotations or metadata) is available besides the text itself, then the problem turns in identifying query terms from plain text. This can be done by using tf–idf, mutual information, natural language processing, or machine learning [Tur00, WPF+99, YBD+09]. Assuming the presence of basic metadata and structure for documents, Tsagkias et al. [TdRW11] use document title, lead paragraph, and body to generate separate queries for a given document. Some of the methods in this chapter follow a similar idea by creating queries using the title and lead paragraph of documents (event descriptions). Similarly to [GPS99], we also explore approaches that assume the availability of manual annotations, possibly given by users while reading the document, as seeds for query formulation. The advantage of having such additional "hints" is that the information needs of the users are made explicit, possibly leading to more effective queries. In our most effective and novel approach we formulate queries by combining annotations via Query Performance Prediction (QPP) [CTZC02], using both pre–retrieval [HHdJ08, SG09] and post–retrieval [CYT10, SKC+12] features. The former are based only on the query and corpus–based statistics, while the latter also analyze the retrieved list of results. In line with the previous work on time–aware performance predictors [KN11], we investigate novel features for QPP that explicitly take the temporal dimension into account. Differently from the previously mentioned approaches, which focus on precision metrics, we consider the performances of queries in terms of recall, which have been recently remarked and considered in different information retrieval scenarios [CG14, LWRM14, CSK13].

4.3 Collecting Complementary Event Sources

Event-related information is ubiquitous in today's digital world and is subject to the search and consumption by different kinds of end users, from people willing to be up-to-date on the latest news to professionals like journalist, librarians, political staff. Collecting events and their descriptions is therefore crucial for easing their access and consumption over time, and different projects indeed exist with this aim (as discussed in Section 4.2.1). The ubiquity of events is also the reason of profound diversities among them, in terms of topics, newsworthiness, complexity (e.g. number of participants, relationships amongst them), temporal granularity, and different event sources (e.g. collaborative event collections, knowledge bases, automatic methods, news agencies) might release events with specific characteristics. For instance, news agencies might give the priority to events with a worldwide impact regarding politics and economics, for instance the 2011 Arab Spring, while events extracted from Social Media might also encompass the release of a new smartphone, a movie premiere, or the acquisition of a football player by a team. It would be convenient to merge events coming from different and potentially diverse sources, to have a comprehensive and yet not redundant set of events.

In this section we spend some more thoughts on this matter, describing the gathering and indexing of three different event sources along with a comparison on the complementarity between them. The overall event set has been created for sake of analysis and is not meant to be a unique global source of events, since we are aware of other event repositories (see Section 4.2.1) that have not been included in our work. Section 4.3.1 describes the way out event set has been constructed, while Section 4.3.2 contains the comparative analysis among the three event sources it is made of.

4.3.1 Event Repository

We now describe the event model of our repository and the three different event sources used to populate it.

Event Model

Our event model is similar to other existing and popular models, namely LODE [STH09] and SEM [vHMS+11]. We describe events in terms of a set of core information, such as participants (named entities like people, places, and organizations) and a time period (start and end date). We also enumerate a set of additional information, e.g. textual description, location, category, which might be available depending on the event source. All the attributes are listed and described in Table 4.1. We regard *participants* as titles of Wikipedia pages involved in the event, which are classified as people, organizations, artifacts, and locations based on the categorization provided by the YAGO2 ontology [HSBW12]).

On the one hand, the core attributes of this event model are general enough to be suitable for events extracted from different sources and methods. This model can also accommodate the events defined in the context of event validation (Chapter 3). On the other hand, it captures all the main aspects that characterize events. Our model does not consider event hierarchy, whose representation and identification is beyond the scope of this study.

Table 4.1 Attributes comprised in our event model and their meanings. The classification of participants in based on the entity categorization of YAGO2.

Attribute	Description
Participants	Named entities participating to the event.
Start	Starting date of the event.
End	Ending date of the event.
Description	Natural language description of the event.
Who	Participants categorized as people or things.
Where	Participants categorized as locations.
Story	High-level storyline the event belongs to.
Category	Main category or topic the event belongs to.
Reference	URL to a website where the event is reported.
Source	The method or repository the event comes from.

Event Extraction

We collect events by exploiting three different methods and sources, which can complement each other to some extent. These are an event detection algorithm that we have previously developed [TCG+14], called *Co-References* hereafter, and two event sources, namely Wikipedia *Current Events* portal[2] and the *YAGO2* knowledge base [HSBW12].

Co-References. This event detection method is part of the approach described in [TCG+14], which has been used in Section 3.4.1 for validation purposes. It works on the edit history of Wikipedia pages, which are regarded as event participants. The main idea of the Co-References method is based on the assumption that the occurrence of a new event triggers new edits in one or more Wikipedia articles, whose corresponding entity might be a participant of that event. By following the links to other Wikipedia pages in such edits, we can connect more entities relevant to an event in a particular time period. Thus, a set of entities is said to participate in a common event if their edits co-refer to each other within a time period σ. The parameter σ represents the time delay until edits of two entities are allowed to refer to each other. We empirically chose $\sigma = 7$ days because this yielded the most meaningful results in our experiments. An example of an event extracted by this algorithm is *{Obama;G8;Deauville} from 26-05-2011 to 27-05-2011* representing the G8 summit held in Deauville, France. We refer the reader to [TCG+14] for more details on the algorithm.

This approach is in principle applicable to any set of Wikipedia pages and time period. However, in order to reduce the computational effort of running our algorithm without compromising the significance of our study, at the time of the study we applied the following restrictions to the dataset handled by the Co-References method.

[2] http://en.wikipedia.org/wiki/Portal:Current_events

First, we chose *politics* as specific topic of interest. While we have already mentioned that different event sources can be biased towards different topics, we believe that other differences can be observed within the same topic. We chose politics because it produces a large amount of newsworthy events. Second, we limit the timespan to the whole 2011. Based on these constraints, we identified $31,998$ Wikipedia pages registered as entities in YAGO2 [HSBW12] and belonging to the *politician* class as seed for event detection, and we discovered 242 events involving them based on their edit history during the whole 2011.

Wikipedia Current Events. Wikpedia's *Current Events* portal contains daily summaries of events and is collaboratively edited and curated by the crowd [TA14]. While covering a fairly wide range of topics, this portal tends to give more attention to events about conflicts, politics, and crime, while others like science and art are less represented. Each event comes with a date, name and description, which spans up to few sentences. Furthermore, events can also have side information like categories, references, and a set of entities annotated within the description, which correspond to hyperlinks to Wikipedia pages. All these information pieces have a direct mapping to our event model. We acquired the event-related information contained in Current Events via the API provided by the WikiTimes project[3], which periodically crawls the website, obtaining over 50,000 events from 2001 to 2013 (the time of this study).

YAGO2. The YAGO2 knowledge base [HSBW12] is an ontology built from Wikipedia infoboxes and combined with Wordnet and GeoNames to obtain 10 million entities and 120 million facts between them. An example of a fact in YAGO2 is <BobDylan> *wasBornIn* <Duluth>, indicating that the entity Bob Dylan was born in Duluth. When possible, facts are augmented with temporal and spatial information, so that it is possible to know when and/or where a fact took place or occurred. We created an event for each temporal fact according to our event model, also including locations (as *where* field) and related entities when available. Due to the lack of textual descriptions for events, we concatenated subject, predicate, object and time of a fact as its description: in the previous example, the description would have been "BobDylan was born in Duluth in 1941". Moreover, in YAGO2 there are also entities representing named events themselves, e.g. *2011_Australian_Open*, with further temporal and spatial information as well as participating entities. We created one event for every entity marked as event, where the description is the name of the entity itself and the other event fields are filled by exploiting the available information.

Remarks

The three event sources described before can release different and, therefore, complementary kinds of events. *Co-References* works on the edit history of Wikipedia and is able to detect events with different duration, complexity (number of entities and their relationships), and newsworthiness (from a wrestling match to Academy Awards

[3] http://wikitimes.l3s.de/

and Egypt Revolution). *YAGO2* encompasses temporal facts regarding entities (e.g. birth, death, political roles) as well as particular entities categorized as events, resulting in a repository that mostly contains high-level and well-known events, often lacking textual descriptions. *Current Events* portal is contributed by Wikipedia users and contains textual descriptions of daily events: they are reliable, thanks to the high level of control within Wikipedia, and endowed with self explanatory textual descriptions. Note that *Co-References* is an event detection algorithm and is subject to errors, i.e. not all the detected events are true events. Instead of manually evaluating the output events, which is time consuming and unfeasible at a large scale (as already discussed in Section 3.1), we employ the automatic event evaluation technique described in Section 3.5.2 and retain only those events judged as true.

Merging these complementary sources would result in a more balanced and diverse event repository, containing events with different granularity, complexity, and time duration. A given event might be present in the repository in form of distinct events coming from different sources, thus providing complementary perspectives of it. This also motivates not performing any duplicate detection on our merged repository: any near duplicate would still provide complementary event-related information.

4.3.2 Analysis

In this Section we analyze some characteristics of the event repository constructed as described in the previous section, especially with the goal of showing how the different sources (Co-References, Current Events, YAGO2) contribute to the overall event set and how they differ from each other. In order to perform a fair comparison between the sources, we center our analysis on the entity set used as input by the Co-References method (31,998 Wikipedia pages registered as entities in YAGO2 and belonging to the *politician* class) because it is the most restrictive one. While a first comparison on the complementarity between Co-References and Current Events in terms of event overlapping has been discussed in [TCG$^+$14], we take into account other aspects such as entity overlap, entity activity, and event complexity.

Number of Events

In Table 4.2 we report the number of events contained in our repository, along with the contributions from the different sources. Note that the sizes of YAGO2 and Current Events refer to their status at the time of the study (2013) and will be different from those at future time points. As already anticipated in Section 4.3.1, the three considered sources cover very different time periods. YAGO2 contains events and temporal facts that occurred even more than 1000 years ago (e.g. Zanzibar <*wasCreatedOnDate*> 1000-01-01), the Current Events method keeps track of events since 2001, and the Co-References method has been restricted to 2011 to get an event set requiring reasonable computational effort while being significant for comparisons.

Table 4.2 Number of events within our repository, split by different sources. For sake of comparison, we also restrict the count to the particular entity set and timespan used by the Co-References method.

Source	Total	Politicians	Politicians 2011
All	2,629,740	50,168	1,401
YAGO2	2,578,547	42,399	360
Current Events	50,951	7,527	799
Co-References	242	242	242

Table 4.3 Number of *active* entities within the event repository, split by different sources. Active entities are those participating to at least one event.

Source	Active Entities
YAGO2	472
Current Events	310
Co-Reference	366

Moreover, the entity sets considered in the sources are different: YAGO2 and Current Events contain almost all the Wikipedia pages, while Co-References has been run on a subset of them belonging to politicians. These facts result in very different contributions from each source, as depicted in Table 4.2. Overall the repository contains more than 2.6 million events, most of them (almost 2.5 million), coming from YAGO2. However, if we restrict the count to those events that occurred in 2011 and involving politicians, then the contributions are more evenly distributed. In order to make a fair analysis, in the following sections we will consider this latter case.

Active and Inactive Entities

A further distinction that we make is between *active* and *inactive* entities: the former participate in at least one event, while the latter do not. Within our politician dataset, composed of 31,998 entities, only 1007 of them resulted to be active. This low number can be attributed to at least two reasons. First, we only consider events that occurred in a particular year (2011), but there might be politicians that were active in other time periods. Second, we noted that our dataset contains many dead politicians, e.g. Vyacheslav Molotov (1890-1986) and Otto von Bismarck (1815-1898), which clearly could not participate in events in 2011. In the rest of our analysis we will consider only active entities.

