ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-071

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Xueliang LIU

le 3 décembre 2012

Fouille d'informations multimédia partagées orienté evénements

Directeur de thèse : **Benoit HUET**

Jury

Mme. Nozha BOUJEMAA, Directeur de recherches, INRIA Paris,	Président
M. Francesco DE NATALE, Professeur, University of Trento	Rapporteurs
M. Hervé GLOTIN, Professeur, Université Sud Toulon	Rapporteurs
M. Bernard MERIALDO, Professeur, EURECOM	Examinateur
M. Benoit HUET, Maître de Conférences, EURECOM	Directeur de thèse

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

Thesis

Event based Social Media Data Mining

Xueliang Liu

Supervisor: Benoit Huet

Dec 3rd, 2012

Acknowledgements

I have spent three years and eight months at EURECOM. During my study at EURECOM, I have met many people who have provided priceless contribution to my research.

First and foremost, I would like to thank my advisor, Prof. Benoit Huet, for his great guidance and timely support during my Ph.D. study. Prof. Huet is an excellent supervisor with abundant experience on research. I am grateful that he gave me invaluable suggestion and advice when I met problems. I especially appreciate the freedom he given me when I was making little progress on exploring the research direction in my first year. Without his continuous support, this thesis can not be achieved. He is also a good friend and shares me lots of ideas on enjoying the life in France.

Many thanks go to the members of my thesis committee, Prof. Nozha BOUJEMAA, Prof. Francesco DE NATALE, Prof. Herve GLOTIN and Prof. Bernard MERIALDO, for their helpful comments and enthusiastic support. Their criticisms, and advice not only were critical to improve the quality of this thesis, but also enlightened me some possible directions to continue the work.

Moreover, I am also indebted to the group members at Multimedia department. I would like to show my appreciation to Prof. Bernard MERIALDO, Prof. Jean-Luc Dugelay, Prof. Nick Evans, Prof. Raphael Troncy: for their daily teaching and the research atmosphere they stimulate at Eurecom's Multimedia Department. Other team members, such as Dr. Feng Wang, Dr. Dong Wang, Dr. Carmelo Velardo, Dr. Nesli Erdogmus, Dr. Antitza Dantcheva, Dr. Yingbo Li, Dr. Simon Bozonnet, Dr. Moctar Mossi, Xuran Zhao, Rui Min, Miriam Redi, Christelle Yemdji, Houda KHROUF, Mathilde SAHUGUET and others, also help me to finish my research successfully. The discussions with them helps me understanding my problems well and inspires me to find a possible solution.

Last, but not the least, I would also like to thank my family for their continuous support, without which I would not have survived during the Ph.D. process.

Abstract

Nowadays the development of media capture devices and social networks makes it easily for users to capture and distribute rich multimedia content. Recent years have witnessed the rapid growth of social media collections available over the Internet. The age of big data provides users facilities to share and access data, while it also demands the revolution of data management techniques. The exponential growth of social media data requires more scalable, effective and robust technologies to manage and index them.

Event is one of the most important cues to recall people's past memory. The reminder value of event makes it extremely helpful in organizing data. With the development of Web 2.0, many event-based information sharing sites are appearing online, and a wide variety of events are scheduled and described by several social online services. The study of the relation between social media and events could leverage the event domain knowledge and ontologies to formulate the raised problems, and it could also exploit multimodal features to mine the patterns deeply, hence gain better performance compared with some other methods.

In this thesis, we study the problem of mining relations between events and social media data. There are mainly three problems that are well investigated. The first problem is event enrichment, in which we investigate how to leverage the social media to events illustration. The second problem is event discovery, which focuses on discovering event patterns from social media stream. We propose burst detection and topic model based methods to find events from the spatial and temporal labeled social media. The third problem is visual event modeling, which studies the problem of automatically collecting training samples to model the visualization of events. The solution of collecting both of the positive and negative samples is also derived from the analysis of social media context.

Thanks to the approaches proposed in this thesis, the intrinsic relationship between social media and events are deeply investigated, which provides a way to explore and organize online medias effectively.

Résumé

Au cours des dernières années, nous avons pu assister à une croissance rapide de la quantité d'information multimédia disponible via les médias sociaux en ligne. La croissance exponentielle des données des médias sociaux a mis en avant un besoin de technologies efficaces, fiables et capables de passer à l'échelle afin de pouvoir les gérer et les indexer. La notion d'"évènement" est une des clés majeures permettant de se remémorer des souvenirs. La valeur mémorielle d'un évènement le rend des plus utiles lorsqu'il s'agit d'organiser des données. Avec le développement du Web 2.0, beaucoup de sites de partage d'information au sujet d'évènements font leur apparition sur internet, et une grande variété d'évènements sont programmés et décrits par plusieurs services et réseaux sociaux en ligne. L'étude des relations entre medias sociaux et évènements pourrait tirer parti des connaissances liées au domaine des évènements et des ontologies afin de formuler les problèmes soulevés ; l'exploitation des caractéristiques multimodales peut aussi permettre d'explorer les caractéristiques en profondeur, et ainsi de gagner en performance par rapport à d'autres méthodes. Dans cette thèse, nous étudions le problème de l'extraction de connaissances quant aux relations entre évènements et données des réseaux sociaux. Trois problèmes sont au centre de notre analyse. Le premier problème porte sur l'enrichissement visuel des évènements : notre recherche vise à comprendre comment utiliser les médias sociaux pour

illustrer des évènements. Le deuxième problème, la découverte d'évènement, se concentre sur la découverte de caractéristiques dans les évènements à partir des flux d'informations provenant des médias sociaux. Nous proposons d'utiliser la détection de niveaux et des méthodes de détection de sujet pour découvrir des évènements grâce aux annotations spatiales et temporelles présentes dans les médias sociaux. Le troisième problème concerne la modélisation visuelle des évènements, dont la problématique est de rassembler de façon automatique des échantillons d'apprentissage, afin de mettre en oeuvre une représentation visuelle des évènements. La solution proposée consiste à rassembler des exemples à la fois positifs et négatifs ; de même, elle est dérivée de l'analyse du contexte des médias sociaux. Grâce aux approches proposées dans cette thèse, la relation intrinsèque entre les médias sociaux et les évènements est étudiée en profondeur, ce qui nous fournit un moyen d'explorer et d'organiser les medias en ligne de manière efficace.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Today's media capture devices and network infrastructures enable users to easily capture and distribute rich multimedia content wherever they are located. Recent years have witnessed the rapid growth of social media collections available for searching and browsing over the Internet. Hence, it is becoming common for people to capture and share images or videos of their everyday life using their mobile phones or digital camera. A plethora of niche mobile applications such as instagr.am[1] or color[2] are connected to large social media applications such as Flickr[3], YouTube[4] and Facebook[5] and contribute to the exponential growth of social media data available online. In YouTube, for example, a statistic of hours videos uploaded per minutes is reported in Figure 1.1. This figure suggests the exponential increasing of videos uploading over the past 5 years. In the middle of 2007, only 6 hours of videos were uploaded per minute, but after 5 years, 72 hours of videos per minutes were uploaded in May 2012. Now over 800 million unique users visit YouTube and over 4 billion hours of video are watched each month. A similar trends could be found in Flickr. As shown in Figure 1.2, the number of photos in Flickr increases from 2 billion to 7 billion in the past 5 years.

The big data era provides users facilities to share and access data, while its advent demands effective data management and indexing technologies. How to search the target efficiently and effectively, how to leverage the big data to solve the large scale problems in industry and research communities, are still open challenges. The ideas of managing and organizing media data can be traced back in the field of multimedia information retrieval(MIR) many years ago. In MIR, there are different types of strategies to manage media data. Concept based methods, also known as "description-based" or "text-based" media indexing/retrieval methods, refer to retrieval from text-based indexing of data. Concept based methods are easy to design and deploy, they are now popularly used in traditional web based image search engines. However, the textual surrounding textual metadata is always missing, or not related to the media content. To overcome this drawback, content-based techniques are proposed to search media by analyzing the actual contents of the media rather than the metadata. They employ computer vision and machine learning tech-

[1] http://instagr.am/
[2] http://color.com/
[3] http://www.flickr.com/
[4] http://www.youtube.com.am/
[5] http://www.facebook.com/

niques to model the content represented by visual features in the format of colors, shapes, textures, or any other information that can be derived from the data itself. Content-based methods are necessary when text annotations are nonexistent or incomplete. Furthermore, they can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media data. In recent years, content based methods made great achievement both in research[6] and academic area [Datta et al., 2008]. Content-based retrieval often fails due to the so call "semantic gap", the gap between low level visual features that are extracted from media data and the high level concepts that would be understood by humans. The research on bridging the semantic gap is still a hot topic in research.

Compared with traditional media data, there are some new characters on social media data. At first, the number of social media data media are typically at large scale, with thousands of millions photos/blogs/videos. Secondly, unlike media data in traditional applications, social media are generally associated with multi-modality features, such as textual description, tags, time, location, visual content. Thirdly, since social media data are always taken by different users in widely varying condition, there is more noise in social media data compared with the traditional media data. The rapid growth of social media data demands more scalable, effective and robust technologies to manage and index them.

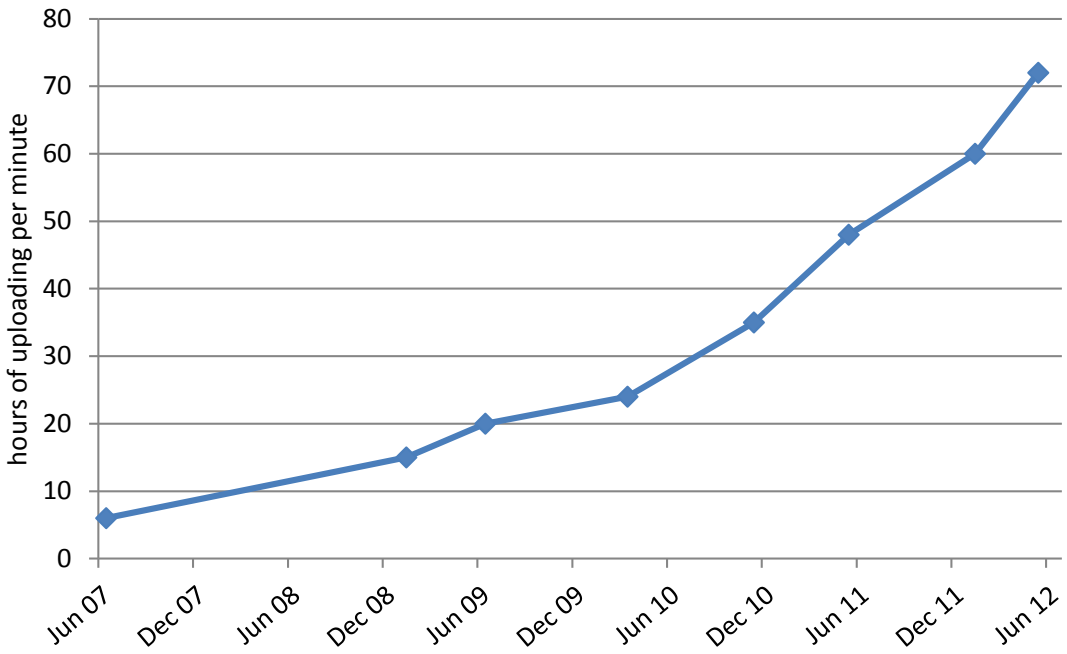


Figure 1.1: Hours of videos uploaded per minute over the past 5 years in Youtube

Event refers to the happening at a specific time and place in real-world phenomena. In history, event is one of the most important cues to recall people's past memory[Tulving, 1984]. The reminder value of event makes it extremely helpful in organizing human life. Before the information technology era, people would like to schedule what to do, or record what happened by events. With the rapid development of information technology, event is

[6] http://image.google.com

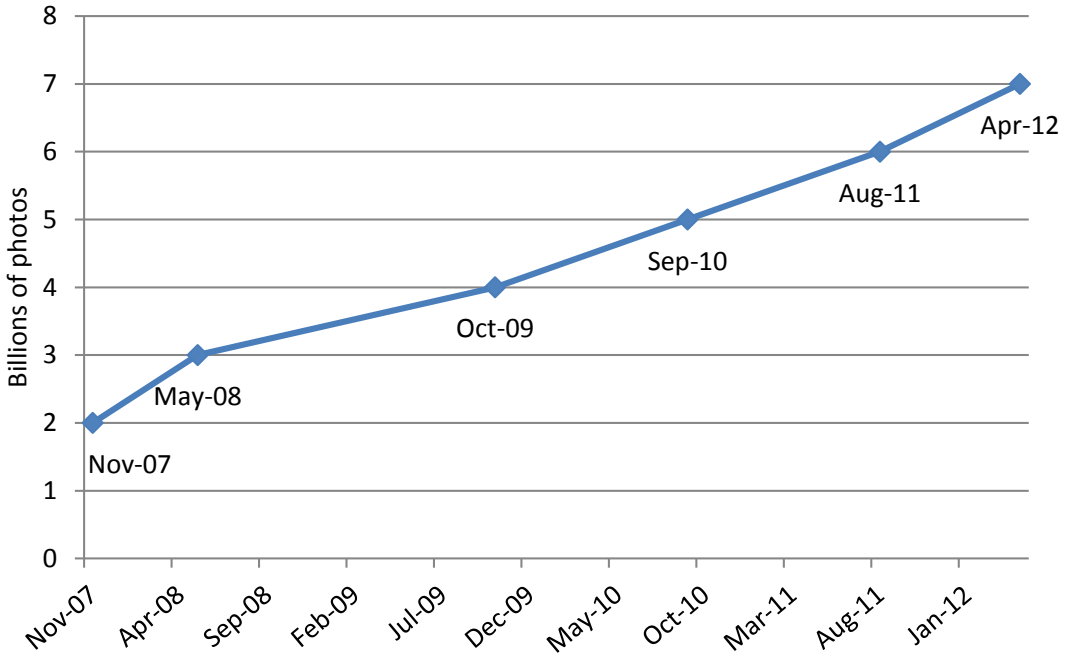


Figure 1.2: Billions of photos in Flickr over the past 5 years

still the natural way for people to organize, browse and visualize media collections, not only for a single user's personal photo album, but also for a global collection of digital media resources. For example, human like to store their daily photos categorized by events, so that the photos can be retrieved easily. With the development of Web 2.0, lots of event-based information sharing sites are built online, which provide common user interaction scenarios. A wide variety of events are scheduled and described by several social online services. Nowadays, people would like to share rich data to user-contributed data taken during a variety of events in the physical world, such as concerts, festivals, conferences and more. Among the data, lots of valuable information could be found in different modalities, such as known event features (e.g., title, time, location) posted on event repositories (e.g., Last.fm events, Upcoming, Facebook events), discussions and reactions related to events shared on different social network sites (e.g., Twitter, Facebook, Blogger), or photos, audio, videos taken during the events on media share sites(e.g., Flickr, YouTube, DailyMotion).

The popularity of social media and social events data makes it imperative to exploit the event semantics from social media data. The study of this problem can derive many applications listed (but not limited) as follows:

- **media data management** Lots of data will be uploaded to social media sharing websites every day by different users. If the event semantics could be inferred from these data, the data can be managed and indexed by event, as the human did to storage their personal photos.
- **social trends monitor** As information spreads in social media sites quickly, social trends monitors become increasingly important in political, business and commercial systems.
- **event illustration** As the saying goes, "A picture is worth a thousand words", hence

social media provides user an easy way to understand events if they are illustrated well.

In the research community, the study on heuristics behind the events and social media has recently drawn much attention [Andrews et al., 2012, Fialho et al., 2010, Westermann and Jain, 2007, Mattivi et al., 2011]. Some related works have been done to study event context on social media sources. For example, GLOCAL[7], an European project, is dedicated to using events as the central concept to search, organize multimedia data. Fialho et al. [2010] focused on the interconnection of social media data and events by semantic web, and they would like to link the data from different domains by the duplicated metadata. In [Quack et al., 2008] the authors worked on clustering the event-related photos by textual and visual features. In their proposal, the event relevant cluster could be detected by the duration and number of users among the photos in the cluster. These previous researches show a promising direction to analyze the relation between social media and events. Compared with the content based approaches, event-based research focuses on the study of social media content that are possibly connected to events. It could leverage the event domain knowledge and ontologies so that problems could be formulated effectively. In addition, event-based research addresses the problem by exploit multimodal features. Besides the textual and visual feature popularly used in concept and content based approaches, geographic, time, owners information in metadata also be incorporated, and the study on these features makes it possible to handle data without any textual description, which is a common problem in social media analysis. Hence the proposed approaches could lead to better performance compared with some other methods. The achievement from latest event-based research on social media show a promising way to alleviate the semantic gap. Additionally the study of inner and deep semantic between them is still under demand. In this thesis, we employ data mining technologies to study the problem of bridging the event and social media modality semantically, report our work on how to illustrate events with rich media data, to discover events from social media streams, and to model event visualization automatically.

1.2 Content of the Thesis

In this thesis, we have thoroughly studied the problem of mining relation between events and social media data with some data mining and analysis approaches. There are mainly three parts of work that are well investigated.

The first part of work studies enriching event illustration with media data. There are two sub parts of work reported in this part. Firstly, a method is presented to combine semantic inference and visual analysis for finding media (photos and videos) illustrating events automatically. We describe a method for finding as much as possible photos and videos relevant for a given event: we start from the media labeled with specific machine tags and that can be used to train classifiers to prune results from general queries. We validate the heuristic with data from media sharing platforms and large event directories. Secondly, we extend the approach in the first part and propose a more general framework to visualize events for a selected topic from social media data by leveraging data mining techniques. In this framework, some essential questions in information retrieval, such as query expanding with Natural Language Process, retrieving multi-modality data from different web service, and depicting result vividly, are well studied.

[7] http://www.glocal-project.eu/

The second part discusses how to mine events from social media streams. We propose two approaches to address the problem. Firstly, we observed that many photos and videos are taken and shared when events occur. We focused on detecting events from the spatial and temporal labeled social media, and proposed a burst detection approach to discover events from social media streams. Secondly, the events were considered as the specific distribution of the topics on human life. We proposed a framework to mine the topics from media taken in a city by Latent Dirichlet Allocation and make the decision rule to discover events. In this part, we also presented our work on MediaEval Social Event Detection task, that is to detect specific events from given dataset. The strategy we investigated is to find the event instances that occurred during this period of time first and then try to match these event instances with photos from the given dataset.

In the third part of work, we present our work to automatically collect training data to model events. In this part of work, we studied the problems of event visualization modeling. In machine learning, data labeling is time consumption task, and it is hard to obtain well labeled data. To address the problem, we propose an automatic way to collect the training samples. Since we already noticed the role by machine tags played in identifying event relevant samples, the positive samples are collected by the media data labeled with machine tags. To collect negative sample, we assumed that the negative samples labeled with common tags refer the common concepts in a given place, we employed the "learning to rank" strategy to discover the negative samples. The automatically collected samples are used to training SVM models, and a competitive performance is achieved.

1.3 Contributions

This thesis is dedicated to the study of the semantic relation between social media and events by data mining techniques. The main contribution of this thesis could be summarized as follows.

1. We have studied the inner spatial and temporal characteristics about users' capture and uploading behavior. We do a statistics on the media data that are labeled with event machine tags, and find that most of the media data is uploaded in the FIVE days which follow the events taken time. We also monitor the geographic distribution of these media data taken during events. These discoveries allow us to find out the proper time window, as well the proper geographic bounding box of venues to query the media data.

2. We have developed a well-designed event illustration system. To build the system, we exploit different query generation strategies by multimodal event features. We also design an effective visual pruner to filter out noisy data from the query results. To improve the final performance, we use the "owner refining" approach to recall the positive samples. We also published a friendly interface to demonstrate the results.

3. We have built a system to retrieve, summarize, and visualize events for a selected topic. The system responds to textual query with events well illustrated by textual and visual description. In this system, we employ Natural Language Process techniques to parse the query and find out the time, location, topic clues. We use the parsed clues to query events from social news website. We also demonstrate the final results with a vivid interface, in the format of textual description, tag clouds, and image collages.

4. We have proposed burst detection approach to discover events from social media data. The burst detection approach was derived from the observation that many media data are uploaded just after the events occurred. To gain a better performance, We studied the different uploading measure such as "the number of media", "the number of media times the number of uploaders". We also evaluated the detection results under different thresholds.

5. We have proposed another event detection approaches based on the topic model approach. We assume that event is one of the common concepts in human life. In this approach, we collect data taken in a city and employ Latent Dirichlet Allocation model to mine the hot topics. We use the inference from validating data to make the decision rule. We also compared the results with the previous approach.

6. To model event visually, we have proposed a framework to collect the training samples automatically, while the collection is done with the analysis of context social media and events. We collect positive samples when they are labeled by event machine tag or event abbreviated name, which precisely refer to the unique event. We collect negative samples by learning to rank approach. We also study in detail the parameters that may impact the final classifier, such as the number of common tags, the time span and spatial bounding box to collect negative samples.

The work presented in this thesis led to several publications in international conferences and workshops, listed as follows.

1. Liu Xueliang, Benoit Huet. "On the automatic online collection of training data for visual event modeling". Multimedia Tools and Applications. (submitted).

2. Liu, Xueliang; Huet, Benoit. "Gathering training sample automatically for social event visual modeling". ACM International Workshop on Socially Aware Multimedia. October 2012. Nara, Japan.

3. Liu, Xueliang, Huet, Benoit. "Social event discovery by topic inference".13th International Workshop on Image Analysis for Multimedia Interactive Services. May 2012, Dublin, Ireland.

4. Liu, Xueliang, Troncy, Raphael,Huet, Benoit. "Using social media to identify events". 3rd ACM Multimedia Workshop on Social Media. October 2011. Scottsdale, Arizona, USA.

5. Liu, Xueliang, Huet, Benoit, Troncy, Raphael. "EURECOM @ MediaEval 2011 social event detection task". MediaEval Benchmarking Initiative for Multimedia Evaluation, September, 2011, Pisa, Italy.

6. Liu, Xueliang, Troncy, Raphael, Huet, Benoit. "Finding media illustrating events". ICMR'11, ACM International Conference on Multimedia Retrieval, April, 2011, Trento, Italy.

7. Liu, Xueliang, Huet, Benoit. "Concept detector refinement using social videos". International workshop on Very-large-scale multimedia corpus, mining and retrieval, October 2010, Firenze, Italy.

8. Liu, Xueliang, Huet, Benoit. "Automatic concept detector refinement for large-scale video semantic annotation". IEEE 4th International Conference on Semantic Computing, September, 2010, Pittsburgh, USA.

1.4 Outline

The following part of this thesis is organized as follows:

Chapter 2 discusses the background and surveys related work, including the social media domains involved, the analysis technologies and evaluation criterion, and state-of-the-art work.

Chapter 3 presents the work of event illustration, and event-based query expansion system. For event illustration, we report the a general overview of LODE ontologies and EventMedia data set in Section 3.2.1 and 3.2.2 respectively. Section 3.2.3 details our approach on how to collect media data to illustrate events. The work of event-based topic expression is detailed in 3.3.

Chapter 4 reports the work of event discovery from social media data. We proposed two methods to discover social events. The burst detection approach is presented in Section 4.2.2, based on our observation that many media data are uploaded when events occur. A topic model based event discovery approaches is presented in Section 4.3.1.

Chapter 5 introduces the work of event modeling from social media data collected automatically. Since the positive samples can be collected easily(Section 5.2.1), most of the discuss focuses on how to collect negative samples by a learning-to-ranking strategy, as described in Section 5.2.2. The final results compared with other approaches are reported in Section 5.3.3.

Conclusions are given in Chapter A.4 which summarizes the major contributions and results obtained in this thesis and suggests to some potential improvement and future work.

Chapter 2

Background and Related Work

When studying a topic, the first steps always involve attentive observation and analysis to better understand the problem, and reviewing the state-of-the-art. This chapter gives a brief introduction to the background knowledge of this thesis. The main task of this work is to analyze social media and events by data mining techniques. Therefore, Section 2.1 introduces the most popular social media and events domains which are involved in this thesis, Section 2.2 gives an overview of the techniques and evaluation criteria used in this thesis, and Section 2.3 reviews the related work.

2.1 Social Media

There has been a tremendous rise in the growth of online social networks all over the world in recent years. According to Wikipedia[1], social media is taken as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content." Thanks to the social media platforms appearing recently, users could share information in diverse formats including blog, photos, videos. In this section, we will review the social media web services that will be involved in this thesis.

2.1.1 Flickr

Flickr[2] is one of the most important online photo and video sharing websites. It was created in 2004, and acquired by Yahoo! in 2005. In the past several years, Flickr is growing at an extremely rapid rate, reaching 3 billion hosted photos in November 2008, 4 billion in October 2009, and 5 billion in September 2010. In March 2012, the number reaches 6 billion. Flickr says that the number of uploads has risen by 20 percent for the past five years, so more photos are added to Flickr today than they were a few years ago.

Its growth and popularity is encouraged by the philosophy of openness. Flickr offers users the ability to either release their images under certain common usage licenses(default) or label them as "all rights reserved". The licensing options primarily include the Creative Commons 2.0 attribution-based and minor content-control licenses. By default, photos uploaded to Flickr are public to everyone. Flickr's functionality includes RSS and Atom feeds and an API that enables independent programmers to query data or expand its services. All of these features make Flickr popular in the research community.

[1] http://en.wikipedia.org/wiki/Social_media

[2] http://www.flickr.com

Due to its easy access and free redistribution, some datasets have been created with the photos downloaded from Flickr. For examples, MIRFlickr [Huiskes and Lew, 2008], it is new image collection consists of 25k images that were downloaded from the social photography site Flickr.com through its public API. It could be used in different purposes, such as visual concept learning, tag propagation/suggestion, and now it is one of the most popular benchmark dataset in social media dataset. Another benchmark dataset, NUS-WIDE was released by NUS [Chua et al., 2009] in 2009. NUS-WIDE is a large scale images dataset, which includes 269K images and 6 visual features for these images that are commonly used in image retrieval. There is lots of research derived from the two Flickr benchmark dataset [Zhu et al., 2010, Tang et al., 2009, Li et al., 2010].

Besides the visual content, Flickr also provides rich metadata for the media data, such as title, tags, uploading/taken time, geographic tag, machine tags .... With these metadata, we can see when where, why the photo was taken, so that the photo content could be understood easily. In research community, there are also some work studying how to leverage these context data to solve the problems in practice, for example [Liangliang et al., 2009, Joshi and Luo, 2008] focusing on geo tags, and Liu et al. [2010] studying textual tags. The metadata that is popularly used in this thesis is "machine tag". According to Flickr, machine tag is "a type of tag that are written in a special format and it could be understood by a machine to automatically perform a special action". There are always 3 parts in a machine tag.

namespace:predicate=value

which includes:

the namespace, i.e. upcoming [who is going to care about this tag]
the predicate, i.e. event [what does this apply to]
the value, i.e. 123456 [which one is this]

Let's take the machine tag "upcoming:event=428084" as an example, according to Flickr when this tag is added to the picture during uploading, it will automatically show on the Upcoming event page. Since Upcoming's servers recognize tags with "upcoming:". If the tags says "event" they know to use it on an event page. Then you tell them which event by adding the number at the end[3]. In conclusion, machine tag can link media across different sites. Figure 2.1 shows an example of photos with this machine tag.

2.1.2 Social Event Sites

An event can be described as a happening that occurs at a given place and time. It is a natural way to refer observable occurrence, and could be documented by people with different media format(e.g. videos and photos) [Westermann and Jain, 2007].

Recent years, with the development of Web 2.0, lots of event-based information sharing sites are built online, which provide common user interaction scenarios. A wide variety of events are scheduled and described by several social online services. Among these websites, Last.FM [4], Eventful [5] and Upcoming [6] are the most popular ones.

Last.FM is the oldest and largest music based social networking site which is created in 2002. In Last.FM, users are allowed to share tracks of their listens and to generate personal listening charts in their profile. Last.fm facilities good interface for user interaction. For

[3] http://www.flickr.com/help/tags/
[4] http://www.last.fm
[5] http://www.eventful.com
[6] http://upcoming.yahoo.com/

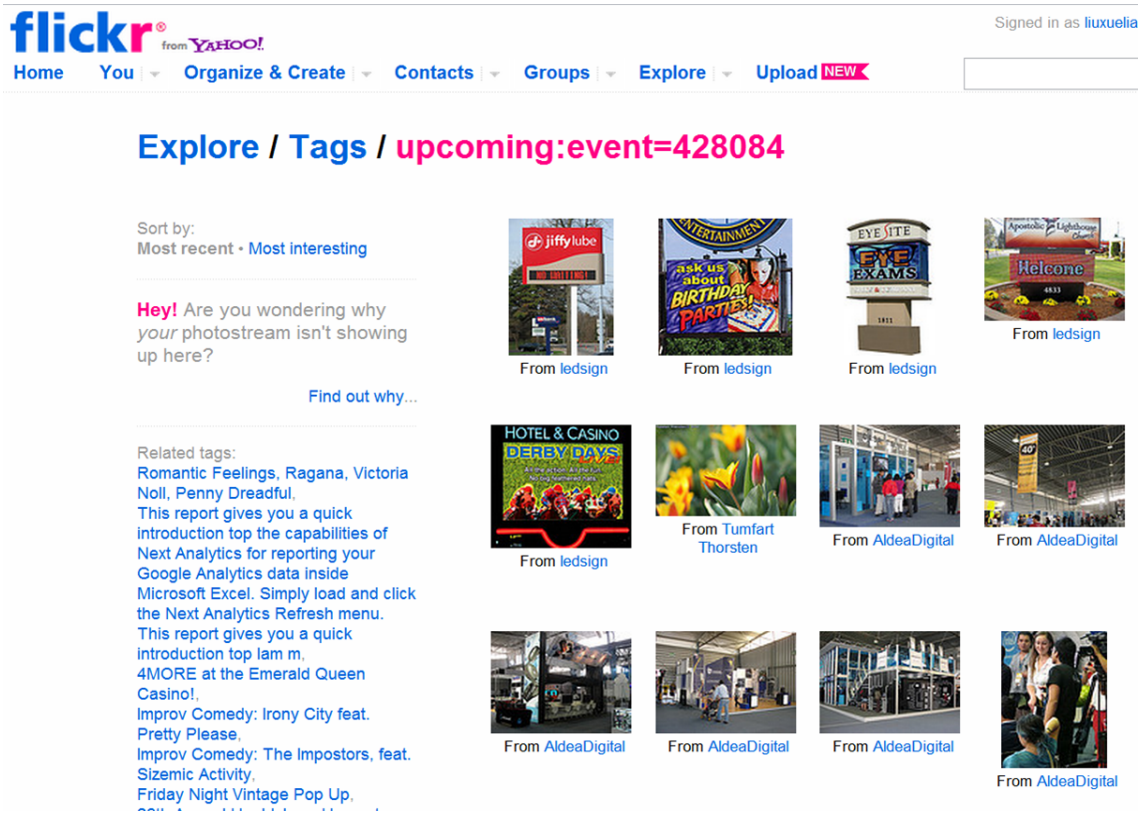


Figure 2.1: An example of photos with machine tags

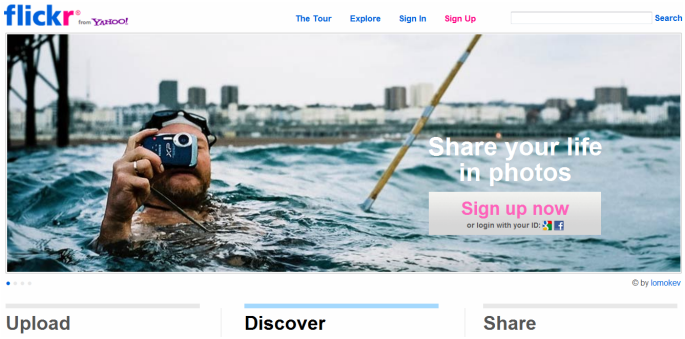
example, users can write self descriptions and blogs in the profile, follow other music listeners, or join a group. Events functionality is well equipped in Last.FM, by which users could search gigs or festivals event on a specified location. In addition, registered users can add new venues or events which will then be listed on the band or artist's main page, together with other details if available. There is also a facility to submit reviews and photographs of past events. Last.FM event machine tag in the format of "lastfm:event=XXX" is popularly used in Flickr, and the media data in Flickr and events in Last.FM are linked by such tags.

Upcoming: Upcoming is another collaborative events database where users can record events (e.g., concerts, exhibits, plays, etc.) and tag them. It is built in 2003 and acquired by Yahoo! in 2005. Event information is contributed by the user community, although an increasing percentage of event data now comes from commercial sources. Upcoming is well capable with social networking functions. It allows users not only to upload, share, favorite events, but also to indicate their plans by marking that they are "going" to or "interested" in an event. Users could also establish "friend" relationships with each other and receive notifications about what their friends are planning.

Eventful: Eventful is a competitor of Upcoming and provides a very similar online service as aforementioned.

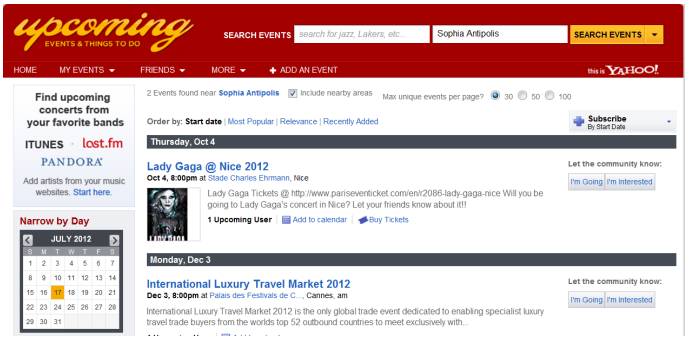
Both of Upcoming and Eventful share data with Flickr by machine tags, and many third party applications are built on the two platforms. The interface of all of the 4 websites discussed before are depicted in Figure 2.2.

There are still some web services used in the thesis, and they will be discussed in the



(a) Flickr

(b) Last.FM



(c) Upcoming

(d) EventFul

Figure 2.2: The social media and event sites

related section.

2.1.3 Data Scraping

In W3C, web service is defined as "a software system designed to support interoperable machine-to-machine interaction over a network". It has an application-programming interface(API) described in a machine-processable format so that other systems could interact with the web service.

Now the dominant API used in Web service is REST(REpresentational State Transfer). REST is an architecture style for designing networked applications. The idea is to use simple HTTP to make calls between machines. REST defines a set of architectural principles by which Web services could be designed in a resource-centered way, including how resource states are addressed and transferred over HTTP by a wide range of clients written in different languages. REST allows developers to use HTTP methods explicitly and in a way that is consistent with the protocol definition. RESTful Web services, which is built under REST protocol, has gained widespread usage across the Web as a simple Web services.

Now most of the social web services make their REST API public, which allow the clients query the data by calling a method, and respond in REST style XML or JSON file, whether they're on the web, the desktop or mobile devices. The data collection in the thesis is performed such a method from web service such as Flickr, Last.FM, and saved on the local disk in XML format. Thanks to the REST API, the social media data could be backed up in local storage and accessed in following process.

2.2 Analysis Technique

In this section, we mainly review the techniques that will be used in analyzing the social media/events data. The overall goal of the thesis is to extract information or pattern from large scale social data set. In computer science, data mining is derived for such purpose. Data mining represents the essential technologies or algorithms that aim to discover patterns in dataset and transform it into an understandable structure. There are two categories of algorithms in the research community, the supervised classification and unsupervised cluster. Supervised classification is a series of algorithms to learn model from labeled data to predict labels on unlabeled data. They are always employed to solve classification problems. While unsupervised cluster is another series of algorithms developed to find patterns among unlabeled data. They are used in the clustering task. There are lots of supervised and unsupervised algorithms. Specifically, the methods used in this thesis are "Support Vector Machine" for classification purpose, and "Topic Model" for clustering purpose. In this section, we will also discuss some criteria to evaluate the performance of proposed approaches.