Entity Overlap

One possible criterion to assess the complementarity between the different event sources is the overlap of their active entities, i.e. how many active entities appear in

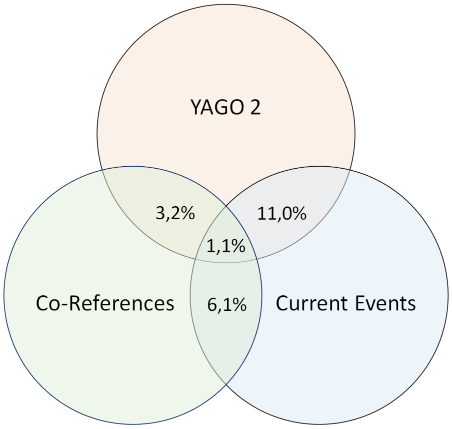


Figure 4.1 Overlap of the event sources in terms of active entities.

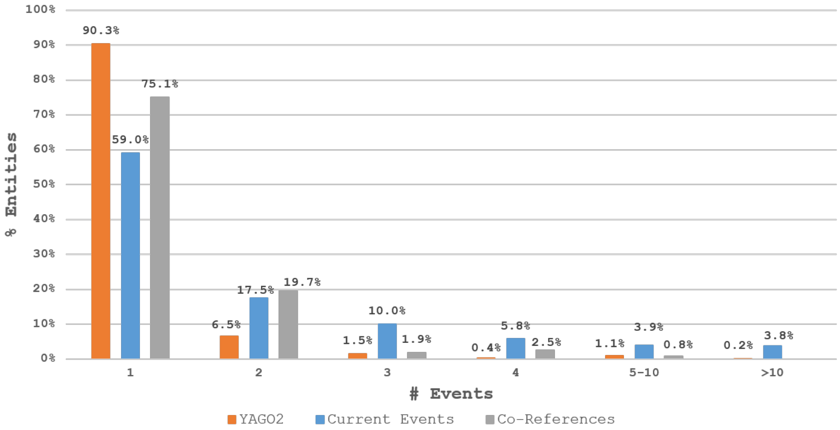


Figure 4.2 Entity Distribution over number of events they participate to.

more than one source. In Table 4.3 we show the number of active entities contained in each different source. The fact that their sum (1148) is slightly greater than the total number of active entities (1007) already reveals that the number of active entities that are present in more than one source is quite low. This fact is confirmed in Figure 4.1, which shows a Venn diagram for the different event sources. It is possible to observe that the number of overlapping active entities over the three sources represents only 1.1% of their union, showing that the different sources can complement each other. Examples of active entities that are present in all the sources are *Gamal Mubarak* (who was involved in the Arab Spring), *Dmitry Medvedev* (who was serving as President of Russia in 2011), and Rahm Emanuel (who was elected as major of Chicago in 2011).

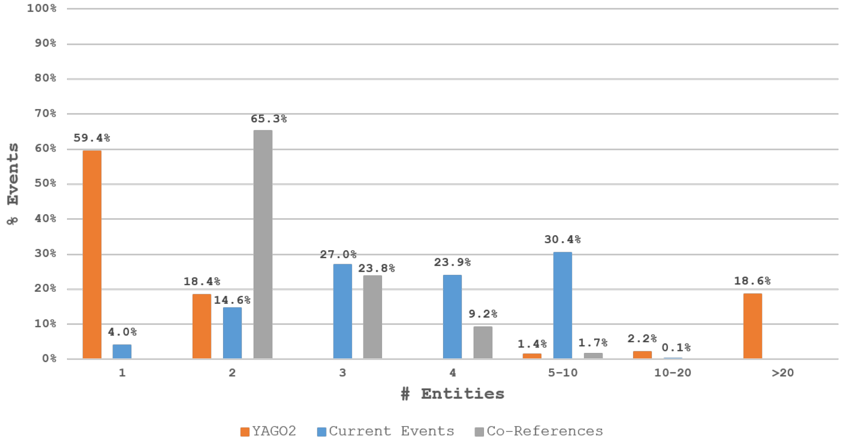


Figure 4.3 Distribution of events over the number of their participating entities (complexity).

Entity Activity

Active entities can have different degrees of activity, based on the number of events they participate in: *strong* entities are those participating in many events, while *weak* entities appear in one or two events. The distribution of entity activity can be a further criterion for describing event sources, since different sources might have entities with different degrees of activity. In Figure 4.2 we show, for every source, the distribution of entities with respect to the number of events they participate in (i.e. their activity). It is possible to observe that YAGO2 mostly has weak entities (more than 90% of them only participate in 1 event), while Current Events and Co-References contain more dynamic entities. Almost 25% of active entities in Co-References appear in at least 2 events, while for Current Events more than 40% of entities appear in at least 2 events and 7,7% participate in at least 5 events. These are other signals of the diversity among the three event sources.

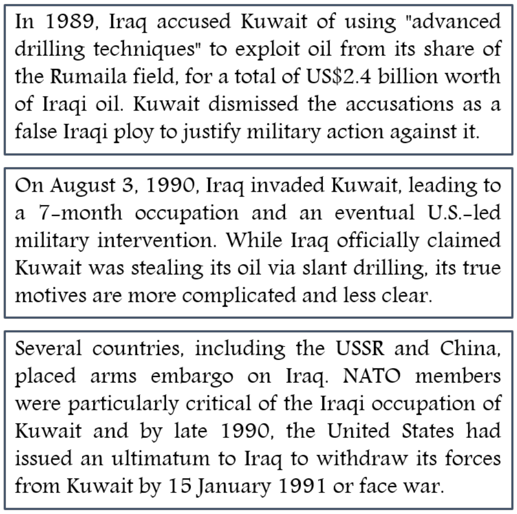
The most active entity is *Barack Obama*, appearing in 84 events in Current Events, while other examples of fairly active entities are *Silvio Berlusconi* (27 events in Current Events), Vladimir Putin (5 events in Co-References), and *George W. Bush* (5 events in YAGO2).

Event Complexity

Finally, another criterion to describe event sources is the number of entities that their events involve, which we call *event complexity*. Intuitively, non-complex events will involve one or few entities, while more complex events will consist of more entities. Although the complexity of an event could be measured in different ways, for instance by considering also their temporal duration and interrelations between participants, we believe that the number of participating entities of an event is a simple but yet

Attribute	Value
Participants	Iraq, Kuwait, Persian Gulf.
Start	02-08-1990
End	04-08-1990
Description	Iraqi troops stormed into the desert sheikdom of Kuwait today, seizing control of its capital city and its rich oilfields, driving its ruler into exile, plunging the strategic Persian Gulf region into crisis and sending tremors of anxiety around the world.

Table 4.4 Example of an old event dating back to 1990.



In 1989, Iraq accused Kuwait of using "advanced drilling techniques" to exploit oil from its share of the Rumaila field, for a total of US$2.4 billion worth of Iraqi oil. Kuwait dismissed the accusations as a false Iraqi ploy to justify military action against it.

On August 3, 1990, Iraq invaded Kuwait, leading to a 7-month occupation and an eventual U.S.-led military intervention. While Iraq officially claimed Kuwait was stealing its oil via slant drilling, its true motives are more complicated and less clear.

Several countries, including the USSR and China, placed arms embargo on Iraq. NATO members were particularly critical of the Iraqi occupation of Kuwait and by late 1990, the United States had issued an ultimatum to Iraq to withdraw its forces from Kuwait by 15 January 1991 or face war.

Figure 4.4 Contextualizing information for the event.

meaningful indicator of event complexity. In Figure 4.3 we show, the distribution of events in terms of the number of their participating entities (their complexity), for each source. Very different patterns can be identified. YAGO2 is highly unbalanced, having almost 60% of events with only one entity and more than 20% with more than 20 entities. The former class corresponds to events representing basic temporal facts like the birth and death of persons, which are composed of only an entity and a date. The latter group represents events that are described by their own Wikipedia page (like *2011 Australian Open* and *Arab Spring*), whose participating entities are those mentioned in the page. Co-References is also unbalanced, but in a different way. It has more than 65% of events with two entities but none with only one, since the event detection method is based on finding entities that mention each other in their edits. The percentage of events with two entities for Co-References is lower than the one reported in [TCG+14] because we relaxed the method by also accepting events when connected components of co-referring entities are found (not only cliques). Current Events is more balanced: more than 95% of its events are quite homogeneously distributed between 2 and 10 participating entities, while very few events consist of only one entity (4%), and more than 10 entities (0.1%). These findings serve as further evidence to how the three event sources are orthogonal and complement each other.

4.4 Retrieving Context for Forgotten Events

In Section 4.3 we have discussed about the importance of collecting and curating events over time, as first step for their inspection and consumption for different purposes and domains. Now we make a temporal jump and question how events stored at the present time would be understood in the future, e.g. after few decades, based on the descriptive information collected initially. Some event-related information, assumed to be generally known at the time of the event and, therefore, not stored along with it, might be forgotten or at least become less vivid over time. This might harm the understanding of the event at a future point in time. To overcome the evaluation of how present events will be remembered and understood in the future, in this section we make the opposite (but equivalent) jump and we investigate how to facilitate and enhance the comprehension of past events by retrieving contextual information related to them. For instance, the event description in Table 4.4, reporting the invasion of Kuwait by Iraq in 1990 (which led to the Golf War), could be better understood with the additional information in Figure 4.4, which provides more context on the possible reasons of the invasion as well as the reactions of other nations. While such side information could have been generally known as part of the context back then, it could not be so obvious for today's average readers. This contextual gap could become broader and broader when moving from worldwide events, such as the Golf War, to those with a smaller impact and echo.

In this scenario characterized by wide temporal and contextual gaps, we present methods to formulate queries from an event description to retrieve contextualizing information from a given document corpus, aiming at high recall. Note that adding context by entity linking based on a knowledge based (e.g. Wikipedia) [MC07, MW08] is not sufficient for different reasons: first, the length of each additional information piece should be substantially smaller than entire Wikipedia pages to be consumed in a short time and minimal disruption; second, only information related to the topic and participants of the event and temporally close to when it happened should be considered, while Wikipedia pages usually cover a much wider set of topics and timespan. Our best and novel method identifies queries expected to produce high recall by means of query performance prediction. Content-wise relevance and temporal proximity of information pieces with respect to input events, guaranteed by the kind of features incorporated into the prediction model, are necessary conditions for being good contextualization candidates. However, there can be other objectives in addition to those just mentioned, for instance simply optimizing precision, which is the reason why we focus on recall in the first place. The set of candidate information pieces retrieved by our models can then be used for re-ranking purposes according to such additional objectives, with the advantage of handling a bigger quantity of relevant results in the first place thanks to the maximization of recall in the retrieval phase. This makes our methods more versatile and employable for a broader range of applications. For instance, they have been used in [TCKN15a] to retrieve high-recall sets of candidate contextualizing results as starting point for a re-ranking that encourages diversity and complementarity between results and event descriptions.

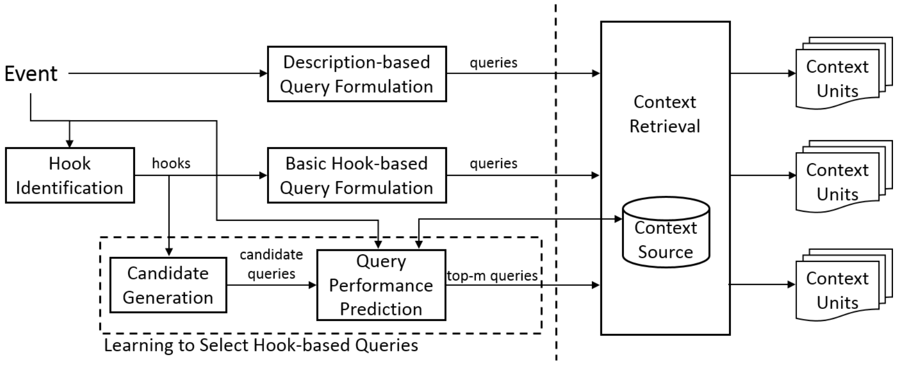


Figure 4.5 Overview of the different methods to formulate queries from event descriptions, which are used to retrieve contextualizing information.