2.2.1 Support Vector Machine

Supervised classification is the task of learning a function from labeled training data. In supervised classification, each example is a pair consisting of an input object (a vector feature) and a desired output value (the supervisory signal). The learning algorithm aims at formulating the training data. The inferred function, which is called a classifier, should predict the correct output value for new coming input object. Support vector

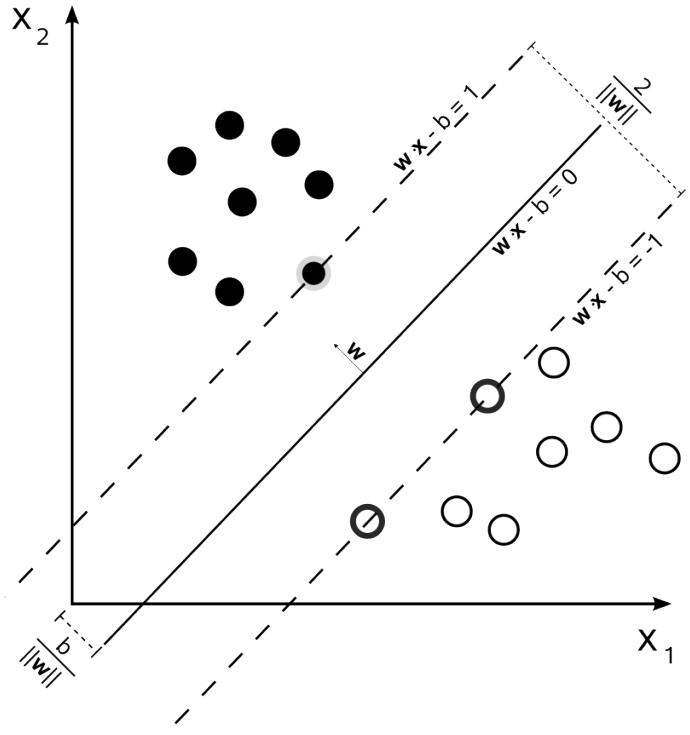


Figure 2.3: Geometry Explanation of SVM

machine(SVM) [Boser et al., 1992] is one of the most effective supervised classification methods and it is commonly used in practices during to its better performance compared with some other classifiers [Caruana and Niculescu-Mizil, 2006].

Mathematically, given some training data D in the form of:

$$D = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$$

where the y_i is either 1 or -1, indicating the class of point $\mathbf{x}_i$. Each $\mathbf{x}_i$ is a "p"-dimensional feature vector. The objective of support vector machines is to find the maximum-margin hyperplane that divides the two classes of points(having $y_i = 1$ vs having $y_i = -1$).

The trained classifier is two hyperplanes that satisfies 1) separates the data and there are no points between them, 2) maximized distance between them. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \phi(\mathbf{x}) - b = 1$$

And

$$\mathbf{w} \cdot \phi(\mathbf{x}) - b = -1.$$

Here $\phi(x)$ is the function to map the data vectors into a higher dimensional space. With optimization theory, the solution is to find the best w and b that satisfy:

$$argmin_{\mathbf{w},b} \|\mathbf{w}^T\mathbf{w}\|$$

subject to:

$$y_i(\mathbf{w} \cdot \phi(\mathbf{x_i}) - b) \geq 1. \forall i \in [1...n]$$

The SVM model can be regards as a hyperplane or set of hyperplanes in a high-dimensional space. Figure 2.3 shows a basic SVM model, which is called binary linear classifier, it could be obtained when $\phi(x) = x$. Binary linear classifier could solve the problems on linearly separable data. In practice there are lots of problems in which the data are not linearly separable. To solve the problems, the kernel function is introduced into SVM model to generate nonlinear classifiers. Kernel function is defined as

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$$

With the kernel functions, the data are mapped into higher dimensional spaces since in higher-dimensional space the data could become more easily separated or better structured. Some of the popular used kernel functions are as follows.

- Polynomial kernel: $k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j})^d$
- radial basis function: $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2)$
- Hyperbolic function: $k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$

SVM model is learned by solving a convex quadratic programming problem. In practice, we used the software package LibSVM [Chang and Lin, 2011] to train the SVM model on our dataset. LibSVM is an integrated and easy-to-use tool for support vector machine, and widely used in various of research direction. It implements many SVM formulations with different kernel functions and provides a simple interface and API where user can easily call it with different type of programming languages.

2.2.2 Topic Model

Unsupervised clustering is a main task of data mining, and a common technique for statistical data analysis used in many fields. It is the task of assigning a set of objects into groups so that the objects in the same group are more similar to each other than to those in other clusters.

Topic model, as a kind of clustering algorithm, is a group of Bayes statistical models for discovering the latent "topics" existing in a collection of documents. In information retrieval, documents collection is represented by the term-document matrix. Latent Semantic Analysis(LSA) [Manning et al., 2008] is the original topic inference approach. It is proposed to find a low-rank approximation to the term-document matrix. The main idea of LSA is to perform Single Value Decomposition(SVD) on the data, so that the concepts can be extracted and the the relation between term and concepts could be discovered. LSA is a linear algebraic model to mine the latent pattern. Probabilistic Latent Semantics Analysis (pLSA) [Hofmann, 1999] model evolved from LSA, which gives a statistical explanation of LSA model. Compared with standard LSA, pLSA defines a proper generative model and has a solid statistical background.

Given the documents collections in the form of co-occurrences of words and documents (w, d), pLSA model, as shown in Figure 2.4 associates an unobserved latent topic variable z between the words and documents, and the probability of each co-occurrence is measured as a mixture of conditionally independent multinomial distribution:

$$P(w, d) = \sum_z P(z)P(d|z)P(w|z) = P(d) \sum_z P(z|d)P(w|z)$$

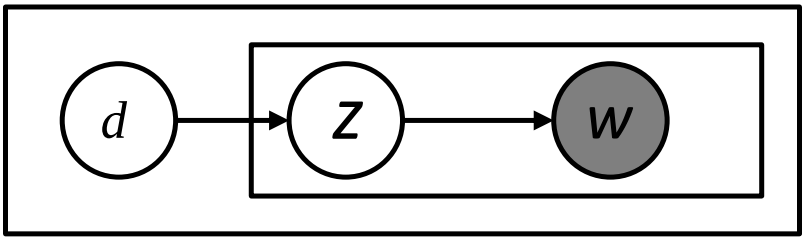


Figure 2.4: pLSA model

The objective function of pLSA is

$$L = p(w, d) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(w_i, d_j)^{n(w_i, d_j)}$$

where $n(w_i, d_j)$ is the frequency of w_i in document d_j. In practice, the function is solved with EM algorithm. In E-step, we calculate the probability of z on given data:

$$p(z|w, d) = \frac{p(z)p(d|z)p(w|z)}{\sum_k (p(z_k)p(d|z_k)p(w|z_k)}$$

and in M-step, the following probability are updated:

$$p(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)p(z_k|d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m)p(z_k|d_i, w_m)}$$

and

$$p(z_k|d_i) = \frac{\sum_{i=1}^{N} n(d_i, w_j)p(z_k|d_i, w_j)}{n(d_i)}$$

Although pLSA is a generative model to the training documents, it is not a generative model and fails to deal with new documents, because the topics are high relevant to documents and there is no prior knowledge on them. Latent Dirichlet allocation (LDA) [Blei et al., 2003], perhaps the most common topic model currently in use, is a generalization of pLSA. In LDA model, the topic distribution is assumed to have a Dirichlet prior. As shown in Figure 2.5. The generative rule is that the documents are represented as multinomial distribution over the latent topics, and each topic is characterized by a distribution on the words dictionary. For a given document w_d, it generates the words in the following process:

1. Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$
2. For each of the words w_{ij}, where $j \in \{1, \ldots, N_i\}$
   (a) Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$.
   (b) Choose a word $w_{i,j} \sim Multinomial(\beta_{z_{i,j}})$.

Where the M denotes the number of documents, N refers to the number of words in documents, and α is the hyper-parameter of the Dirichlet distribution on the latent topics. Given the observed documents, the likelihood of the model can be calculated as

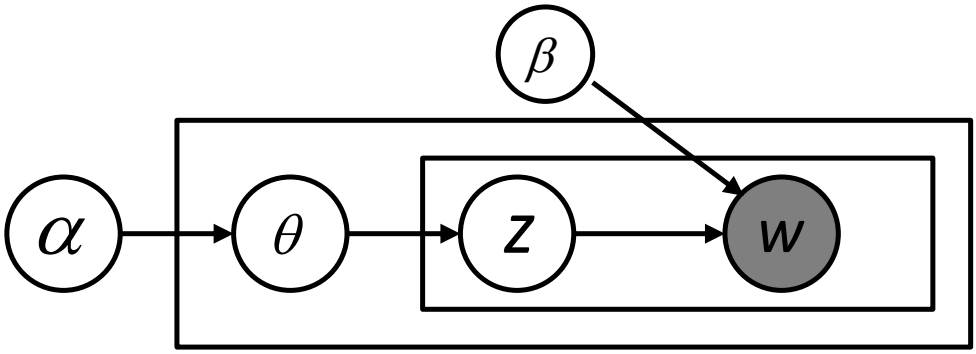


Figure 2.5: LDA model

$$P(W, z, \theta|\alpha, \beta) = \prod_{d=1}^{M} \int P(\theta_d|\alpha) \prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(W_{d,n}|z_{d,n}, \beta) d\theta_d$$

The model parameters can be learned by some Bayes inference methods, such as variable inference, Gibbs Sampling or EM algorithm.

2.2.3 Evaluation

In this section, we shall discuss the methodologies of measuring the performance of proposed algorithm. Nowadays the common used evaluation criteria for classification and clustering is Precision, Recall, Accuracy, F1-score, Purity, and Normalized Mutual Information.

In classification, the precision is ratio of the number of true positives on the total number of elements labeled as belonging to the positive class. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class. Accuracy is defined as the number of true predicted divided by the total number of all elements.

To be more clearly, we use four value *true positives*(tp), *true negatives*(tn), *false positives*(fp), and *false negatives*(fn) to define the precision, recall and accuracy mathematically. The terms *positive* and *negative* refer to the results that are predicted by a system, and *true* or *false* refer to whether the prediction is right corresponding to the ground truth. The relationship of the four value is depicted in Figure 2.6.

Precision, recall, accuracy are mathematically then defined as:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$Accuracy = \frac{tp + nt}{tp + np + np + nt}$$

In practice, F1 is also a evaluation measure and popularly used. It is combined measure that weight the precision and recall, and defined as follows:

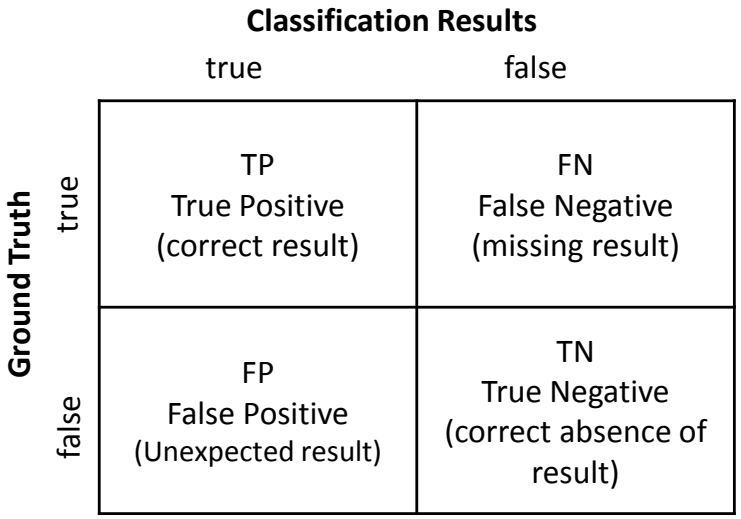


Figure 2.6: Classification VS Ground Truth

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

In a clustering task, the commonly used evaluation criteria is purity and normalized mutual information (NMI). To compute purity, for each cluster the class with most frequency is assigned. The purity of this cluster is then measured by counting the number of correctly assigned documents and dividing by total number of documents in the cluster. And the purity of the whole results is the mean of purity on each cluster.

$$Purity(\Omega, C) = \frac{1}{N}\sum_{k}(max_j(w_k \cap c_j))$$

Where Ω is the set of cluster$\{w_i\}$ and C is the set of classes $\{c_j\}$, and N is the number of clusters.

In the definition, it could be observed that a high purity is easy to obtain (for example, when the number of cluster is equal to the number of documents, purity =1). To make the trade-off between the quality of clustering and the number of cluster, NMI is proposed and defined as follows:

$$NMI(\Omega, C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2}$$

where I is mutual information and defined as:

$$I(\Omega; C) = \sum_{k}\sum_{j} P(w_k \cap c_j) log \frac{P(w_k \cap c_j)}{P(w_k)P(c_j}$$

where $P(\bullet)$ is the probability that documents belonging set $\bullet$.

H is information entropy that could be defined as:

$$H(\Omega) = -\sum_{k} P(w_k) log P(w_k)$$

NMI combine the quality of clustering and number of clusters, hence is a good evaluation criteria in cluster task.

2.3 Related Work

The objective of the thesis is to mine the patterns between social media and events data, and to find an effective way to manage the social data. Some of the problems have been addressed in multimedia retrieval, social media analysis, multimedia illustration, and event discovery. In this section, we mainly survey the state-of-the-art work that are proposed in these fields.

2.3.1 Multimedia retrieval

Information retrieval is an essential task in computer science that has been extensively studied for many years [Manning et al., 2008]. The idea of using computers to search for relevant pieces of information was popularized by Vannevar Bush in 1945 [Singhal, 2001]. In the past several decades, lots of researches have been done concerning searching for documents, for information or metadata within documents. Recently, with the popularity of multimedia data, such as audio, image and video, multimedia retrieval begins to draw the attention in the research community. It studies the problem of how to extract semantic information for multimedia data source that are captured by digital equipment but hard to present by text directly. The methodology of the multimedia retrieval is to extract some features from the perceivable data directly and then solve the problems with machine learning techniques. TRECVid is an annual benchmarking activity focusing on a list of different multimedia information retrieval research areas about content based video retrieval. The goal of the workshop is to promote research in multimedia retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. The usage of low-level visual features for improving content-based multimedia retrieval systems has made great progress in the past ten years [Datta et al., 2008]. Figure 2.7 shows a general framework of multimedia retrieval. Firstly features are extracted from the labeled images or video key-frames of videos, then the classifiers are trained after a learning process. When a new dataset comes, the classifiers can be used to give a proper estimation to concepts.

There are two main problems addressed in multimedia retrieval research. At first, feature extraction studies the problems of what features to extract in order to improve the system performance. Features such as color histogram, texture, shape and edge abstraction have then been discussed [Lew et al., 2006]. Recently, SIFT [Lowe, 1999] feature makes great process when representing the content. SIFT feature is a local descriptor, and it is extracted from some key points on an image detected by algorithms such as HoG. Nowadays SIFT is popularly used in content based retrieval since it is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination change.

Once image features are extracted, the question remains as to how they could be formulated in a retrieval model. The methods essentially aim to reduce the semantic gap as much as possible. Numerous learning approaches have been developed to solve the retrieval problem. For example, Bayesian classification was used to build the image retrieval model by Vailaya et al. [2001], and Supervised classification based on SVMs has been applied to model the visual feature in Zhang et al. [2001]. Some researchers focus on solve the problem with rich media context. In Tang et al. [2012], the authors proposed

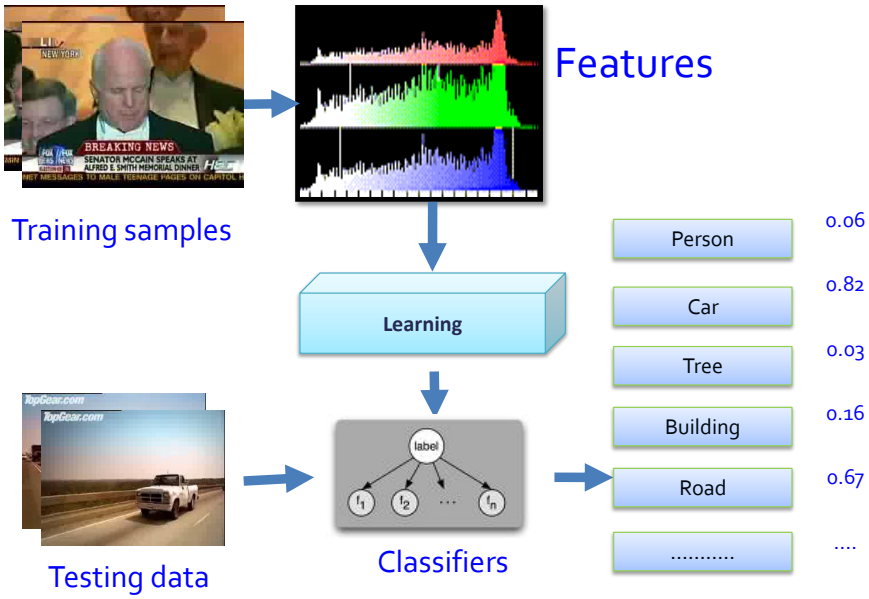


Figure 2.7: The framework of multmedia retrieval system

a semantic-gap-oriented active learning method, which aims at bridging the semantic gap with user-interaction. In [Glotin et al., 2010], a divide and conquer approach is proposed to address the image retrieval problem. The proposed method, matrix modular support vector Machine, could partition a task into small subtasks, and fuse results to produce a good decision. Among the research, some applications are also built and published online. ALIPR[7] is an automatic image annotation system [Li and Wang, 2008], which is probably the most successful online system. In ALIPR, an effort to automatically validate computer generated tags with human-given annotation is being used in an attempt to build a very large collection of searchable images.

In Multimedia retrieval, as the difficulty of recognizing the displayed scenes from the raw image/video data, most of the proposed approaches suffer from the semantic gap problems, that is the gap between low level visual features and high level concepts. In the thesis, we focus on analyzing social media data leveraging event knowledge and ontologies, which shows a possible way to alleviate semantic gap.

2.3.2 Multimedia DataSet

In order to make the multimedia data searchable by its content or context, various methods of mapping the multimedia data into high-dimensional spaces have been introduced. To evaluate the proposed methods, some datasets are released as popularly used as the benchmark.

The TREC Video Retrieval Evaluation(TRECVid) [Over et al., 2011] dedicates to promote progress in content-based analysis and retrieval from digital video by providing a large scale collection. It is funded by NIST and other US government agencies. Many researchers worldwide contribute significant effort over the last ten years. And this effort published in its annual workshop has yielded a better understanding of how systems can effectively

[7] http://www.alipr.com/

accomplish such processing and how one can reliably benchmark their performance.

PASCAL dataset [Everingham et al., 2009] is another dataset popularly used in computer vision research. It is supported by the PASCAL challenge which aims at recognizing objects in realistic scenes. It provides a well-labeled dataset: all of the objects in training data are annotated within rectangles. Now PASCAL challenge has been an important benchmark in objects recognition. The annual results of this challenge are reported in its annual workshop.

The LabelMe dataset [Russell et al., 2007] from MIT is quite similar to the PASCAL dataset since it contains general photographs containing multiple objects. LabelMe provides both local and web-based annotation interface, and it encourages casual and professional users to contribute and share annotations. A well designed toolkit is also provided by LabelME. With the toolkit, the researchers can access the data and visual features easily, hence they can focus on developing and deploying their algorithms.

All of the mentioned datasets are traditional datasets which have been used in the research community for many years. With the popularity of social media, huge amount of photos and videos are uploaded online at an exponential growth. Besides the visual content that also exists in traditional dataset, the social media data also brings many metadata in different modality, such as “time”,”GPS coordinate”, ”text description” and ”tags”. To facilitate the research on social media, many well designed datasets are also released. For example, NUS-WIDE [Chua et al., 2009] is one of the most popularly used dataset which was created by National University of Singapore. The dataset includes about 270K images downloaded from Flickr. In this dataset, some low level visual features are extracted for these photos and ground-truth of 81 concepts are labeled in the dataset by manual process. MIRFlickr [Huiskes and Lew, 2008] is another well-known dataset, which is offered by the LIACS Medialab at Leiden University in Netherlands. MIRFlickr has two versions with different scale, MIRFLICKR-1M collection and MIRFLICKR-25000 collection. Similar to NUS-WIDE, it also supplies a number of content-based visual descriptors as well as rich metadata for the entire set of images. The ground truth of the dataset is labeled with main concept and subconcept hierarchically. On those social media datasets built very recently, a number of recent research has shown acceptable results [Zha et al., 2009, Hong et al., 2010].

It is very time-consuming and labor-intensive to build a dataset since much effort is needed to label data manually. In the thesis, we study the problem of how to collect the training samples in an automatic way.

2.3.3 Ontology and Semantic Web

Data structuring and reusing is a challenge in big data. Web based knowledge sharing demands that human and/or machine agents agree on common and explicit rule so as to exchange to share knowledge across different communities and domains. In computer science and information management, knowledge reuse is facilitated if they are structured by ontology. Ontologies are the structural frameworks for organizing information and are used in artificial intelligence. It defines a set of representational primitives with which to model a domain of knowledge or discourse. In other words, ontology describes a domain with a knowledge base.

Ontologies are defined and used in semantic web applications today to structure data. In traditional WWW, the data is unstructured and hard to reuse when cross domain. To enabling users to find, share, and combine information more easily, the semantic Web

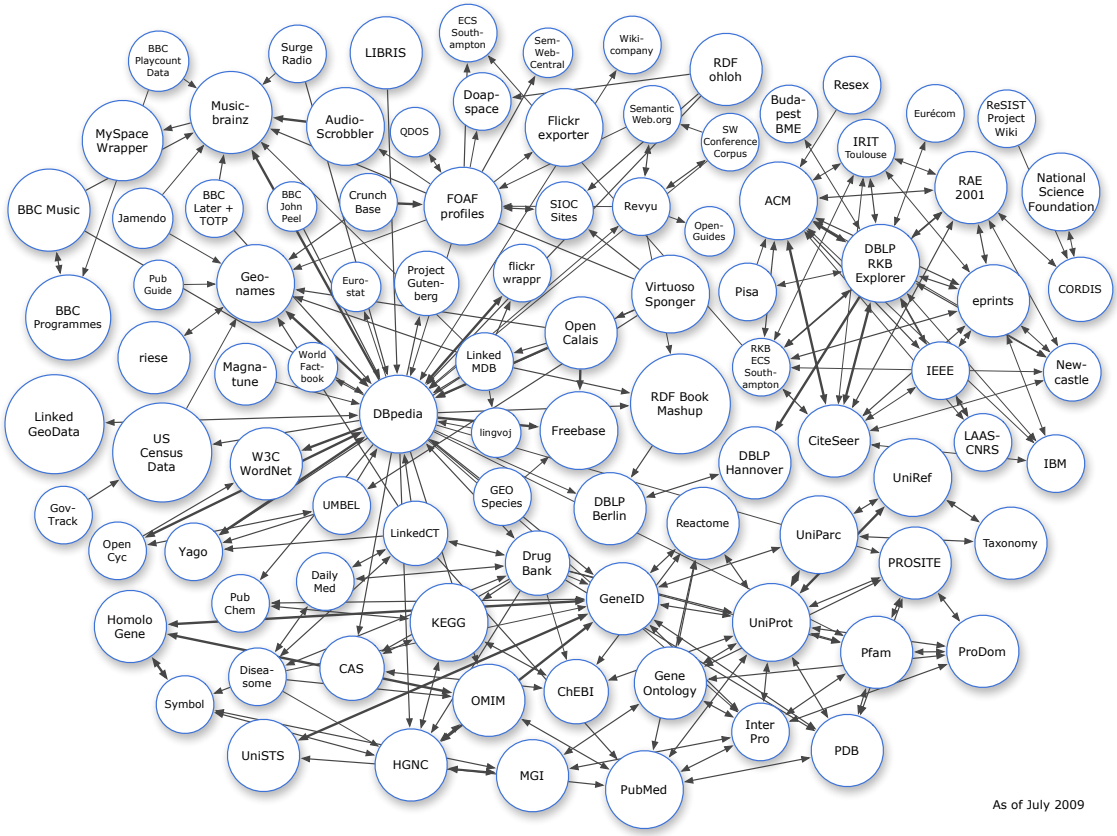


Figure 2.8: The Linking Open Data cloud diagram, the version of July 2009 with 95 datasets, and the latest version with 295 datasets could be downloaded from `http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19.pdf`

is proposed by W3C to link data among different sites. The Semantic Web is a system that enables machines to "understand" and respond to complex human requests based on their meaning. Such an "understanding" requires that the relevant information sources be semantically structured. According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries."

In the semantic web, data is represented in Resource Description Framework (RDF) format. RDF is a framework for representing information about resources in a graph form. It is designed to represent structured metadata of WWW resources, such as the title, author, time, location. RDF links enable users to navigate data from an item within one data source to related data items within other sources using a Semantic Web browser. RDF links can also be followed by the crawlers of Semantic Web search engines, which may provide sophisticated search and query capabilities over crawled data. As query results are structured data and not just links to HTML pages, it is possible to reuse them within other applications.

The availability of tools and systems has contributed to the increasingly development of semantic web. For example, the W3C Linking Open Data community project is launched with the objective "to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different

data sources". Figure 2.8 shows some dataset examples that have been published and interlinked by the Linking Open Data project. In the latest results(until September 2011), there 295 data sets consisting of over 31 billion RDF triples, which are interlinked by around 504 million RDF links.

The ontologies and semantic web technologies have been applied in many different fields. For example, LSCOM[8] is the large scale concept ontology used in content based multimedia research. LODE [9] is an ontology for linking open descriptions of events. The structure knowledge could facilitate the related research work. In Athanasiadis et al. [2005], the authors studied the use of knowledge for the automatic extraction of semantic metadata from multimedia content, they extended and enriched current general-purpose ontologies to include low-level visual features to represent the knowledge. In Bertini et al. [2010], a novel web-based tool was presented that allowed a user friendly semantic browsing of video collections based on ontologies and concepts.

Semantic web is a very useful tool to structure data. In this thesis, we present a large dataset composed of semantic descriptions of events, photos and videos interlinked with the larger Linked Open Data cloud and we show the benefits of using semantic web technologies for integrating multimedia metadata.

2.3.4 Social Media analysis

With the ubiquitous availability of media sharing web service, lots of online media data has been shared online, such as most notably community photo collection Flickr, and the video collection YouTube. These collections contain not only large scale image/videos data, but also rich metadata such as text(in the form of tags, titles, description, comments), time(which refers when the medias are taken and uploaded), location(which suggests where the medias are taken at). In the research community, working with data from these media collections to extract high-quality information has received increasing attentions recently.

To facilitate the research on social media data, NUS-WIDE [Chua et al., 2009] was created by National University of Singapore. The large scale dataset includes images associated tags from Flickr. Some low features are extracted for these photos and ground-truth of 81 concepts could be found in the dataset for evaluation purpose. Besides NUS-WIDE, there are also some other benchmark collection popularly used for research. For example, MIRFlickr [Huiskes and Lew, 2008] or MSRC [Schroff et al., 2011] are alternative choices for visual concept analysis.

In recent years there has been considerable interest for social media analysis in research community. As we know, tags are commonly used on media sharing web sites, but tags are also very diverse as they are labeled by different users with various purpose. In [Liu et al., 2010], the authors proposed a social image retagging approach that aimed to assign better content descriptor to the social images and remove noise description. Xu et al. [2009] proposed a tag refinement algorithm named as regularized latent Dirichlet allocation (rLDA) to jointly model the tag similarity and tag relevance. Liu et al. [2009]proposed a bi-layer sparse coding methods to infer how an image or semantic region to be reconstructed from the over-segmented image patches of the entire image dataset. In [Tang et al., 2009], the authors took tags as a knowledge source and they studied the problem of inferring semantic concepts from associated noisy tags of social images. Besides these works that focus on studying tags in social media, some joint methods which integrate both the visual content

[8] http://www.lscom.org/index.html

[9] http://linkedevents.org/ontology/

and the tags are proposed. Gao et al. [2012] proposed an united hyper-graph learning approach that simultaneously utilizes both visual and textual information to estimate the relevance of user tagged images. To annotate the images more accurately, Tang et al. [2011] proposed a novel kNN-sparse graph-based semi-supervised learning approach for harnessing the labeled and unlabeled data simultaneously.

In the recent work of social media, much achievement has been gained to mine the graphical metadata in social media data. In [Hays and Efros, 2008], the authors studied the problem of estimating geographic information from an image, they proposed a simple algorithm for estimating a distribution over geographic locations from a single image using a KNN approach. In [Quack et al., 2008] an approach was described to retrieving geotagged photos from these web-sites using a grid of geo-spatial tiles and to mine images of objects from community photo collections in an unsupervised fashion. Arase et al. [2010] proposed a method to segment photo collections into trips, categorize them based on their trip and detect people's trip patterns based on their research of geo-tagged photos.

2.3.5 Multimedia illustration

In recent years, research on how to better support the end-user experience when searching and browsing multimedia content has drawn lots of attention. It is well known that vivid photos could attract human attention compared with text description. In [Delgado et al., 2010], to improve the users' attention when reading news articles, a system was proposed to help people reading news by illustrating the news story. The application provides mechanisms to automatically select the best illustration for each scene and to select the set of illustrations to improve the story sequence. In [Joshi et al., 2006], an unsupervised approach was presented to describe stories with automatically collected pictures. In this framework, semantic keywords are extracted from the story, and used to search an annotated image database. Then a novel image ranking scheme automatically choose the most important images. In [Zhu et al., 2007], a Text-to-Picture system was developed that synthesizes a picture from natural language text without limitation. The system firstly identified "picturable" text units by natural language processing, then searched for the most likely image parts conditioned on the text, and finally optimized picture layout conditioned on both the text and image parts. Besides the works that illustrate text work with photos, some studies also have been done to generate video representation from textual content. For example, in [Schwarz et al., 2010], a system was presented to create a visual representation for a given, short text. In this system, the authors also used some techniques to query images by the given text string, and the novelty is that the final images are selected in a user-assisted process and automatically used to create a storyboard animation. All of these approaches or systems studied how to demonstrate text content with multimedia data.

Our scheme presented in this thesis attempts to enrich set of images/videos to illustrate social events, with studying users' uploading behaviors on Flickr and matching concert events with photos based on different modalities; such as text/tags, time, and geo-location, and resulting in an enriched photo set which better illustrates events. A similar work was presented in [Diakopoulos et al., 2010], where a strategy was proposed to studies how to extract the valuable information from the overwhelming amount of content on social media on given broadcast news. But their work focused on filtering noise information and outline the summarization, while illustrating events with media addresses the problem of how to leverage vivid multi-modal content to share experience.

2.3.6 Event Illustration and Modeling

In our work, we defined events as the public happening taken on a given location and time. The earlier relevant research focused on the study of news, since the data about news are abundant and easy to collect [Billsus and Pazzani, 1999, Toda and Kataoka, 2005]. With the popularity of social media sites, these repositories are used by users to share their experiences and interests on the Web. These sites also host substantial amounts of user-contributed materials (e.g., photographs, videos, and textual content) for a wide variety of real-world events of different type and scale. How to mine the events information has gathered recent attention. In [Becker et al., 2009, 2010], the authors follow a very similar approach, exploiting the rich "context" associated with social media content and applied clustering algorithms to identify social events. In [Mattivi et al., 2012], a demonstration was proposed to categorize photos by events/subevents by visual content analysis. Diakopoulos et al. [2010] analyzed Twitter messages corresponding to large scale media events to improve event reasoning, visualization, and analytic. Some other research has been done to find events directly from Twitter post [Weng and Lee, 2011, Sakaki et al., 2010]. In [Weng and Lee, 2011], the authors studied how to employ a wavelet-based techniques to detect events from Twitter stream. A similar method can be found in [Chen and Roy, 2009] to detect events from Flickr time series data. In [Sakaki et al., 2010], the authors investigate how to filter the tweets to detect seismic activity as it happens. In [Trad et al., 2011], a method is introduced to retrieve events-related photos in a collections.

A natural extension of our work would benefit from [Kennedy and Naaman, 2009]. In this paper, the authors proposed a system to present the media content from live music events, assuming a series of concerts by the same artist such as a world tour. By synchronizing the music clips with audio fingerprint and other metadata, the system gives a novel interface to organize the user-contributed content. In this thesis, we did not yet consider audio fingerprint for tracking down series of events but rely only on semantic metadata so far.

A web service with similar illustration functionalities can be found in EventBurn[10]. It creates a summary of a given hot event from popular services like Twitter, Facebook, and Flickr, but fails to extract events automatically from social media streams.

2.3.7 Social Event Detection

In computer vision, event detection concerns the recognition and localization of special spatial-temporal patterns from image sequence [Ballan et al., 2010], which is a challenging topic to computer vision/video surveillance researchers [Aggarwal and Cai, 1997]. Recent years are witnessing the success of content-sharing web site such as Wikipedia, Flickr, YouTube, Last.fm etc. The Web 2.0 has effectively exploded the number of data sources. Social media data in the format of text documents, images, audio, video, accomplished with rich metadata data are now easily created, and shared. Research related to and using social media to facilitate the traditional work the has become a hot topic in the past years [Joshi and Luo, 2008, Kennedy et al., 2007, Hays and Efros, 2008, Papadopoulos et al., 2011b]. Some work studied the problems of how to explore the rich data to make progress on traditional event detection problem. In [Joshi and Luo, 2008], the authors proposed a method to summarize spatial neighborhoods around the taken place of photos as bag of geo tags, incorporate geographical information for events categorization. In [Duan et al.,

[10] `http://www.eventburn.com/`

2010], a visual event recognition framework was proposed for consumer domain videos by leveraging a large amount of loosely labeled web videos. These work incorporated social media data to model the short activities recognition (such as walk, jump et. al).

However, the events addressed in the chapter is different compared with these work: we define event as something happening in a specific place and time, for example, a concert held in a building during night, a international conference or meeting taken in an exhibition center. And our objective is to mine the social event pattern from social media stream by robust learning approaches. Generally, typical event discovery methods maybe include clustering, statistical, probabilistic model based methods and some other approaches.

The event detection by clustering focuses on clustering social media data by visual or spatial-temporal metadata features. It clusters the data firstly and then identify the cluster as event/non-event in the following process. For example, Quack et al. [2008] presented methods to mine events and object from community photo collections by clustering approaches. In their system, the photos are clustered according to several different modalities(visual and textual features), and the clusters are then classified as objects or events by their durance and users, since events are usually characterized by a short duration. A very similar framework is proposed to classify the events and landmarks in [Papadopoulos et al., 2011b].

Statistical methods for event detection involve to find the reproduced pattern among the time series data. Among the various methods that have been developed, static threshold method is the simplest and straightforward one. Events are found when the monitor parameters are achieved a given threshold. For example, In [Rattenbury et al., 2007], focusing on the problem of extracting place and event semantics for tags that are assigned to photos on Flickr, a burst scan approach was proposed to extract semantics of tags, unstructured text-labels assigned to resources on the Web, based on each tag usage patterns. The static threshold method could meet the timeline requirement and provide immediate output results, but it is hard to find out a proper value as the threshold point and fails to provide robust results.