4.4.1 Overview

Input events e are represented based on the model introduced in Section 4.3.1. Among the event attributes listed in Table 4.1, in the context of this section we particularly make use of their descriptions, lifespan (start-end date), and participants (as named entities automatically extracted from the description). In order to work with events that are sufficiently old, newsworthy, and well described, we use articles from the New York Times Archive[4] within the period 1987-1992 to represent the events to be contextualized. Each article is assumed to refer to one event, whose description is made of the article's title and lead paragraph. There is, however, one further information that is assumed to be provided as input to our methods: the *contextualization hooks* for each event. Each hook h for an event e consists in an aspect of e that requires further contextualizing information for its proper interpretation and can be, for example, an entity or concept mention, but also other terms or a short phrase [CTKN14]. Referring to the sample event in Table 4.4, hooks could be *Iraq*, *Kuwait*, *oil fields*, *Persian Gulf*. Since automatically identifying the aspects that would require contextualizing information is not a trivial task, which also depends on the background knowledge of the user, we decided to leave this task out from our problem definition by assuming their availability as known input. In other words, a user inspecting an event is expected to mark the parts of its description that he/she finds difficult to understand. Another key component of our problem definition is the context source, which contains a set of contextualization units cu that can be retrieved as additional information to better understand a given event. We chose Wikipedia as context source because it contains a large amount of knowledge at different levels of specificity, spanning diverse topics and a wide temporal horizon. A Wikipedia dump has been pre-processed as described in Section 4.4.6 to come up with a set of annotated and indexed Wikipedia paragraphs to be used as contextualization units.

An overview of the methods described in the rest of this section is shown in Figure

[4] http://catalog.ldc.upenn.edu/LDC2008T19

4.5. It also encompasses the retrieval of context units as final step exploiting our generated queries, although this part is out from the contributions of this section. Given an event and a set of contextualization hooks, we consider three families of methods for formulating queries to retrieve ranked lists of candidate contextualization units from a given context source. One uses the description of the event to be contextualized itself as a "generator" of queries (Section 4.4.2), while the others also include contextualization hooks (Sections 4.4.3 and 4.4.4). The results retrieved by these methods can be further re-ranked according to different and complementary aspects. Our main contribution is a procedure that constructs queries based on recall-oriented Query Performance Prediction (Section 4.4.4). Note that this component interacts with the context source due to the exploitation of post-retrieval features within the prediction model. Since some of these methods can generate more than one query for an input event, we experimented with two procedures to merge the ranked result lists retrieved by different queries. These are a round-robin strategy, which picks one result from each ranked list (skipping those results already encountered before), and CombSUM [FS94], which sums the scores of the same result from all the ranked lists where it was retrieved. In our experiments the round-robin method showed better performances than CombSUM, especially in terms of recall (which has been also observed by Tsagkias et al. [TdRW11]). Therefore, we will only report results obtained when using round-robin as strategy for merging ranked lists.

4.4.2 Description-based Query Formulation

The first and simpler family of query formulation methods exploits the textual description of input events. As explained at the beginning of Section 4.4.1, the event description field is filled with the title and lead paragraph of New York Times articles. We kept the distinction between these two pieces of text and, similarly to [TdRW11], we derive three methods to formulate queries from event descriptions: *title*, *lead*, and *title+lead*. *Title* formulates a query consisting of the article's title, which is indicative of its main topic. *Lead* uses the lead paragraph of the event description, representing a concise summary and including its main participants. *Title+lead*, as a combination of the previous two methods, formulates a query consisting of both the title and the lead paragraph of the description. This one stays applicable also in case there was no distinction between title and lead paragraph. Before being performed, all the queries are pre–processed by tokenization, stop-word removal, and stemming. We did not investigate further information extraction approaches for query formulation, since it has been already proven in [TdRW11] that the methods described above perform better than more complex information extraction techniques, e.g., keyphrase extraction.

4.4.3 Basic Hook-based Query Formulation

As already introduced in Section 4.4.1, the input events in our model come with a set of hooks explicitly representing the information needs of the reader or, in other words, what requires contextualization to be understood and interpreted. The analysis done in [CTKN14] showed that contextualization hooks are not only entity mentions, concept mentions, but also general terms and even short phrases. We consider two basic hook–based query formulation methods: *all_hooks* and *each_hook*. *All_hooks* includes all the hooks for an event in a single query, representing a tailored perspective of the user's combined information needs for the document. *Each_hook* queries each hook separately, focusing on specific information about single participants, aspects, or sub–topics of the event. The results retrieved by each query are merged in a round-robin fashion as described at the end of Section 4.4.1. The queries generated by these methods are augmented with the title part of the event description, under the assumption that it is a good representative of the event's topic.

We also experimented with more advanced methods based on identifying hook relationships, for instance considering their co–occurrence in a document collection. However, since these approaches did not perform better than the *all_hooks* method described before, we will not discuss them further.

4.4.4 Learning to Select Hook-based Queries

The basic methods described above might represent too drastic and opposite ways of combining contextualization hooks, which are either queried together or kept isolated from each other. We believe that intermediate and more flexible strategies should be further investigated. Different methods based on ranking and selection of query terms from an initial query might be employed [BC08, LCKC09, MC13], considering the entire set of hooks for an event as the initial query. We explore an adaptive method which formulates queries based on the characteristics of the input event and hooks. Our approach consists of predicting the performances of candidate queries representing subsets of hooks for a given event, ranking them according to the predicted performance, and selecting the top-m of them to be actually performed for the event. The value of m is identified through experiments. In contrast to previous works in query performance prediction, our prediction model is trained on *recall* instead of *precision*. Furthermore, we define novel features for query performance prediction that explicitly take the temporal dimension into account. Finally, our method assesses performances of subsets of query terms (hooks) and can generate more than one query (subsets of hooks), while most of the available approaches predict performances of single query terms separately and generate only one subset of them.

Candidate Queries

Given a event e and the set of its hooks H_e, we compute its power set $\mathcal{P}(H_e)$ and we create a candidate query for each set of hooks $p \in \mathcal{P}(H_e)$. Again, candidate queries are augmented with the title of the event description. The effort of the computation of features for each element in the power set is not critical in our scenario for two reasons. First, working with short text like the lead paragraph limits the number of hooks within it. Second, the features employed to predict the query performances (described hereafter) are either pre-retrieval measures, which can be computed offline, or do not require heavy post-retrieval computation.

Features

For each set of hooks $p \in P(H_e)$, we want to predict the performance of the query q_p containing all the hooks in p (along with the the title in the event description). We measure the performances of each candidate query in terms of its recall because, as already explained, our method is meant to serve as starting point for any subsequent re-ranking step, retrieving as many useful contextualization units as possible in the first place. In this work we predict query performances with a regression model learned via Support Vector Regression (SVR) [DBK+96]. In this model, each learning sample $s = \left(\boldsymbol{f}_q, r_q\right)$ consists in a feature vector $\boldsymbol{f}_q$ describing query q (as well as the event it refers to) and its recall r_q, i.e., the label to be predicted. Note that different numbers of top–l results can be used to compute the recall, i.e., the labels, and the choice is discussed in Section 4.4.7. The feature set that we use to represent queries and the event they belong to are described in the rest of this section. It is composed of novel temporal features for query performance prediction, along with more standard ones [CK12, HO04, MT05].

Linguistic Features. We compute a family of linguistic features [MT05] for a query by considering its text and the event it refers to. This results in a set of features both at query and event level: the length of the query, in words; the number of duplicate terms in the query; the number of entities (people, locations, organization, artifacts) in the query; the number of nouns in the query; the number of verbs in the query; the number of hooks in the query; the length of the title in the event description; the length of the lead paragraph in the event description; the number of participants in the event (entities extracted from its description, i.e. title and lead paragraph); the number of nouns in the event description; the number of verbs in the event description; the number of hooks for the event; the number of duplicates in the event description.

Document Frequency. The Document Frequency of a hook h represents the percentage of contextualization units in the corpus containing h and it is computed as:

$$df(h) = \log \frac{N_h}{N} \tag{4.1}$$

where N_h is the number of contextualization units in the corpus containing h and N is the size of the corpus. At event level, we compute the document frequency for every hook of the event the query belongs to, i.e., $df(h)\ \forall h \in H_e$, and then we derive aggregate statistics such as average, standard deviation, maximum value, minimum value. Similarly, at query level, we compute $df(h)$ for every hook in the query and we derive the same aggregate statistics as before. In the following, we will refer to average, standard deviation, maximum value, and minimum value simply as *aggregated statistics*.

Temporal Document Frequency. In order to restrict the popularity of a term to a particular time period $T = [t_0 - w; t_0 + w]$, we compute Eq. 4.1 only for those contextualization units having at least one temporal reference contained in T. This can be done efficiently since contextualization units in our corpus have been annotated with the temporal references mentioned in them. The time period we are interested in is centered around the time of the event (publication date of the related document) and the parameter w determines the width of the interval. After experimenting different values of w, we set $w = 2$ *years* for our study.

Scope. The scope of a query has been defined in [HO04] as the percentage of documents (contextualization units in our case) in the corpus that contain at least one query term. Besides the scope of the query itself, we also compute the scope of the event title and the scope of the event hooks H_e when queried together.

Temporal Scope. We define the temporal scope of a query as the percentage of contextualization units in the corpus that contain at least one query term and at least one temporal expression within a given time period. The time period that we consider is the same as the one considered for the computation of temporal document frequency, i.e. a period centered around the time of the event and with a temporal window equal to w. Again, we experimented different values of w and we set $w = 2$ *years*.

Relevance. For a given query q, we retrieve the top-k contextualization units and we compute aggregated statistics of their relevance scores given by the underlying retrieval model (described in Section 4.4.5). The value of k has been empirically set to 100 after experimenting different candidate values. We also computed relevance features at event level, using both event's title and all event's hooks as two distinct queries.

Temporal Similarity. For a given query q generated from an event e and every retrieved contextualization unit c in its top-k result set (again, $k = 100$), we compute the temporal similarity between q and c and we derive aggregated statistics over the elements in the result set. Temporal similarity between time points t_1 and t_2 is computed through the time-decay function [KN10]:

$$TSU(t_1, t_2) = \alpha^{\lambda \frac{|t_1 - t_2|}{\mu}} \tag{4.2}$$

where α and λ are constants, $0 < \alpha < 1$ and $\lambda > 0$, and μ is a unit of time distance.

We set $\lambda = 0.25$, $\alpha = 0.5$, and $\mu = 2$ *years* in our experiments. The temporal similarity between a query q and a result c is computed as $\max_{t \in T_c}\{TSU(t, t_e)\}$, where T_c is the set of temporal references mentioned in c and t_e is the time of the event q refers to. This can be done efficiently since temporal references mentioned in contextualization units have been extracted and stored at indexing time.

We also computed temporal similarity features (aggregated statistics) at event level, using both its title and concatenated hooks as two separate queries. The computation of these features is the same as the ones described above.

Query Ranking and Result Merging

Once we have a model that predicts the recall of an input query, we can use it to rank the candidate queries available for an input event. Given an event e, its hooks H_e, and its power set of candidate queries $P(H_e)$ described before, we predict the recall of each candidate $p \in P(H_e)$ by using our regression model. Then, the candidates are ranked by the predicted values and the top-m candidates are selected as queries to be actually performed for the event. The selection of the value of m will be discussed as part of our experiments (Section 4.4.7). Finally, the ranked result lists retrieved by each query are merged using a round-robin strategy as described in Section 4.4.1.

4.4.5 Retrieval Model

We use the query-likelihood language modeling [PC98] to rank and retrieve context units given a query. We briefly summarize it here for the sake of the reader. Given a query q generated for an input event by one of the methods described in the previous sections, the similarity of a context unit c (i.e. its relevance) with respect to q is computed based on the likelihood of generating the query q from a language model estimated from c:

$$P(c|q) \propto P(c) \prod_{w \in q} P(w|c)^{n(w,q)} \tag{4.3}$$

where w is one of the query terms in q, which are assumed to be independent from each other, $n(w, q)$ is the term frequency of w in q, and $P(w|c)$ is the probability of w estimated using Dirichlet smoothing:

$$P(w|c) = \frac{n(w, c) + \mu P(w)}{\mu + \sum_{w'} n(w', c)} \tag{4.4}$$

where μ is the smoothing parameter and $P(w)$ is the probability of each term w in the collection (context source).

All the query formulation methods described before have been evaluated using this retrieval model to retrieve context units. Also the post-retrieval features of our query performance prediction (Section 4.4.4) have been computed based on context units retrieved and ranked by this model.

4.4.6 Experimental Setup

In this section we describe the different ingredients of our experimental setup, namely the dataset, evaluation metrics, and baselines.

Document Collections. The problem that we tackle demands for sufficiently old events as input. In order to meet this requirement, we used the New York Times Annotated Corpus as source of events to be contextualized, which contains 1.8 million documents from January 1987 to June 2007. This also let us work with newsworthy and well described events. We employed Wikipedia as context source because it is considered the most comprehensive and up-to-date online encyclopedia, covering a wide topical and temporal range of both general and specific knowledge. We downloaded the Wikipedia dump of February 4, 2013 and we split each of the 4,414,920 Wikipedia articles within this snapshot in *paragraphs*. We got a total of 25,708,539 paragraphs, which we use as contextualization units. For each paragraph, we used Stanford CoreNLP [MSB+14] for tokenization, named entity annotation and temporal expression extraction. Additionally, we also extracted the *anchor* texts found in the hyperlinks of each paragraph. The whole set of annotated paragraphs have been indexed by using Apache Solr[5].

Ground truth. In order to obtain ground truth data for the evaluation of our methods, we manually selected a set of 51 eventful articles published in the first years of the New York Times dataset (29 articles in 1987, 2 articles in 1988, 6 articles in 1990, 7 articles in 1991, and 7 articles in 1992) and belonging to a diverse set of topics (politics, education, science, business, technology, sport). Six human annotators made use of an annotation interface to inspect pairs of event descriptions (title and lead paragraph of each article) and candidate context units for it. They were asked to annotate each event/context pair based on the following guidelines: a pair should be marked as *relevant* if the context unit both provides *additional information* complementing the one in the associated event description and enhances their understanding and interpretation of (at least) one of the aspects composing the event; otherwise, the pair should be declared as *irrelevant*. For each event (article), candidate context to be evaluated have been collected by manually casting a set of queries and retrieving the top-100 contextualization units for each of them. Additionally, we retrieved up to 20 candidate contextualization units using every query generated by each method (Sections 4.4.2, 4.4.3, 4.4.4). We merged the retrieved units and removed duplicates afterwards. In total, our annotated dataset consists of 9,464 event/context pairs, with an average of 26.9 relevant contextualization units evaluated for each event description. We measured the inter-annotator agreement by averaging the pairwise Cohen's kappa values of all the possible combinations of annotators who has an overlap of evaluated event/context pairs. Overall, we observed an agreement of 0.37, which can still be considered as fair given the high complexity and subjectivity of this contextualization task. This ground-truth has been made publicly available[6] to encourage

[5] https://lucene.apache.org/solr/

[6] http://www.l3s.de/~ntran/contextualization/

further research on this challenging task.

Parameter Settings. For query performance prediction, the regression model described in Section 4.4.4 was trained by using the Support Vector Regression implementation of LibSVM[7]. In particular, we trained a n–SVR model with Gaussian Kernel through 10–fold cross validation at the event-level. The open parameters were tuned via grid search to $C = 3$, $\gamma = 0.5$ and $\nu = 0.75$. Linguistic features were extracted using Stanford CoreNLP [MSB+14]. Regarding the Dirichlet smoothing, we set $\mu = 2000$ and $\lambda p(w_i) = 0.5$. For computing temporal similarity feature, we set $\lambda = 0.25$, $\alpha = 0.5$, and $\mu = 2$ *years* in our experiments. We also observed that changing those parameters did not affect the correlation capabilities of the feature.

Evaluation Metrics. We center our evaluation on recall because our main goal is to retrieve large result sets including as many *relevant* contextualization units as possible, which can be used as starting point for re-ranking (which rather focuses on precision). We measure the recall of each query formulation method at different size of result set, from $k = 10$ to $k = 500$. Note that recall is measured only with respect to the relevant contextualization units collected as described before. This means that, especially for large values of k, other relevant units might have been retrieved by our methods without being counted for the computation of recall (since they were not evaluated). For sake of completeness, we also report precision at $k = 1, 3, 10$ (P@1, P@3, P@10, respectively) and mean average precision (MAP). All the reported scores are averaged over the test sets of the splits in the cross validation. The statistical significance of improvements was measured using a two-tailed paired t-test and is marked as $^{\blacktriangle}$ and $^{\vartriangle}$ (with $p < 0.01$ and $p < 0.05$, respectively).

Baselines. Our main contribution is the method in Section 4.4.4, which formulates queries through a recall-based query performance prediction. Therefore, we will regard both the description-based (Section 4.4.2) and basic hook-based (Section 4.4.3) query formulation as baselines. The description-based approach has been also adopted by Tsagkias et al. [TdRW11], which is another reason to consider this query formulation as a baseline.

4.4.7 Results

We now evaluate and compare the performances of the different query formulation approaches focusing on the recall metric. Since these methods generate different numbers of queries, we allow each of them to retrieve the same number of results k. The strategy chosen to create a single result set of k elements from different ranked lists have been discussed in Section 4.4.1. The query-likelihood language model has been used for retrieval as described in Section 4.4.5.

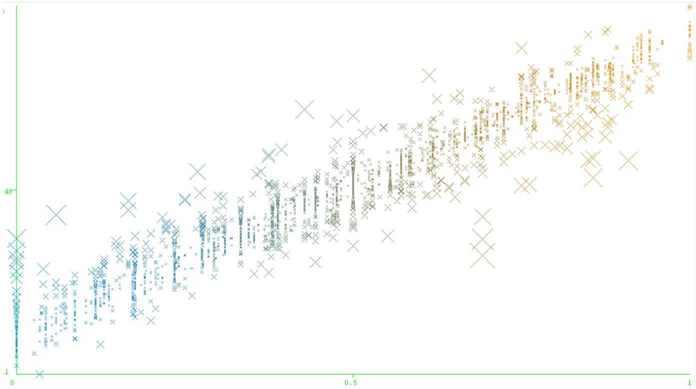


Figure 4.6 Correlation plot between labels and predictions.

Performance Prediction Model

The query formulation method described in Section 4.4.4 is based on predicting the performances (recall in our case) of candidate queries, ranking them according to the prediction, and then using the top-m queries to retrieve results. Thus, the quality of the query performance prediction itself has to be evaluated before assessing and comparing the performances of the whole query formulation method. The regression model has been trained via 10–fold cross validation at event-level, and the results reported hereafter have been averaged over the test sets of the 10 folds. The Correlation Coefficient is equal to 0.973, the Root Mean Squared Error equals to 0.056, and the Mean Absolute Error equals to 0.037. We also show the correlation plot between true labels and predictions in Figure 4.6. The low error values and high correlation value, if compared with the performances in predicting query precision reported in previous works (e.g. [CYT10, RK14]), show that the recall of queries in our task can be predicted quite accurately by using the features described in Section 4.4.4.

Feature Analysis. In order to analyze which are the most important features in our model, we identified the top–10 features according to their absolute correlation coefficient. Referring to Section 4.4.4, these are: *max query relevance*, *number of hooks for event*, *min event's hooks df*, *max event's hooks temporal df*, *event's hooks scope*, *avg query temporal similarity*, *event's title temporal scope*, *std query relevance*, *avg event's title temporal similarity*, and *std query temporal similarity*. The presence of temporal document frequency, temporal similarity, and temporal scope shows that the temporal features that we introduced play an important role in the model. We can also note that both query–level and event–level features are important, since the set is made of 4 features from the former and 6 features from the latter class. Finally, there is only one linguistic feature in the set, namely the number of hooks in the event description (even barely linguistic), showing that this class of features alone does not correlate well with query performances. This fact was pointed out already in [CK12].

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

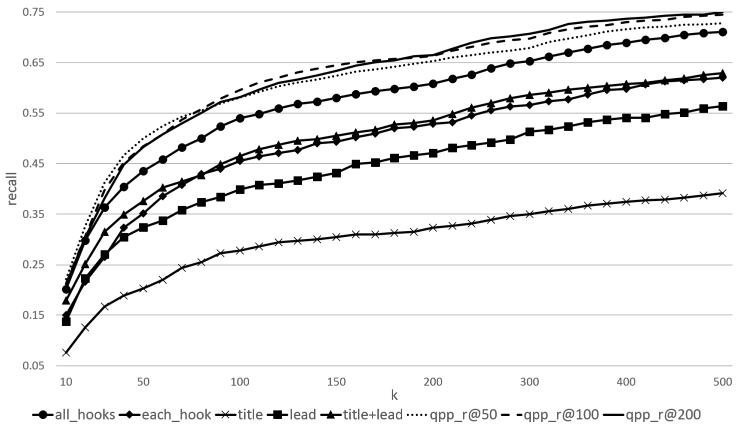


Figure 4.7 Recall curves of description-based and hook-based methods.

Recall-based Comparison

We compare recall values for the description–based methods (*title*, *lead*, *title+lead*), the basic hook–based methods (*each_hook*, *all_hooks*), as well as the method based on query performance prediction, hereafter called *qpp*. For the latter method, we report the performances achieved when using prediction models trained with different labels: we experimented with different l values, namely $l = 50, 100$ and 200, for the computation of the recall at l to be used as label. For instance, the method *qpp*_r@100 uses a query performance prediction model that has been trained with samples (i.e. queries) having the value of recall that they achieved at $k = 100$ as label. These three methods will be named *qpp_r@50*, *qpp_r@100* and *qpp_r@200*, respectively, in the rest of the experiments. Note that each *qpp* method considered here uses the top-2 queries, according to their predicted performances, to retrieve the results. The choice of selecting $m = 2$ queries will be explained later on.

The recall curves of the different methods, for different values of top–k results, are shown in Figure 4.7. The curves of *title* and *lead* are the lowest ones, while their combination ($title + lead$) becomes comparable with $each\_hook$. Querying using all the hooks of an event together, i.e., all_hooks, leads to higher recall values than all the aforementioned methods, showing that performing hook–based queries does bring better performances in terms of recall with respect to description–based methods. The difference in performances between $each\_hook$ and all_hooks is due to the fact that querying all the hooks together prefers contextualization candidates that contain many hooks. These are potentially more relevant, as they refer to different aspects (hooks) of the same event. Regarding the *qpp* methods, for $k > 20 - 30$, the achieved recall values are between 3% and 7% higher than the ones obtained by all_hooks. For larger values of k, e.g. $k > 400$, the difference between the *qpp* methods and all_hooks reduces because the prediction models used by the *qpp* methods have been optimized for lower values of k (recall that $l = 50, 100, 200$). It is possible to note that the choice of the recall value used as label affects the shape of the recall curve:

Table 4.5 Recall of *all_hooks* and *qpp* methods over different classes of events grouped by their retrieval difficulty. Significant improvement is assessed between values in different columns (methods) for the same metric and retrieval difficulty.

	R@50		R@100		R@200	
	qpp	*all_hooks*	*qpp*	*all_hooks*	*qpp*	*all_hooks*
easy	0.6208$^{\vartriangle}$	0.5666	0.7361$^{\vartriangle}$	0.6969	0.7951$^{\blacktriangle}$	0.7686
hard	0.3837$^{\blacktriangle}$	0.3094	0.4606$^{\blacktriangle}$	0.3892	0.5391$^{\blacktriangle}$	0.4550

a model trained with recall at k as label is optimized for that k and should perform better than other models in the neighborhood of k. For instance, *qpp_r*@50 is the best around $k = 50$ and then it is beaten by both *qpp_r*@100 and *qpp_r*@200 for increasing values of k. This means that, if the number of k results to be retrieved is known and fixed in advance, this information can be exploited early in the training of the query performance prediction model by setting $l = k$, leading to higher recall values for that particular k.