Wavelet-based spatial analysis is another statistical method used in detecting events due to its robustness to noise. In [Chen and Roy, 2009], a wavelet based approach is proposed to detect events from social media data. At first, the temporal and spatial distributions of tag usage are analyzed by a wavelet transform to suppress noise, so that the tags could be identified if they are related with events.

Probabilistic model based approaches consist of these methods in which the event or other related probabilities are modeled with Bayes inference framework. In particular, the topic models such as LSA, pLSA or LDA [Blei et al., 2003] have shown encouraging results in natural language processing. Specific to event detection, Pan and Mitra [2011] developed a system to combine the popular LDA model with temporal segmentation and spatial clustering for automatically identifying events from a large document collection. In [Firan et al., 2010], the authors focused on building a Naive Bayes event models which classify photos as either relevant or irrelevant to given events.

The prototype presented in [Gao et al., 2011] attempts to identify public event using both spatial-temporal context and photo content. In their system, event information are collected as event database, and then photo content model are built for different types of events. With the assistance of such as system, users could find information about events or activities when they travel in foreign place. In some aspects, their work shows interesting complementarity with our work. The idea of employing Twitter data to search for additional information about an event could be used in our work to extend the set of

relevant words describing events. As a result, the enrichment process would retrieve an even more diverse set of candidate images.

In this thesis, we analyze the trends in social media data and leverage burst detection and topic model based approaches to discovery events.

2.3.8 Automatic Data Collection

Dataset annotation is a time and labor consuption process. The problem of collect data with high quality labels has been addressed for many years. For example, In [Fergus et al., 2005], an approach was proposed to learn an object classifier from just its name, by utilizing the raw output of image search engines available on the Internet. The developed model,TSI-pLSA was extended from pLSA to include spatial information in a translation and scale invariant manner. Berg and Forsyth [2006] proposed a system to collects images containing categories of animals from internet. Their system combines of both the text surrounding and the visual feature of the images to collect a large number of animal images. In [Li and Wang, 2007], a model named as OPTIMOL was built to create well labeled dataset automatically. The system worked in an incremental learning framework, it could be updated by the newly accepted images in current round and be used in the next round. In [Schroff et al., 2011], a two-step approach was proposed. Firstly a Bayes posterior estimator is trained on the surrounding metadata of images to rerank the initial list. Then the top ranked images are used to learn a SVM classifier and further refine the ranking. In [Li et al., 2011], the authors presented work related to the collection the negative training samples from the semantic analysis of tags and visual features. There method is also done in two steps. At first, tagging vocabularies are built to label a set of reliable negatives, and then visual classifier are trained to obtain the real negatives.

The work presented in this thesis tackles the problems arising when attempting to organize social media at large scale, which gives new challenge on the traditional multimedia retrieval approaches. Events are a natural way for people to organize, browse and visualize media collections. The research on event could leverage the event domain knowledge and ontologies to solve practical problems. And the latest pilot results on event semantic show a possible way to alleviate the semantic gap [Fialho et al., 2010, Becker et al., 2010]. In thesis, we focus on exploiting the inner semantic between social media and events in several aspects. Firstly, to explore the information with flexibility and depth, we exploited the wealth of social event context, and proposed a framework to illustrate events with rich media data. Secondly, to discover the event patterns from media data, we study media uploading trends and the latent topics among them, and developed several approaches to mine events. Thirdly, to model events in visual aspect, we presented a method to collect training data automatically leveraging the social event context.

2.4 Conclusion

In this chapter, we review the essential knowledge that will be involved in the thesis. At first, we introduce the basic of social media website Flickr and 3 social event websites. The content of these service and how to query the data are discussed. Then we review the classification and clustering techniques to be used in analyzing social data in the following section, which include SVM and Topic Model. We also give a brief discuss on how to evaluate the performance of a proposed algorithms. In the end, we review the related work

in the field of multimedia data analysis. In the following chapter, we will see the details of mining the social media data with described techniques.

Chapter 3

Media Enrichment for Social Events

Organizing media data according to events in the real world is the natural way for human to recall his experience. Exploiting event context to solve the management and retrieval problems raising from the social media draws lots of interest in the multimedia community. In this chapter, we present our work to infer the semantics behind the events and explore social media to illustrate events. With the study of users uploading behavior, we extend the set of illustrating images and videos for a particular event by querying social media with diverse and multi-modality features, and pruning the results with content based visual analysis. In addition, a more general framework is also proposed to visualize events for a selected topic from social media data by leveraging data mining techniques. In this chapter, how to illustrate events with rich media data is studied in details.

3.1 Introduction

Events are one of the most important cues to recall people's past memory [Tulving, 1984], and a natural way for human to organize their life. Traditionally, people would like to group their data by events, for example, to put the photos taken from the same event together so that they could be found easily. Motivated by such phenomena, this chapter focuses on building the link between events and social media, so that the social media data could be indexed by events. In other words, we exploit the techniques to infer the heuristics behind the events and study how to explore social media to illustrate events. There are two parts of work reported in this chapter. At first, a method is presented to combine semantic inference and visual analysis for automatically finding media (photos and videos) illustrating events. The overall goal is to design a web-based environment that allows users to explore and select events, to inspect associated media, and to discover meaningful, surprising or entertaining connections between events, media and people participating in events. The heuristic is validated with data from media sharing platforms and large event directories. We present a large dataset composed of semantic descriptions of events, photos and videos interlinked with the large Linked Open Data cloud and we show the benefits of using semantic web technologies for integrating multimedia metadata. Secondly, we extend the approach in the first part and proposed an event-based query expansion framework to visualize events for a selected topic from social media data by leveraging data mining techniques. In this framework, some essential questions in information retrieval, such as query expanding with Natural Language Process, retrieving multi-modality data from different web services, and depicting result vividly, are studied carefully.

3.2 Our Event Illustration Framework

Our goal is to aggregate these heterogeneous sources of information using linked data, so that we can explore the information with the flexibility and depth afforded by semantic web technologies. Furthermore, we investigate the underlying connections between events to allow users to discover meaningful, entertaining or surprising relationships amongst them. We also use these connections as means of providing information and illustrations about future events, thus enhancing decision support. In this section, we present a method for automatically finding medias hosted on Flickr and YouTube that can be associated to public events. We show the benefits of using linked data technologies for enriching semantically the descriptions of both events and media data.

3.2.1 LODE Ontology and Event Directories

Large numbers of web sites contain information about scheduled events, of which some may display media captured at these events. This information is, however, often incomplete and always locked into the sites. In previous research, user study has been carried out in order to collect end-user experiences, opinions and interests while discovering, attending and sharing events, and user insights about potential web-based technologies that support these activities. The results of this study support the development of an environment that merges event directories, social networks and media sharing platforms [Fialho et al., 2010]. We argue that linked data technology is suitable for doing this integration at large scale given they naturally based on URIs for identifying objects and a simple triple model (RDF) for representing semantic descriptions. In this section, we revise the LODE event model and describe the techniques to populate this ontology by scraping three large event directories: last.fm, eventful and upcoming.

The LODE ontology[1] is a minimal model that encapsulates the most useful properties for describing events [Shaw et al., 2009]. The goal of this ontology is to enable interoperable modeling of the "factual" aspects of events, where these can be characterized in terms of the *four Ws*: *What* happened, *Where* did it happen, *When* did it happen, and *Who* were involved. LODE is not yet another "event" ontology *per se*. It has been designed as an *interlingua* model that solves an interpretable problem by providing a set of axioms expressing mappings among existing event ontologies. Hence, the ontology contains numerous OWL axioms stating classes and properties equivalence between models such as the Event Ontology [Raimond et al., 2007], CIDOC-CRM, DOLCE, SEM [van Hage et al., 2009] to name a few.

Figure 3.1 depicts the metadata attached to the event identified by id=`1380633` on Last.fm according to the LODE ontology. More precisely, it indicates that an event categorized as a `Concert` has been given on the `24th of January 2010 at 20:00 PM` in the `Henry Fonda Theater` featuring the `Radiohead` rock band. The link between the media and the event is realized through the `lode:illustrate` property, while more information about the `sioc:UserAccount` can be attached to his URI. Hence, we see that the video hosted on YouTube has for `ma:creator` the user `aghorrorag`. We use the Last.fm, Eventful and Upcoming APIs to query the online events and then convert each event description into the LODE ontology. We mint new URIs into our own namespace for events (`http://data.linkedevents.org/event/`), agents (`http://data.linkedevents.org/agent/`) and locations (`http://data.linkedevents.org/location/`). A graph representation of an event is com-

[1] `http://linkedevents.org/ontology/`

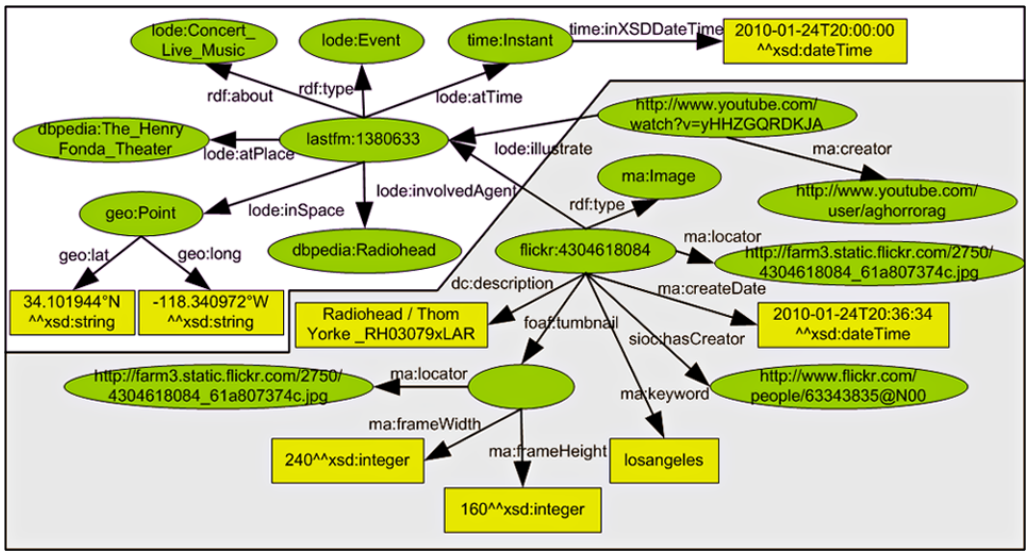


Figure 3.1: The *Radiohead Haiti Relief Concert* described with LODE (*top*) and illustrated with media described by the Media Ontology (*bottom*)

posed of the type of the event, a full text description, the agents (e.g. artists) involved, a date(instant or interval represented with OWL Time [Hobbs and Pan, 2006]), a location in terms of both geographical coordinates and a URI denoting the venue and users participation. A graph representing an agent or a location is composed of a label and a description (e.g. the artist's biography). Event directories have overlap in their coverage. We interlink these events descriptions when they involve the same agents at the same date or when they happen at the same venue at the same date. We invoke additional semantic web lookup services such as DBpedia and freebase, or foursquare and geonames in order to enrich the descriptions of the agents and the locations. Hence, the venue has been converted into a foursquare URI (`http://foursquare.com/venue/185188`) that provides additional information such as the number of different users that have *check in* at this place and the current virtual mayor while the wikipedia URI (`http://fr.wikipedia.org/wiki/Nouveau_casino`) provides the history of this venue in French.

The agent URI, which has for label "Róisín Murphy" has also been interlinked with the DBpedia URI (`http://dbpedia.org/resource/RóisínMurphy`) which provides additional information about the solo singer such as its complete discography. This URI is declared to be `owl:sameAs` another identifier from Freebase (`http://rdf.freebase.com/ns/guid.9202a8c04000641\discretionary{-}{}{}f80000000004a1685`) which provides information about the 2 bands she has been part of. The linked data journey can be rich and long. One of the challenges we want to address is how to visualize these enriched interconnected datasets while still supporting simple user tasks such as searching and browsing enriched media collections.

3.2.2 The EventMedia Dataset

As described in Section2.1.1, explicit relationships between scheduled events and photos hosted on Flickr can be looked up using special machine tags such as `lastfm:event=XXX` or `upcoming:event=XXX`. The work of [Troncy et al., 2010b] has explored the overlap in metadata between four popular web sites, namely Flickr as a hosting web site for photos and Last.fm, Eventful and Upcoming as a documentation of past and upcoming events. A

large dataset called "EventMedia" is presented which is composed of events descriptions together with media descriptions associated with these events and interlinked with the larger Linked Open Data cloud. In this dataset, more than 1.7 million photos are linked by nearly 110.000 events in total (Table 3.1).

	Event	Agent	Location	Photos	User
`Last.fm`	57,258	50,151	16,471	1,393,039	18,542
`Upcoming`	13,114		7,330	347,959	4,518
`Eventful`	37,647	6,543	14,576	52	12

Table 3.1: Number of event/agent/location and photo/user descriptions in the dataset published in [Troncy et al., 2010b]

In this section, we consider a subset of this events dataset that corresponds to the intersection of Last.fm, Flickr and YouTube to discover meaningful, surprising or entertaining connections between events, media and people participating in events. In other words, we consider the set of last.fm events for which there is at least one photo and one video shared respectively on Flickr and YouTube that has been tagged with the `lastfm:event=xxx` machine tags. Since machine tags are actually not popular in YouTube, the number of YouTube videos that actually contains such a machine tag is unsurprisingly much smaller than the number of Flickr photos. Hence, this intersection yields a dataset of 110 events, 4790 photos and 263 videos.

The Ontology for Media Resource currently developed by W3C is a core vocabulary which covers basic metadata properties to describe media resources[2]. It provides properties for describing the duration of a video, its target audience, copyright, genre, rating or the various renditions of a photos. Media fragments can also be defined in order to have a smaller granularity and attach keywords or formal annotations to parts of a media. The ontology contains a formal set of axioms defining mapping between different metadata formats for multimedia. We use this vocabulary together with properties from SIOC, FOAF and Dublin Core to convert into RDF the Flickr photo and YouTube video descriptions (Figure 3.2). The link between the media and the event is realized through the `lode:illustrate` property, while more information about the `sioc:UserAccount` can be attached to his URI. In Figure 3.2, we see that both the video hosted on YouTube and the photo hosted on Flickr has the same `ma:creator`: the user `cartoixa`.

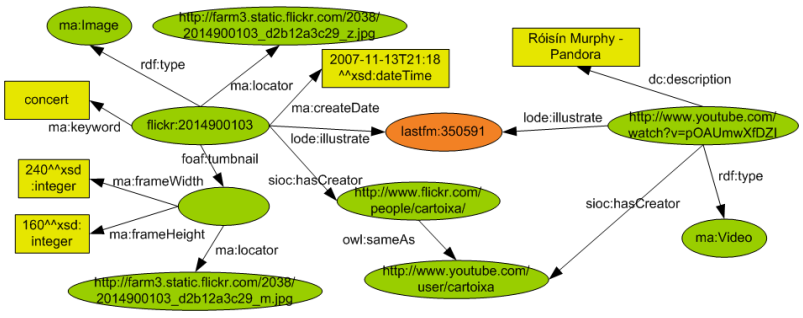


Figure 3.2: A photo and a video taken by the same user at the *Róisín Murphy Concert* described with the Media Ontology

[2] `http://www.w3.org/TR/mediaont-10/`

3.2.3 Find Media Illustrating Events

The set of photos and videos available on the web that can be explicitly associated to an event using a machine tag is generally a tiny subset, lots of media data that are actually relevant for this event are out of the scope. Our goal is to find as much as possible media resources that have **not** been tagged with a `lastfm:event=xxx` machine tag but that should still be associated to an event description. In the following, we investigate several approaches to find those photos and videos to which we can then propagate the rich semantic description of the event improving the recall accuracy of multimedia query for events.

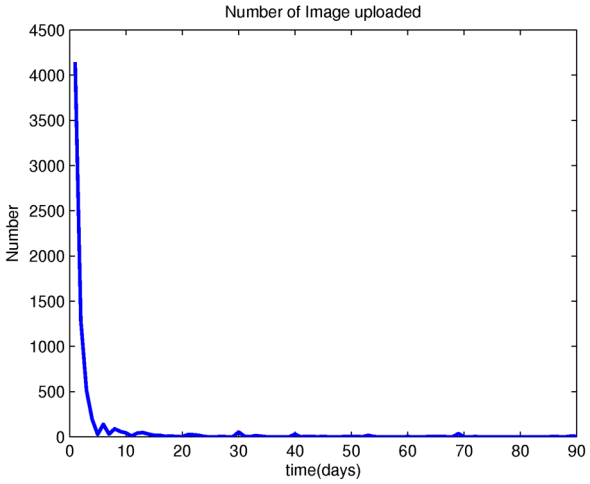


Figure 3.3: Image uploading tendency along time

Starting from an event description, four dimensions from the LODE model can easily be mapped to metadata available in Flickr and YouTube and be used as search query in these two sharing platforms: the *what* dimension that represents the title, the *where* dimension that gives the geo-coordinates attached to a media, the *when* dimension that is matched with either the taken date or the upload date of a media, and the *who* dimension that suggests the artists involved in the events. Querying Flickr or YouTube with just one of these dimensions brings far too many results: many events took place on the same date or at nearby locations and the title is often ambiguous. Consequently, we will query the media sharing sites using at least two dimensions. We also find that there are recurrent annual events with the same title and held in the same location, which makes the combination of "title" and "geo tag" inaccurate. In addition, we also discard the *who* because of its inconvenience to perform the media query. Actually, there are always too many artists joining an events, and nothing could be found if all of the name unionized as the query parameters. In addition, the artists likely fill the "stage name", other than his/her real name, which are either no meaning at all (for example "Yr Ods","Yeah Yeah Yeahs"), or with misunderstood meaning(for example "Beach House" "Blue Roses"). So querying with artists names will bring more noisy media another than relevant ones. In the following, we consider the two combinations "title" + "time" and "geotag" + "time" for performing search query and finding media that could be relevant for a given event. It should be noticed that the query is not very specific and some irrelevant media data will be retrieved. To prune the noisy media, a visual content analysis technique is developed, which aims at removing

the noise images if the visual difference is remarkable enough. Since we know that the media data labeled with machine tag is highly relevant to events and could be obtained easily, they are the best choice as the training samples for filtering noise. However in many events, only few images labeled with machine tags could be queried, and it also be found in these cases, noisy images from the query results with geo tags are hardly found. Hence we use these data to build a visual model to filter the erroneous medias, as described in section 3.2.3.4. The whole framework to enrich event with media data is described in Figure 3.4.

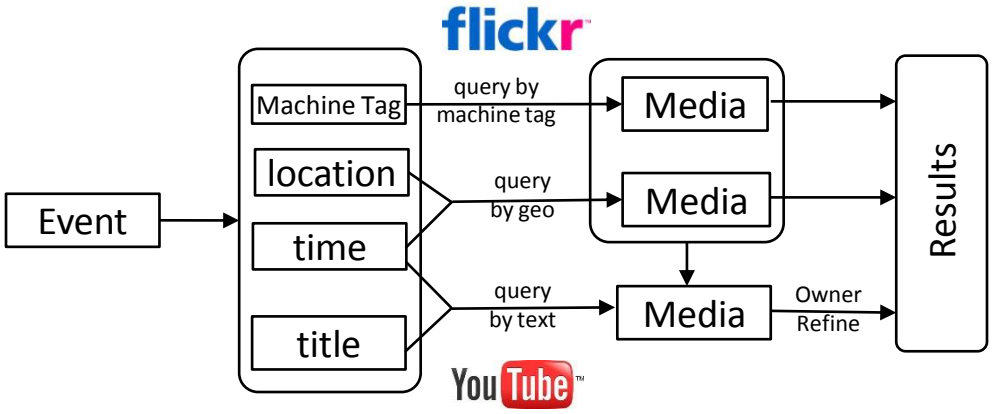


Figure 3.4: The proposed framework to enrich event with photos/videos

3.2.3.1 Media Context Analysis

We would like to collect high quality social media data by online query with geographical, temporal, and textual parameters. How to choose the query parameters plays an important role in the process. If loose parameters are given, many irrelevant media will be obtained and pollute the results. However, querying with parameters that are too strict will reduce the number of highly relevant media data. To make the tradeoff between quality and number, we should study the time and location trends of the media with machine tags, to infer the proper time and location window corresponding to events.

Since the media documents labeled with machine tags are taken at events, we do temporal-spatial statistics on these data to find out underlying principles. Time is one of the most key components of event, and there are more than one time measurement in events corresponding with media: event taken time, media taken time, media post-process time, media uploading time and so on. To find out a reasonable time window to fit our query, we first investigate the time difference between the start time of an event and the upload time of Flickr photos attached to this event. For the 110 events composing our dataset, we analyze the 4790 photos that are annotated with the Last.fm machine tag in order to compute the time delay between the event start time and the time at which the photos were captured according to the EXIF metadata[3]. Figure 3.3 shows the result: the y-axis represents the number of photos uploaded on a day to day basis, while the x-axis represents the time (in days) after the event occurred.

[3] http://en.wikipedia.org/wiki/Exchangeable_image_file_format

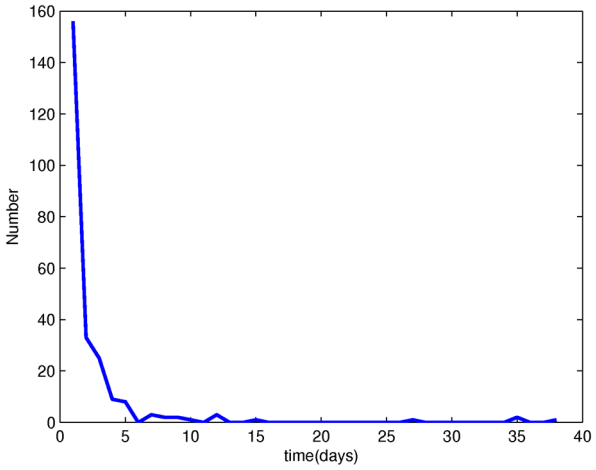


Figure 3.5: Video uploading tendency along time

The trend is clearly a long-tail curve where most of the photos taken at an event are uploaded during or right after the event took place and within the first 5 days. After ten days, only very few photos from the event are still being uploaded. In the following, we choose a threshold of **5 days** when querying the photos using either the title or the geotag information. We conduct a similar analysis with the 263 YouTube videos that are annotated with the Last.fm machine tag. The "taken time" metadata not available for videos in YouTube, we use the"upload time" instead. Figure 3.5 shows the results and we observe the same long tail: most the videos are uploaded within the first 5 days following an event.

Following we would like to model the venue location. The Flickr API allows to query photos based on their geographical location. Given region parameters, in the form of center and radius, or rectangle bounding box, the photos taken within a specified location can be retrieved. However, it is not so easy to obtain the geographical area covered for a place, since there are no public data for the size of a venue. We address this issue by leveraging on the event context provided by Last.fm and used by Flickr users. On a given venue(VenueID $= V$), all of the past events ($\{eid\}$) which took place there could be retrieved using the Last.fm API. Then the machine tags "lastfm:event=eid" is used to search for geo-tagged media on Flickr. Following a bounding box is computed using the GPS coordinates of the retrieved photos. The basic idea is to compute the bounding box with photos taken near the location, and to filter the ones which are far from the bounding box. The final bounding box is estimated as the minimized rectangle of the GPS coordinates after removing the outliers (photos which are located further than twice the variance of the set in either direction (longitude or latitude)). Algorithm 1 details the processing steps leading to the venue's location estimation.

Figure 4.2 shows the result of our bounding box estimation approach for the venue Koko (London, UK). The blue marker is the GPS location of Koko according to Last.fm, and the red markers are the places where photos were taken and labeled by machine tags of past event ids shared on Flickr. The red rectangle corresponds to the learnt bounding box for the venue Koko. We see (top part of Figure 4.2) that some photos taken too far away from the venue have been appropriately discarded.

Algorithm 1 Estimate the bounding box for a venue

```
INPUT: VenueName
OUTPUT: BoundingBox
PhotoSet = [ ]
EventSet=GetPastEvent(VenueName)
for each eventid in EventSet do
    photos = GetFlickrPhotos(eventid, hasGeo = True)
    PhotoSet.append(photos)
end for
GeoSet = GetGeoInfo(PhotoSet)
GeoSet.filter()
return MinRect(GeoSet)
```

3.2.3.2 Query by Geotag

Nowadays, geographical metadata is a common and key component in social media data. It could be labeled by an automatically extracting process if the media is captured by GPS-equipment devices, or be labeled manually when users sharing their media online. The metadata, named as geotags, usually are described in different format. For example, it is always composed as latitude and longitude coordinates, though it can also include altitude, bearing, distance, accuracy data, or place names. Geotags provide information to retrieve and manage media data. They are extremely valuable for application to structure the data according to location and it is also helpful for users to find a wide variety of location-specific information [Arase et al., 2010, Zheng et al., 2009]. Since we have already known that many photos/videos are captured during events, and some of them likely are labeled with geotags indicating event taken places, these media data could be retrieved if querying with geotags parameters. Considering that a place is generally a venue, we assume that at any given place and time there is a single event taking place. For all events in our dataset, we extract the latitude and longitude information from the LODE descriptions and then perform geographical based query using the Flickr API applying a time filter of 5 days following each event date. We perform the same query using the YouTube API although the number of videos that are geotagged is much smaller than photos. Figures 3.6(a) and 3.6(b) show the distribution of the number of retrieved photos and videos for the 110 events in our dataset. We observe that the data is centralized in the left bins which means that for most of the events (n=95), the number of photos (resp. videos) retrieved with geotags is within the 0-100 range (resp. 0-20 range). The largest bin is composed of 45 events that have each between 1 and 50 photos retrieved.

3.2.3.3 Query by Title

Title is the most describable and readable information for events. Similarly to geo-tagged queries, we perform full text search queries on Flickr and YouTube based on the event titles that are extracted from the LODE description. The retrieved photos and videos are also filtered using a time interval of five days following the event taken time. When performing search query using the Flickr API query, we use the "text mode" rather than the "tag mode" since the latter is more strict and many photos will miss. The number of photos retrieved at this stage is however in an order of magnitude greater than with

geo-tagged queries. Due to the well-known polysemy problems of textual-based query, the title-based query brings lots of irrelevant photos. We describe in the Section 3.2.3.4 an heuristic for filtering out irrelevant media. In contrast, we do not observe this noise when querying the YouTube API with only the event title (filtered by the time of the event) using a strict match mode. Hence, the number of videos retrieved per event is rather small and most of the them are relevant. The distribution of the number of retrieved photos and videos for the 110 events in our dataset is depicted in Figures 3.7a and 3.7b. Generally, the results of query by title have a similar distribution than the result of query by geotag. For most of the events, a lower number of photos is obtained. Out of the 110 events under investigation, there are 80 events with less than 150 photos, and 83 events with less than 25 videos. However, for some events, a large number of media is retrieved: 12 events (resp. 15) with more than 500 photos (resp. 50 videos). Compared with Figure 3.6, we can clearly see that the standard deviation of Figure 3.7 is larger and that again photos are more readily available than videos.

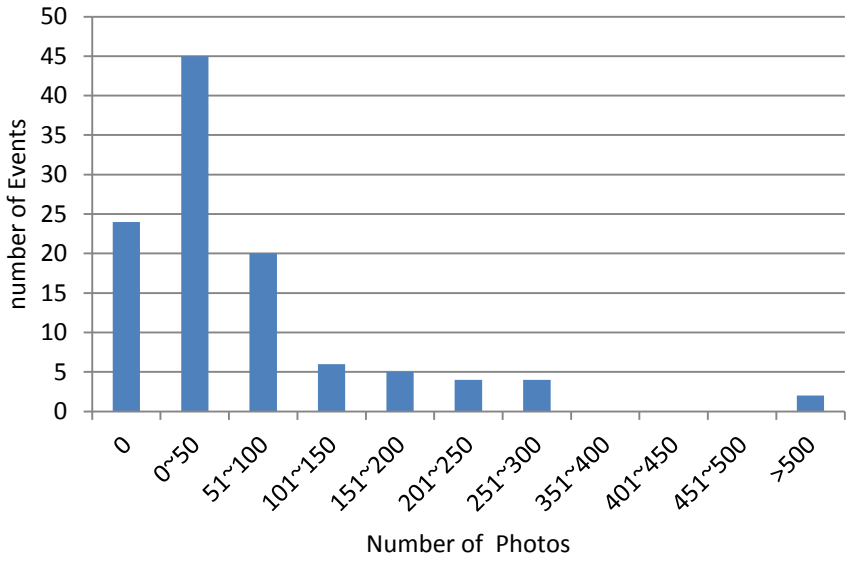
3.2.3.4 Pruning Irrelevant Media

Images and videos with specific machine tags such as `lastfm:event=207358` can be unconditionally associated to events. We consider that media retrieved with geotag queries during a correct time frame should also be relevant for those events. The problem arises with the media retrieved with text-based queries (using the event title) where one can find many irrelevant media. For example, the event identified by 207358 has for title `Malia`. However, a search on Flickr or YouTube with this keyword returns photos about cities, different people (Malawian singer, French swimmer, daughter of the US president Barack Obama) or even hotels with this name.

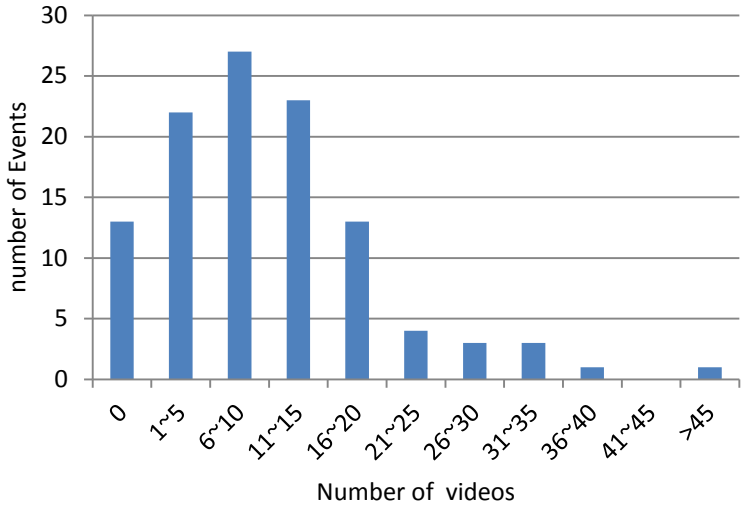
In order to filter out this noise and to avoid propagating rich event descriptions to these media documents, we propose a method for pruning the set of candidates photos using visual content based analysis. The photos captured at a single event are very diverse, depicting the artist, the scene, the audience or even the tickets. The diversity of the data makes it difficult to remove all the noisy images that should not be associated with the event considered, while keeping as much as possible the relevant ones. We address this issue in two steps to ensure high precision and recall ratio.

The main idea of our algorithm is to measure the visual similarity of media documents, that is, to learn a threshold from training set and used to filter noisy data in testing data. First, we build a training dataset composed of the media containing either the event machine tag or a combination of geo-coordinates and time frames corresponding to the event dimensions. The photos resulting from query by title compose the testing dataset. The visual features used in our approach are 225D color moments in Lab space, 64D Gabor texture, and 73D Edge histogram. For each image pairs in the training data, the nearest neighbors algorithm using the ***L1*** distance measure in the training set is performed and the smallest distance is taken as threshold. Second, images originating from the title query are composed of training images. Images for which the distance to images in the test set is below the threshold are candidates for illustrating the event. Mathematically, let E as the training photos set, and F as the testing photos set. The objective is to select the photos from F which are similar to the photos in E, to additionally enrich the set E illustrating an event. The visual similarity between two images is computed as follows:

$$L_1(F_j, E_i) = \sum_k |F_j(k) - E_i(k)| \tag{3.1}$$



(a) Number of photos per event in geotag based query



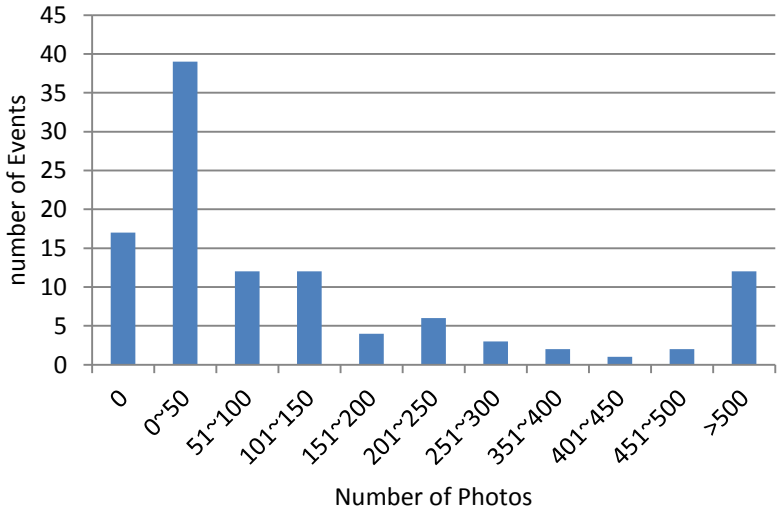
(b) Number of videos per event in geotag based query

Figure 3.6: Statistics for geotag based query

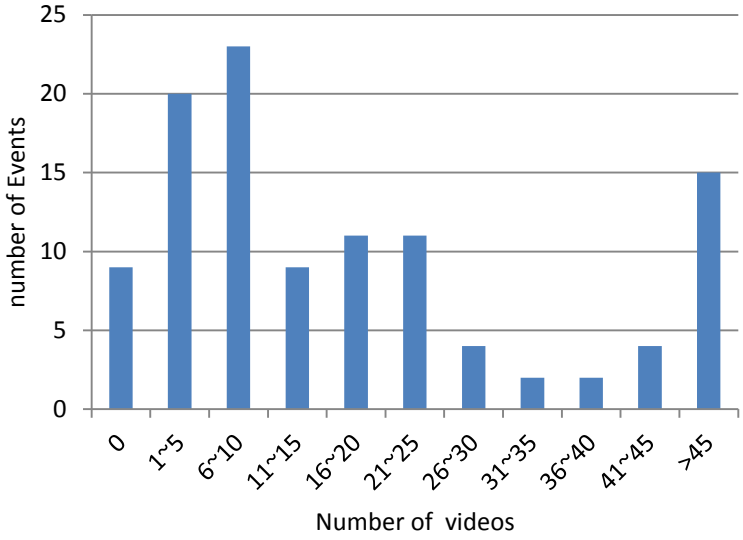
where $F_j(k)$ and $E_i(k)$ are normalized concatenating low level feature vector of the images. F_j is added to the set of media illustrating the event when

$$\exists E_i \in E : L_1(F_j, E_i) < THD_i$$

where THD_i is the threshold which is also learned from the E data. As shown in Equation 3.2, we use a strict strategy to decide the threshold, which is chosen as the minimal value of similarity of images pairs in training set. And the threshold is also adaptive to



(a) Number of photos per event in title based query



(b) Number of videos per event in title based query

Figure 3.7: Statistics for title based query

different events because of the visual diversity within the training dataset. In order to remove noisy images in the testing data, the threshold should be adjusted respectively. Figure 3.8 shows the value of threshold used in the experiments which range from 0.01 to 0.346.

$$THD_i = \min_{\{j\}\setminus i} \sum_k |EventMedia_j(k) - EventMedia_i(k)| \tag{3.2}$$

The algorithm can be formalized as follows in Algorithm 2:

Algorithm 2 Pruning function

```
INPUT: TrainingSet,TestingSet
OUTPUT: PrunedSet
for each img in TrainingSet do
    D = [ ]
    for each imgj in TrainingSet-{img} do
        D.append(dist_L1(img,imgj))
    end for
    Threshold = min(D)
    for each imgt in TestingSet do
        if dist_L1(imgt,img)< Threshold then
            PrunedSet.append(imgt)
        end if
    end for
end for
return PrunedSet
```

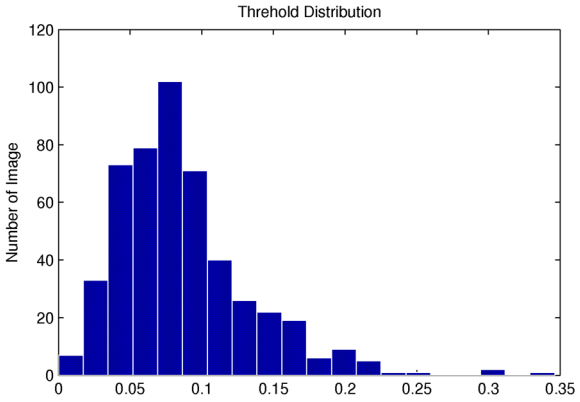


Figure 3.8: The distribution of threshold

In visual pruning, in order to filter out most of the irrelevant photos, a strict threshold strategy is applied, and some relevant ones are also be discarded, which leads to a lower recall ratio. In order to recover these photos and improve the recall ratio, we exploit the "owner" property in social media and proposed a refinement method. We assume that a person cannot attend more than one event simultaneously. Therefore, all the photos that have been taken by the same owner during the event duration should be assigned to the event. So if the owner has shared more photos during this period, they are automatically added as illustrative media for the event. Using this heuristic, it is possible to retrieve photos which do not have any textual/geographical description. As far as we know, "owner refinement" is the only effective approach to match events and media data when no enough metadata(such as textual, graphical metadata) is available.