Another comparative analysis between *qpp* methods and *all_hooks* can be done by categorizing the events according to their *difficulty*, which we define in terms of the amount of relevant context that can be retrieved for a given input event. This means that difficult events are those for which few relevant context can be retrieved. We categorize input events in *easy* and *hard* with respect to the *all_hooks* method, since it represents a baseline in this comparative analysis with *qpp* methods. The recall curves obtained by *all_hooks* for the different events are shown in Figure 4.8 (dotted curves), where we can observe a wide variability with respect to the average one (continuous curve). The splitting of the events in easy and hard was performed based on the recall at $k = 200$ of *all_hooks*. Since the recall values associated to the different events exhibited a uniform distribution, we split them in two equal parts, one representing *easy* events and the other representing *hard* events. Table 4.5 shows the performances of *qpp_r*@50, *qpp_r*@100, and *qpp_r*@200 compared to the ones of *all_hooks* for the different categories of difficulty. The comparison between each *qpp* method and *all_hooks* is done considering the recall at those k values used to train the prediction model (i.e. $k = l$, $l = 50, 100, 200$). For instance, the comparison between *qpp_r*@100 and *all_hooks* is done considering recall at k=100. Besides *qpp_r*@50, *qpp_r*@100, and *qpp_r*@200 are on average better than *all_hooks* both for easy and hard events, their improvements are greater for the hard ones. In case of *qpp_r*@100, for instance, the relative improvement with respect to the recall value achieved by *all_hooks* is 5.6% for easy events and 18.3% for the hard ones. We believe that the capability of getting higher recall improvements for events whose relevant context units are difficult to retrieve is a valuable characteristic of the *qpp* methods.

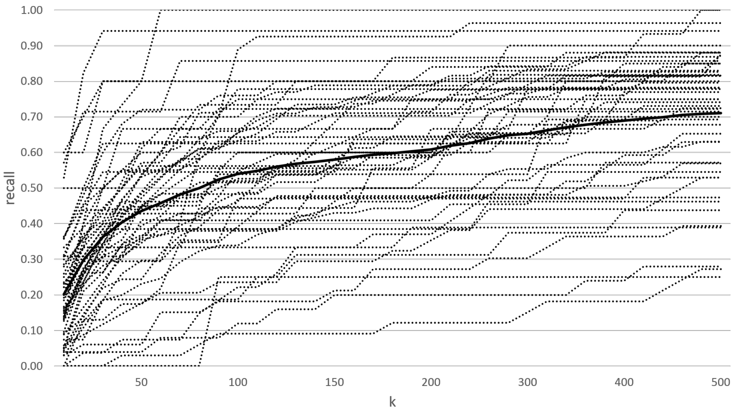


Figure 4.8 Recall curves of *all_hooks* over events.

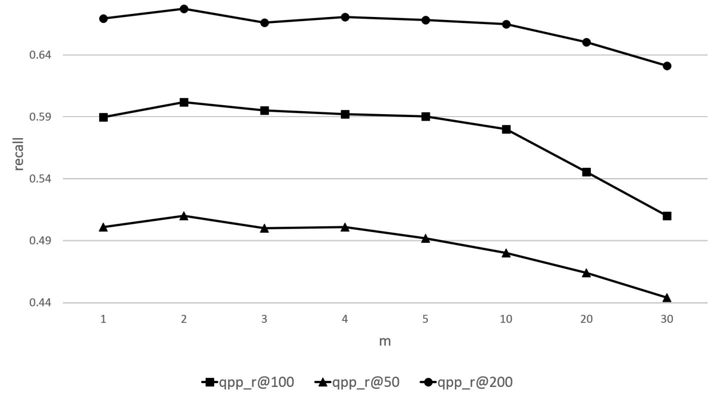


Figure 4.9 Recall values of *qpp_r*@50, *qpp_r*@100, and *qpp_r*@200 by varying the number of top–m queries.

As a summary of this section, we proved that exploiting hooks in query formulation is more effective, in terms of recall, than description–based query formulation methods. Moreover, we showed that learning to select candidate hook-based queries can be better (again in terms of recall) than the basic hook–based query formulation methods, which either query each specified hook separately or group all of them in a single query.

Number of Queries

The number of top ranked queries that *qpp* methods perform is an open parameter, which we tuned via an empirical analysis observing the recall performances when selecting different numbers of top–m ranked queries. Remember that, for sake of fair comparison, we allow each method to pick the same number of results k from the result lists retrieved by the queries that it generated for a given event (the two meth-

Table 4.6 Precision of description-based and hook-based query methods. Significant improvement is assessed with respect to Row 1 (within the first group) and Row 3 (for the second and third groups).

	P@1	P@3	P@10	MAP
Description-based query models				
title	0.2156	0.1895	0.1745	0.2446
lead	0.4902$^{\blacktriangle}$	0.4641$^{\blacktriangle}$	0.3333$^{\blacktriangle}$	0.4908$^{\blacktriangle}$
title+lead	0.5294$^{\blacktriangle}$	0.4705$^{\blacktriangle}$	0.3901$^{\blacktriangle}$	0.5161$^{\blacktriangle}$
Basic hook-based query models				
each_hook	0.3333	0.3464	0.2745	0.4003
all_hooks	0.5490	0.5098	0.4137	0.5640
Query performance prediction model				
qpp_r@100	0.5882	0.5490$^{\blacktriangle}$	0.4529$^{\blacktriangle}$	0.5802$^{\blacktriangle}$

Table 4.7 Precision of *all_hooks* and *qpp_r@100* on the subset of events categorized as *hard* according to their retrieval difficulty.

	P@1	P@3	P@10	MAP
all_hooks	0.5000	0.3462	0.2885	0.4487
qpp_r@100	0.5000	0.4743$^{\vartriangle}$	0.3730$^{\vartriangle}$	0.5048$^{\vartriangle}$

ods that we considered to merge ranked lists have been described in Section 4.4.1). This means that increasing the number of queries to be selected and performed does not necessarily lead to higher recall. Figure 4.9 shows the recall values of *qpp_r*@50, *qpp_r*@100 and *qpp_r*@200 (computed at top–50, top–100 and top–200 results, respectively) for different numbers of top–m selected queries. A common trend over the different curves can be observed: they stay quite stable for small values of m, exhibiting a little peak for $m = 2$, and then they decrease for increasing values of m. After observing this behavior, we decided to fix the number of performed queries to $m = 2$.

Precision-based Comparison

After having centered our evaluation on recall so far, as most important metrics for our goal, we briefly compare the query formulation methods in terms of precision metrics. Their performances are summarized in Table 4.6. The best method among the description-based ones (first group from top) is *title+lead*, which also outperforms *each_hook*. The other two hook-based methods, namely *all_hooks* and *qpp_r@100*, do obtain higher scores than *title+lead*. The best method overall is again *qpp_r@100*, whose improvement with respect to *title + lead* resulted to be statistically significant (except for P@1). The comparison in Table 4.7 between *all_hooks* and *qpp_r@100* on the set of *hard* events introduced before shows that the latter obtains significant improvement over the former. These results corroborate the hierarchies among the methods that came out already for recall. The only exception is that *title+lead* is clearly better than *each_hook* in terms of precision, while they were comparable when considering recall.

Re-ranking. We have mentioned many times in the whole Section 4.4 that the main goal of our query formulation is to retrieve result sets containing as many relevant information as possible, in order to provide a better basis for any re-ranking approach. We conclude our evaluation mentioning that the query formulation methods described in this section have been indeed used in [TCKN15a] to effectively re-rank high-recall sets of contextualization units for the re-contextualization of past news articles. This work proposes a re-ranking model that balances the content-based and temporal relevance of the contextualization units with their complementarity to the content of the input news article, to provide relevant and diverse information to a reader. This intuition is enacted by defining a set of novel features, which are combined together in different Learning to Rank models. We do not describe this method further, since it is not part of the contributions of this chapter, but we refer the reader to [TCKN15a] for more details.

4.5 Conclusion

In this chapter we have touched upon two aspects regarding the curation and reuse of event-related information over time. First, we have discussed how the ubiquity of events can make them deeply differ from each other, e.g. in terms of topic, duration, complexity, granularity of description. We therefore made a comparative analysis among three different event sources that highlighted the differences and potential complementarity of the events they contain. This served as a motivation for gathering and maintaining events from different sources into a global, more comprehensive, but yet not redundant event repository. Then we made a temporal jump and investigated how such stored events and their descriptions could be properly understood after few decades. In this scenario, the understanding of events could be harmed due to the fading of important contextual information commonly known at the time of the event. We proposed methods to cast queries from old event description to retrieve candidate contextual information, which should be topically and temporally relevant to the input event. Our best performing approach was based on query performance prediction maximizing recall and encompassing a novel set of features. We aimed at high-recall instead of other metrics (e.g. precision) to make our query formulation methods reusable for other re-ranking purposes, which would target additional objectives while still taking advantage of an initial result set including as many relevant results as possible.

5 Conclusion and Future Work

Events play a fundamental role in shaping today's world. Nowadays, thanks to the technological advances of the digital age, event-related information is digitally embedded in any form of media and serves several purposes. While the tasks of detecting and describing events have drawn research and technological effort for decades, still several issues need to be addressed for supporting the effective exploitation of event-related information at subsequent stages. Solutions to those issues have to consider a wide and diverse range of scenarios due to the ubiquity of events in many settings of the physical and digital world. In this conclusive chapter, we summarize the ones that have been developed and evaluated in this thesis and we provide indications for possible future directions of research regarding each of them.

5.1 Summary

In Chapter 2 we considered personal events captured in form of photo collections and we investigated how to keep their revisiting enjoyable at future points in time, since the mere size of such collections makes sorting pictures or just going through them cumbersome processes. To overcome this problem, we aimed at automatically identifying those photos that are perceived as most important by the collection owner, where the notion of image importance is driven by user expectations and represents what photos users perceive as important and would select for the purposes of long-term preservation and revisiting. This concise set of highly important pictures would make the future reminiscence of the related event more enjoyable and less demanding. First, in order to better understand the selection process in this scenario as well as to acquire evidences of user expectations, we performed a user study where participants selected, from their own photo collections, those pictures that they would like to preserve for future revisiting, namely those most important to them. The fact that many hidden and subjective selection criteria (memory evocation, personal importance, image typicality) were rated as highly important already signaled the difficulty of automatizing the selection task. The participants accredited an important role also to the aspect of coverage, which consists in the set of selected pictures being a representative sample of the original collection. Interesting however, their selections actually exhibited only a poor degree of coverage. Afterwards, we presented an expectation-oriented method, relying on Machine Learning and an extensive set of visual and semantic features, to estimate the aforementioned photo importance based on user expectations and to identify subsets of most important pictures according to it. Since the concept of coverage resulted to be an important criterion in our user study and it is also the dominant ingredient of a wide part of state of the art methods, we also investigated how to combine it with our expectation-oriented selection. In our experiments encompassing real-world photo collections and considering human selections as evaluation criterion, our method outperformed such state of the art works and comparable results to it were achieved only when not considering coverage as a primary selection aspect.