Table 3.2: Number of photos and videos retrieved for 110 events using the event machine tag (ID), the geo-coordinates or the event title

	QueryByID	**QueryByTitle**	**QueryByGeo**	**ID∩Title**	**Geo∩Title**	**Geo∩ID**	**Geo∩ID∩Title**
Photos	4790	32583	6933	2350	494	484	405
Videos	263	4237	1163	103	39	115	29

Table 3.3: Number of photos for 20 events, results of the pruning algorithm and results of the simple heuristic extension

ID	**DataSet (nb of photos)**			**Pruning Result**			**Extended Heuristic**	
	TrainingData	TestingData	GroundTruth	Pruned	Precision	Recall	Extend	NewRecall
346054	2	24	2	1	1	0.500	1	0.500
158744	3	48	48	23	1	0.479	44	0.917
371981	4	16	6	4	1	0.667	4	0.667
341832	7	0	0	0	1	1.000	0	1.000
362195	7	0	0	0	1	1.000	0	1.000
235445	10	1	1	0	1	0.000	0	0.000
42644	13	85	81	13	1	0.160	13	0.160
165697	23	1	1	0	1	0.000	1	1.000
137530	24	9	4	0	1	0.000	1	0.250
517159	24	0	0	0	1	1.000	0	1.000
222241	36	204	180	33	0.97	0.183	72	0.400
234649	45	35	4	1	1	0.250	1	0.250
207358	54	68	4	4	1	1.000	4	1.000
429517	60	171	169	27	1	0.160	41	0.243
437747	65	144	142	8	1	0.056	13	0.092
117886	68	99	97	4	1	0.041	11	0.113
150390	71	16	16	1	1	0.063	1	0.063
350591	79	85	85	6	1	0.071	66	0.776
472733	93	500	478	8	1	0.017	18	0.038
176257	97	260	255	47	1	0.184	147	0.576
Avg	785	1766	1573	180	0.998	0.114	438	0.278

3.2.4 Results

Table 3.2 shows the overall number of photos and videos retrieved for each strategy for the 110 events that composed our dataset. We first observe that these two strategies are effective to retrieve an order of magnitude more media than using solely machine tags. Hence, while 4790 photos are tagged with the `lastfm:event=xxx` machine tag, 6933 photos can be retrieved using the geo-location of the event and 32583 photos can be retrieved using the event title. After removing the duplicated ones, we obtain 36412 photos that are candidate to illustrate an event which is 7,6 times more than the ones labeled by a machine tag. For the videos, the number of candidates is 19,6 times more than the ones with machine tags. Unsurprisingly, most of the media uploaded and shared on the web do not have machine tag.

For evaluating our pruning algorithm, we take the top 20 events from our 110 events dataset. For these 20 events, there are 785 images in the training set (photos containing either an event machine tag or a geotag) and 1766 photos in the testing set (photos retrieved by event title). We build manually the ground truth for those 1766 photos selecting which ones should be attached to an event and which ones should not (Table 3.3). The 20 events were all concert events and photos are often depicting artists, venues, stages or audience. Some photos were, however, sometimes hard to judge but the manual assessor used all metadata available around each photo such as the entire list of tags or the albums in which the photos were gathered to decide whether the photo should be discarded or not. In the

end, we manually remove 193 irrelevant images by their visual appearance and metadata. The remaining 1573 images are used as ground truth dataset.

The results of the pruning algorithm detailed in the Section 3.2.3.4 applied to the 1766 photos shown in the Table 3.3. The threshold used is quite strong in order to guarantee a precision of **1** for most of the events. However, this causes about 80% of the candidate images to be excluded, besides many relevant photos. In order to increase the recall ratio, we extend the resulting images by our pruning algorithm with all the ones uploaded by the same uploader. The reason is that if one photo can reliably be attached to an event, we infer that this participant indeed attended the event and that all the others photos taken by this person during this time frame are likely to be illustrative media for this event. This simple heuristic allows to significantly improve the recall ratio (from 0.114 to 0.278) without sacrificing to the precision.

Figure 3.9: Tag Clouds of photos associated to the event 1097166: *Alela Diane at Tivoli De Helling (Utrecht) on 14 Jul 2009*

Figure 3.10: Photo Collage for event 1097166: *Alela Diane at Tivoli De Helling (Utrecht) on 14 Jul 2009*

We also investigate the concept shift as the set of relevant media increases by looking at the tag cloud associated to these media attached to a particular event. Figure 3.9 depicts tag cloud examples for the event `1097166` that corresponds to the live concert of Alela Diane which took place on Tuesday 14 July 2009 at 7:30pm at the venue `Tivoli De`

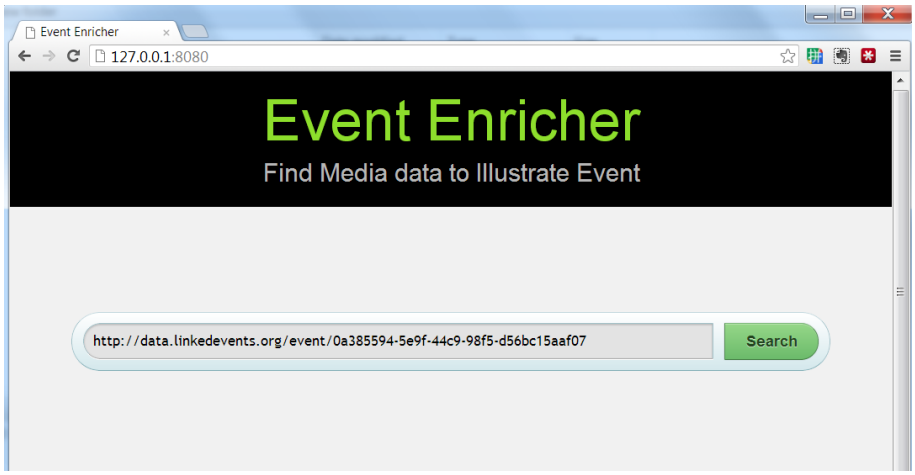


Figure 3.11: Interface of illustrating social event in EventMedia

`Helling` in Utrecht, The Netherlands. From the three sub figures, we can clearly see the topic shift when new metadata is added from a another source. Figure 3.9(a) is built from all the tags of all photos retrieved when using the event machine tag: the most frequent keyword in this cloud is, as expected, `lastfm:event`. However, in Figure 3.9(b), when the tags set are enlarged by the metadata from the photos retrieved by query by title (and pruned with our algorithm), the topic shifts to `aleladiane` who is the artist performing during this event. A similar observation can be made after looking at Figure 3.9(c), where some metadata from query by geotag are added, and the most significant keyword changes to the location of the event `utrecht`.

Besides tag cloud, we also provide another vivid visualization called photo collage [Mei et al., 2009], to illustrate events. Photo collage is a visual clustering technique that can depict the event with different points of view. With the enriched photos, the method proposed in [Mei et al., 2009] is used to create the photo collage for an event. And Figure 3.10 demonstrates the video collages for event `1097166`.

3.2.5 Demonstration

Finally, we briefly present an online web service that we have developed to incorporate the work present in this section to search and browse media illustrating events. As shown in Figure 3.11, the web service named as EventEnricher and developed with Python + web.py. On a given event URI defined in EventMedia, the demo could show the enriching results with several tables. In details, with the event URI as the query parameter, the service firstly query the event information on EventMedia dataset, and parse the event context such as event title, taken time, taken place. Then the approaches presented in this section is employed to query media data in Flickr. After the pruning and refinement process, the final data is presented in four tables named "query by machine tag", "query by title", "query by geo", and "final results" in a new web page, as shown in 3.12.

3.2.6 Discussion

Event directories are largely overlapping, providing multiple identifiers for the same venues, artists, and events. We argued that linked data technology helps to integrate at large scale

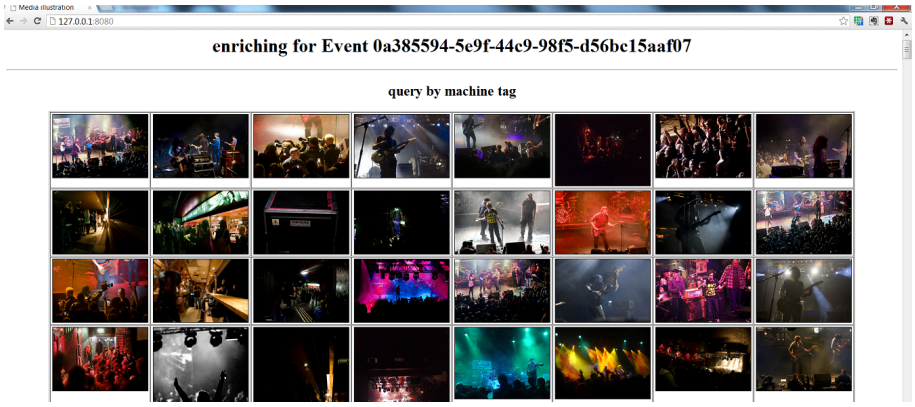


Figure 3.12: Enriching results for event: 0a385594-5e9f-44c9-98f5-d56bc15aaf07. Many photos are not shown here because of the limited space

all data sources because of the use of URIs for identifying objects and a simple triple model for representing all metadata yielding a giant graph. Rich semantic descriptions of events can then be propagated to the media to which they are attached. Hence, for the dataset[4] presented in the Section 3.2.2, 1,248,021 photos (that is 73 %) have been geo-tagged for free since Flickr had no geo-tagged information for those photos but only knowledge of an event machine tag that points to a rich description of an event including venues that are geo-localized. Similarly, the propagation of semantic metadata enables to detect inconsistencies between data sources such as the misplacement of a venue.

In this section, we have shown how linked data technologies can be used for integrating information contained in event and media directories. We used the LODE and Media Ontology respectively for expressing linked data description of events and photos. We described a method for finding as much as possible photos and videos relevant for a given event: we start from the media that contain specific machine tags and that can be used to train classifiers that will prune results from general queries. We evaluated our approach against a manually built gold standard and we show that we are able to increase significantly the recall with a very conservative approach that does not scarify the precision. Ultimately, we aim at providing an event-based environment for users to explore, annotate and share media and we present an initial user interface that we continue to develop. We are currently consolidating and cleaning our dataset with more sources and more linkage. We intend to provide soon user participation at events from public Foursquare check-in and live Tweets. Our priority is also to express the right licensing and attribution information to the data that has been rdf-ized. We truly believe that multimedia will then be finally added back to the Semantic Web.

In our research, we also studied the problems of graphical metadata used in Flickr and other web service thoroughly. We found that Flickr can not provide high quality geo tags service, and one or two kilometers error is very common in Flickr. For examples, the following Figure 3.13 shows the location of “Parc Del Forum”, which is located in Barcelona Spain. The blue marker (with GPS coordinates: (latitude=“41.385" longitude=“2.170”)) is queried from Flickr, and the red one is queried from Last.fm, and the GPS coordinates is latitude=41.383796,longitude=2.191429. There is about 4 kilometer distance between

[4]The entire dataset is composed of more than 30 million RDF triples and is available as a dump at `http://www.eurecom.fr/~troncy/ldtc2010/`

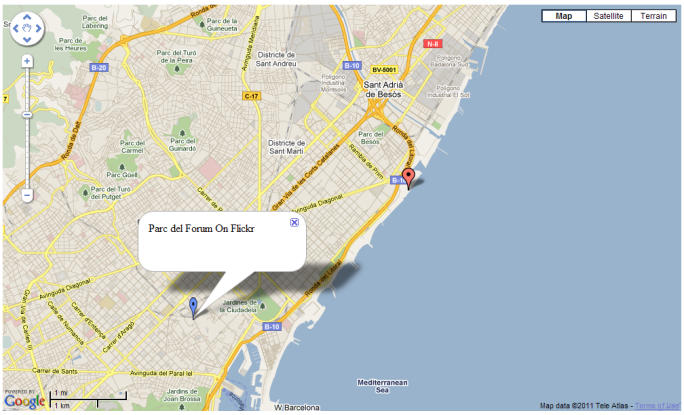


Figure 3.13: Parc Del Forum in different web service

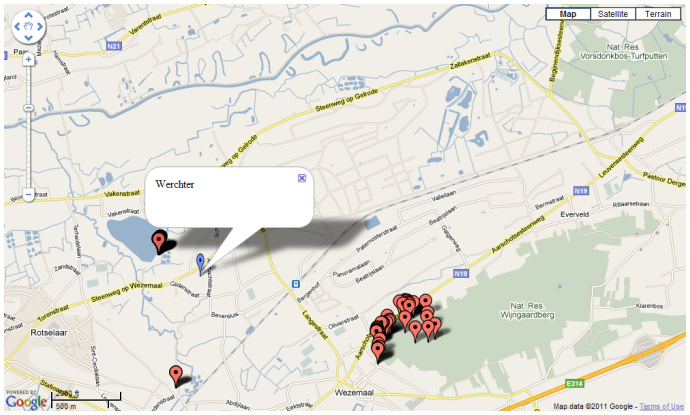


Figure 3.14: Photos taken in venue werchter, labeled by blue marker

them.

In addition, though highly accurate GPS information could be obtained from photos which are taken by GPS equipped capture, there is a problem on some other data if the photo is labeled to Flickr map manually. The users always lack of enough background knowledge on the venue and it is difficult for a user to find the precise location when they want to give a GPS label to their photos, and an approximate location is probably targeted finally. Figure 3.14 shows the photos taken during the event held in venue wercheter, Belgium. It could be noticed that though most of the photos are closed with each other, there are still some photos which are very far from the cluster. By the way, the blue marker is the venue place provided by Flickr, which is not useful because of the inaccuracy problem as we just discussed. The bounding box estimation algorithm in this section is used to remove the noisy photos and to find out a reasonable venue place and span window.

3.3 Event-based Topic Expression

Social media has been playing a vital role in our digital life. Its massive content contributed by users can derive many interesting applications. With the exponential increase of the online resources contributed by social media users, how to fully leverage such vast amount of information to benefit all the online users with interesting services, is still an open

Figure 3.15: The snapshot of a query result example

and challenging problem. In this section, we address a more general problem: discovering and illustrating events on any given query, and propose a novel framework to extract and illustrate social events automatically by leveraging the massive social media content. The two main difficulties of this challenge are: 1) how to extract the events according to the given query; 2) how to illustrate the events properly with sufficient media data. We answer the first question by parsing the query terms semantically using a natural language processing algorithm, we interpret the query semantically to identify the related topic, location and time, and then extract the most relevant popular events from social news web service, based on the semantic interpretation of query. We tackle the second question by aggregating relevant information from various different media sources, for every event, we query relevant tweets from Twitter and compelling pictures provided by Google image search. Finally each event is illustrated in a multi-modality format – rich textual description, tag cloud and photo collage. Our work makes two main contributions:

- We extract events from social news websites instead of identifying them from social media directly as others. This benefits our framework by avoiding any additional time cost, extra computation or storage requirements. Relevant events data is retrieved online at real time. Thanks to the built-in collective intelligence of these social news web services, events returned by our framework are mostly the most popular ones.
- To illustrate an event well, we collect multiple types of media documents from different media sources. Thus, end users are provided with not only multi-modality data, but also different views on an event. Using several popular presentation techniques (e.g. tag cloud, photo collage, etc.), we display the search results with a rich user-friendly interface.

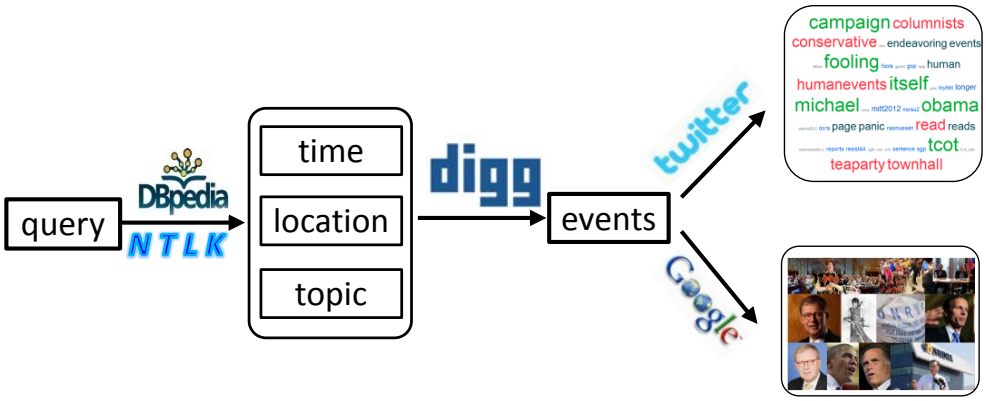


Figure 3.16: Overview of the proposed framework

With our novel proposal, a list of related events could be retrieved for an textual query, as shown in Figure 3.15.

3.3.1 Our Proposal

Our framework, shown in Figure 3.16, extracts and illustrates public events by leveraging the social media data directly. Since events can be defined as something happening at certain time in a given location, we start by parsing the query to identify the topic, location and time information with natural language processing algorithm. Rather than detecting events from Twitter data directly, events can be obtained by crawling and scraping from social news web service. This saves time, computation and storage compared to alternative events detection processes. To provide a vivid illustration for each event, we retrieve the relevant tweets from Twitter, and show the collected data with tag cloud. In addition, we also retrieve photos with Google image search engine, and summarize the results with photo collage/montage.

3.3.1.1 Semantic Query Parsing

The three basic properties of event are location, time, and topic, as stated in [Liu et al., 2011]. To identify the meaning of a given query input, we would like to extract the information in the three dimensions. Here, we assume that the query input is a noun phrase headed or tailed with complements, such as "The news of the past three days in New York". We extract the structured data from this noun phrase, where there is a predictable organization of entities and relationships. The issue of extracting structured data from text has been well studied in Natural Language Processing (*NLP*). This process, composed of 3 steps is performed using *NLTK*[5], a well-known *NLP* package. First, the input text is segmented into words using a tokenizer. Then, each word is tagged with part-of-speech tag (*POS*), which provides the lexical categories for words. With the *POS* tag, we use the *RegexpParser chunker* in *NLTK* to create the chunker tree and identify each sub noun phrases in the input string. The process is depicted in Figure 3.17

Then, to determine the semantic meaning of each noun phrase, different techniques are employed to extract the location, time, and topic information from the parsed noun

[5] `http://nltk.org/`

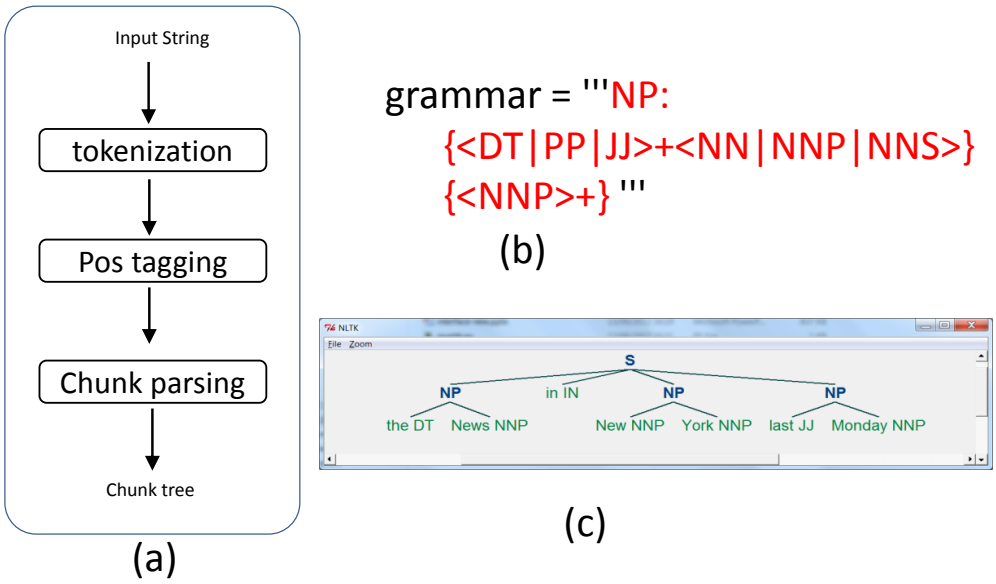


Figure 3.17: Semantic parsing using NLP. (a) the flowchart; (b) the grammar used for chunker parsing; (c) an example for input query "the News in New York last Monday";

phrases. The location is obtained by a query of the DBpedia[6] knowledge database, which provides structured information extracted from Wikipedia. If geographical metadata, such as "geo:point", "geo:area", are found for a noun phrase, the corresponding location is kept as reference for the events to collect. For determining the time information, we develop a script which parses and converts the human readable string, such as "tomorrow","last week", "Monday" to a time structure. We use the DBpedia API and our script to parse the sub noun phrases in order to obtain the location and time information addressed by the query. Since it is hard to model the topic in sophisticated way, we assign the nouns as the topic keywords of the event to search for, if neither time or location can be determined from it.

3.3.1.2 Events Extraction

Recently, there has been some research focusing on detecting events from social media data, such as detecting events from Twitter stream [Weng and Lee, 2011], or Flickr [Chen and Roy, 2009]. Indeed, useful information can be mined from community-contributed data. However, in these methods, huge amount of data have to be downloaded from the web service previously and lots of computation should be spent for the following process. Therefore, it takes many resources to store the data and a lot of time to process it.

Currently, the world's biggest happenings are collected as News by communication services and broadcasted publicly to the mass audience through various channels. Some web services, such as Google News[7] or Digg[8] have been developed to organize the news data in a structured data. In this section, rather than detecting events from social streams, we

[6] http://dbpedia.org
[7] http://news.google.com
[8] http://digg.com/

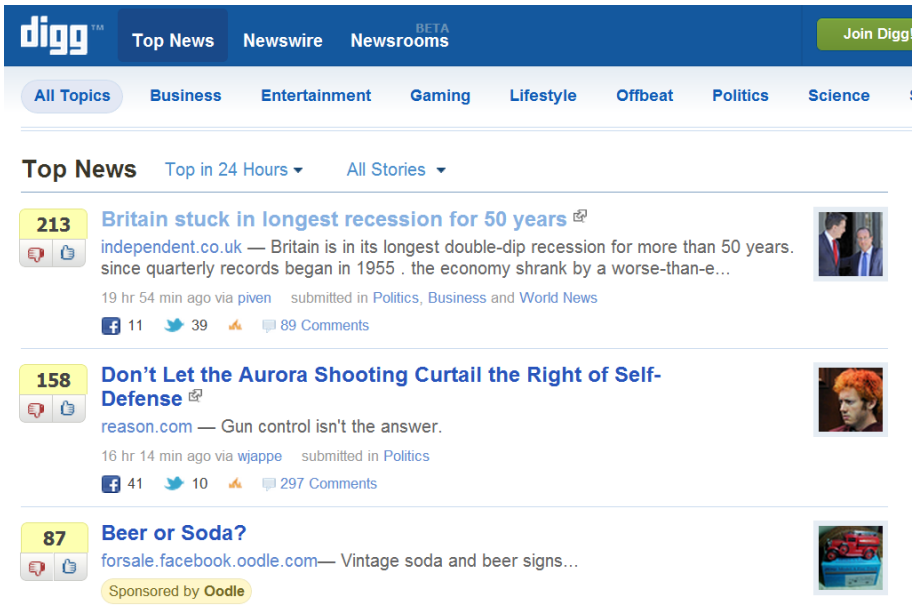


Figure 3.18: Social News Web Service, Digg

aim at querying the events from the social news web service Digg, as shown in Figure 3.18. Digg features user-posted news that are ranked based on the user popularity and comments. With the ranking, popular news can easily be found, leveraging from users' collective intelligence. Thanks to its public API, the news in Digg can be easily accessed. We take the time, location and topic keywords as the query parameters to retrieve popular events, ranked by the popularity. In the querying results, many useful metadata can be found for each event, such as "title","submit_date" to describe the property of events. We also extract the "link" metadata that refers the original content of the news, so that the original content could be retrieved and be used in the text description in the final result.

Though the events are extracted from Digg in this section, the whole framework is flexible and can be extended to handle more events directories, such as Google News, Last.FM easily.

3.3.1.3 Media Illustration

Years of multimedia research have witnessed that it is easier and more accurate for the computer to identify specific pattern compared with abstract concept. To find media illustrating events, a query is specifically tailored for each event. To illustrate events using text and images, we retrieve the original news content, related tweets from Twitter and images from Google image search, and show them in proper format.

The events are fully described in the original news web page. For each event, we extract the "link" metadata in Digg and crawl the original web page. To extract the main body of the web page, the method proposed in [Chengjie and Yi, 2004] is used ,which recognizes main content by character number on the assumption that characters accumulate more in the main content than in other part of web page.

For each event, we would like to collect other textual description and comments from different users. The textual data is retrieved from Twitter, the famous online microblog service that enables the users to send instance posts known as "tweets". As there are

hundreds of thousands tweets posted every minute in a city, tweets are also taken as the data source to mine events/trends [Weng and Lee, 2011]. In this section, we retrieve related tweets to illustrate events. The event title is often the most useful information for describing an event. We perform full text search queries on Twitter based on the event title extracted from Digg. The tweets retrieved are also filtered using a time interval from the time of posting the event in Digg to the end of querying period .

To provide a nice and meaningful visualization, we use a tag cloud to organize the textual data. Tag cloud is a form of histogram which can represent the amplitude of over a hundred items. In tag cloud, the importance of each tag is shown with different format. This format is useful for quickly perceiving the most prominent terms. For each event, we segment each tweet into tags, count the frequency of each tag and generate the tag cloud with tags in different font size and color.

Besides tag cloud, we also provide another vivid visualization called photo collage [Mei et al., 2009], to illustrate events. Photo collage is a visual clustering technique that can depict the event with different points of view, and it is well used to depict or summarize concept/object/event. To generate the photo collage, we retrieve photos from the Google image search engine with the event title as parameter and filter out those photos for which the cosine distance between its textual metadata and the event title is below a given threshold. We decided, after experimentation, not use Flickr as photo source because most of photos are taken during tours/vacations and the dominant concepts are scene, landmark, and building. With the selected photos, the method proposed in [Mei et al., 2009] is used to create the photo collage for each event.

In the end, to assist the users viewing the system well, we would like to show the events according to the importance. In details, we measure the importance of each event by the entropy of its tag cloud [Aouiche et al., 2009]. In details, Let $t \in T$ be a tag from a tag cloud T, the entropy of T can be calculated as

$$entropy(T) = \sum_{t \in T} p(t) log(p(t))$$

where $p(t) = \frac{weight(t)}{\sum_{x \in T} weight(x)}$. The tag cloud entropy quantifies the disparity of weights between tags. The lower the entropy, the more interesting the corresponding tag cloud is. After ranking, the events with more information will be shown first and events with less information shown later.

3.3.2 Results

Our system can be used to extract and illustrate events from a query input without limitation. For a given query, we firstly parse it to find out the topics, location, and time information. It should be noticed that for some query, one or more semantic is missing, and a default value will be taken. For example, in query " the Olympic in London", no time information could be found. The default value are "last week" for time, to ensure the timeliness of events, and "worldwide" for location, and "" for topic, to reduce the limitation of event querying. We then use the parsed topic, time, and location to query events from Digg. The top 20 events are kept as hot candidates according to their ranking promoted by the users.

For example, from the query "New York in the last 3 days", the following events in Table 3.4 are found.

Figure 3.19: An example of event visualization, to illustrate the event "Companies fire back at proposed NYC big soda ban", (A)Event title; (B)Event abstract, a link to the original news; (C)Tag cloud; (D)Photo collage; (E) Navigation

For all of the queried events, we provide an indexing interface as shown in Figure 3.15, that provides the thumbnails for all of the events with the combination of photo and event title. All of the events are ranked by their importance as described in Section 3.3.1.3. When navigating to a specific event, the event illustration as shown in Figure 3.19 will be provided. There are five parts to help people understand the event well.

The event title is shown in part A, which is the highest level description for the events.

To help the users understand the event content well, we also parse the original news web page, and mine the main textual content part, as shown in part B.

We parse the title and time metadata from the obtained events, use the parameters to query from Twitter the comments from different users, then use the tag clouds to show the results. The tag cloud is presented in part C. We can clearly see that the larger size content such as "back", "ban", "companies", "fire", matches the event topic "companies fire back at the big soda ban" very well.

Besides text visualization, we also query photos with Google image search, and filter the ones that can not be matched with textual metadata. All of the matched photos are collaged in the same layout. The photo collage is shown in part D. From the figure, it could be found that most of photos are relevant to the event, the ban of large-sized soft drinks at food service outlets.

To assist user navigating between events, the link to the previous/next event and the index page is located at the bottom of the webpage, as shown in part E.

To conclude, while tag cloud and photo collage in part C and D provide attractive interface with abstract content, the textual content in part B gives concrete description that would assist the users to understand the event well.

Table 3.4: Events found for query "New York in the last 3 days"

Date	Event Title
01/06/2012	Companies fire back at proposed NYC big soda ban
01/06/2012	Motorcyclist clocked at 193 mph? in the rain
01/06/2012	New York Bill Proposes Mandatory Wearable ID Tags @ skewnews.com
31/05/2012	Carmelo Anthony Becomes VP Of PowerCoco Energy Drinks
01/06/2012	Cheap Dresses Online Australia - On Sale From $10
02/06/2012	A Powerful Interview with Former Guant?namo Prisoner Lakhdar Boumediene
01/06/2012	NYPD vs. CPD: How Police Deal With Press and Protesters - New York - Slideshows
31/05/2012	Obama Campaign May Be Fooling Itself
03/06/2012	6 New York cops claim they were demoted for being Republicans
02/06/2012	Mets Pitch First No-Hitter in 50-Year History
31/05/2012	2013 Mercedes-Benz GL63 AMG Makes Debut
01/06/2012	Shadows Meet The Clouds, Gray On Gray, Like Dusty Charcoal On An Ashen Brow
31/05/2012	Scientific Proof That Men Have the Dirtiest Desks
03/06/2012	Space shuttle hardware is on the move in Houston and NYC
31/05/2012	Obama Presides Over Secret 'Kill List'
01/06/2012	Facebook Forced To Let You Vote On Privacy Changes
02/06/2012	Paul Krugman Pompously Insults Ron Paul And His Supporters Mediaite
02/06/2012	NBA Arrested For Marijuana Possession
31/05/2012	Bar Refaeli gets close to Olympic star

3.3.3 Discussion

In this section, we proposed an original framework leveraging on social media data (News, Media Sharing Platform and Microblog) to extract and to depict public events. The process is done without any human assistance on a given textual query. At first, natural language processing is employed to parse the input query. Then the social news web service is taken as the source to extract relevant events, finally the metadata from obtained events is extracted and used to query textual and visual content from different online sources (Twitter and Google). We present the results in an attractive visual format combining tag clouds and photo collage, which are generated online without any additional time cost, extra computation or storage requirements.

3.4 Conclusion

Organizing media data according to real-life events is attracting interest in the multimedia community. Event illustrating not only provides vivid explanation to event, so that people can understand event well, but also is a promising method to organize and index the huge

amount of data taken by different users. In this chapter, the problem of illustrating events with rich media data are well studied. Mainly, two problems are researched: the first problem is how to find the underlying media data for a given event. We study the user uploading behavior, and enrich the media data with query by event title, held place and taken time. We also propose a heuristics rule to prune noisy data by visual feature based analysis. The second problem is more general: how to discover events towards a given text query and illustrate them with media data. To solve the problem, we firstly parse the query with Natural Language Techniques to find out the time, location, and topic semantic, then we query social events from a social news website with the parsed results. Finally the multi-modality media data are obtained by event context queried from different source to illustrate social events. The novel proposals make it possible to combine diverse and complementary features such as time, location, text, visual feature while removing the noise of irrelevant samples hence vivid and high relevant illustrating results are provided.

Chapter 4

Events Discovery from Social Media

In this chapter, we discuss how to mine social events from media stream. Based on the context analysis on social media data, we propose two approaches to address the problem. At first, we consider much media data is uploaded when events occur, and propose a burst detection approach to target event. Secondly, event is regarded as relevant with the latent topics in human life, and we integrate topic model and make decision rule with validating data to identify events. In this chapter, events discovery from different perspectives are exploited.

4.1 Introduction

In Chapter 3, we have discussed the problem of how to leverage social media illustrating events. We have seen the role that events are playing in managing social media data. However, though many social events are scheduled on some event repositories, there are still lots of events out that are not recorded online.

In this chapter, we present our work on automatically discovering and identifying events from social media data. There will be two parts of work introduced in this chapter. The first part is event discovery from social media stream. Our approach is based on the observation that many photos and videos are taken and shared when events occur. We focus on detecting events from the spatial and temporal labeled social media, and propose our approaches to retrieve those media, perform event detection and identification, and finally enrich the detected events with visual summaries. We select 9 venues across the globe that demonstrate a significant activity according to the EventMedia dataset and we thoroughly evaluate our approach against an official ground truth obtained directly from the event venues' web sites. The results show our approaches can not only detect events with high accuracy but also mine and identify events that have not been published in popular event directories such as Last.fm, Eventful or Upcoming. In addition to the textual identification of events, we show how we can build visual summaries of past events providing viewers with a more compelling feeling of the event's atmosphere.

In the second part, we propose another solution to the event discovery problem. It is well known that there are lots of constant concepts in the real world which do not vary along time, and some of them are highly relevant to events, while some others are not. Hence events can be regarded as the specific distribution over these topics. We mine these topics with a topic model called *Latent Dirichlet Allocation*(LDA). In order to build our decision rule, we project validating data into the latent space and find out the event

boundary with the measure of KL divergence.

In this chapter, our work on social media detection in MediaEval 2011 will also be presented. This task aims at discovering events and detecting media items that are related to either a specific social event or an event-class of interest on a given dataset. We solve the problem leveraging the knowledge from cross web service: to retrieve the relevant events from popular repositories and then to match the events and media data according to their context.

4.2 Social Event Discovery

In this section, we address the problem of structuring social media activity into events and in particular how to automatically detect those events and their properties (location, time and participation). Event directories such as Last.fm, Eventful and Upcoming publish information about scheduled events in order to help users in tagging and structuring multimedia material and ease their future retrieval. While such services are becoming increasingly popular they are often incomplete, sometimes inaccurate in terms of the information that they provide. On the other hand, they largely overlap in terms of coverage, but they generally fail to give a good feeling of the atmosphere of an event, while this feature is considered as of primary importance to support users in deciding to attend or not an upcoming event [Fialho et al., 2010]. The problem we tackle is therefore how can we make use of metadata attached to media and events (tags, description, geographic location) to discover events by analysis the context of social media data.

4.2.1 Data Acquisition

Large numbers of web sites contain information about scheduled events, of which some may display media captured at these events. Using the public API offered by the web sites, several datasets containing events and media descriptions have been recently published [Troncy et al., 2010b]. We use part of the EventMedia dataset which is composed of more than 1,7 million photos explicitly associated to more than 110,000 events(Section 3.2.2). This dataset has been built from the overlap in metadata between four popular web sites, namely Flickr as a hosting web site for photos and Last.fm, Eventful and Upcoming as a documentation of past and upcoming events. Explicit relationships between scheduled events and photos are looked up using special machine tags such as `lastfm:event=XXX` or `upcoming:event=XXX`.

Descriptions of events are scraped from those three event directories using their API and represented according to the LODE ontology[1], which provides a minimal model that encapsulates the most useful properties for describing events [Shaw et al., 2009]. The goal of this ontology is to enable interpretable modeling of the "factual" aspects of events, where these can be characterized in terms of the *four Ws*: *What* happened, *Where* did it happen, *When* did it happen, and *Who* was involved. Furthermore, the W3C Ontology for Media Resource[2] is used to describe the media resources. This model provides properties for describing media properties such as the duration of a video, its target audience, copyright, genre, rating or the various renditions of a photo. Media fragments can also be defined in order to have a smaller granularity and attach keywords or formal annotations to parts of

[1] `http://linkedevents.org/ontology/`

[2] `http://www.w3.org/TR/mediaont-10/`

a media. This ontology contains a formal set of axioms defining mapping between different metadata formats for multimedia.

The EventMedia dataset has knowledge of more than 13,000 different venues for which at least one description of event explicitly associated to at least one photo is available. From this very large dataset, we selected 9 venues that proved to have a significant activity the last three years. Table 4.1 shows the number of events, photos and distinct users for those venues during this period according to the EventMedia dataset[3]. The ranking value corresponds to the popularity of the venue in the entire dataset when the sorting criteria is the number of distinct users that have uploaded at least one photo on Flickr taken during an event hosted at those venues.

Table 4.1: Number of events, photos and distinct users for 9 venues in the EventMedia dataset

Venue	**NbEvents**	**NbPhotos**	**NbUsers**	**rank**
Melkweg	352	6912	266	1
Koko	151	3546	155	5
HMV Forum	106	2650	130	8
111 Minna Gallery	24	1369	105	14
Hammersmith Apollo	79	2124	96	20
Circolo degli Artisti	784	2590	86	29
Circolo Magnolia	79	2190	76	40
AncienneBelgique	212	7831	56	83
Rotown	204	3623	49	101

These 9 venues are diverse in terms of the type of events they host and location. They generally host ranging from music concerts and festivals, to circus or exhibitions. They are mainly situated in Europe (UK, The Netherlands, Italy, Belgium) except the venue `111 Minna Gallery` which is located in San Francisco (USA). More importantly, for all these venues, it is possible to get ground truth of the events they host since they all publish an up-to-date program on their web sites. We have developed different scrapers of these web sites in order to get accurate ground truth of past events.

4.2.2 Detecting Events Based on Social Media Activity

Media sharing websites are regularly providing novel means for users to semantically enrich their photo and video collections. Geo-tagging adds location information to medias in the form of latitude and longitude coordinates. Such information is either captured by the device itself (when featuring GPS functionalities) or through user input (via textual input or location identification on a digital map). It is expected that when an event takes place, there will be many persons taking picture or videos and later uploading them on sharing platforms. Our solution starts from this point to address the event detection problem. As depicted in Figure 4.1, there are mainly 3 steps in the whole framework. At first, we collect data on a given location, and then use the time series analysis technique to find out the burst points. Finally, we summarize and show the found events with the same methods as proposed in Section 3.2.3.

[3]These numbers can be reproduced querying the public SPARQL endpoint exposed by EventMedia, `http://semantics.eurecom.fr:8080/sparql`.

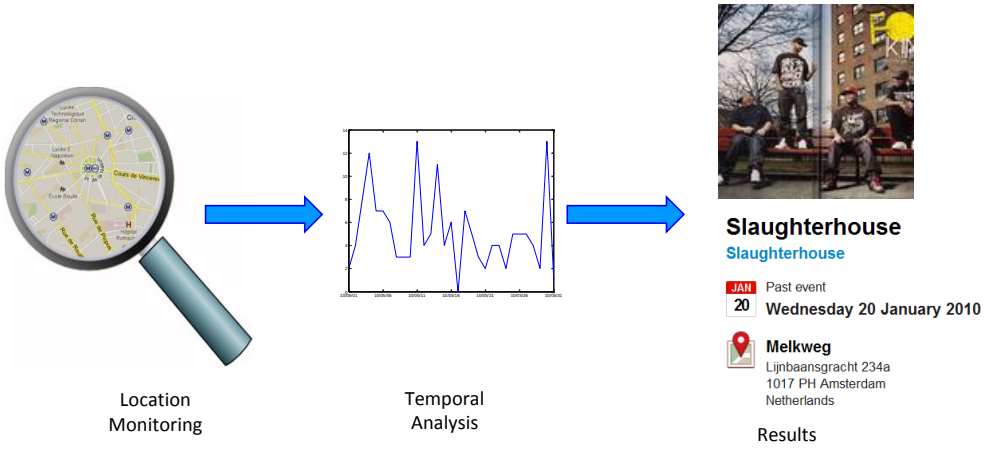


Figure 4.1: The proposed framework to discover events

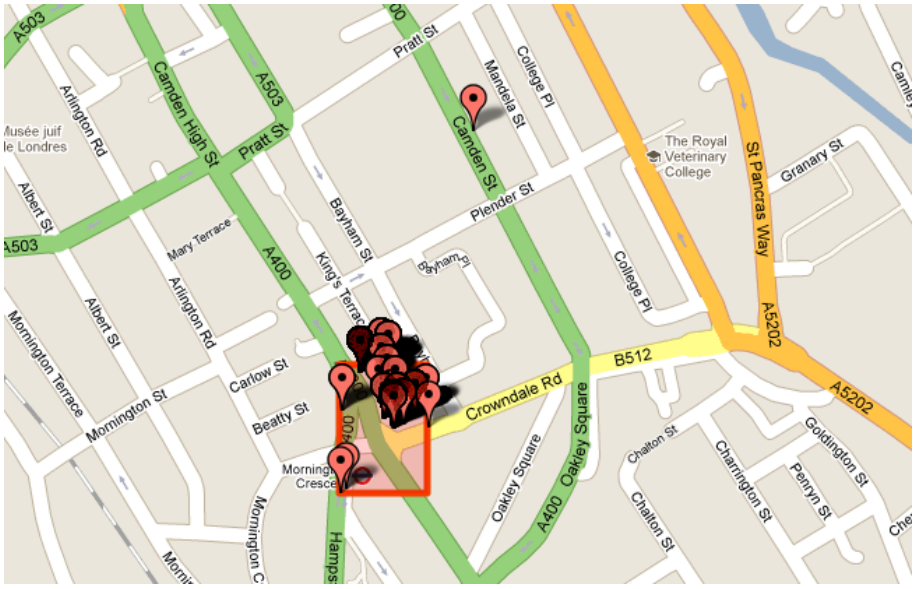


Figure 4.2: Bounding Box for the venue Koko in London (UK)

We use the approach presented in 3.2.3.1 to estimate the bounding box for the 9 venues selected (Table 4.1) and we crawl all photos taken at those locations during a one month period from May 1st 2010 to May 31st 2010. Hence, a collection of 4604 geo-tagged photos has been obtained. The number of geo-tagged photos hosted on Flickr represents still a tiny fraction of all photos shared on this web site. Many other relevant photos will not retrieved using a pure location-based search. In order to obtain more relevant photos, we use the venue name as the keywords and query photos without geo-tags during the same period. We extend the dataset with another 4589 relevant photos. To sum up, there are 9178 photos taken from the 9 venues in total(Table 4.2).

Table 4.2: Number of photos taken in the 9 selected venues during May 2010

Name	Geotagged Photos	Venue Name tagged Photos	Duplicate	Total
Koko	372	2040	3	2409
Rotown	90	273	1	362
Melkweg	363	700	8	1055
HMV Forum	184	412	0	596
111 Minna Gallery	937	3	0	940
Ancienne Belgique	2206	288	2	2492
Circolo degli Artisti	70	553	1	622
Circolo Magnolia	95	236	0	331
Hammersmith Apollo	287	84	0	371
All	4604	4589	15	9178

4.2.2.1 Ground Truth Collection

To evaluate the performance of our approach, we make two types of ground truth from the online data. The first type of ground truth is acquired from the three event repositories. The events records corresponding the 9 venues in Last.fm, Upcoming and Eventful during May 2010 are queried. It could be found that only Last.fm hosts events in May 2010, and no relevant events could be found in the other two repositories. In addition, we also taken the events records in the official agenda for each venues. We know that the records in the three event repositories are provided by the users and it is possible to miss valuable data. For the 9 well chosen venues, they all publish up-to-date programs on their web sites. For example Figure 4.3 shows the interface of venue Melkweg, and lots of events are scheduled there. These official agenda are well maintained hence a more accurate ground truth could be obtained there. To collect the ground truth, we develop specific script for each of the venues. In details, we use BeautifulSoup [4] to parse the HTML downloaded from these web services and extract the structured event information.

Table 4.3 reports the statistic of ground truth collected for the 9 venues, grouped by official agenda and the three event repositories. It could be found that for most the venues, more events are collected for the official agenda compared with the event repositories. For example, melweg, there are 69 events founded in its official web sites, while the number of events from the event directories are 45. However, a few events that are missing from the official web sites are accessible in the event repositories. Hence we can see the difficulty to collect the complete ground truth for a venue.

4.2.2.2 Analyzing the Flickr Activity around Venues

We are interested in detected events by monitoring the social media sharing activity at specific locations. The 9178 photos gathered by crawling Flickr during May 2010 using either the geographical bounding boxes or the venues name will be the basis for the event detection experiments. Table 4.2 shows the number of photos obtained for each location using both geo-tag based query and tag based query (where the tag is the venue name). Surprisingly, we notice that there are very few photos common to both queries. This

[4] `http://www.crummy.com/software/BeautifulSoup/`

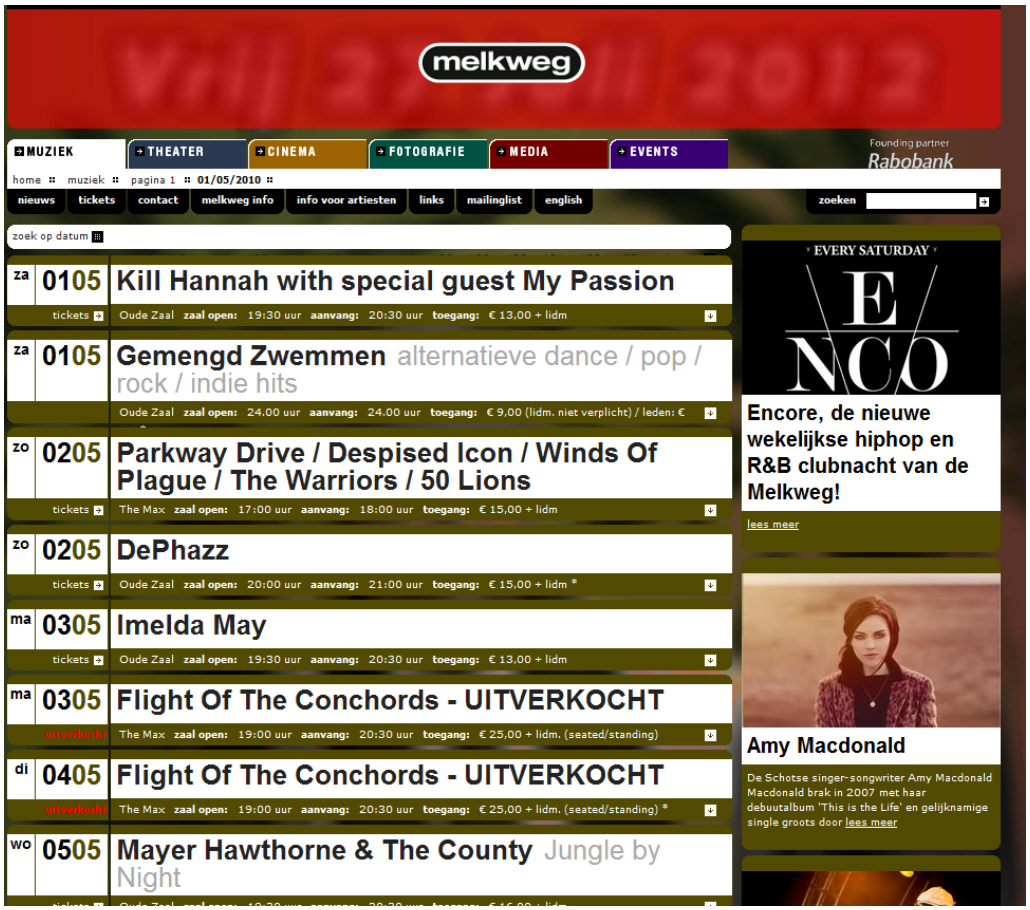


Figure 4.3: The inferface of melkweg web sites

Table 4.3: The ground truth for the 9 venues

Venues	official agenda	event repositories	Matched
melkweg	69	45	44
koko	20	0	0
HMV Forum	14	13	13
111 Minna Gallery	23	0	0
Ancienne Belgique	38	29	29
rotown	16	13	13
Circolo degli Artisti	22	18	11
Circolo Magnolia	25	15	13
HMV Hammersmith Apollo	15	17	13
Total	242	150	136

shows that users seldom label their photos with a location name when geo-coordinates are available.

We aim at mining events automatically based on photo upload activity at particular locations. Our objective in terms of events detection is to identify the date and title of the event given its venue or location. Figure 4.5 and 4.4 depict the Flickr uploading activity during May 2010 for two of the nine venues selected, Melkweg and Koko. Looking at both

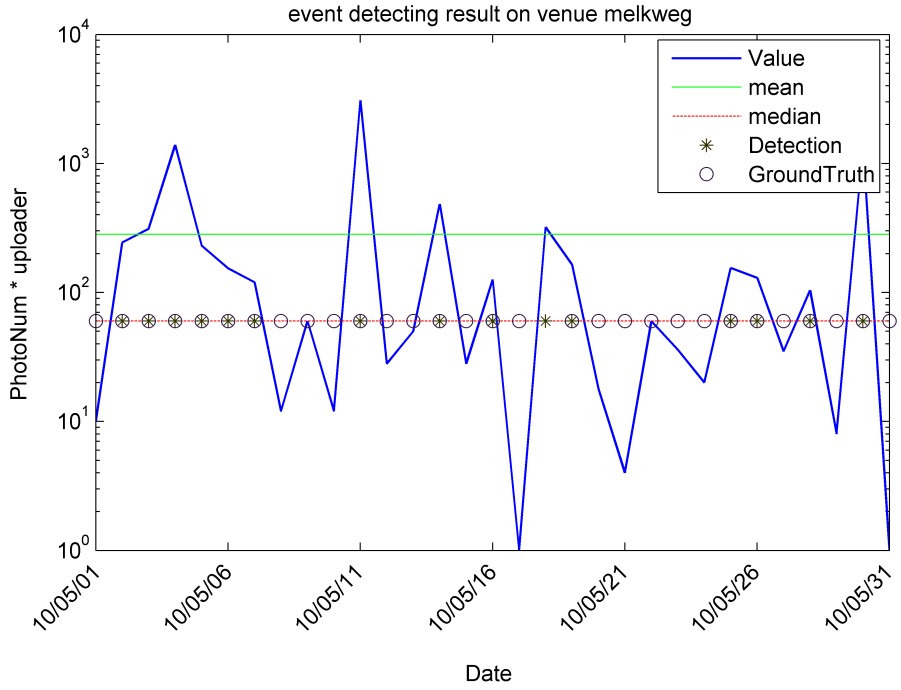


Figure 4.4: Event detection during May 2010 in the Melkweg (Amsterdam, NL)

curves, one can see that the number of pictures taken and uploaded varies temporally over the month. Our approach consist in carefully selecting the dates with high number of uploads and consider those candidate date as events. More formally, let us consider the time series $\{d_i, i \in [1, T]\}$ that represents the temporal evolution of the photo upload characteristic at a given venue v. The event e starting at time t is detected when the photo upload characteristic is greater than a given threshold.

$$e_t = \arg_i(t_i > THRESHOLD) \tag{4.1}$$

4.2.3 Detection Results and Evaluation

In this experiment, we firstly compare three photo upload characteristics. The first is the number of photo uploaded $d_i = ||p_i||$. The second attempts to incorporate the social dimension of the event and accounts for the number of different photo uploaders u_i. The third characteristic combines the previous ones by pondering the number of photo uploaded with the number of different person that uploaded those photos $d_i = ||p_i * u_i||$.

The detection method described in section 4.2.2.2 is run on the nine selected venues. The Figure 4.4 and 4.5 show results for 2 of these venues: Koko and Melkweg respectively. In those figures, the main blue curve shows the number of photos multiplied by the number of photo uploaders per day, the green curve corresponds to the mean while the red one is the median of the blue curve, the stars is the date for which events are detected (based on median thresholding). The ground truth (obtained as detailed in section 4.2.1) is shown using circles over the median. One can see that many events would be missed using the mean threshold while the median is able to capture many more events while keeping the false detection rate low.

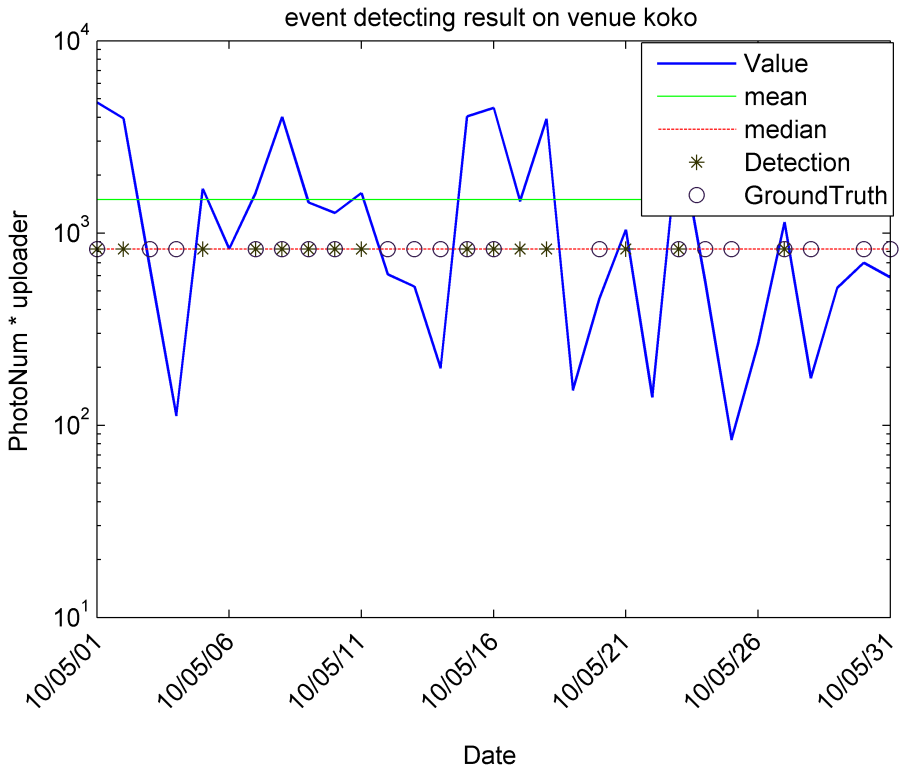


Figure 4.5: Event detection during May 2010 in the Koko (London, UK) venues

In May 2010, a total of 242 events are reported from the official schedules of the nine selected venues. In order to evaluate the performance and accuracy of our proposed approach, we need to align the detected events with official events. Once event start times are identified, we use the tags of the corresponding photos to deduce the topic of each event. All words from tags and titles of photos taken on that day are parsed and sorted by their occurrence. The top 15 keywords are kept to infer the topic of individual events. Detected events are manually matched with ground truth events based on date and title. Matching events are those sharing a common starting date and for which at least one non stop-word can be found in both the top 15 keywords list of detected events and the title of the ground truth events.

Beside the ground truth from official web site, we also compare the detection results with other event directories, such as Eventful, Upcoming and Last.fm. Last.fm identifies 150 events which took place in May 2010 in our 9 venues, while Eventful and Upcoming did not return any events. Of the 150 Last.fm events, 136 can be matched to the ground truth data, based on our starting date and title matching rule. When matching the event in Last.fm to our detection result, we identify 17 events that are not present in Last.fm but are present in the ground truth. In other words, using our approach, it is possible to extend of more than 10% the Last.fm event directory with new events.

Table 4.4 shows the positive detection on the Melkweg and Koko venues. The detection time and tags are reported in this table. Due to limited width, only relevant tags are shown. For each event, the ground truth and the Last.fm information (when applicable) are also given.

A first look upon Figure 4.5 and 4.4 seemed to favor the use of the median over the

Table 4.4: Event Detection Results on Melkweg and Koko

Venue	Detection Results		Ground Truth		LastFM	
	Date	Tags	Date	Title	LastFM	Title
melkweg	03/05/2010	parkwaydrive drive parkway	03/05/2010	Parkway Drive / Despised Icon / Winds Of Plague / The Warriors / 50 Lions	1336473	Parkway Drive
melkweg	02/05/2010	flight conchords flightoftheconchords	02/05/2010	Flight Of The Conchords - UITVERKOCHT	1439320	Flight of the Conchords
melkweg	04/05/2010	flightoftheconchords	04/05/2010	Flight Of The Conchords - UITVERKOCHT	1439407	Flight of the Conchords
melkweg	05/05/2010	mayerhawtorne mayer hawthorne	05/05/2010	Mayer Hawthorne & The County	1416229	Mayer Hawthorne The County
melkweg	11/05/2010	bonobo	11/05/2010	Bonobo - UITVERKOCHT	1398102	Bonobo
melkweg	14/05/2010	paulweller paul	14/05/2010	Paul Weller - UITVERKOCHT	1406677	Paul Weller
melkweg	18/05/2010	brokensocialscene	18/05/2010	Broken Social Scene - UITVERKOCHT	1334429	Broken Social Scene
melkweg	19/05/2010	mikestern richardbona	19/05/2010	Mike Stern band with special guest Richard Bona featuring Dave Weckl & Bob Malach		
melkweg	25/05/2010	beattimemelkweg	24/05/2010	Beattime - The Kika Edition		
melkweg	26/05/2010	beattime	24/05/2010	Beattime - The Kika Edition		
melkweg	28/05/2010	offcentre	28/05/2010	Off Centre - day 3 - night met Kode 9 / Falty DL / Gold Panda / Kelpe		
melkweg	30/05/2010	joannanewsom	30/05/2010	Joanna Newsom	1425481	Joanna Newsom
Koko	02/05/2010	camdencrawl crawl	02/05/2010	The camden crawl		
Koko	05/05/2010	mrhudson	05/05/2010	Mr Hudson		
Koko	08/05/2010	shehim	07/05/2010	She HIM		
Koko	07/05/2010	shehim sheandhim	07/05/2010	She HIM		
Koko	11/05/2010	houses kids glass	11/05/2010	kids in glass houses		
Koko	18/05/2010	rita maria	18/05/2010	maria rita		
Koko	23/05/2010	metric	23/05/2010	metric		
Koko	27/05/2010	yeasayer	26/05/2010	yeasayer		

mean as threshold condition for detecting events. We have identified three uploading characteristics upon which detection could be performed: number of image, number of uploaders and number of image times number of uploaders. Table 4.5 gives the results for these 6 conditions. It can be seen that when the median is used for threshold, more positive events are detected compared with the ones using the mean. Unfortunately, the number of false detection also increases. Of the three upload characteristics, the combined "Image*Owner" achieves the best overall performance with an F1-Measure of **0.289** with a median based threshold. Based on these results, we are using the median threshold approach on the "Image*Owner" upload characteristic for event detection.

The choice of the threshold is an important factor of the events detection performance our approach can achieve. The venues hosting events have very different size and popularity. It is therefore unlikely a common threshold would provide a good and generic performance. In experiments, we also study another heuristic threshold to detect events, that is to learn a threshold from past events. For each venues, we query the past events(before May 2010), and then for each of the event, we query photos with geo and text methods as proposed in data collecting section. In the collected photos dataset, we make the statistics about the number of owners and uploading for each events, and also employ the median value as the threshold.

Figure 4.6 shows the statistics results on Melkweg. It could be found that there are about 50 events reported in this figure, while the range of monitored parameter gives a very broad of the variability, from very few to about 1800. The similar conclusion could also be found in some another venues, such as Rotown shown in Figure 4.7. These results also suggest the difficulty of solving the problems with single threshold based approach.

We use the same "median value" strategy, which is found as the best method to select static threshold, to decide the threshold and use it to discovery events, and the result is reported in the last line of Table 4.5. It could be found that with the heuristic threshold, 69 events are found in total, which leads to the best performance under the evaluation

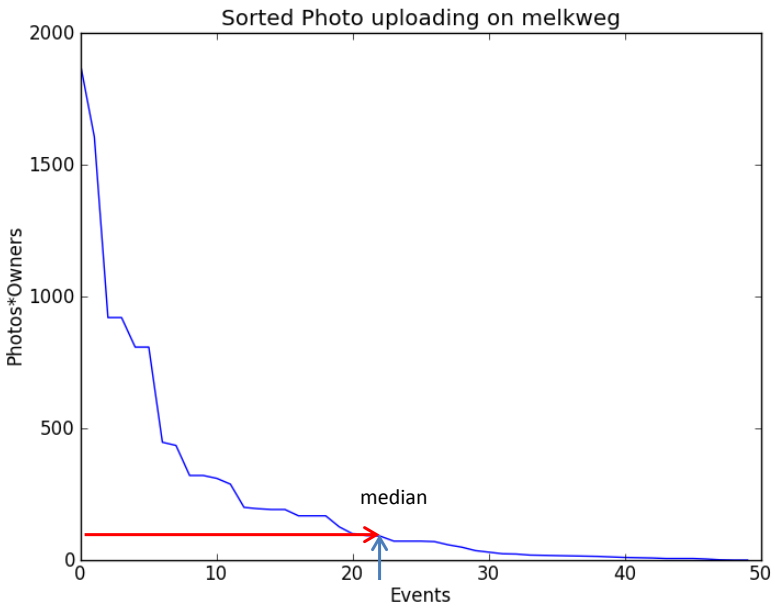


Figure 4.6: The photos taken in the past events on Melkweg

of Recall. However, with such threshold, the Precision and F1 measure degrades a lot compared with the threshold "median", since to much noise is brought meanwhile. Finally, we choose the statical threshold "median" to discovery the social events from media stream.

The details of the detection results for each venues are shown on Table 4.6. In this table, the Recall column requires further discussion. Since the detection of events is based on photo upload characteristics, any events for which no or too few photos have been uploaded on Flickr cannot be detected. The recall values reported are in fact currently floored and should be understood as such. As the current uploading trends continues to evolve, we expect the number of "detectable" events using our proposed approach to grow accordingly.

After careful examination, we find that most false matches occur in following two scenarios:

(1) the detection is technically correct, but the ground truth is missing. In previous section, we have described that the ground truth comes from the official scheduled events. If there are some events which happened indeed but are not recorded in their websites, a false positive error will raise. An example of this scenario is illustrated in Figure 4.8, where a great uploading phenomena could be seen on 2010/05/06, and our approach detected an event happened on that day, but no event is found in ground truth. However, when digging into the media data, lots of photos could be found that provide strong clue of happening. One of them are depicted in Figure 4.9. Both of the photo visualization(a guitar show) and the tag distribution (such as "live,concert,guitar") imply an concert held there, which is discovered by our approach, but missing in the ground truth.

(2)the detection is false, due to the huge amount of noise images uploading, especially when they are uploaded by a few users. In our method, an event is reported when the number of uploading is above on the given threshold, but it could be affected by the noise data while there is no events taken on that date in the venue. This is actually is a problem of

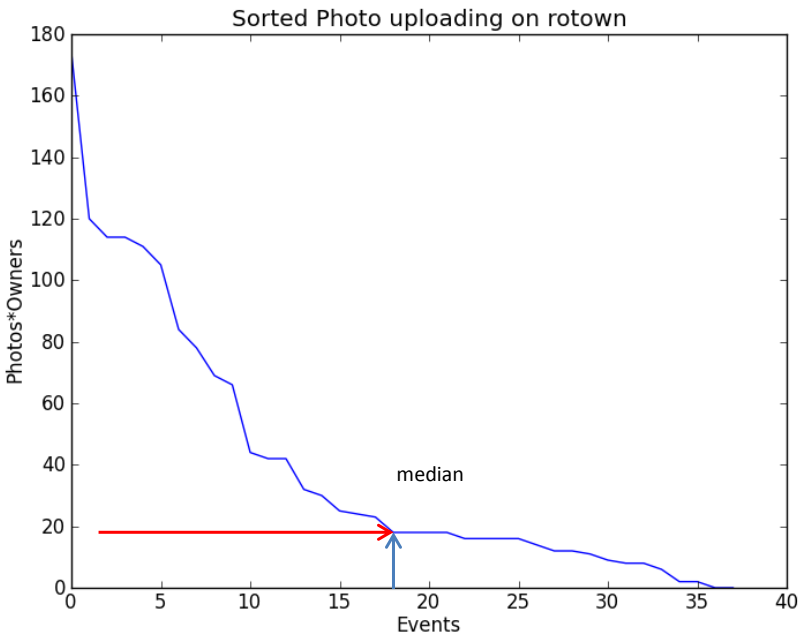


Figure 4.7: The photos taken in the past events on Rotown

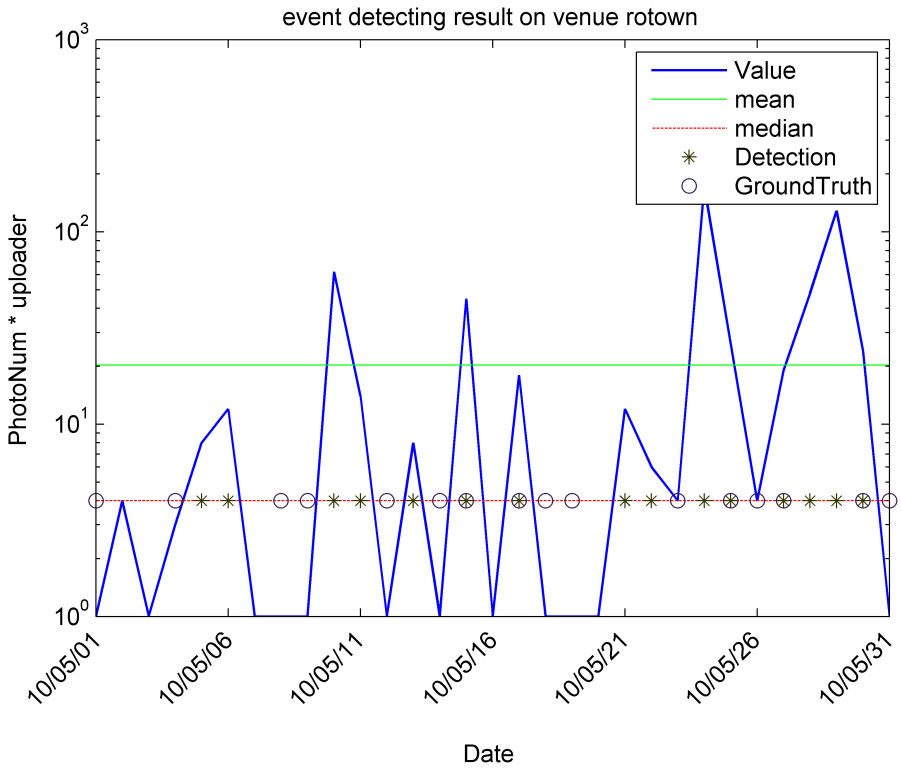


Figure 4.8: False Positive in venue "rotown"

<photo datetaken="2010-05-06 20:13:37" owner="42501630@N06" tags="blue brussels azul concert belgique guitar mark live stage escenario concierto guitarra bruxelles bleu singer guitare unplugged ancienne cantante chanteur acustico scène acoustique lanegan" title="Mark Lanegan" url_m="http://farm5.static.flickr.com/4136/4814482293_c5ce559ef6.jpg"/>

Figure 4.9: A photo taken in rotown on May 6th

Figure 4.10: False positive samples in HMV

noisy images and detecting mechanism. For example, on 2010/05/08, 77 photos were taken by a Flickr user(id = "27978006@N04") near the venue "Hammer Smith Apollo". However, these photos, as shown in Figure 4.10, capture many different views in the street, and no events could be found. Ideally, a more robust event discovery mechanism is demanded.

Besides the wrong decision made by the system, there are some reasons that lead to False Negative ratio. At first, no event can be found without the media source, and it is hard for the burst detecting approach to discover events if no photos are taken during an event and shared online thereafter. This is the major reason that leads to poor recall ratio in our approach. We forecast that the issue could be solved with further development of social media, and when more media could be captured in the future events, enough media source can be used to discover events. In addition, our event discovery algorithm works on time series with discrimination. For some popular venues, events are taken with high frequency and lots of photos are take along date every day. And missing detection may happen if there is no discrimination among the data, which demands the analysis techniques that dive into document content.

We have presented an approach to detect events based on photo upload characteristics

Table 4.5: Event detection on different conditions

Source	Threshold	True Predict	False Predict	F1
Image	mean	43	21	0.281
	median	64	51	0.359
Owner	mean	56	56	0.316
	median	58	62	0.320
Image*Owner	mean	34	**18**	0.231
	median	**67**	53	**0.370**
	hieratical	**69**	90	0.327

Table 4.6: Event detection results for the 9 selected venues

Venues	Ground Truth	Detection	Matched	Precision	Recall
Melkweg	69	15	12	0.800	0.174
Koko	20	15	8	0.533	0.400
HMV Forum	14	12	9	0.750	0.643
111 Minna Gallery	23	15	2	0.133	0.087
Ancienne Belgique	38	15	9	0.600	0.237
Rotown	16	15	8	0.500	0.533
Circolo degli Artisti	22	15	8	0.533	0.364
Circolo Magnolia	25	3	1	0.333	0.040
Hammersmith Apollo	15	15	10	0.667	0.667
Total	**242**	**120**	**67**	**0.558**	**0.277**

at a given location. Event directories fail to provide viewers a good feeling about the atmosphere of events [Troncy et al., 2010a]. Showing related photos and videos about a past event is a natural way to provide more insight to viewers. In order to bring more diversity to the final illustration results, we use the approach presented in Section 3.2 to automatically select illustrations for detected events.