We then moved towards more public settings and in Chapter 3 we introduced the problem of validating the verity of real-world events by looking for evidences of their occurrence in a corpus of textual documents, which serves as ground truth. In scenarios where events are detected by automatic methods, which are subject to errors, being able to handle true events and to ignore the false ones becomes fundamental to ensure a reliable exploitation of such event-related information for any other purpose. Also, event validation can be used in a more exploratory way to find documents that corroborate and explain the occurrence of manually specified events. Since manually inspecting large amounts of text in search of proofs of event occurrences is a cumbersome procedure, we aimed at automatizing the validation process by retrieving from the ground truth a set of candidate documents related to an event to be validated, and checking whether any of them contains evidences of the occurrence of the event. We employed supervised Machine Learning to carry out this task, reaching substantial agreement when comparing the decisions of our automatic validation with those made by human evaluators. We also explored the effects of performing event validation as a post-processing step of event detection and we observed an increase of precision within the set of detected events while preserving recall. Finally, we presented and released a novel crowdsourced dataset for benchmarking the performances of event validation methods.

Finally, in Chapter 4 we discussed two other matters that affect public events after they have been detected and validated. The first one regards the process of storing events and their descriptions into repositories for easing their access and inspection. In this regard, one significant fact is that the ubiquity of events makes them differ from each other under several aspects, e.g. topic, duration, newsworthiness, complexity. By comparing three event sources, we highlighted their differences to motivate the importance of collecting events from complementary sources and merging them into a global and more comprehensive one, so that it accommodates events with different characteristics and can satisfy information needs in a broader range of contexts. Then, we delved deeper into the temporal dimension and questioned how events could be properly understood after relatively long time periods (such as decades), only based on their descriptions originally collected when they occurred. Properly understanding old events usually requires the knowledge of contextual information belonging to the time of the events. However, while such information might have been believed to be commonly known at that time, and therefore not explicitly stated within the event descriptions, it might no longer be common background knowledge given the presence of large temporal and contextual gaps. In order to cope with these issues, we explored different methods to formulate queries from event descriptions as seeds for retrieving additional context. By exploiting recall-based Query Performance Prediction, the generated queries were able to retrieve topically and temporally relevant contextualization candidates with high recall. We explicitly targeted recall as query performance criterion so that the set of retrieved results can be used as a favorable basis within different subsequent computations (e.g. re-ranking), since any of them could pursue additional objectives while still benefiting from having more topically and temporally relevant results in the first place.

5.2 Future Work

While presenting approaches and analyses to address issues regarding the usage and exploitation of event-related information in different scenarios, the work of this thesis also disclosed challenging aspects that have not been explored yet. In this section we indicate promising research directions to be pursued in the near future.

The problem of selecting important pictures from photo collections of personal events, discussed in Chapter 2, still contains many aspects to be investigated. First, while the current selection model estimates the long-term importance of each single photo in isolation, the selection or unselection of one image might affect the selection probabilities of other images in the collection. One possible approach to take this into account consists in generating subsets of photos as candidate selections, either from scratch or starting from the one created by our method, and then predicting the probability of each candidate to be the set that the user would have actually selected. Second, user input could be required to get evidences of hidden information related to pictures, e.g. context, memories, background knowledge regarding people or places, which are difficult (if not impossible) to be recognized from the mere visual content. In order to still keep the amount of user investment as low as possible, the selection model should recognize those pictures whose importance estimation is uncertain and ask for further information only for them, similarly to the active learning paradigm. Third, it would be interesting to investigate how the selection outcomes vary when targeting different selection recipients (e.g. relatives, friends, colleagues) or preservation horizons.

Expansions of our approach to Event Validation (Chapter 3) regard both the retrieval of candidate documents for a given event to be validated and the model for inferring its actual occurrence in text. The former case encompasses the experimentation of other document corpora as ground truth, for instance Social Networks and Wikipedia, each one being different in terms of trustworthiness as well as length and content of documents in them. Also the estimation of the newsworthiness or impact of an event would influence the retrieval process, for instance by performing more specific queries and retrieving more results for less newsworthy events, whose evidences might be more difficult to be found. Regarding the latter case, promising expansions consist of including more semantic representations of both input events and documents, either by means of deep learning or feature engineering, and explaining the relationships and roles among event participants within the same event. Also the estimation of the credibility of a document (or corpus), more recently studied by Samadi et al. [STVB16], could play an important role in the validation process.

The most effective methods for formulating queries from event descriptions (Chapter 4) assumed the aspects requiring further contextualizing information to be specified as input by the reader, which is perhaps the greatest limitation of our approaches and the one that should be tackled first in the future. Identifying what requires additional information to be properly understood would require, for instance, the recognition of obsolete terms, topics whose meaning and common opinion has changed over time, and any relevant people or fact that might have been forgotten over time. Even more challenging, there might be a dependence on the readers, because the set of aspects for which they need further information might change based on their background knowledge (e.g. different education, profession, age, etc.).

We conclude by pointing to two general future research directions. Given the recent advances introduced by Deep Learning in many domains, e.g. vision, speech, text, one direction is definitely its experimentation within all the problems discussed in this thesis. This would consist in both employing Deep Neural Networks for feature extraction, as already done for the image representations in Chapter 2, and designing network architectures tailored for each of our tasks. The second direction regards more synergistic exploitations of visual and textual event-related information, going beyond their independent processing. While the usage of both text and images to summarize and contextualize events has been previously considered (e.g. in [LWL+11, LTW+16, LLW+13, LLZL10]), we believe that a deeper understanding, mining, and exploitation of their cross-modal relationships is needed, as also suggested by other works [Bat14, HE17]. Such synergies would be valuable not only for public events, which perhaps have received more attention due to the large availability of textual and visual data about them, but also for events shared on social networks and for even more personal settings.

Bibliography

[ABK$^+$07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735, 2007.

[AC14] Jun Araki and Jamie Callan. An annotation similarity model in passage ranking for historical fact validation. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1111–1114, 2014.

[ACBDN17] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. Event recognition in personal photo collections via multiple instance learning-based classification of multiple images. *Journal of Electronic Imaging*, 26(6):060502, 2017.

[ACDN18] Kashif Ahmad, Nicola Conci, and FGB De Natale. A saliency-based approach to event recognition. *Signal Processing: Image Communication*, 60:42–51, 2018.

[AG17] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[AHES09] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 1597–1604, 2009.

[AK15] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[All02] James Allan. *Topic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publishers, 2002.

[ANP07] Panagiotis Antonopoulos, Nikos Nikolaidis, and Ioannis Pitas. Hierarchical face clustering using sift image features. In *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*, CIISP '07, pages 325–329, 2007.

[APL98] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 37–45, 1998.

[APM14] K. Apostolidis, C. Papagiannopoulou, and Vasileios Mezaris. CERTH at MediaEval 2014 synchronization of multi-user event media task. In *Proceedings of MediaEval 2014 Workshop*, 2014.

[AS12] Charu C Aggarwal and Karthik Subbian. Event detection in social streams. In *Proceedings of the 2012 SIAM international conference on Data Mining*, SDM '12, pages 624–635, 2012.

[AZ13] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 1578–1585, 2013.

[Bat14] John Bateman. *Text and image: A critical introduction to the visual/verbal divide.* Routledge, 2014.

[BC08] Michael Bendersky and W. Bruce Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 491–498, 2008.

[BCJC13] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 459–460, 2013.

[BEWB05] Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6(Sep):1579–1619, 2005.

[BGL$^+$93] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS '93, pages 737–744, 1993.

[BGVG13] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Event recognition in photo collections with a stopwatch HMM. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '13, pages 1193–1200, 2013.

[BH10] Cosmin Adrian Bejan and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1412–1422, 2010.

[BHDR11] Marc Bron, Bouke Huurnink, and Maarten De Rijke. Linking archives using document enrichment and term selection. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries*, TPDL '11, pages 360–371, 2011.

[BPAK17] Christina Boididou, Symeon Papadopoulos, Lazaros Apostolidis, and Yiannis Kompatsiaris. Learning to detect misleading content on twitter. In *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval*, ICMR '17, pages 278–286, 2017.

[BPK$^+$14] Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 743–748, 2014.

[Cer18] Andrea Ceroni. *Personal Photo Management and Preservation. In: Mezaris V., Niederée C., Logie R. (eds) Personal Multimedia Preservation – Remembering or Forgetting Images and Videos.* Springer, 2018.

[CF14] Andrea Ceroni and Marco Fisichella. Towards an entity-based automatic event validation. In *Proceedings of the 36th European Conference on Information Retrieval*, ECIR '14, pages 605–611, 2014.

[CFGW05] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(3):269–288, 2005.

[CG14] Gordon V. Cormack and Maura R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 153–162, 2014.

[CG16a] Jorge E Camargo and Fabio A González. Multimodal latent topic analysis for image collection summarization. *Information Sciences*, 328:270–287, 2016.

[CG16b] Mário Cordeiro and João Gama. Online social networks event detection: a survey. In *Solving Large Scale Learning Tasks. Challenges and Algorithms*, pages 1–41. Springer, 2016.

[CGF15] Andrea Ceroni, Ujwal Kumar Gadiraju, and Marco Fisichella. Improving event detection by automatically assessing validity of event occurrence in text. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1815–1818, 2015.

[CGF17] Andrea Ceroni, Ujwal Gadiraju, and Marco Fisichella. JustEvents: A crowdsourced corpus for event validation with strict temporal constraints. In *Proceedings of the 39th European Conference on Information Retrieval*, ECIR '17, pages 484–492, 2017.

[CGG$^+$14] Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini, and Marco Fisichella. Information evolution in wikipedia. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 1–10, 2014.

[CGM$^+$16] Andrea Ceroni, Ujwal Gadiraju, Jan Matschke, Simon Wingert, and Marco Fisichella. Where the event lies: Predicting event occurrence in textual documents. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1157–1160, 2016.

[CHV99] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[CK12] David Carmel and Oren Kurland. Query performance prediction for IR. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1196–1197, 2012.

[CL08] Wei-Ta Chu and Chia-Hung Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 829–832, 2008.

[CME18] Andrea Ceroni, Chenyang Ma, and Ralph Ewerth. Mining exoticism from visual content with fusion-based deep neural networks. In *Proceedings of the 8th ACM International Conference on Multimedia Retrieval*, ICMR '18, pages 37–45, 2018.

[CMP11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web*, WWW '11, pages 675–684, 2011.

[CN16] Chen Chen and Vincent Ng. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI '16, pages 2913–2920, 2016.

[Cop14] Andrea Copeland. The use of personal value estimations to select images for preservation in public library digital community collections. *Future Internet*, 6(2), 2014.

[CP00] Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS '00, pages 388–394, 2000.

[CSF$^+$15] Andrea Ceroni, Vassilios Solachidis, Mingxin Fu, Nattiya Kanhabua, Olga Papadopoulou, Claudia Niederée, and Vasileios Mezaris. Investigating human behaviors in selecting personal photos to preserve memories. In *2015 IEEE International Conference on Multimedia & Expo Workshops*, ICME Workshops '15, pages 1–6, 2015.

[CSK13] David Carmel, Anna Shtok, and Oren Kurland. Position-based contextualization for passage retrieval. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1241–1244, 2013.

[CSN$^+$15] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. To keep or not to keep: An expectation-oriented photo selection method for personal photo collections. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 187–194, 2015.

[CSN$^+$17] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, and Vasileios Mezaris. Expo: An expectation-oriented system for selecting important photos from personal collections. In *Proceedings of the 7th ACM International Conference on Multimedia Retrieval*, ICMR '17, pages 452–456, 2017.

[CTKN14] Andrea Ceroni, Nam Khanh Tran, Nattiya Kanhabua, and Claudia Niederée. Bridging temporal context gaps using time-aware re-contextualization. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1127–1130, 2014.

[CTZC02] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 299–306, 2002.

[CWX$^+$07] Jingyu Cui, Fang Wen, Rong Xiao, Yuandong Tian, and Xiaoou Tang. Easyalbum: An interactive photo annotation system based on face clustering and re-ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 367–376, 2007.

[CXL$^+$15] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, Jun Zhao, et al. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL '15, pages 167–176, 2015.

[CYT10] David Carmel and Elad Yom-Tov. Estimating the query difficulty for information retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 911–911, 2010.