For each of the 67 events detected using our proposed approach (see section 4.2.3), we query Flickr for photos containing event specific words w_e and uploaded within a five day window after the event. This results in a potential enrichment for 23 events (Table 4.7). This table reports four different information about each individual event. The first part *Detection Dataset* corresponds to the data used for detecting events and from which a subset of relevant photos will be used for creating the visual model later used for pruning. The second part *Enriching Source* reports the number of newly retrieved images which will be considered for enriching the event illustration set. The third part shows the visual pruning results. Finally, the last part gives the total number of photos that illustrate the event after enrichment and visual pruning.

Let us take the first event as an example. For this event, there are 101 photos taken on that day, and 66 of them are relevant to the event (based on textual match). 32 photos are obtained on Flickr through the enrichment process. A manual visual check identified 31 of those images to be relevant to the associated event. After automatic visual pruning 20 photos are left, and all of them are relevant photos, indicating that the pruning process has discarded the non relevant photo from the enrichment set. There is a final total of 86

(66+20) photos obtained to illustrate this events. A visual representation of this event is depicted in Figure 4.11.

In summary, for the 23 events, there are 1678 photos used for event detection, out of which 1138 are thought to be sufficiently relevant based on tag similarity to be used to model the events visually. The enriching Flickr search retrieved 632 new photos. 464 of those photos are relevant to event according to manual visual check. The automated visual pruning algorithm identifies 584 relevant photos, from which 417 are judged relevant through human inspection. The proposed approach is therefore able to populate illustrative sets of media representing particular events with a precision of 0.714 and recall rate 0.898. In addition, the EventIDs featuring a (*) are events for which no entry can be found in Last.fm but which have been detected by our algorithm.

Table 4.7: Illustrating events with photos

EventID	**VenueID**	**Prediction DataSet**		**Enriching Source**		**Pruning Results**			Final Number
		Photos	Related	Photos	Related	Photos	Recall	Precision	
(*)1	1	101	66	32	31	20	0.645	1	97
2	9	44	43	48	48	48	1	1	91
3	1	116	97	5	5	3	0.6	1	102
(*)4	2	161	107	21	20	20	1	1	127
5	2	61	17	1	0	1	NULL	0	17
6	6	7	1	1	1	1	1	1	2
7	1	238	201	4	0	2	NULL	0	201
(*)8	2	95	44	9	9	9	1	1	53
9	9	17	17	24	24	24	1	1	41
10	6	4	3	5	5	5	1	1	8
11	3	22	20	4	4	4	1	1	24
12	3	3	2	5	5	4	0.8	1	7
13	9	173	161	11	11	11	1	1	172
14	5	70	37	5	5	4	0.8	1	42
15	5	66	11	249	88	220	0.636	0.255	99
16	9	9	2	1	1	1	1	1	3
(*)17	2	135	53	14	14	14	1	1	67
18	6	43	41	2	2	2	1	1	43
(*)19	2	95	78	4	4	4	1	1	82
20	3	50	29	67	67	67	1	1	96
(*)21	5	53	34	10	10	10	1	1	44
22	5	52	11	54	54	54	1	1	65
23	3	63	63	56	56	56	1	1	119

To provide the vivid interface, a visual cluster is generated for each events. From the final photos set identified in the previous section, a visual cluster is created following the method proposed in [Wang et al., 2007]. Figure 4.11 shows the visual cluster result for the first event in Table 4.7.

During the enrichment phase, we expect to bring more diverse photos into the collection. For example, the Figure 4.12 depicts the event 2, which is held in `Hammersmith Apollo` with the title `iggy stooges`. Figure 4.12(A) is generated from the relevant photos from the detection set that corresponds to photos that are either geo-tagged or tagged with a venue name. Figure 4.12(B) shows the collection of images resulting from our enriching and visual pruning method. We can clearly see the increased visual diversity of the scenes between the two sets. The final set of images illustrating the `iggy stooges` event will be composed of both sets.

Figure 4.11: Visual cluster for the event 1, which was held on 07/05/2010, in the venue Koko with the title She&Him

4.3 Topic Based Event Detection

In this section, we propose another method for discovering specific events from social media data. It is well known that there are lots of concepts in the real world, and some of them are high relevant to events, while some others are not. Here we take the events as special distribution over these topics, and discover events by their distribution on these topics. As shown in Figure 4.13, first the topics are learned from large quantities of data captured at a given location. Then, we use the least mean square algorithm to estimate the events distribution on a group of validated data samples. We detect the events, from a test dataset, if they fit the distribution over latent topics well. Importantly, unlike the some previous work on the event classification, the object of this section is to discover specific events such as Lady Gaga concert, the wedding of Prince William, etc. To the best of our knowledge, this is the first attempt to discover specific events from latent topics point of view.

4.3.1 Event Detection

The goal of the proposed work is to detect events from social media data for a given location automatically. To solve the problem, we consider inferring latent topics existing in localised media data. These topics are stationary with respect to a location and can be learned from a large scale dataset. Documents originating from an event have a specific distribution over these topics, and we would like to infer the distribution to detect events. Our framework consists of two steps: (i) the distribution of latent topics for the given social media documents, which are solved by the LDA model. (ii) Estimating the distribution of events over the topics, which could be solved by the least mean square optimization on the KL divergence measure. All of the details will be given in the following section.

4.3.1.1 Topics Learning

For a given place (a city for example), the set of topics associated with a period of time can be seen as stationary. The semantic of events can be regarded as special distribution over

(A) Before enrichment

(B) After enrichment & Pruning

Figure 4.12: Visual cluster for the event 2, which was held on 03/05/2010, in the venue Hammersmith Apollo with the title iggy stooges

those topics. There have already been lots of approaches to infer topics among documents. The method used in this paper is the LDA model [Blei et al., 2003] which is a generative probabilistic graphical model to discover topics in documents. In practice, we collect the geo-tagged Flickr photos for a given city (or location), and choose stem words from the title and tags from each photo as representation of the social media to train the LDA model. Here, we should argue that the textual feature is not the only feature which could be used, other representations (Bag of visual words[Jiang, Yu-Gang And Ngo, Chong-Wah And Yang, 2007], for example) also fit our framework. When the models are obtained, they are used to infer distribution over these topics on the validated data and to estimate

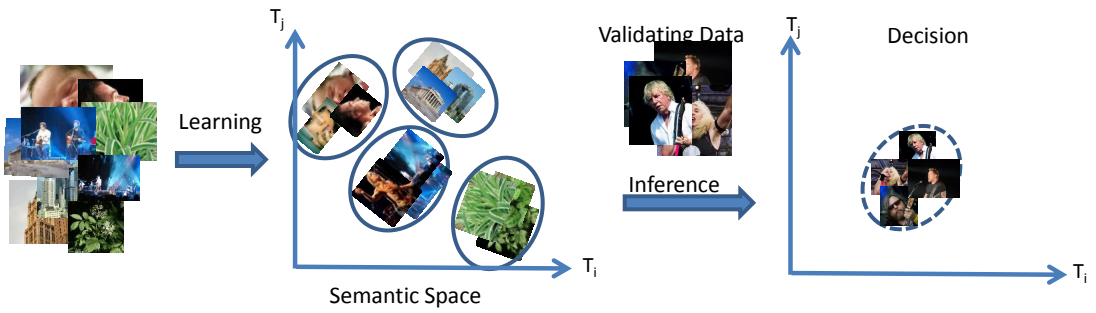


Figure 4.13: The proposed framework

the distribution of event, which will be detailed in the following section.

4.3.1.2 Events Estimation

From the LDA model(Section 2.2.2), the distribution of a document d over latent topics can be inferred by equation 4.2.

$$P(\theta_d|\alpha,\beta,d) = \int\int p(w,z,\theta|\alpha,\beta)dwdz \tag{4.2}$$

While w is the words in documents, z is the topic for words in a document, and θ is the distribution for topics.

To infer the distribution of event relevant documents, there is still a binary classification problem to solve: assign a social media document with an event versus no event. Here, we estimate the distribution of events over the inferred topics using validation data, which are the positive media documents of an event. The details concerning the validation data acquisition will be introduced in the experiment section.

Suppose D is the inference of the validation data over the latent topics, the event distribution $\mathbf{e}$ can be estimated by least mean square optimization theory. The objective is to minimize equation 4.3.

$$\mathbf{e} = \underset{e\in R^N}{\operatorname{argmin}} \sum_i Dist(D_i, e) \tag{4.3}$$

Where the function $Dist$ measures the distance between a validating instance D_i and the events estimation $\mathbf{e}$. It is well known that the best measure between two distributions is Kullback Leibler divergence. However KL divergence is not a symmetric measure. We use the following standard symmetric version as the distance measure

$$\begin{aligned} Dist(p,q) &= D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p) \\ &= \sum_i p(i)\log\frac{p(i)}{q(i)} + \sum_i q(i)\log\frac{q(i)}{p(i)} \end{aligned} \tag{4.4}$$

When the event distribution is estimated from equation 4.3, it can be used to verify if a new document d is event related or not, according to the rule defined in equation 4.5.

$$d \;\; is \begin{cases} event, & \text{if } Dist(d,\mathbf{e}) \le T \\ noevent, & otherwise \end{cases} \tag{4.5}$$

Where the value T is the threshold of the decision function, which is used to decide if a document is relevant or not in the detection process. In practice, the value of T can be inferred from the validation dataset D as follows:

$$T = k\max_i\{Dist(D_i, \mathbf{e})\} \tag{4.6}$$

where k is used to suppress the influence of noise contained in the validation dataset. It will be studied in the Experiment section 4.3.2.

4.3.2 Experiments and Results

To validate the proposed approach, we collected a large dataset from the Internet. The events we aim to detect in this study are concerts. There are two reasons for this choice. On one hand, concerts are popular and important social activities of our life and there are significant amounts of photos taken during concerts and shared on Flickr. On the other hand, it is easy to generate the ground truth on convert events, thanks to event directories such as LastFM, Upcoming or directly from the venue's agenda. In this section, we concentrate on 7 venues that are located in 6 cities around the world during May 2010 for concert event detection. Other types of events could also be detected with our approach given the appropriate dataset / ground truth combination.

First, we introduce the data collection used in the experiments, and then show a walk through for how our approach works on event detection. The basic idea of our approach is to learn the concept on large scale data, and find event distribution on these concepts. So we need a dataset to train the LDA model, and a validation dataset to estimate the events, we also need a testing to evaluate the performance. Most of the dataset is reused from previous section, which is described as follows.
(1) **Training Data** The training photos set is crawled from Flickr based on its public API. We have chosen 6 cities, which are located in Europe and America. The geographic information of these cities (such as the cities geo-coordinates and size) is obtained from Wikipedia. The cities shape is approximated to a circle for simplicity reasons. Although such assumption is reasonable since previous research has studied the distribution rule of social media and shown that most photos are taken in the center of cities [Hollenstein and Purves, 2010]. Using geo-coordinates based queries, we gather a collection containing about 49K photos during the month of May 2010 in these 6 cities.
(2) **Validation Data** The validation set includes the photos taken during events which have been held in a target venues in the past. They are used to estimate the event distribution as described in section 4.3.1.2. The validation data is collected by Flickr API with event machine tags, as proposed in 4.2.1.
(3) **Test Data** The test data includes the social documents that we aim to mine events from. This dataset is collected using the Flickr API with queries combining location (venue coordinate) or text (event title) and time (May 2010), as in 4.2.1.

As we use the bag of textual words to represent the media documents, some preprocessing is performed before learning the models. At first, we remove all of the photos without any textual words from media data. Then we filter the stopwords for each documents. We also filter the words that refer to the corresponding location.For example, for the data collected in Amsterdam, we remove the words such as "Amsterdam","Netherland". In the precessing, lots of documents are removed. A summary of the photos on the three dataset can be found in table 4.8. Since the training set is collected at the city level, the venues "Koko" and "HMV Forum" share the same training set but different validation set.

Table 4.8: Photos Collections over the Venues.

Venue	City	Training Set	Validating Set	Testing Set
Melkweg	Amsterdam	3786	179	355
Koko	London	23384	194	724
111 Minna Gallery	Chicago	11725	175	313
Ancienne Belgique	Brussel	2120	321	496
Rotown	Rotterdam	1575	71	118
Circolo degli Artisti	Rome	6551	107	167
HMV Forum	London	23384	189	97

Besides the dataset, we also create the ground truth on these venues for May 2010, based on the events that are listed in the agenda of the venues' website, as we did in Section 4.2.2.1. However, it is important to mention that not all of events are represented in social media data; It is likely that some for some events no photos were captured or shared on Flickr. And our objective is to find out as many events as possible. The number of ground truth events for each venue is reported in table 4.10.

4.3.3 Results

To evaluate the proposed approach, we perform the LDA model on the training data and employ the trained models to infer the topics distribution on both validation data and test data. Then, the decision rule can be learned after the inference process on the validation data. In the events distribution estimation, the ratio k in equation 4.6 plays an important role in balancing the recall and precision rate. Obviously, lower k value will lead to high recall but lower precision, and vice versa.

To obtain the optimized value of k, we calculate the ratio

$$k = \frac{Dist(D_i, e)}{\max(Dist(D_i, e))}, D_i \in D$$

Figure 4.14 illustrates the effect of k on the validation data. From the figure, it is clear that the number of photos decreases as k increases and a noticeable drop occurs for $k \geq 0{,}3$. Based on this result, we choose $k = 0.3$ in equation 4.6.

Then, we choose the venues "melkweg" to explain how to discover the event by the proposed approach. At first, we use the data collected from Amsterdam in the target month to train the LDA model. The Figure 4.15 depicts the most important topics learned in Amsterdam. From the keywords, the topics can be recognized as "transport", "park", "building", "brige", "concert", and so on.

With the LDA model, we also infer the distribution of each documents in validating and testing set over these topics. With the validating set, the decision rule could be made as defined in 4.6. In the end, the rule is employed on the testing data to find out the event relevant media documents. Figure 4.16 shows some examples of the decision results. The photos in the red box(the top part of figure) are taken as irrelevant to events while the photos in the blue box(the bottom part of the figure) are taken as relevant to events. The

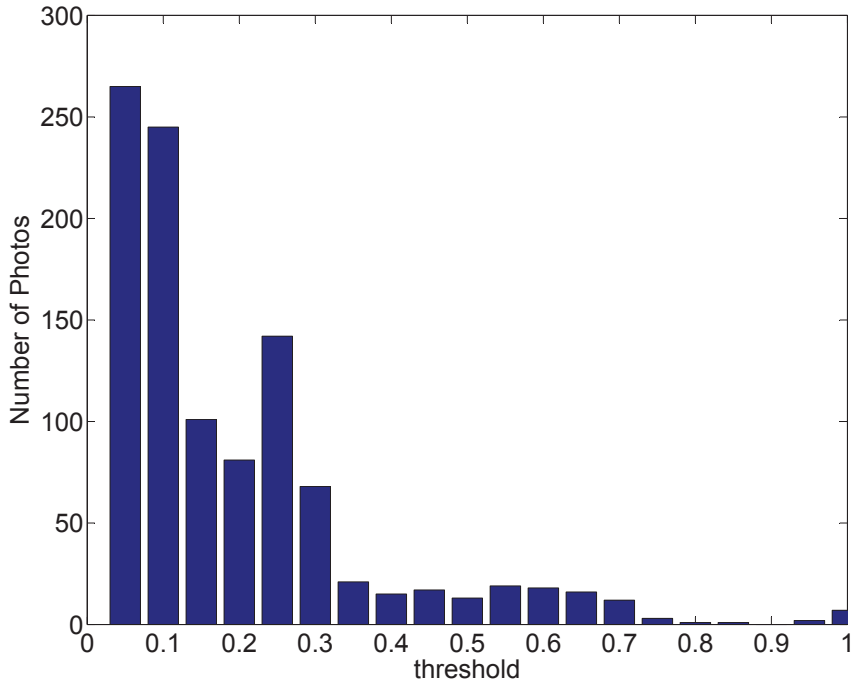


Figure 4.14: The histogram of threshold value

final event detection process is performed on the test data and the results are manually and individually checked, based on the matching between the textual descriptions of images and ground truth event. Since we detect events at the media document level, more than one document is inferred to the same event. Therefore, we evaluate precision at the documents level and the recall at events level respectively. Table 4.9 report the statistics of media data on event detection for the 7 venues. In this table, the number of documents that are detected as event-related is represented. In total, 265 out of the 2270 photos are identified as event-related and 160 photos (out of the 265) are assigned to the right event, leading to an average precision of 0.60.

Table 4.9: Social Media Data statistics over Event Detection

Venue	**Total Number**	**Detection**	**Positive**	**Precision**
Melkweg	355	42	32	0.76
Koko	724	95	44	0.46
111 Minna Gallery	313	26	10	0.38
Ancienne Belgique	496	32	19	0.59
Rotown	118	6	4	0.67
Circolo degli Artisti	167	46	36	0.78
HMV Forum	97	18	15	0.83
Total	2270	265	160	0.60

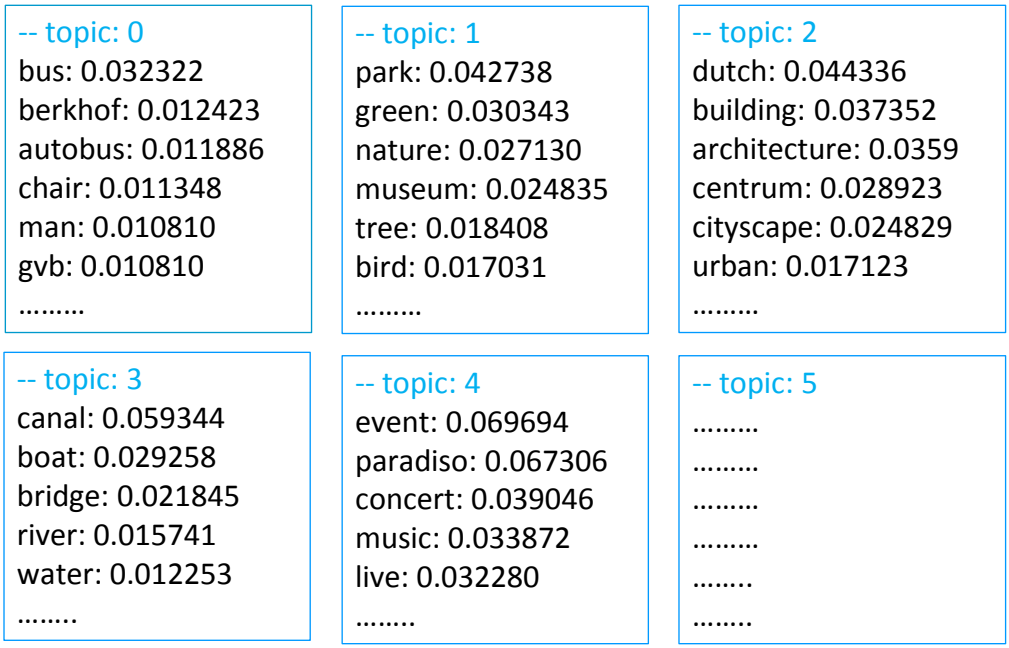


Figure 4.15: The LDA topics learned in Amsterdam

Table 4.10 reports the performance of our event detection approach in terms of recall. In total, out of the 203 events available in the dataset (according the meta data), 63 of them are detected by our approach. This corresponds to a recall of 0.31.

Table 4.10: Social Event Detection Performance

Venue	**GroundTruth**	**Detection**	**Recall**
Melkweg	69	14	0.20
Koko	21	12	0.57
111 Minna Gallery	23	4	0.17
Ancienne Belgique	38	10	0.26
Rotown	16	2	0.13
Circolo degli Artisti	22	15	0.68
HMV Forum	14	6	0.43
Total	203	63	0.31

In addition, our proposed approach is robust when handling the semantic on social data. In our selected venues, "KoKo" and "HMV Forum" are located in the same city "London". The results on these two places are obtained from the same topics model, which is trained on the photo documents captured all over London. Nonetheless, acceptable results are achieved on the two places, as shown in Table 4.9 and 4.10. Those findings strongly support the assumption that event semantics can be taken as special distribution over latent topics which could be learned from the media collection of an entire city.

We also compare the topic based methods with the one proposed in Section 4.2.2. The result is reported in Figure 4.11. It is found that compared with the threshold based approach which finds 52 events in these venues, more 12 events could be discovered by the methods stated in this section.

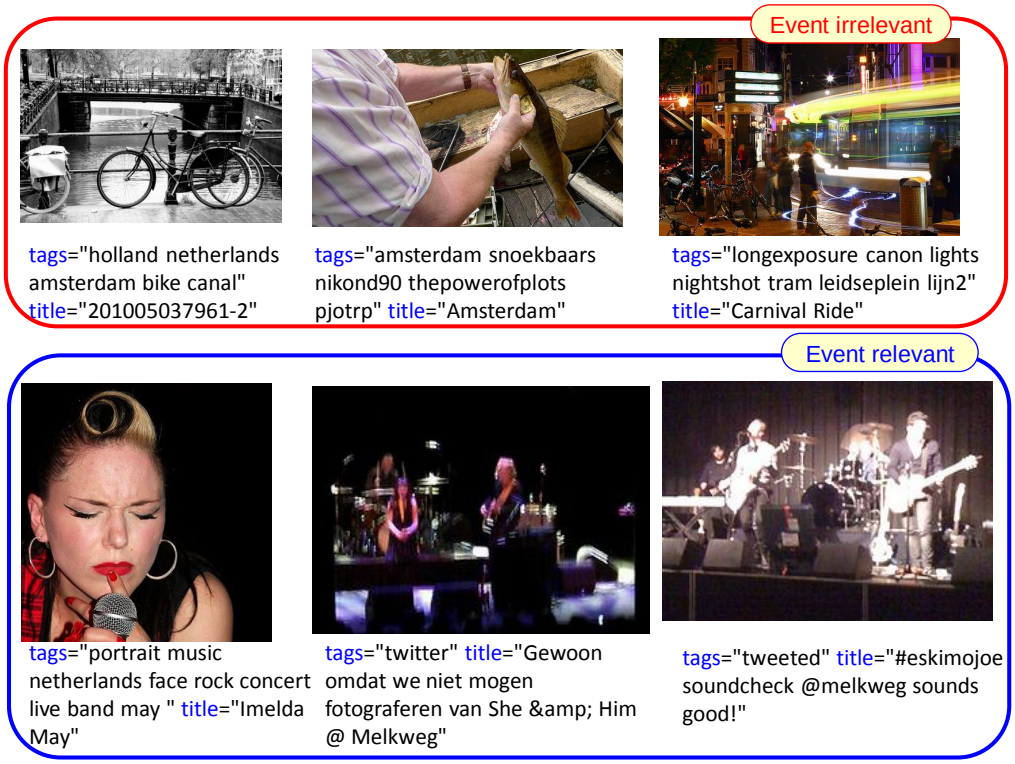


Figure 4.16: The decision made in melkweg

Table 4.11: Event Discovery Results on Topic-based and threshold-based approaches

Venue	Topic-based Approach	Threshold-based Approach	Overlap
Melkweg	14	12	6
Koko	12	8	7
111 Minna Gallery	4	2	1
Ancienne Belgique	10	8	4
Rotown	2	7	2
Circolo degli Artisti	15	7	7
HMV Forum	7	8	3
Total	64	52	30

4.3.4 Discussion

In this section, we present another method for automatically detecting events taking place at given location and time. The events are taken as special distribution over latent topics. We mine the topics by LDA model from a large data collection of shared social media and use the venue specific validation data to infer the event distribution. The experimental results using Flickr photo data demonstrate the effectiveness of the proposed approach.

4.4 Event Detection on MediaEval Challenge

The Social Event Detection task at MediaEval 2011 aims at detecting social events that occurred during May 2009 from a dataset composed of images shared on Flickr [Papadopoulos

et al., 2011a]. The strategy we investigate is to find the event instances that occurred during this period of time and then try to match these event instances with photos from the Flickr dataset. We also study how to employ the visual features and "owner" metadata from the photos to improve the performance. We first detail our approach (Section 4.4.2) before presenting and discussing our results (Section 4.4.3). Finally, we conclude the section in Section 4.5.

4.4.1 Problems Statement

4.4.1.1 Two Challenges

The SED task is composed of two challenges with a common test dataset of images with their metadata (time-stamps, tags, geotags for a small subset of them). Participants are invited to submit results to either one of the challenges, or to both of them.

The first challenge reads: **Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection. For each event provide all photos associated with it.** Soccer events, for the purpose of this task, are soccer games and social events centered around soccer such as the celebration of winning a cup. In contrast, a single person playing with a soccer ball out in the street is not a soccer event under the task's definition.

The second challenge reads: **Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it.** For both these venues, more than one event took place in May 2009. We consider that multiple bands playing the same evening are not distinct events, but a lineup of multiple artists (i.e. we consider that two different events cannot happen the same day at the same location). Some events (e.g. a festival) can last several days with a lineup of artists and will be considered as a single event.

4.4.1.2 Dataset

A collection of 73.645 photos is provided by the SED organizer. The collected photos represent the complete set of geotagged photos that were available for five different cities (i.e., Amsterdam, Barcelona, London, Paris and Rome, based on the geotags) and were taken in May 2009, further augmented with a few non-geotagged photos for the same cities and time period [Troncy et al., 2010b]. However, before providing the XML photo metadata archive (including any tags, geotags, time-stamps, etc.) to the task participants, the geotags were removed for 80% of the photos in the collection (randomly selected). This was done for simulating the frequent lack of geotags in photo collections on the Internet (including the Flickr collection). The exploiting on the photos metadata that are not provided in the dataset are prohibited.

The participants are expected to submit a set of photo clusters, each cluster comprising only photos associated with a single event (thus, each cluster defining a retrieved soccer event). The ground truth is also provided by the organizer and publicly available from the MediaEval website. But we should argue that the ground truth is incomplete, and lots of relevant photos are out of the scope of the ground truth.

4.4.2 Approach Description

The challenge of the social event detection task is to find the photo clusters that are relevant to events held on a given location during a particular period of time. However, the event to mine is so specific and is hardly to model by the photos metadata simply. We tackle this problem in two steps: first, we attempt to retrieve all of the events that occurred at a given place and time; second, we use the extracted information about these events and attempt to match them to the photos metadata in the dataset. All of the photos that are matched to the same event can be grouped in one cluster. Besides these two main steps, we also improve the detection results with visual feature and "owner" metadata.

4.4.2.1 Prior knowledge acquisition

We know that it is easier and more accurate for the computer to identify specific pattern compared with abstract concept. To find concerts or soccer events that may be hidden in the dataset, we first look for all instances of these two types of events held in a given place and time.

Soccer games and concerts are types of favorite activities in people's daily life and one can find substantial information online about such scheduled events. For example, FBLeague[5], as depicted in Figure 4.17, provides the official football games that registered in FIFA[6] and UEFA[7]. From this web site, we obtained 461 football games that occurred in May 2009,among which 6 took place in Roma and Barcelona. These 6 soccer events are our prior knowledge for the challenge 1, as described in Table 4.12.

Table 4.12: The 6 Soccer games

Date	Team A	Team B	Title
May 10	Barcelona	Villarreal	Primera Division
May 23	Barcelona	Osasuna	Primera Division
May 27	Barcelona	Manchester United	Campions Liga Champions League
May 3	Roma	Chievo	Serie A
May 16	Roma	Catania	Serie A
May 31	Roma	Torino	Serie A

For challenge 2, we extract concerts information from event directories such as Last.fm, Eventful, and Upcoming . After manual check, only Last.fm contains descriptions of events held on the given conditions. Using its public API, we found 68 events that took place in the *Paradiso* and 3 events in *Parc del Forum* in May 2009. Some of the events samples are reported in Table 4.13.

4.4.2.2 Event Identification Model

With the prior knowledge of scheduled events description, the event detection task changes to a matching problem where a model can be used to measure the relationship between events and photos. Here, we consider events as *something* happening in *some place* during *sometime*. Therefore, the title, time and location are three key factors that identify an

[5] `http://www.fbleague.com`
[6] `http://www.fifa.com`
[7] `http://www.uefa.com`

Figure 4.17: Interface of FBleage, where most of the football games could be found

Table 4.13: Concerts event samples in *Paradiso* and *Parc del Forum*

Venues	Date	Title
Paradiso	May 31	Malajube
Paradiso	May 31	Band Of Heathens
Paradiso	May 30	Earth: Crisis Editie - 100% NL
Paradiso	May 30	Luciano & The Jahmessenjah Band
Paradiso	May 4	Shearwater
Paradiso	May 2	LondonCalling
Parc del Forum	May 9	Boikot + Soziedad Alkoholika
Parc del Forum	May 30	Ezra Furman & the Harpoons live
Parc del Forum	May 28	Primavera Sound

event. The corresponding photo metadata are text description, taken time and place. Since these three factors are independent, we can measure the probability of a given photo P to be relevant to an event E by

$$p(E|P) = p(E.text|P.title)p(E.time|P.time)p(E.geo|P.geo) \tag{4.7}$$

where: The first item measures the similarity of a photo text description with an event title. Since both of them are short and sparse, the most straightforward way to measure

them is:

$$p(Text1|Text2) = \frac{|Text1 \cap Text2|}{|Text2|} \tag{4.8}$$

Where the function $|\cdot|$ is the total number of words in a text vector.

The second item in Equation 4.7 measures the difference between photo taken time and event held time. Here, we measure the difference using the Dirac function.

$$p(Time1|Time2) = \delta(\frac{date(Time2 - Time1)}{N}) \tag{4.9}$$

Where the function $date(\cdot)$ calculates the number of days for a time span, δ is the Dirac delta function that takes the value 1 when and only when the input parameter is zero, and N is used for scaling (its value will be discussed in the Section 4.4.3).

The third item in Equation 4.7 measures the distance between photo geo tags and event location. The best distance measure to use seems the L2 distance between the two locations. However, an important amount of photos do not have geo tags and when provided, GPS data in the Flickr dataset can be inaccurate. Consequently, we just use the city/venue name to measure the location feature and we use the textual metric formalized in the Equation 4.8.

This method finds many photos with a clear description and association to events. However, text-based matching brings also noise and it can not deal with photos without any text description. We employ visual features analysis techniques to remove the noisy photos and exploit "owner" metadata to find out relevant photos without text description.

4.4.2.3 Visual Filter

Visual pruning is employed to remove the noisy photos from the results of the Event Identification Model. We assume that the photos that are corresponding to the same event should be similar visually. The method used here is quite straightforward. Given a set of the photo feature $\{f_i, i \in [1, N]\}$, the distance between each feature f_i and its mean vector m is measure by the $L1$ distance.

$$d_i = sum(|f_i - m|) \tag{4.10}$$

Photos are then sorted according to the distance d_i. The bigger the distance and the less similar the photo is with the photo cluster, so we prune the photos with such a large distance. Experimentally, we remove the 5% photos that are far from the center in the visual feature space.

4.4.2.4 Owner Refinement

Owner refinement, as stated in Section 3.2.3.4, is another way to improve the detection results. We assume that a person can not attend more than one event simultaneously. Therefore, all the photos that have been taken by the same owner during the event duration should be assigned to the same cluster. Using this heuristic, it is possible to retrieve photos which do not have any textual description.

4.4.3 Experiments and Results

Based on the proposed approach and the events instances obtained previously, we design our runs as follows:

Challenge 1:

- **run1** The parameter N in Equation 4.9 is set to 3, and the basic *Event Identification Model* is run.
- **run2** *Owner Refinement* is performed on the results of **run1**.

Challenge 2:

- **run1** the parameter N in Equation 4.9 is set to 1, and the basic *Event Identification Model* is run.
- **run2** *Owner Refinement* is performed on the results of **run1**.
- **run3** the parameter N in Equation 4.9 is set to 3 to reduce the impact from erroneous taken time, and the basic *Event Identification Model* is run.
- **run4** *Owner Refinement* is performed on the results of **run3**.
- **run5** *Visual Pruning* and *Owner Refinement* are performed on the results of **run3**.

The evaluation of the submissions to the SED task is performed with the use of the ground truth EventMedia associations [Troncy et al., 2010b].

Two evaluation measures are used:

- Harmonic mean (F1-score) of Precision and Recall for the retrieved images. This measures only the goodness of the retrieved photos but not the number of retrieved events, nor how accurate the correspondence between retrieved images and events is.
- Normalized Mutual Information (NMI). This compares two sets of photo clusters (where each cluster comprises the images of a single event), jointly considering the goodness of the retrieved photos and their assignment to different events.

Both evaluation measures receive values in the range [0, 1] with higher values indicating a better agreement with the ground truth results.

A summary of the results is detailed in the Table 4.14. As shown in the Table 4.14,

Table 4.14: Event Detection Results

Run	Results		Evaluation			
	Events	**Photos**	**P(%)**	**R(%)**	**F(%)**	**NMI**
run 1.1	2	216	97,69	41,21	57,97	0,2420
run 1.2	2	222	97,75	42,38	59,13	0,2472
run 2.1	18	1133	70,79	48,90	57,84	0,4516
run 2.2	18	1172	71,13	50,49	59,06	0,4697
run 2.3	24	1502	70,51	64,57	67,41	0,5987
run 2.4	24	1556	70,99	67,01	68,95	0,6171
run 2.5	24	1546	71,00	66,59	68,72	0,6139

2 events are found for challenge 1 with 216 photos identified by the Event Identification Model. 6 additional photos are found by the "Owner Refinement" approach. For the challenge 2, there are mainly two groups of runs. The first group **(run2.1,run2.2)** used the parameter N=1, and 18 events are found from the 69 events set previous obtained. In the second group (**run2.3, run2.4, run2.5**), 24 events are found with the parameter N=3. In general, the results for the challenge 1 are just average since only 6 football games

were found as prior knowledge and we suppose that several other games have been missed. For the challenge 2, the results are more promising and competitive.

In the two groups of results, we could conclude that the *Owner Refinement* approach can improve the recall ratio greatly without losing precision, compared with (run1.1 vs run1.2), (run2.1 vs run2.2) and (run2.3 vs run2.4). In addition, from the challenge 2, it can be seen that when N is set to 3, more events are discovered. We could conclude that though an event are held on a single day, the time record from the attached media has a wider span. It is consistent with our previous work(Section 3.2.3.1)

We also compare the our results with other teams. According to its official report, the best performance of each team is reported in Table 4.15 and 4.16 respectively. It could be found that our methods gain the best performance measured with Precision in challenge 1 and F1 in challenge 2.

Table 4.15: SED evaluation for challenge 1

	ANU	**CERTH**	**EURECOM**	**LIA**	**NTNU**	**QUML**
P	84.86	90.58	**97.75**	7.49	92.47	76.81
R	52.54	**67.58**	42.38	15.62	43.16	62.11
F1	64.9	**77.4**	59.13	10.13	58.85	68.68
NMI	0.237	**0.618**	0.247	0.026	0.475	0.414

Table 4.16: SED evaluation for challenge 2

	ANU	**CERTH**	**EURECOM**	**LIA**	**NTNU**	**QUML**	**VTT**
P	70.79	54.31	70.99	14.37	**78.85**	42.14	73.79
R	43.9	**80.61**	67.01	37.99	56.83	17.13	64.21
F1	50.44	64.9	**68.95**	12.44	66.05	33.01	68.67
NMI	0.448	0.385	0.617	0.013	0.645	0.498	**0.678**

In the end, we visualize some photos clusters as Figure 4.18 and 4.19.

4.5 Conclusion

We presented two novel methods for automatically detecting events and their properties (location, time and participation). The key idea of our first approach consists in temporally monitoring media upload on social sharing web sites, registered to a specific location or referring to a specific venue. We show that using the appropriate thresholding function, it is possible to detect events in an efficient way. The result of our work can therefore be used to automatically structure online media. It was shown that the approach identified a number of events uncovered by the most popular event directories (Last.fm, Eventful and Upcoming).