[DB07] Nicholas DiFonzo and Prashant Bordia. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35, 2007.

[DBK$^+$96] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS '96, pages 155–161, 1996.

[DDN14] M. S. Dao, A. D. Duong, and F. G. B. De Natale. Unsupervised social media events clustering using user-centric parallel split-n-merge algorithms. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '14, pages 4798–4802, 2014.

[DDNDN14] Minh-Son Dao, Duc-Tien Dang-Nguyen, and Francesco GB De Natale. Robust event discovery from photo collections using signature image bases (SIBs). *Multimedia Tools and Applications*, 70(1):25–53, 2014.

[DGM06] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Proceedings of the 1st International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW '06, pages 177–190. 2006.

[DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In

Proceedings of the European Conference on Computer Vision, ECCV '06, pages 288–301, 2006.

[DOB11] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1657–1664, 2011.

[DSJY11] A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 207–216, 2011.

[Dug13] Maeve Duggan. Photo and video sharing grow online. Pew Research Center, 2013.

[EdV13] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2):121–137, 2013.

[FBC12] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 171–175, 2012.

[FBJ15] John Foley, Michael Bendersky, and Vanja Josifovski. Learning to extract local events from the web. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 423–432, 2015.

[FCDN14] Marco Fisichella, Andrea Ceroni, Fan Deng, and Wolfgang Nejdl. Predicting pair similarities for near-duplicate detection in high dimensional spaces. In *Proceedings of the 25th International Conference on Database and Expert Systems Applications*, DEXA '14, pages 59–73, 2014.

[FCS$^+$15] Mingxin Fu, Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. Learning personalized expectation-oriented photo selection models for personal photo collections. In *Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops*, ICME Workshops '15, pages 1–6, 2015.

[FS94] Edward A Fox and Joseph A Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference*, TREC-2, pages 243–252, 1994.

[FYYL05] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 181–192, 2005.

[GEL+15] Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, and René Speck. Defacto – temporal and multilingual deep fact validation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:85–101, 2015.

[GGMPW02] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '02, pages 326–335, 2002.

[GLD12] Wei Gao, Peng Li, and Kareem Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1173–1182, 2012.

[GLO+16] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[GNC+14] Ujwal Gadiraju, Kaweh Djafari Naini, Andrea Ceroni, Mihai Georgescu, Dang Duc Pham, and Marco Fisichella. Wikipevent: Temporal event data for the semantic web. In *Proceedings of the 13th International Semantic Web Conference (Posters & Demonstrations Track)*, ISWC '14, pages 125–128, 2014.

[GPS99] Gene Golovchinsky, Morgan N. Price, and Bill N. Schilit. From reading to retrieval: Freeform ink annotations as queries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 19–25, 1999.

[Har79] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[HCL07] Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, 2007.

[HCMB03] Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-free news search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 1–10, 2003.

[HE17] Christian Andreas Henning and Ralph Ewerth. Estimating the information gap between textual and visual representations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ICMR '17, pages 14–22, 2017.

[HFG$^+$12] Dirk Hovy, James Fan, Alfio Gliozzo, Siddharth Patwardhan, and Chris Welty. When did that happen?: linking events and relations to timestamps. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 185–193, 2012.

[HFK$^+$16] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12–22, 2016.

[HHdJ08] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1419–1420, 2008.

[HO04] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *Proceedings of the 11th International Conference on String Processing and Information Retrieval*, SPIRE '04, pages 43–54, 2004.

[HR04] Peter Howarth and Stefan Rüger. Evaluation of texture features for content-based image retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, CIVR '04, pages 326–334, 2004.

[HSBW12] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2012.

[HSZ11] Xianpei Han, L Sun, and J Zhao. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 765–774, 2011.

[HXW$^+$11] Lei Huang, Tian Xia, Ji Wan, Yongdong Zhang, and Shouxun Lin. Personalized portraits ranking. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 1277–1280, 2011.

[HYB$^+$11] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, 2011.

[HYL$^+$03] Jeffrey Ho, Ming-Husang Yang, Jongwoo Lim, Kuang-Chih Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of the 2003 IEEE conference on Computer Vision and Pattern Recognition*, CVPR'03, pages 11–18, 2003.

[HZM$^+$11] Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 1127–1136, 2011.

[Itt73] Johannes Itten. *The art of color: the subjective experience and objective rationale of color*. John Wiley New York, 1973.

[Joh67] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[JPD$^+$12] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(9):1704–1716, 2012.

[KCJ$^+$13] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *Proceedings of the 13th IEEE International Conference on Data Mining*, ICDM '13, pages 1103–1108, 2013.

[KN10] Nattiya Kanhabua and Kjetil Nørvåg. Determining time of queries for re-ranking search results. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'10, pages 261–272, 2010.

[KN11] Nattiya Kanhabua and Kjetil Nørvåg. Time-based query performance predictors. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1181–1182, 2011.

[KNS13] Nattiya Kanhabua, Claudia Nieder and Wolf Siberski. Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the 10th International Conference on Preservation of Digital Objects*, iPres '13, 2013.

[KSFS05] Tim Kindberg, Mirjana Spasojevic, Rowanne Fleck, and Abigail Sellen. The ubiquitous camera: An in-depth study of camera phone use. *IEEE Pervasive Computing*, 4(2):42–50, 2005.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS '12, pages 1097–1105. 2012.

[KSRW06] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. Understanding photowork. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 761–770, 2006.

[KT16] Houda Khrouf and Raphaël Troncy. EventMedia: A LOD dataset of events illustrated with media. *Semantic Web*, 7(2):193–199, 2016.

[KVW14] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. A fresh look on knowledge bases: Distilling named events from news. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1689–1698, 2014.

[KW14] Erdal Kuzey and Gerhard Weikum. Evin: Building a knowledge base of events. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 103–106, 2014.

[LCKC09] Chia-Jung Lee, Ruey-Cheng Chen, Shao-Hang Kao, and Pu-Jen Cheng. A term dependency-based approach for query terms ranking. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1267–1276, 2009.

[LD04] Brian Lavoie and Lorcan Dempsey. Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, 10(7/8), 2004.

[Leb17] Julien Leblay. A declarative approach to data-driven fact checking. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI '17, pages 147–153, 2017.

[LGMN12] Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. Defacto-deep fact validation. In *Proceedings of the 11th International Semantic Web Conference*, ISWC '12, pages 312–327, 2012.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, pages 74–81, 2004.

[LJH13] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 73–82, 2013.

[LK77] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[LLC10] Congcong Li, Alexander C. Loui, and Tsuhan Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 827–830, 2010.

[Llo82] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LLT03] Jun Li, Joo Hwee Lim, and Qi Tian. Automatic summarization for personal digital photos. In *Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia*, pages 1536–1540, 2003.

[LLW$^+$13] Zechao Li, Jing Liu, Meng Wang, Changsheng Xu, and Hanqing Lu. Enhancing news organization for convenient retrieval and browsing. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1, 2013.

[LLZL10] Zechao Li, Jing Liu, Xiaobin Zhu, and Hanqing Lu. Multi-modal multi-correlation person-centric news retrieval. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management*, CIKM '10, pages 179–188, 2010.

[LM02] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings of the 2002 International Conference on Image Processing*, ICIP '02, pages 900–903, 2002.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LS13] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, pages 1–49, 2013.

[LSK09] C. Liensberger, J. Stottinger, and M. Kampel. Color-based and context-aware skin detection for online video annotation. In *Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing*, MMSP '09, pages 1–6, 2009.

[LT08] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th European Conference on Computer Vision*, ECCV '08, pages 386–399. 2008.

[LTM03] Joo-Hwee Lim, Qi Tian, and Philippe Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE MultiMedia*, 10(4):28–37, 2003.

[LTW+16] Zechao Li, Jinhui Tang, Xueming Wang, Jing Liu, and Hanqing Lu. Multimedia news summarization in search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):1–20, 2016.

[LWL+11] Zechao Li, Meng Wang, Jing Liu, Changsheng Xu, and Hanqing Lu. News contextualization with geographic and visual information. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 133–142, 2011.

[LWLM05] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 106–113, 2005.

[LWRM14] Cheng Li, Yue Wang, Paul Resnick, and Qiaozhu Mei. ReQ-ReC: High recall retrieval with query pooling and interactive classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 163–172, 2014.

[MB10] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *Signal Processing Letters, IEEE*, 17(5):513–516, 2010.

[MB15] Arunav Mishra and Klaus Berberich. ExposÉ: Exploring past news for seminal events. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 223–226, 2015.

[MB16] Arunav Mishra and Klaus Berberich. Event digest: A holistic view on past events. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 493–502, 2016.

[MBK16] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, and Amin Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *Proceedings of the 33nd International Conference on Machine Learning*, ICML '16, pages 1358–1367, 2016.

[MC07] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, 2007.

[MC13] K. Tamsin Maxwell and W. Bruce Croft. Compact query term selection using topically related text. In *Proceedings of the 36th International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 583–592, 2013.

[ME00] M. A. McDaniel and G. O. Einstein. Strategic and automatic processes in prospective memory retrieval: A multiprocess framework. *Applied Cognitive Psychology*, 14:127–144, 2000.

[ME07] Mark A. McDaniel and Gilles O. Einstein. *Prospective Memory: An Overview and Synthesis of an Emerging Field*. SAGE Publications Inc., 2007.

[Mee16] Mary Meeker. Internet trends 2016. Technical report, Kleiner Perkins Caufield & Byers, 2016.

[MG14] Yunqian Ma and Guodong Guo. *Support vector machines applications*. Springer Science & Business Media, 2014.

[MGM$^+$16] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI '16, pages 3818–3824, 2016.

[MGW$^+$15] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1751–1754, 2015.

[MGW17] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 708–717, 2017.

[MHE14] Ine Mols, Elise van den Hoven, and Berry Eggen. Making memories: A cultural probe study into the remembering of everyday life. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, pages 256–265, 2014.

[MM14] E. Mavridaki and V. Mezaris. No-reference blur assessment in natural images using fourier transform and spatial pyramids. In *Proceedings of the 2014 IEEE International Conference on Image Processing*, ICIP '14, pages 566–570, 2014.

[MM15] Eftichia Mavridaki and Vasileios Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *Proceedings of the 2015 IEEE International Conference on Image Processing*, ICIP '15, pages 887–891, 2015.

[MMJ13] Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 409–418, 2013.

[MMS$^+$14] Suguru Matsuyoshi, Yusuke Miyao, Tomohide Shibata, Chuan-Jie Lin, Cheng-Wei Shih, Yotaro Watanabe, and Teruko Mitamura. Overview of the ntcir-11 recognizing inference in text and validation (rite-val) task. In *NTCIR*, 2014.

[MMT$^+$15] Fotini Markatopoulou, Anastasia Moumtzidou, Christos Tzelepis, Kostas Avgerinakis, Nikolaos Gkalelis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. ITI-CERTH participation to TRECVID 2015. In *Proceedings of TRECVID 2015 Workshop*, 2015.

[MPP$^+$15] Foteini Markatopoulou, Nikiforos Pittaras, Olga Papadopoulou, Vasileios Mezaris, and Ioannis Patras. A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection. In *Proceedings of the 21st International Conference on MultiMedia Modeling*, MMM '15, pages 282–293, 2015.

[MRRG$^+$13] Vuk Milicic, Giuseppe Rizzo, José Luis Redondo Garcia, Raphaël Troncy, and Thomas Steiner. Live topic generation from event streams. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 285–288, 2013.

[MS13] Catherine C. Marshall and Frank M. Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 234–243, 2013.

[MSB$^+$14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 55–60, 2014.

[MSM11] David McClosky, Mihai Surdeanu, and Christopher D. Manning. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 1626–1635, 2011.

[MT05] Josiane Mothe and Ludovic Tanguy. Linguistic features to predict query difficulty. In *ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*, 2005.