In our second methods, the events are taken as special distribution over latent topics. We mine the topics from a large data collection of shared social media and use the venue specific validation data to infer the event distribution. The experimental results using Flickr photo data demonstrate the effectiveness of the proposed approach.

In this chapter, we also report our work on "Social Event Detection" task in MediaEval 2012. Our solution combines the methods proposed in this chapter and a promising result is achieved.

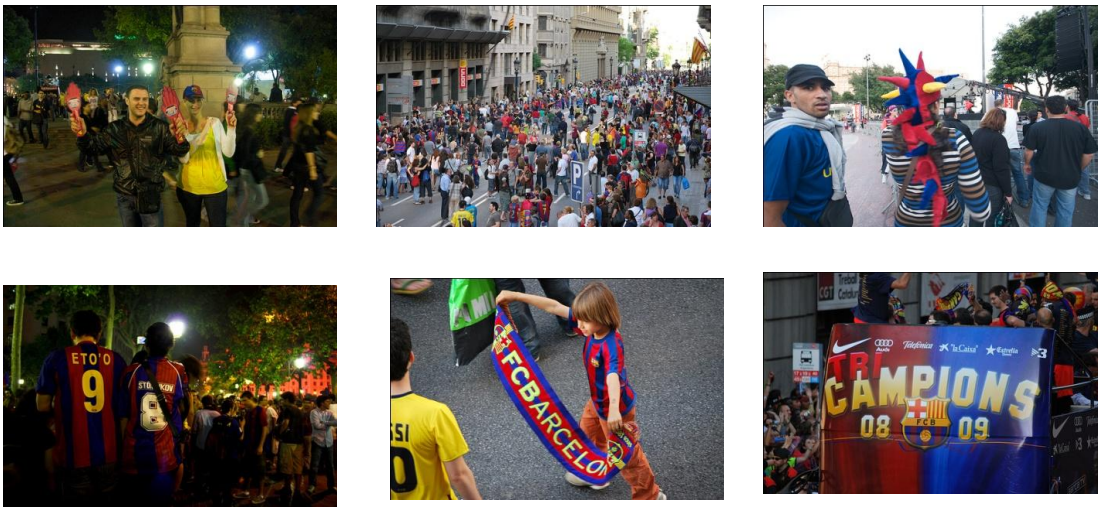


Figure 4.18: Photo cluster for soccer "Barcelona vs Manchester United, Campions Liga Champions League"

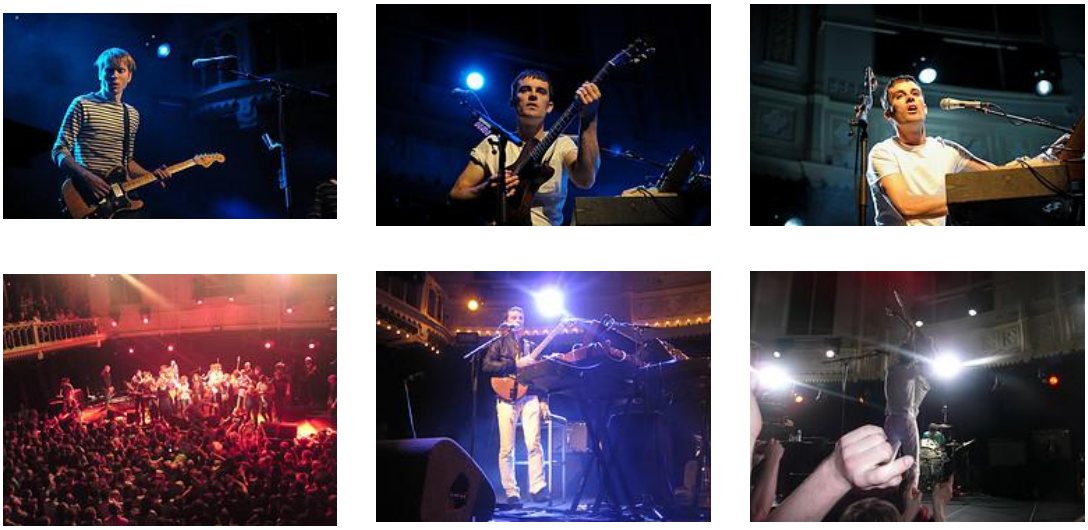


Figure 4.19: Photo cluster for soccer "Franz Ferdinand" in Paradiso

Chapter 5

Training Sample Collection for Social Event Modeling

In the past chapter, we mainly studied events and social media according their metadata. In this chapter, we will focus on modeling events by social media visual content. To model events visually, a framework is presented to collect the training samples automatically, while the collection is done with the analysis of context social media and events. We collected positive samples when they are labeled as event machine tags or event abbreviated names, which precisely refer to the unique event. We collected negative samples by a learning-to-rank approach. The automatically collected samples are then used to train classifiers and promising results are achieved. In this chapter, how to collect training data to automatically to model event visually is studied.

5.1 Introduction

With the popularity of digital capture equipments and the easy use of media sharing web services, many photos taken during public happenings are uploaded and shared by participants. The resulting social media flooding gives rise to new challenge for these websites in terms of data query and management system. How to leverage these user contributed data in research is an open and challenging problem. Recently a new field of study concerning how to index the media data by social events has began to emerge, and some researchers have proposed possible solutions for this problem(Troncy et al. [2010b], Becker et al. [2012]). These works aim at associating media data with events by exploring their rich context. However, it is well known that data missing, or inaccuracy is a frequent issue in collaborate contributed data, which limits the application of these methods.

Besides metadata, the main content in social media is the visual content, in the format of photos or videos (audio being only present in videos). State of the art visual concept modeling methods can be used for event based analysis. However, substantial labeled training data is required for learning the models and creating such a collection is a particularly expensive and tedious task.

In this section, we propose a novel framework for automatically collecting high quality training data and use them to model social events. The data is acquired based on the analysis of rich event context. The positive samples are obtained based on specific tags which identify the events accurately, in the form of machine tags and abbreviation of event title. The negative samples are selected by ranking the localized photos candidates. Finally

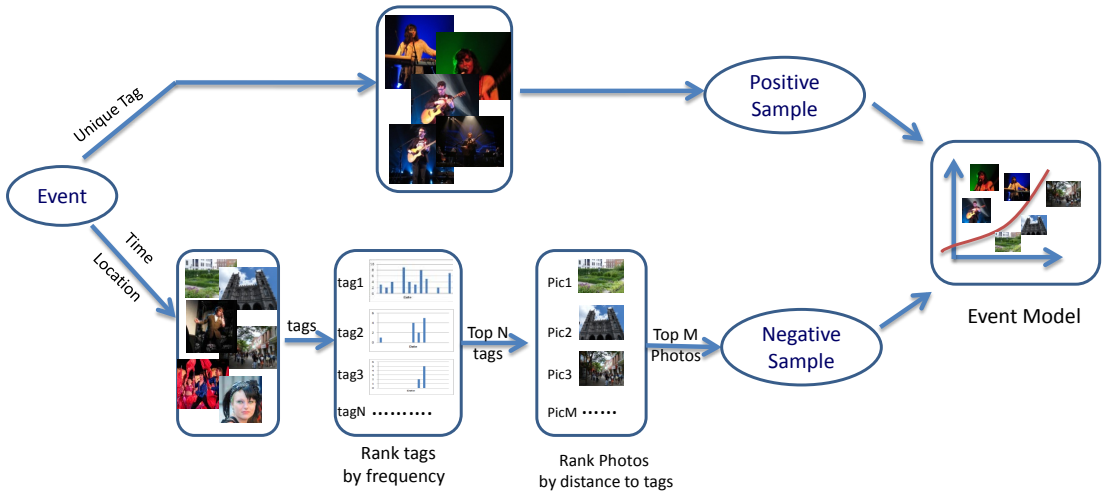


Figure 5.1: Overview of the framework for modeling events semantic

both positive and negative samples are employed to train event models, which are verified with the manually labeled ground-truth, and an acceptable accuracy is achieved. The contributions of this section is twofold:

- We propose framework to collect the training sample automatically, while the collection is done with the analysis of context social media and events. It shows a possible solution on how to use the social media data to visual content based analysis.
- Compared with latest work that use KNN to classify photos in event-driven research since no negative samples are available, we take the event visual modeling as the real classification problem. The visual presentation of each event is learned using SVM , which high accuracy has been proved in many past practice and verified again.

5.2 The Visual Event Model

We propose an original scheme for collecting the training samples for modeling social events visual semantics without human assistance. We define a social event as the specific happening that takes place at a given location and time and involves many persons (i.e. concerts, conferences, exhibitions, etc...). Figure 5.1 depicts the automated steps leading to the creation of the dataset to learning individual event models. Thanks to the machine tags used by event repositories, such as Last.FM[1], Upcoming[2] and Eventful[3] and media sharing platforms, such as Flickr, which give explicit and accurate links between events and social media, it is possible to automatically obtain the set of media documents that refer to specific events and can be used as the positive samples when learning event's visual model.

However, positive samples are not sufficient to construct accurate models. Albeit techniques such as one-class SVM could be used to learn solely using positive samples, these algorithms suffer frequently from under-training or overfitting problems and are therefore

[1] http:/www.last.fm
[2] http:/www.upcoming.org
[3] http:/www.eventful.com

not easy to use in practice. Here, we propose a framework for obtaining relevant negative samples from online social media data. During our previous research, we have observed that when querying for photos originating from an event based on its date and location, the negative samples (those photos which do not correspond to this particular event and should therefore not be associated with this event) are photos depicting general concepts in daily life, such as building, object and portrait. In addition, we noticed that when containing tags these photos are frequently labeled with common tags for this location. For example, the city name is among the popular tags in most cities. In the work presented here we consider those photos captured at similar location to events and containing common tags as negative samples for this specific event. Common tags, along with their corresponding photos, are identified based on a novel approach inspired from learning to rank [Liu, 2011], which we detail in 5.2.2.

Figure 5.2: Machine Tags Used in Last.fm(Top) and Flickr(Bottom)

5.2.1 Positive Samples Collection

We collect social event visual positive samples by querying social media platforms with event identification tag. The identification tag of an event is that tag that precisely refers to this unique event. There are different kinds of tags to identify events in social media

data. The machine tag is a overlap metadata that is available from some events repositories (such as LastFM[4], Upcoming[5], Facebook[6]) and can be used to refer the event when they upload media data taken during the event, so it is popularly used to connect events and photo/video in media sharing platforms, such as Flickr[7]. In these social event websites, machine tags are formatted as "$DOMAIN:event=$XXX", where "$DOMAIN" is the name of website, and "$XXX" is unique event id provided by the events sites, for example, "lastfm:event=1842684" is an event registered in Last.FM whose id is 1842684, and "facebook:event=108938242471051" is a public event in facebook whose id is 108938242471051. When users take photos during the event, they could upload them to media shared websites with such as a tag that provides the background of the photos. The machine tags can be recognized by both kinds of web service and give explicit and accurate links between events and multimedia documents. The media documents containing the appropriate machine tag are taken as positive samples for the corresponding event.

Machine tag is frequently and commonly used nowadays. In the Flickr namespace, there are about 2 million Last.fm event tags, and 400 thousand Upcoming event tags[8]. However, many real world events still do not feature such metadata. To overcome this issue, we use the abbreviated event name to identify such events. Such events abbreviations are well known and popular among the attendees. For example,"ACMMM10" is short for ACM conference on Multimedia 2010, without any ambiguity. All photos with such tag are assumed to be positive samples of this social event.

5.2.2 Negative Samples Collection

Since social events are characterized by a grouping of people at a given time and place, the most relevant negative samples are those images taken around the same period and location as the event but do not originate from the event. Here is an example to motivate our assumption. Given an event held in a city near a famous landmark, it is likely that among the photos taken by attendees some will show the landmark. As a famous landmark, it is expected to be captured frequently by tourists. It is important that such photos are included in the negative samples in order to differentiate between the event and its surrounding. Based on the assumption, we collect negative samples with tags referring to the commonest concepts in the location. We measure the commonness of a tag by its frequency over a given period, and our approach aims at collecting negative samples from localized data. Intuitively, the negative samples refer to a unknown concept for each location. And the tags can be integrally considered as carrying such a latent concept. Let C as target concept, and $T = T_i$ as the tag list of an image I containing n tags. The probability of C in I is defined as:

$$P(C|I) = \frac{P(T|C) * P(C)}{P(T)} \tag{5.1}$$

where the prior probability $P(C)$ and $P(T)$ can be viewed as a constant for the purpose of ranking the images. We assume that the concept C is dominant in the location but different from the event. And our solution to estimate C and to calculate $P(C|I)$ can be solved in 3 steps.

[4] http:/www.last.fm
[5] http:/www.upcoming.org
[6] http://www.facebook.com/events/
[7] http:/www.flickr.com
[8] http://www.husk.org

The first step consists in gathering the photo candidates. For each event, online services are used to identify the location and date. These parameters are then employed to query the Flickr API for a photo set (P). The location is defined by a circle, whose center is determined by the GPS coordinates of the event venue and radius value (R). The time interval is the period of D days before and after the event's date. In order to obtain a large set of candidate photos, appropriate values should be set for both D (days) and R (kms). The influence of those two parameters will be studied in the experiment section 5.3.2.

The second step is to represent concept C with the "common tags". Here, we define "common tags" as tags that represent the most general concepts associated with photos taken in a location. They are commonly and frequently associated with a set of photos, but are different from event. In this chapter, we use two strategies to measure the commonness of tags. Firstly, the commonness of a tag can be influenced by the number of days it appears within a given period. More formally, the commonness of tag t can be calculated as:

$$Score_1(t) = \sum_{i=1}^{D} SD(t,i)/D \tag{5.2}$$

where the value of $SD(t,i)$ is 1 if tag t appears on day i, and 0 if not.

Secondly, since our objective is to find out the negative samples which are irrelevant to an event, the tags that are used to label relevant samples should be punished. Obviously, the more frequently a tag used in labeling media samples from events, the less relevant it is to common concepts on a location. Hence the commonness of a tag t can also be defined as follows:

$$Score_2(t) = \frac{\sum_{i=1}^{D} SD(t,i)/D}{1 + lg(Freq_e(t))} \tag{5.3}$$

Where function $Freq_e$ measures how many times is a tag used to label the positive media sample from the same event.

We rank the tags according to the two types of common scores decreasingly. The top N tags are kept as a group of common tags $CTags$. These tags are prevalently used and highly relevant to the location but do not represent an event due to the fact that they cover a too large time-span. The effect of N, the number of common tags kept to represent the location, is also studied in the experiment section 5.3.2.

The last step is to calculate the probability of C in I so that the negative photo samples could be selected based on commonness ranking. For each photo p of P, we extract the title and tags as their text description $Text(p)$, and compute the similarity between those terms and the common tags obtained previously. The measure used here is the cosine distance.

$$P(C|I) = \frac{CTags \cdot Text(p)}{\|CTags\|\|Text(p)\|}$$

All of the negative candidate are ranked by their textual similarity to the common tags set ($CTags$) and the top M photos are kept as negative samples for training the visual model.

Having collected both positive and negative visual examples of a particular event, machine learning approaches can be employed to learn the visual model. The methodology used to train the Support Vector Machines used in this work is detailed in 5.2.3.

5.2.3 Model Training

The visual model of individual event is learned as follows. We represent each photo in the training data as *Bag of Words*(BoW) feature. For each image, 400-dimensional bag-of-

visual-words feature are extracted in three steps. First, the Difference of Gaussian (DoG) filter is performed on the gray scale images to find key-points and scales respectively; then 128D Scale Invariant Feature Transform (SIFT) feature is computed over the local region defined by the key-points and scales. Finally, we cluster the visual feature with K-means for each event, and the SIFT description is quantized to generate 400D bag of visual words.

The event visual model is learned by *Support Vector Machine* with *Radial Basis Function* (RBF) kernel implemented by the latest LIBSVM [Chang and Lin, 2011]. Cross-validation is used to optimize the SVM model parameters.

5.3 Experiments

5.3.1 Data Set and Experiment Setting

Our proposed algorithm is evaluated on a variety of events, including 10 concerts from LastFM, 3 scientific conferences and 1 popular carnival. Photos from the Flickr social sharing websites are automatically collected for training the corresponding visual models using the approach described previously, as well as for verifying (optimising) the model parameters, and in order to test the performance of the approach for associating media with events.

For our experiments, three photo sets are created. The first set contains all the Flickr photos which match the machine tag of one of the 10 selected events. The result of querying the Flickr API for each event ID in our dataset generates the set of positive samples. We randomly split the positive photos originating from each event into three parts according to usage: 40% for training, 30% for verifying/optimising, and 30% for testing.

The second set contains the negative candidates. Photos that are taken within a given spatial distance (less than R Kms) and a given temporal interval (less than D days) of each selected event are retrieved from Flickr. The impact of the two parameters (R and D) on the visual model quality will be reported shortly. The process of common tags generation and photos ranking, described in section 5.2.2, is performed on this photo set in order to retain only the 200 most common photos for each event as negative samples for training the model.

The last set of media is called Real Online data (**RO**) and is used to evaluate our approach in a real life situation. Here, the candidate illustrative media are those Flickr photos which have been taken within five days of the event (D=5) and for which either the geo-location is within one kilometer of the event (R=1Kms) or at least one tag matches the event title. The ground truth on this collection is provided by manual human labeling.

The number of photos for each event of the three set can be found in Table 5.2 while some photos samples can be seen in Table 5.4.

We use half the positive samples and 200 negative samples to train SVM model for each event, and optimize the parameters D, R and common vocabulary size N using the verification data.

5.3.2 Location Distance, Time Interval and Tags Size

We investigate the impact of parameter R, and D, the location distance and time interval between photo taken and event held, to the final event model. We change the two parameters gradually and test the trained model on the verification dataset. Specifically, R is chosen from 4 to 20 with step of 4 kms, and D is set from 5 to 30 with step 5 days.

Table 5.1: Event DataSet with metadata

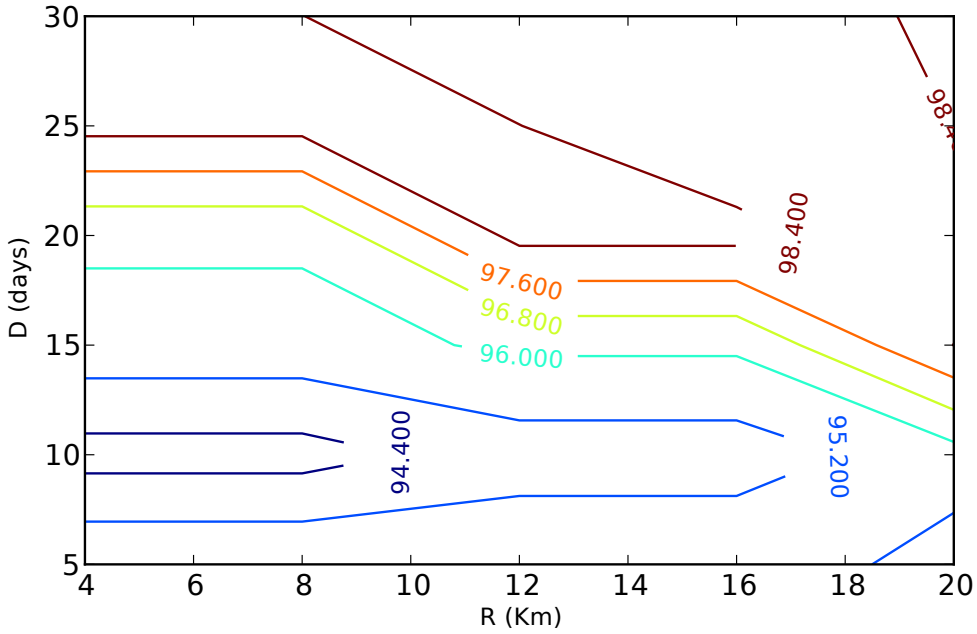
EventID	Title	Date	Latitude	Longtitude
804783	Metallica	03/03/2009	54.964053	-1.62214
1830095	Hole in the Sky Bergen Metal Festival XII	24/08/2011	60.389585	5.323773
1858887	Duran Duran	23/04/2011	41.888098	-87.6294
1499065	Osheaga en Ville	28/07/2010	45.509788	-73.5634
1787326	The Asylum Tour: The Door	03/03/2011	34.062496	-118.349
1351984	Bospop 2010	10/07/2010	50.788893	5.708738
1842684	Buskers Bern	11/08/2011	46.947232	7.452345
2020655	Lacuna Coil - Darkness Rising Tour	18/11/2011	50.72309	-1.86497
1301748	End Of The Road Festival	10/09/2010	50.951341	-2.08262
1370837	Into The Great Wide Open	03/09/2010	52.0333333	4.433333

Table 5.2: The media collection

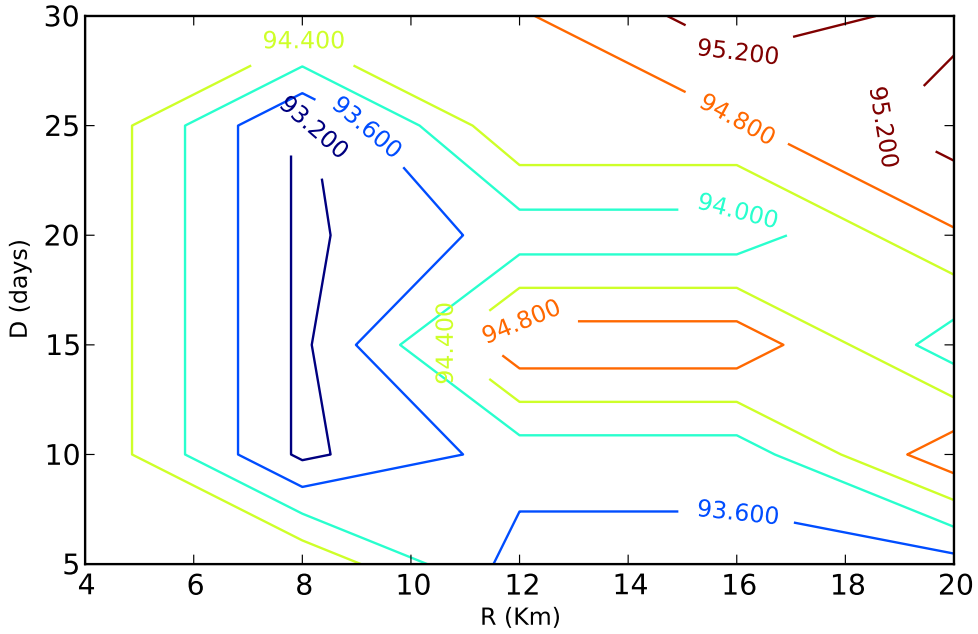
EventID	Positive Samples	Negative Candidate	RO	
			Pos	Neg
lastfm:804783	441	1063	466	64
lastfm:1830095	716	748	398	134
lastfm:1858887	408	745	431	266
lastfm:1499065	348	712	16	153
lastfm:1787326	446	913	0	313
lastfm:1351984	307	584	498	19
lastfm:1842684	602	1125	535	78
lastfm:2020655	538	745	750	6
lastfm:1301748	944	541	1157	80
lastfm:1370837	592	1025	592	115
ACMMM10	100	557	178	23
SIGIR2010	30	525	0	201
ACMMM07	118	64	15	44
NICECarnival2011	52	848	60	209
Total	5642	10195	5096	1705

Grid searching on the two parameters is performed in the process. Figure 5.3 and 5.4 show some examples of resulting classification accuracy averaged over the different size of common tags vocabulary under the two different commonness measurement(Equation 5.2 and 5.3). Results for all selected events from the two groups of measurement favor the use of loose parameters for both time interval and location distance. This finding is supported by the fact that the larger the values of D and R, the more photos are retrieved from Flickr and this results in increased diversity within the selected negative samples.

We also evaluate the influence of N, the number of common tags employed, with respect to the resulting event model accuracy. For each combination of parameters R and D, we optimize the model with vocabulary size varying from 5 to 50 tags. The results, presented in Figure 5.5 and 5.6, clearly indicate that the best performance is obtained for a vocabulary of 10 tags under the two measures.



(a) Event 804783

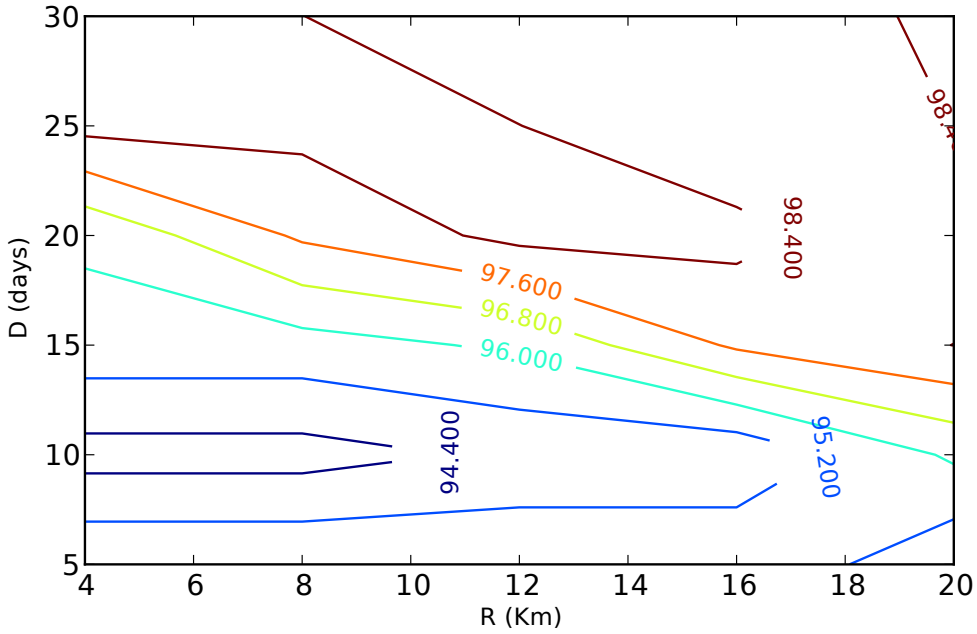


(b) Event 1351984

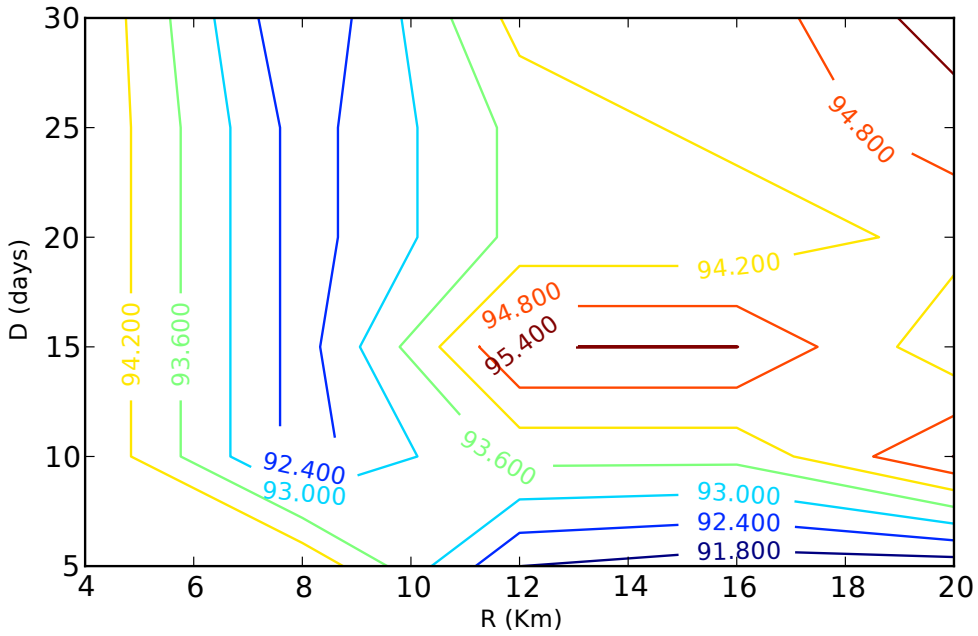
Figure 5.3: Cross Validation on R and D for two Events with $Score_1$

5.3.3 Performance Evaluation

Having carefully chosen parameters (R, D and N), we evaluate the optimized visual models on manually labeled real online data (**RO**). The results of the evaluation runs are measured



(a) Event 1351984



(b) Event 1351984

Figure 5.4: Cross Validation on R and D for two Events with $Score_2$

in terms of classification accuracy (Acc) [Manning et al., 2008] and presented in Table 5.3. Our automatically learned visual event models are compared with four other approaches at the task of mining online media illustrating events and collecting training sample effectively. The first and also the most basic approach, consists in simply running a Flickr query (as

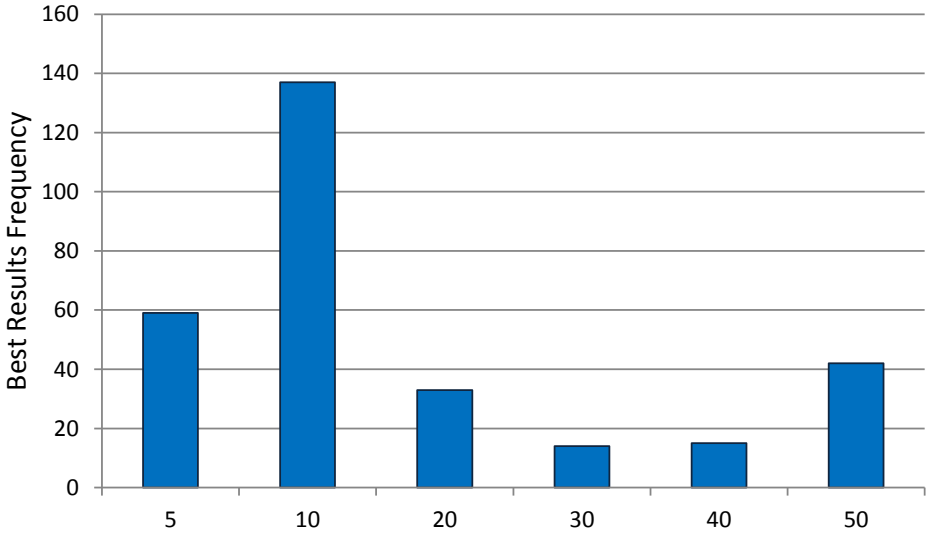


Figure 5.5: Performance vs size of common tag vocabulary with $Score_1$

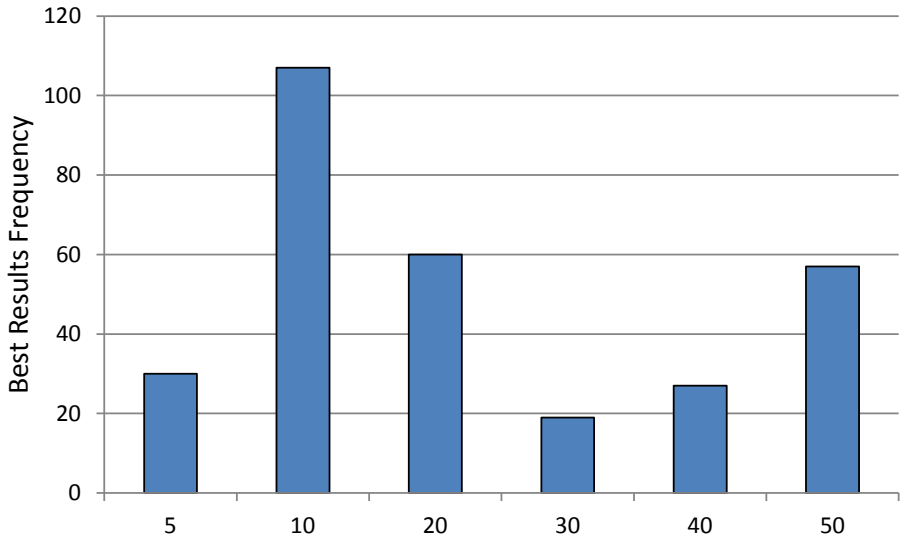


Figure 5.6: Performance vs size of common tag vocabulary with $Score_2$

the one used to create the real online data) and assuming all returned media are positive. In other words, the accuracy value reported in the column **Query**, indicates the proportion of correct photo event associations in (**RO**). The second approach reported for comparison is one where the SVM model is replaced by a K-NN visual filter proposed in chapter 3. In addition, we compare different negative sample collection methods. In the third approach (column **Random Sample**), the negative samples are randomly selected from the localized negative candidates. In order to evaluate the influence of "location", a unique set of 200 negative samples is randomly selected from the entire set of (200 * 14 events) negative samples and used to train all SVM models (column **Uniform Negative**).

Table 5.3: Performance Evaluation (Accuracy)

EventID	Query	$Score_1$	$Score_2$	Pruning	Random Sample	Uniform Negative
lastfm:804783	87.92	88.68	65.47	46.98	50.00	75.85
lastfm:1830095	74.81	78.38	96.43	80.26	96.62	84.96
lastfm:1858887	61.84	63.41	73.17	63.56	76.47	73.89
lastfm:1499065	9.47	90.53	89.94	89.94	92.90	89.35
lastfm:1787326	0.00	98.40	94.77	92.65	97.12	42.49
lastfm:1351984	96.32	96.32	89.36	55.32	86.65	93.81
lastfm:1842684	87.28	87.93	84.34	67.86	79.28	87.11
lastfm:2020655	99.21	91.80	90.34	71.69	75.00	94.58
lastfm:1301748	93.53	93.53	68.63	73.73	64.83	93.21
lastfm:1370837	83.73	85.15	82.32	73.83	60.25	80.62
ACMMM10	88.56	91.04	89.05	87.56	86.57	89.05
SIGIR2010	0.00	60.19	60.19	42.28	16.41	22.38
ACMMM07	25.42	57.62	57.62	46.61	28.81	27.18
NICECarnival2011	22.30	76.58	78.06	59.10	55.39	56.51
Average	69.41	83.31	79.98	68.64	70.07	73.42

From the results presented in table 5.3, it is interesting to note that the approach proposed in Section 3.2.3.4 for analyzing visual content achieves, on average, almost the same performance as the Flickr **Query**. When compared with the approach in [Liu et al., 2011], the our learned visual model performs significantly and consistently better (83.3% (measured with $score_1$) and 79.98% (measured with $score_2$) vs 68.6% on average over all 14 events). This result shows the importance of modeling visual content.

In addition, compared with our approach, the models trained using random negative samples expose degraded accuracy (from 83.3% to 70.1%). Moreover, the performance of models trained with the uniform negative dataset is better than when random negative event sample are used, but not as accurate as our approaches. Those results confirm our hypothesis, "location" information plays an important role in negative samples collection and our approach is effective in collecting such negative samples.

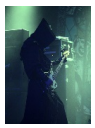
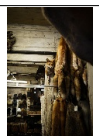
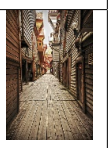
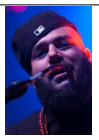
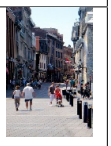
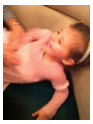
Overall, the experiments have clearly shown the value of using visual analysis to model social events content. Furthermore, we have demonstrated that the construction of the event model can be automated without compromising the resulting performance.

5.4 Conclusion

In this chapter, we proposed a novel framework leveraging on the huge number of media documents available on social media websites to gather the training data collection necessary to learn social event models. The positive samples are collected using photos with identification tags explicitly referring to the event. The negative samples correspond to those photos taken at the same period and in the vicinity of the event but for which the tags are identified as being common (repeatedly appearing over time). We evaluate the trained visual models on a manually labeled dataset, study the effect of the methodology related parameters and the final result reports an better performance compared with some proposals in latest research on a real world scenario. The main contribution of this work is

to automatically collect media samples to train the event models. In this work, the impacts of different parameters, such as the location box and time windows used in collecting data, are well studied.