[MTY$^+$14] Andrew J. McMinn, Daniel Tsvetkov, Tsvetan Yordanov, Andrew Patterson, Rrobi Szk, Jesus A. Rodriguez Perez, and Joemon M. Jose. An interactive interface for visualizing events on twitter. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1271–1272, 2014.

[MW08] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, 2008.

[MW13] M. Meeker and L. Wu. Internet trends. Technical report, Kleiner Perkins Caufield and Byers, 2013.

[NG15] Thien Huu Nguyen and Ralph Grishman. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL '15, pages 365–371, 2015.

[NPD17] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. Building entity-centric event collections. In *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 1–10, 2017.

[OAM$^+$13] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quéenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2013*, 2013.

[ODOO10] Pere Obrador, Rodrigo De Oliveira, and Nuria Oliver. Supporting personal photo storytelling for social albums. In *Proceedings of the 18th ACM international conference on Multimedia*, MM '10, pages 561–570, 2010.

[OWJ17] Charles Otto, Dayong Wang, and Anil K. Jain. Clustering millions of faces by identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017.

[PC98] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, 1998.

[Plu80] Robert Plutchik. *Emotion: A psychoevolutionary synthesis.* Harpercollins College Division, 1980.

[PM14] C. Papagiannopoulou and Vasileios Mezaris. Concept-based image clustering and summarization of event-related image collections. In *Proceedings of the 1st ACM Workshop on Human Centered Event Understanding from Multimedia at ACM Multimedia*, HuEvent '14, 2014.

[PMS+16] Olga Papadopoulou, Vasileios Mezaris, Vassilios Solachidis, Bahaa Beih Eldesouky, Heiko Maus, and Mark A. Greenwood. D4.4: Information analysis, consolidation and concentration techniques, and evaluation - Final release. 2016. http://www.forgetit-project.eu/fileadmin/fm-dam/deliverables/ForgetIT_WP4_D4.4.pdf.

[PVZ+15] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, BMVC '15, pages 1–12, 2015.

[QRRM11] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, 2011.

[RA12] Devendra Singh Raghuvanshi and Dheeraj Agrawal. Human face detection by using skin color segmentation, face features and regions properties. *International Journal of Computer Applications*, 38:14–17, 2012.

[RASC14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14 (Companion Volume), pages 512–519, 2014.

[RB13] Sunita Roy and Samir K Bandyophadyay. Face detection using a hybrid approach that combines HSV and RGB. *International Journal of Computer Science and Mobile Computing*, 2(3):127–136, 2013.

[RBDN15] Andrea Rosani, Giulia Boato, and Francesco GB De Natale. Eventmask: A game-based framework for event-saliency identification in images. *IEEE Transactions on Multimedia*, 17(8):1359–1371, 2015.

[RCCC16] Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, NAACL-HLT '16, pages 7–17, 2016.

[RH00] Yong Rui and T. Huang. Optimizing learning in image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '00, pages 236–243, 2000.

[RK14] Fiana Raiber and Oren Kurland. Query-performance prediction: Setting the expectations straight. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 13–22, 2014.

[RL15] Victoria L Rubin and Tatiana Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917, 2015.

[RSB10] Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. Automatic creation of photo books from stories in social media. In *Proceedings of the 2nd ACM SIGMM Workshop on Social Media*, WSM '10, pages 15–20, 2010.

[RW03] Kerry Rodden and Kenneth R Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, pages 409–416, 2003.

[SAC$^+$17] Emanuele Sansone, Konstantinos Apostolidis, Nicola Conci, Giulia Boato, Vasileios Mezaris, and Francesco GB De Natale. Automatic synchronization of multi-user photo galleries. *IEEE Transactions on Multimedia*, 19(6):1285–1298, 2017.

[SBS14] Boon-Siew Seah, Sourav S. Bhowmick, and Aixin Sun. Prism: Concept-preserving social image search results summarization. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 737–746, 2014.

[SCWT14] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS '14, pages 1988–1996, 2014.

[SEL00] Andreas E Savakis, Stephen P Etz, and Alexander CP Loui. Evaluation of image appeal in consumer photography. In *Human Vision and Electronic Imaging*, pages 111–120, 2000.

[SG09] Falk Scholer and Steven Garcia. A case for improved evaluation of query difficulty prediction. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 640–641, 2009.

[SG12] Jannik Strötgen and Michael Gertz. Event-centric search and exploration in document collections. In *Proceedings of the 12th Joint Conference on Digital Libraries*, JCDL '12, pages 223–232, 2012.

[SGG$^+$16] Andreas Spitz, Johanna Geiß, Michael Gertz, Stefan Hagedorn, and Kai-Uwe Sattler. Refining imprecise spatio-temporal events: a network-based approach. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, GIR '16, pages 1–10, 2016.

[SKC$^+$12] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):1–35, 2012.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '15, pages 815–823, 2015.

[SLJ$^+$15] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '15, pages 1–9, 2015.

[SM16] Zahra Riahi Samani and Mohsen Ebrahimi Moghaddam. A knowledge-based semantic approach for image collection summarization. *Multimedia Tools and Applications*, 76(9):1–23, 2016.

[SMJ11] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 4:1–4:8, 2011.

[ST17] Rachele Sprugnoli and Sara Tonelli. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506, 2017.

[STH09] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking open descriptions of events. In *Proceedings of the 4th Asian Conference on The Semantic Web*, ASWC '09, pages 153–167. 2009.

[STVB16] Mehdi Samadi, Partha Pratim Talukdar, Manuela M Veloso, and Manuel Blum. ClaimEval: Integrated and flexible framework for claim evaluation using credibility of sources. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI '16, pages 222–228, 2016.

[SVDHG11] Eefke Smit, Jeffrey Van Der Hoeven, and David Giaretta. Avoiding a digital dark age for data: why publishers should care about digital preservation. *Learned Publishing*, 24(1):35–49, 2011.

[SWD05] G. Sharma, W. Wu, and E. N. Dalal. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, 30(1):21–30, 2005.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[TA14] Giang Binh Tran and Mohammad Alrifai. Indexing and analyzing wikipedia's current events portal, the daily news summaries by the crowd. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 511–516, 2014.

[Tan13] Yichuan Tang. Deep learning using linear support vector machines. In *Proceedings of the 30th International Conference on Machine Learning*, ICML '13 (Companion Volume), 2013.

[TCG$^+$14] Tuan Tran, Andrea Ceroni, Mihai Georgescu, Kaweh Djafari Naini, and Marco Fisichella. WikipEvent: Leveraging wikipedia edit history for event detection. In *Proceedings of the 15th International Conference on Web Information Systems Engineering*, WISE '14, pages 90–108. 2014.

[TCKN15a] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 339–348, 2015.

[TCKN15b] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. Time-travel translator: Automatically contextualizing news articles. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion Volume, pages 247–250, 2015.

[TdRW11] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 565–574, 2011.

[TIWB14] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS '14, pages 1413–1421, 2014.

[TLZ$^+$04] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. Classification of digital photos taken by photographers or home users. In *Proceedings of the 5th Pacific Rim Conference on Advances in Multimedia Information Processing*, PCM '04, pages 198–205, 2004.

[TMM$^+$16] Christos Tzelepis, Zhigang Ma, Vasileios Mezaris, Bogdan Ionescu, Ioannis Kompatsiaris, Giulia Boato, Nicu Sebe, and Shuicheng Yan. Event-based media processing and analysis: A survey of the literature. *Image and Vision Computing*, 53:3–19, 2016.

[TMY78] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978.

[Tur00] Peter D. Turney. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336, 2000.

[UK05] S. Uchihashi and T. Kanade. Content-free image retrieval based on relations exploited from user feedbacks. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*, ICME '05, pages 1358–1361, 2005.

[vdSGS10] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[VdVBM15] Jeroen B.P. Vuurens, Arjen P. de Vries, Roi Blanco, and Peter Mika. Online news tracking for ad-hoc queries. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1047–1048, 2015.

[vdWSV07] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '07, pages 1–8, 2007.

[vHMS$^+$11] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.

[VJ04] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[VM94] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology*, 123(4):394–409, 1994.

[vTP+13] Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. Automatic selection of social media responses to news. In *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 50–58, 2013.

[Wan05] Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

[Wan17] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 422–426, 2017.

[Wat05] CRW Watch. Leap of faith: Using the internet despite the dangers. *City: Princeton Survey Research Associates: Washington, DC*, 2005.

[WBT10] Yaowen Wu, Christian Bauckhage, and Christian Thurau. The good, the bad, and the ugly: Predicting aesthetic image labels. In *Proceedings of the 20th International Conference on Pattern Recognition*, ICPR '10, pages 1586–1589, 2010.

[WLC07] Ming-Ni Wu, Chia-Chen Lin, and Chin-Chen Chang. Novel image copy detection with rotating tolerance. *Journal of Systems and Software*, 80(7):1057–1069, 2007.

[WLS+16] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 4810–4819, 2016.

[WNL14] Maria K Wolters, Elaine Niven, and Robert H Logie. The art of deleting snapshots. In *Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems (Extended Abstracts)*, CHI EA '14, pages 2521–2526, 2014.

[WNR+15] Maria K. Wolters, Elaine Niven, Mari Runardotter, Francesco Gallo, Heiko Maus, and Robert H. Logie. Personal photo preservation for the smartphone generation. In *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems (Extended Abstracts)*, CHI EA '15, pages 1549–1554, 2015.

[WPF+99] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, DL '99, pages 254–255, 1999.

[WSL$^+$14] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1386–1393, 2014.

[WSS14] Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2065–2074, 2014.

[WWS$^+$06] Zhou Wang, Guixing Wu, H. R. Sheikh, E. P. Simoncelli, En-Hui Yang, and A. C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, 2006.

[WYZ15] Ke Wu, Song Yang, and Kenny Q Zhu. False rumors detection on sina weibo by propagation structures. In *Proceedings of the 31st International Conference on Data Engineering*, ICDE '15, pages 651–662, 2015.

[XLY$^+$16] Zheng Xu, Yunhuai Liu, Neil Yen, Lin Mei, Xiangfeng Luo, Xiao Wei, and Chuanping Hu. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 2016.

[XWL$^+$15] Zheng Xu, Xiao Wei, Xiangfeng Luo, Yunhuai Liu, Lin Mei, Chuanping Hu, and Lan Chen. Knowle: a semantic link network based system for organizing large scale online news events. *Future Generation Computer Systems*, 43:40–50, 2015.

[YBD$^+$09] Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. Query by document. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, 2009.

[YBO14] Che-Hua Yeh, Brian A. Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system considering positive and negative user feedback. *ACM Trans. Multimedia Comput. Commun. Appl.*, 10(4):36:1–36:20, 2014.

[YDYX14] Hongliang Yu, Zhi-Hong Deng, Yunlun Yang, and Tao Xiong. A joint optimization model for image summarization based on image content and tags. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, AAAI '14, pages 215–221, 2014.

[YHBO10] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 211–220, 2010.

[YLLT15] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '15, pages 3676–3684, 2015.

[YLYY12] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 13:1–13:7, 2012.

[YSCG08] Shengye Yan, Shiguang Shan, Xilin Chen, and Wen Gao. Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '08, pages 1–7, 2008.

[ZBNT04] Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.

[ZLLT11] Wengang Zhou, Houqiang Li, Yijuan Lu, and Qi Tian. Large scale image search with geometric coding. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1349–1352, 2011.

[ZTL$^+$06] Ming Zhao, Yong Wei Teo, Siliang Liu, Tat-Seng Chua, and Ramesh Jain. Automatic person annotation of family photo album. In *Proceedings of the 5th International Conference on Image and Video Retrieval*, CIVR '06, pages 163–172, 2006.

[ZWS11] Chunhui Zhu, Fang Wen, and Jian Sun. A rank-order distance based clustering algorithm for face tagging. In *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 481–488, 2011.

[ZZSH10] Jing Zhang, Li Zhuo, Lansun Shen, and Lin He. A personalized image retrieval based on user interest model. *Int. J. Patt. Recogn. Artif. Intell.*, 24(3), 2010.