Table 5.4: Event Training and Testing Samples, some data could not be obtained from social media sites.

EventID	Training				Testing			
	Positive		Negative		Positive		Negative	
804783								
1830095								
1858887								
1499065								
1787326								
1351984								
1842684								
2020655								
1301748								
1370837								
ACMMM10								
ACMMM07								
SIGIR10								
NICE carnival								

Chapter 6

Conclusion

In this chapter, we mainly conclude the content and contribution of this thesis and discuss the future perspectives.

6.1 Achievements

Social media (including blogs, photos, videos, events) hosting and sharing websites, such as Flickr, Facebook, YouTube, Last.FM are gaining popularity around the world. The online data growing at exponential ratio gives new challenges to the traditional management and retrieval technologies. The big data era demands scalable, effective, robust technologies to manage and index them. Recently organizing media data according to real-life events is attracting interest in the multimedia community. Event-based multimedia research focuses on the studying of social media content that are possible connected to events. It could also leverage the event domain knowledge and ontologies during the research, hence gain better performance compared with some other methods.

In this thesis, we have studied the inner semantic and connection between events and social media in many different directions.

1. **General Uploading Analysis**
   We have studied the inner characteristics of media taken during events in two aspects(Section 3.2.3.1). The statistics is exploited on the media data labeled with event machine tags. In the research, we have found that most of the media data is uploaded in five dates after the events. We also discovered the distribution of these data on geographic. These clues are very helpful for use to find out the proper time window, as well the proper geographic bounding box of venues to collect the media data that are relevant to events.

2. **Events Illustration**
   We have developed an event illustration and an event-based query expanding system. In event illustration system(Section 3.2.3), we query media to enrich event illustration set by exploited different multimodal event features. We also designed an effective visual pruner to filter the noisy data in the query results. We also employ the "owner refining" approach to recall the positive samples To improve the final performance. Finally a friendly interface to demonstrate the results is proposed. In the event-based query expansion system(Section 3.3), we use the Natural Language Process techniques to parse the query input and find out the time, location, topic clues.

Then we query events from social news sites by the parsed clues. Hence, our system could respond the query string with the events well illustrated by textual and visual description in vivid interface.

3. **Events Discovery**
   We have studied serval approaches to discover events from social media data. The burst detection(Section 4.2) approach is a simple event detection approach, based on fact that much media data uploading could be observed after the events occurred. In the research, we studied the threshold factors in order to gain a better performance. In the topic model based approach(Section 4.3), based on the assumption that event is one of the common concepts in human life, we collect data taken in a city and employ Latent Dirichlet Allocation model to mine the hot topics. We employ the validating data to make the decision rule. We also compared the results between the two approaches.

4. **Event Modeling**
   We have proposed a framework to collect the training samples automatically to model event in visual aspect(Section 5.2). And the collection is performed on the analysis of context social media and events. The positive samples are collected if they are featured with event machine tag or event abbreviated name, which precisely refer to the content of event. Negative samples are obtained by learning to rank approach. In the research work, the parameters that may impact the final classifier, such as the number of common tags, the time span and spatial bounding box to collect negative samples are well studied.

To summarize, in this thesis, we presented event trends analysis, illustration, discovery, and visual modeling techniques. All of the proposed approaches contribute to mine the interact between events and their associated social media documents. Hence the relation between social media and events is well studied.

6.2 Perspectives

Recently years, some research has been done to study the relation of events and social media data, besides this thesis. Though rich achievement has been gained in these research, there are many remaining research for future work, to leverage the event based analyzing techniques to solving the real world problems. Some directions and possible improvements are as follows:

1. **Personal Events**
   Most of the previous research addresses the problems of public events. In public events, there are much social media data contributed by different users. But in the media shared web services, most of the left media are taken during personal events with less and closed people, such as birthday, wedding. The exploit of these personal events could progress the event based research greatly. However, the characteristics of personal events leads it a more challenge problem compared with public events. First, personal events is private and only public to a few of(or even single) people, which leads that not so much media taken during the events, and collective intelligence can not be utilized in data processing. Second, to protect the personal data and privacy right, the data taken in personal events are always protected and only shared in a

closed group. Hence it is a bit hard to collect plenty of personal data for research purpose. How to study the problem of personal events is still an open challenge.

2. **Large Scale**
   Large scale is a new trend in big data era. With the rapid increase of social media data in recent, there is a growing demands for software solutions to extract relevant information from them. How to apply the traditional machine learning and statistical analysis to large-scale problems is a new challenge both in research and industry. Now a possible solution to large scale problem is cloud computation. Cloud computing is a framework that makes use of thousands of computing resources that are delivered as a service over a network to solve the large scale problem. Its advantages are the ability to horizontally scale to petabytes of data on thousands of commodity servers, easy-to-understand programming semantics, high degree of fault tolerance. Today cloud computation has widespread adoption for large-scale analytic, and owes a large part of its success when solving text analysis, web indexing, and graph processing problems.

3. **Solid Ground Truth**
   Generating ground truth for developing and evaluating various information retrieval systems is an essential process. However, in event based research, thought some large scale event datasets are released, there is still no public dataset to provide solid ground truth on the relation of event and media data, since it is well known that obtaining well labeled data is an expensive and tedious task. To obtain the abundant and high quality ground truth with social media data, it is meaningful to leverage some platform such as Amazon's Mechanical Turk [1]. Amazon Mechanical Turk (MTurk) is a crowdsourcing platform that provides 24/7, on-demand access to workers able to complete tasks requiring human-intervention. For task requesters, MTurk makes it possible to quickly collect human data in a cost-effective way. Some related work [Kittur et al., 2008, Lee and Hu, 2012] shows the consist results could be provided by MTurk platform.

6.3 Conclusion

In this chapter, we conclude the work in this thesis. We summarized the work that has been done to analyze the relation between social media data and events, including general uploading analysis, event illustration, event discovery and visual modeling. We also discuss some new trends and perspectives to continues the work.

[1] `https://www.mturk.com`

Appendix A

Résumé en Français

Les dispositifs actuels de capture de médias et les infrastructures réseaux permettent aux utilisateurs de facilement visualiser et de diffuser des contenus multimédias riches, où qu'ils se trouvent. Ces dernières années ont vu la croissance rapide de sites de partage de médias sociaux disponibles pour la recherche et la navigation sur Internet. Par conséquent, il est devenu courant pour les gens de capturer et de partager des images ou des vidéos de leur vie quotidienne en utilisant leurs téléphones mobiles ou appareils photo numériques. L'ère des données de grande echelle offre des services aux utilisateurs pour partager et d'acceder des données, tandis que son avènement exige une gestion efficace de ces mêmes données et des technologies d'indexation. Comment rechercher un document multimedia de façon efficace, comment exploiter les données pour résoudre les problèmes de passage a l'échelle dans les communautés de l'industrie et de la recherche, les défis restent ouverts. L'idée de gérer et d'organiser les données multimédias peut être retracée dans le domaine de la recherche d'information multimédia (MIR) a il y a plusieurs années. En MIR, les méthodes basées sur le concept reposent sur la récupération des données grace a l'indexation textuelle. Elles sont faciles à concevoir et à déployer et couramment utilisés dans les moteurs de recherche d'images traditionnels que l'on peut trouver sur le Web. Toutefois, le texte entourant les métadonnées textuelles est parfois manquant, ou non liée au contenu des médias. Pour pallier cet inconvénient, des techniques basées sur le contenu sont proposées pour rechercher les médias en analysant le contenu réel des médias plutôt que les métadonnées. Ces techniques emploient la vision par ordinateur et des techniques d'apprentissage automatique pour modéliser le contenu représenté par des caractéristiques textuelles et visuelles. Toutefois, le methode de recherche d'information multimedia basée sur le contenu échoue souvent à cause du, communément appelé, ń fossé sémantique ż, ou autrement dit l'écart entre les caractéristiques visuelles de bas niveau qui sont extraites à partir des données des médias et les concepts de haut niveau qui serait compris par les humains. La recherche visant à combler le fossé sémantique est encore un sujet de recherche brûlant.

La croissance rapide de la quantité de données sur les médias sociaux exige des technologies plus évolutives, efficaces et robustes pour les gérer et les indexer. Les chercheurs commencent à étudier ces problèmes sous des angles différents. Un événement réfère à la survenance à une heure précise et en un lieu donné de phénomènes du monde réel. En histoire, l'événement est l'un des indices les plus importants pour se remettre en mémoire le passé [Tulving, 1984]. La valeur remémorative d'un événement, est extrêmement utile dans l'organisation de la vie humaine. Avant l'ère de l'informatique, les gens aimaient planifier ce qu'il faut faire, ou enregistrer ce qui s'est passé par lebiais d'événements. Avec le développe-

ment rapide des technologies de l'information, l'événement est une nouvelle fois le moyen naturel pour les gens d'organiser, de parcourir et de visualiser leurs collections multimédia. Avec le développement du Web 2.0, un bon nombre de sites de partage d'information liés aux évènements sont construits en ligne. La popularité des médias sociaux et des données partagées, rendent impératif l'exploitation de la sémantique des événements contenue dans les données de médias sociaux.

Dans le milieu de la recherche, l'étude sur d'heuristiques applicable aux événements et aux médias sociaux a récemment attiré beaucoup d'attention [Andrews et al., 2012, Fialho et al., 2010, Westermann and Jain, 2007, Mattivi et al., 2011]. Certains travaux connexes ont été réalisées pour étudier contexte d'événement dans les médias sociaux. Par exemple, GLOCAL[1] http://www.glocal-project.eu/, est un projet européen qui se consacre à l"événements comme le concept central pour rechercher et organiser les données multimédias. Fialho et al. [2010] met l'accent sur l'interconnexion de données de médias sociaux et les événements de web sémantique, et ils voudraient lier les données de différents domaines par les métadonnées dupliquées dans [Quack et al., 2008]. Les auteurs ont travaillé sur la mise en cluster du liées aux événements de photos par des caractéristiques textuelles et visuelles Dans sa proposition, le groupe d'événement pertinent n'a pu être détectée par la durée et le nombre d'utilisateurs parmi les photos du groupe Dans [Becker et al., 2009, 2010].., les auteurs suivent une approche très similaire, en exploitant les riches "" contexte associé au contenu des médias sociaux et appliqué des algorithmes de clustering pour identifier les événements sociaux. Quack et al. [2008]présente une méthode pour fouiller des informations de type événements et objets dans des collections de photos communautaires par des approche de regroupement. Dans leur système, les photos sont regroupées selon plusieurs modalités différentes (caractéristiques visuelles et textuelles), et les groupes sont alors considérés comme des objets ou des événements en fonction de leur durée et des utilisateurs, puisque les événements sont généralement caractérisés par une courte durée.

Ces recherches antérieures montrent une voie prometteuse pour analyser les relations entre les médias sociaux et les événements. Par rapport aux approches basées sur le contenu, la recherche orientée vers les événements porte sur l'étude du contenu des médias sociaux qui sont éventuellement liés à des événements. Cela permettrait de tirer parti de la connaissance du domaine des événements ainsi que des ontologies afin que les problèmes puissent être formulées de manière efficace. De plus, la recherche d'information basé sur les événements aborde le problème en exploitant des caractéristiques multimodales. Outre les caractéristiques textuelles et visuelles couramment utilisées par les approches d'analyse de contenus et de detection de concepts, les données géographiques, temporelles, les informations concernant le propriétaire figurants dans les métadonnées peuvent aussi être intégré. Grace a l'étude de ces caractéristiques, il est possible de traiter des données sans aucune description textuelle, ce qui est un problème commun lors de l'analyse des médias sociaux. Par conséquent, les approches proposées pourraient conduire à une meilleure performance par rapport auxautres méthodes. Les resultats obtenus par les methodes les plus recentes d'analyse d'évènement dans les médias sociaux montrent une voie prometteuse pour réduire le fossé sémantique. Par ailleurs l'étude de la sémantique intérieure et profonde entre évènements est toujours d'actualité. Dans cette thèse, nous employons des techniques d'extraction de données, pour étudier le problème, et tenter de combler fossé sémantique entre les caracteristiques multimodales des médias provenants d'événements et l'événement lui-même. Notre travail contribue, à illustrer des événements avec des medias enrichis, à

[1] http://www.glocal-project.eu/

découvrir des événements grâce a l'analyse du flux de médias sociaux, et à modéliser les propriétés visuelles des événements automatiquement Les principales contributions de la thèse peuvent être résumées comme suit:

1. Nous avons étudié les caractéristiques spatiales et temporelles intrinsèques concernant la capture et le téléchargement de media par les utilisateurs des plateformes de partage de données. Nous avons fait une étude statistique sur les données concernant les médias étiquetés avec des codes machines d'événements, et avons constaté que la plupart des données multimedia sont téléchargées dans les cinq jours qui suivent l'événement. Nous avons également surveillé la répartition géographique de ces données multimédia prises lors d'événements. Ces resultats nous permettent de connaitre le laps de temps approprié, ainsi que la délimitation géographique appropriée pour effectuer des requêtes visant les données des médias.

2. Nous avons développé un système pour créer l'illustration d'événements. Pour construire l'illustration, le système exploite différentes stratégies de recherche utilisant des caractéristiques multimodales. Nous avons également concu un filtre visuel efficace pour eliminer les images ou videos ne correspondant pas a l'evenement à partir des résultats de la requête. Pour améliorer la performance finale, nous utilisons l'approche "raffinage par le propriétaire" afin de concerver les échantillons positifs mais visuellement différents. Nous avons également mis au point une interface conviviale pour présenter les résultats.

3. Nous avons construit un système pour récupérer, résumer et visualiser des événements pour un sujet choisi. Le système répond à la requête textuelle avec des événements bien illustrées par la description textuelle et visuelle. Dans ce système, nous employons des techniques de traitement du langage naturel afin d'analyser la requête et d'en extraire la date, le lieu, le sujet ou le titre de l'évènement. Nous utilisons les indices extraits pour rechercher les événements sur des site Web communautaires de nouvelles. Nous présentons également les résultats finaux avec une interface attrayante, grace a une description textuelle, des nuages de mots clefs, et des collages d'images.

4. Nous avons proposé une approche de détection de pics afin de découvrir les événements à partir de données extraites de médias sociaux. L'approche de détection de pics résulte de l'observation que de nombreuses données multimédia sont téléchargés juste après les événements se soit produits. Pour obtenir une meilleure performance, nous avons étudié différents paramètres, tel que " le nombre de médias" ou " le nombre de médias multiplié par le nombre de téléchargeurs". Nous avons également évalué les résultats de détection en fonction de divers seuils.

5. Nous avons proposé une autre approche pour détection d'événements basée sur une approche de modélisation de sujet. Dans cette approche, nous recueillons des données prises dans une ville et d'employons le modèle d'allocation latente de Dirichlet pour en extraire les sujets d'actualité. Nous utilisons l'inférence à partir de la validation des données pour faire la règle de décision. Nous avons également comparé les résultats obtenus avec l'approche précédente.

6. Pour modèle visuel d'événement, nous avons proposé un cadre de travail pour recueillir les échantillons d'apprentissage automatiquement. La collecte se fait grâce

à l'analyse des médias sociaux et du contexte des événements. Nous collectons des échantillons positifs quand ils sont étiquetés par un "machine-tag" événement. Nous collectons des échantillons négatifs avec une approche inspiree de l'approche "learning to Rank". Nous étudions également en détail les paramètres qui peuvent influer sur le classificateur final, comme le nombre d'annotationscommunes, le laps de temps et la taille de la zone de collecte des échantillons négatifs.

A.1 Enrichissement d'événement

Organiser les données en fonction d'événements médiatiques dans le monde réel est un moyen naturel pour l'homme de se rappeler de son expérience. L'utilisation du contexte de l'évènement pour résoudre les problèmes de gestion et d'indexation dans les médias sociaux attire beaucoup d'intérêt de la part de la communauté de recherche du multimédia. Dans cette section, nous présentons notre travail dont l'objectif est de déduire la sémantique derrière les événements, d'explorer les médias sociaux pour illustrer des événements, et nous nous concentrons sur le renforcement des liens entre les événements et les médias sociaux, de sorte que les données des médias sociaux puissent être indexées par les événements.

A.1.1 EventMedia

Les relations explicites entre les événements prévus et des photos hébergées sur Flickr peuvent être étudié à l'aide des balises de type machine-tag spéciales telles que `lastfm:event=XXX` ou `upcoming:event=XXX`. Le travail de [Troncy et al., 2010b] a exploré le recouvrement des métadonnées entre les quatre sites web les plus populaires, à savoir Flickr comme site d'hébergement web pour les photos et Last.fm qui catalogue et documente des événements passés, présents et à venir. Un vaste ensemble de données appelé " EventMedia" est présenté. Il se compose de descriptions d'événements ainsi que des descriptions des médias associés à ces événements et est interconnecté avec le Linked Open Data. Dans cet ensemble de données, plus de 1,7 millions de photos sont reliées a près de 110,000 événements au total (tableau A.1-ensemble de données).

	Event	**Agent**	**Location**	**Photos**	**User**
`Last.fm`	57,258	50,151	16,471	1,393,039	18,542
`Upcoming`	13,114		7,330	347,959	4,518
`Eventful`	37,647	6,543	14,576	52	12

Table A.1: Descriptions des données dans l'ensemble de données publiées dans EventMedia

Dans cette section, nous considérons un sous-ensemble de cet ensemble de données d'événements qui correspond à l'intersection de Last.fm, Flickr et YouTube pour découvrir des liens significatifs, surprenant et divertissant entre les événements, les médias et les personnes participant à ces manifestations publiques. En d'autres termes, nous considérons l'ensemble des événements last.fm pour lesquels il existe au moins une photo et une vidéo partagée respectivement sur Flickr et YouTube qui a été étiquetée avec une balise machine `LastFM:event=xxx`. Par conséquent, cette intersection donne un ensemble de données de 110 événements et 4790 photos.

A.1.2 Trouvez des illustrations pour des événements

L'ensemble des photos et des vidéos disponibles sur le Web qui peuvent être explicitement associée à un événement en utilisant une balise machine est généralement un minuscule sous-ensemble. Beaucoup de document multimédias qui sont réellement pertinents pour cet événement sont hors de portée, car pas explicitement associé. Notre objectif est de trouver autant que possible les ressources des médias qui n'ont **pas** été marqués avec une balise machine `lastfm:event=xxx`, mais qui doit encore être associée à une description de l'événement. Dans ce qui suit, nous étudions plusieurs approches pour trouver les photos et les vidéos descrivant sémantiquement un évènement .

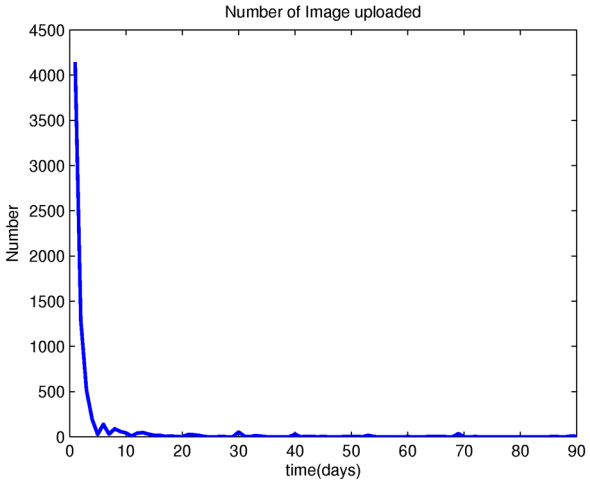


Figure A.1: Proportion de photos enligne en fonction du temps écoulé depuis la capture pendant un évènement.

A partir de la description de l'événement, quatre dimensions du modèle LODE peut facilement être mis en correspondance avec les métadonnées disponibles dans Flickr et YouTube et être utilisé comme parametre de requête pour la recherche dans ces deux plates-formes de partage: le *Quoi* représentant le titre, le*Où* renseignant les coordonnées géographique , le *Quand* qui va de pair avec soit la date de prise de vue ou de la date de la mise en lign d'un média, et le *Qui* informant des artistes ou acteurs impliqués dans les événements. Les requêtes Flickr ou YouTube comportant un ou une seul(e) de ces dimensions retournent beaucoup trop de résultats: car de nombreux événements ont eu lieu à la même date ou à des endroits proches et le titre est souvent ambiguë. Par conséquent, nous allons interroger les sites de médias de partage en utilisant au moins deux de ces dimensions. Nous constatons également qu'il y a des événements récurrents annuellement avec le même titre et tenu au même endroit, ce qui rend la combinaison du "titre" et "géo-tag" inexactes. Nous avons également écarté le *qui* en raison de sa faible précision lors de l' exécution de requêtes concernant des médias. Dans ce qui suit, nous considérons les deux combinaisons "titre" +"date" et "géo-tag" + "date" pour effectuer les requêtes et trouver des documents qui pourraient être pertinents pour un événement donné. Il convient de noter que la requête n'est pas très spécifique et certaines données multimédia non pertinentes seront récupérées. Pour eliminer les médias non pertinents, une technique d'analyse du contenu visuel est développée, afin de supprimer les images indesirable si la différence visuelle est assez importante. Puisque nous savons que les données de médias éti-

quetés avec une balise machine sont très pertinent à des événements et peuvent être obtenu facilement, ils sont le meilleur choix quant aux échantillons de reference pour lefiltrage du bruit. Cependant, dans de nombreux événements, seules quelques images étiquetées avec des balises machines peuvent être retrouvées, et dans le même temps il existe des images pertinentes à partir des résultats de requête avec geo tags. C'est pourquoi nous utilisons les données geo-tag pour construire un modèle visuel dans le but de filtrer les médias erronées, telles que décrites dans la section A.1.2.3. L'ensemble de notre méthode pour enrichir l'événement avec les médias de données est décrite dans la figure A.2.

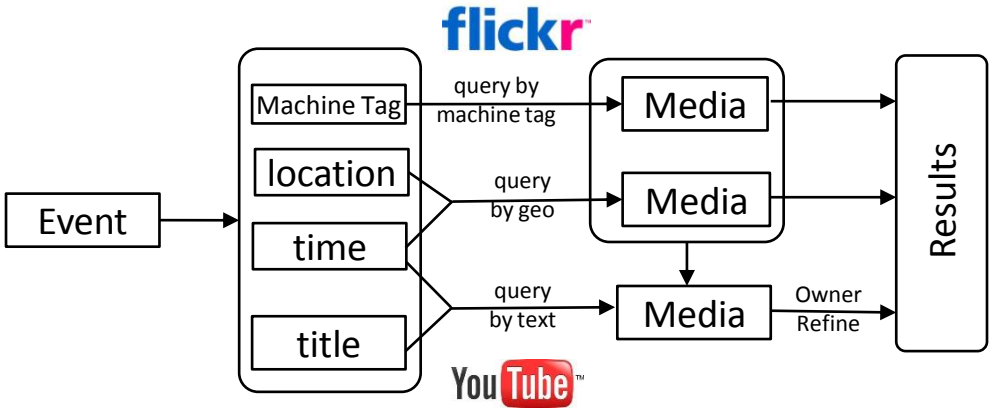


Figure A.2: Le cadre de travail proposé pour enrichir l'événement avec des photos/vidéos

A.1.2.1 Analyse du contexte des médias

Etant donné que les documents multimedia étiquetés avec des balises machine sont prises lors d'événements, nous faisons des etudes statistiques spatio-temporelles sur ces données afin de déterminer les principes suivant lequels les utilisateurs téléchargementleurs medias sur les plateformes de partage. Nous avons d'abord étudier la différence de temps entre l'heure de début d'un événement et le temps de téléchargement des photos sur Flickr rattachées à cet événement. Pour les 110 événements qui composent notre base de données, nous analysons les 4790 photos qui sont annotés avec la balise machine à Last.fm afin de calculer le délai entre l'heure de début d'événement et le moment où les photos ont été capturées conformément aux métadonnées EXIF. La figure A.1 montre le résultat: l'axe des ordonnées représente le nombre de photos téléchargées sur une base quotidienne, tandis que l'axe des abscisses représente le temps (en jours) après l'événement.

La tendance est clairement une courbe à longue queue où la plupart des photos prises lors d'un événement sont téléchargés pendant ou juste après que l'événement ait eu lieu et dans les 5 premiers jours. Après dix jours, seuls très peu de photos de l'événement sont encore partagées par les participants. Dans ce qui suit, nous choisissons un seuil de **5 jours** lors de l'interrogation des photos en utilisant soit le titre ou l'information de géocodage.

Ensuite, nous nous attaquons à modéliser l'emplacement des sites ou les évènements ont lieu. Il n'est pas si facile d'obtenir la zone géographique couverte par un batiment ou une salle, car il n'existe pas de données publiques pour la taille d'un lieu. Nous abordons cette question en nous appuyant sur les informations contextuelles de l'événement fournies par Last.fm et utilisés par les utilisateurs de Flickr. Sur un site donné (VenueID = V),

tous les événements passés ($\{eid\}$) sont récupéré à l'aide de l'API Last.fm. Ensuite, les balises machine "lastfm:event=*eid*" sont utilisés pour rechercher les médias comportants des informations géo-spatiales sur Flickr. Il est alors possible de calculer la boite englobante en utilisant les coordonnées GPS des photos récupérées. L'idée de base est de calculer la zone de délimitation avec des photos prises à proximité de l'emplacement et de filtrer celles qui sont loin de la zone de délimitation. La boite de délimitation finale est estimé par le rectangle minimisant les coordonnées GPS après le retrait des valeurs aberrantes (photos qui sont situés plus loin que le double de la variance de l'ensemble dans les deux sens (longitude ou la latitude)).

A.1.2.2 Requête en ligne

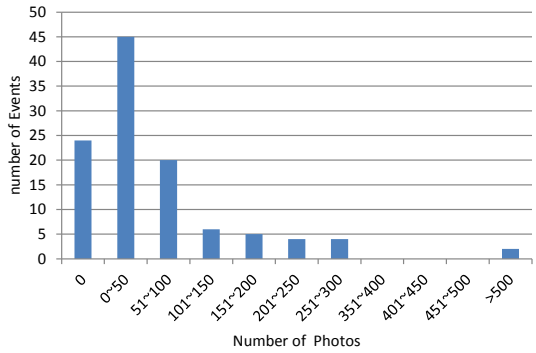
Lorsque nous avons calculé l'information spatiale et temporelle d'un événement, la requête avec les paramètres combinant ń emplacement + date ż et ń date + le titre ż peut être effectuée pour recueillir les medias candidats pour l'enrichissement visuel de l'évènement.

Comme nous avons déjà fait part, de nombreuses photos/vidéos sont capturées lors d'événements, et qu'un certains nombre d'entre eux sont étiquetés avec des balises de géo-localisation indiquant le lieu d'événements, ces données multimemédia peuvent être récupérées grâce a une requette basees sur l'interrogation des paramètres de type géotags. Considérant qu'en unendroit est généralement associé un seul lieu (supposition qui ignore les différents étages d'un batiment), nous supposons que, à un endroit et à un moment , un seul événement se déroule.

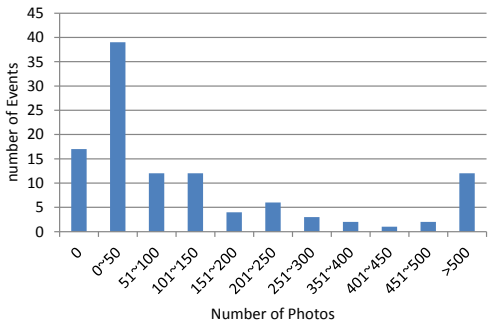
Pour tous les événements dans notre base de données, on extrait la latitude et la longitude des informations à partir des descriptions LODE puis effectuons une requête géographique et temporelle utilisant l'API de Flickr. La fenêtre temporelle utilisée est fixée a 5 jours après la date de l'evenement. Nous effectuons la même requête avec l'API de YouTube, bien que le nombre de vidéos géolocalisées est beaucoup plus petit que les photos. Les figures A.3(a) montrent la répartition du nombre de photos et de vidéos récupérées pour les 110 événements dans notre base de données. Nous observons que les données sont centralisées dans les elements de gauche. Cela veut dire que pour la plupart des événements (n = 95), le nombre de photos extraites avec géotags se situe dans la plage 0-100. L'élément le plus grand est composé de 45 événements qui ont chacun entre 1 et 50 photos récupérées.

Le titre est l'information la plus descriptive et visible pour les événements. De la même manière que pour les requêtes géo-localisées, nous effectuons des requêtes sur le texte complet sur Flickr basé sur les titres d'événements qui sont extraites de la description LODE. Les photos récupérées sont également filtrés en utilisant un intervalle de temps de cinq jours suivant la date événement. En raison des problèmes, bien connus, de polysémie textuelle, la requête basée sur le titre uniquement retourne beaucoup de photos non pertinentes. Nous décrivons dans la section A.1.2.3 une heuristique pour filtrer les médias pertinents.

En général, les résultats de recherche par le titre ont une distribution similaire à celle des résultat de la requête par géotag. Pour la plupart des événements, une baisse du nombre de photos est obtenue. Sur les 110 événements de notre base de donnée, il y a 80 événements avec moins de 150 photos et 83 événements avec moins de 25 vidéos. Toutefois, pour certains événements, un grand nombre de médias sont récupérées: 12 événements avec plus de 500 photos et 15 évènements avec plus de 50 videos. En comparaison avec les figures A.3(a), nous pouvons clairement voir que l'écart type des figures A.3(b) est plus grande et aussi que les photos sont plus faciles à obtenir que les vidéos.



(a) Géolocaliser basée sur requête



(b) Titre basé sur requête

Figure A.3: Statistiques de la requête en ligne

A.1.2.3 Médias d'élagage non pertinents

Les images et vidéos avec des balises spécifiques tels que `lastfm:event=207358` peuvent être inconditionnellement associé à des événements. Nous considérons que les médias récupérés avec des requêtes géo-localiser pendant un laps de temps approprié sont également pertinents pour ces événements. Le problème se pose avec les médias récupérés avec des requêtes basées sur le texte (en utilisant par exemple le nom de l'événement) car il est possible de trouver de nombreux médias non pertinents.

Afin de filtrer ces medias indesirable et d'éviter de propager les riches descriptions des événements de ces documents multimédia, nous proposons une méthode pour eliminer de l'ensemble de photos les candidats en analysant le contenu visuel.

L'idée principale de notre algorithme est de mesurer la similitude visuelle des documents multimedia média. Tout d'abord, nous construisons un ensemble de données d'entrainement composé de documents contenant soit la balise machine d'un événement ou la combinaison de coordonnées géographiques et temporelles correspondant a l'événement. Les photos résultant de la requête par le titre, compose l'ensemble de données de test. Les caractéristiques visuelles utilisées dans notre approche sont les moments de couleurs Lab dans l'espace 225D, 64D Gabor texture, et l'histogramme de contour 73D. Les images dont la distance par rapport à aux images dans l'ensemble de test est inférieure à un seuil sont des candidats pour illustrer l'événement. Mathématiquement, soit E l'ensemble des photos d'entrainement, et F l'ensemble des photos de test, l'objectif est de sélectionner les photos à partir de F qui sont similaires aux photos de E, afin d'enrichir l'ensemble E avec de nouvelles illustrations pour un événement. La similitude visuelle entre deux images est calculée comme suit:

$$L_1(F_j, E_i) = \sum_k |F_j(k) - E_i(k)| \tag{A.1}$$

où $F_j(k)$ et $E_i(k)$ sont respectivement la concaténation des vecteurs de caracteristiques bas niveau normalisées des images. F_j est ajouté à l'ensemble des medias illustrant l'évènement dans le cas où

$$\exists E_i \in E : L_1(F_j, E_i) < THD_i$$

où le seuil THD_i est appris à partir des données E. Comme le montre l'équation A.2, nous utilisons une stratégie rigoureuse pour décider du seuil, qui est choisie comme la valeur minimale de la similitude des paires d'images dans les données d'entrainement.

$$THD_i = \min_{\{j\}\setminus i} \sum_k |EventMedia_j(k) - EventMedia_i(k)| \tag{A.2}$$

Toutefois, certains médias pertinents sont également mis au rebut, ce qui conduit à une performence inférieur en terme de rappel. Afin de récupérer ces photos et d'améliorer le taux de rappel, nous exploitons l'information concernant le propriétaire du media dans les platforme sociale et avons proposé une méthode de raffinement basée sur cette dernière. Nous supposons qu'une personne ne peut pas participer à plus d'un événement à la fois. Par conséquent, toutes les photos qui ont été prises par la même personne (propriétaire) pendant la durée de l'événement doit être attribué à l'évènement lui-même. Grace à l'utilisation de cette heuristique, il est possible de récupérer des photos qui n'ont pas de description textuelle ou géographique. Pour autant que nous le savons, le "raffinement propriétaire" est la seule approche efficace pour mettre en correspondance des données multimédias avec des événements lorsque qu'aucune quantité suffisante de métadonnées (telles que textuelle, graphique métadonnées) est disponible.

A.1.3 Results

Pour évaluer notre approache, nous prenons les 20 meilleurs événements de notre jeu de données (110 événements). Pour ces 20 événements, il y a 785 images dans la base d'apprentissage (photos contenant soit une étiquette événement soit des données de géolocalisation) et 1766 photos dans l'ensemble de test (photos récupérées par titre de l'événement). Nous construisons manuellement la vérité terrain pour les 1766 photos sélectionnées, celles qui doivent être rattachés à un événement et celles ne devraient pas (tableau A.2). Les 20 événements sont tous des événements de type concert et les photos représentent souvent des artistes, des lieux, des scènes ou des spectateurs. Certaines photos étaient cependant parfois difficile à juger, mais l'évaluateur a utilisé toutes les métadonnées disponibles autour de chaque photo, comme par exemple la liste complète des balises ou les albums dans lesquels les photos ont été réunis pour décider si la photo doit être conservée ou pas. Finalement, nous avons supprimer manuellement 193 images non pertinentes de par leur aspect visuel et les métadonnées. Les 1573 autres images sont utilisées comme vérité terrain.

Les résultats de l'algorithme de filtrage, détailé dans la section A.1.2.3, appliqué sur les 1766 photos est présenté dans le tableau A.2. Le seuil retenu est suffissament fort pour garantir une précision de **1** pour la plupart des événements. Toutefois, cela provoque environ 80% des images candidates sont exclues, incluant de nombreuses photos pertinentes. Afin d'augmenter le taux de rappel, nous étendons les images obtenues par notre algorithme de filtrage visuel avec toutes celles mises en ligne par des propriétaires présents à l'évènement. La raison est que si une photo peut être attaché à un événement de manière fiable, nous en déduisons que ce participant en effet assisté à l'événement et que toutes les autres photos prises par cette personne au cours de cette période sont susceptibles d'être d'autres illustrations pour cet événement. Cette heuristique simple permet d'améliorer significativement le taux de rappel (de 0,114 à 0,278) sans pour autant sacrifier à la précision.

Table A.2: Nombre de photos pour 20 événements, les résultats de l'algorithme d'élagage et les résultats de l'extension heuristique simple

ID	DataSet (nb of photos)			Pruning Result			Extended Heuristic	
	TrainingData	TestingData	GroundTruth	Pruned	Precision	Recall	Extend	NewRecall
346054	2	24	2	1	1	0.500	1	0.500
158744	3	48	48	23	1	0.479	44	0.917
371981	4	16	6	4	1	0.667	4	0.667
341832	7	0	0	0	1	1.000	0	1.000
362195	7	0	0	0	1	1.000	0	1.000
235445	10	1	1	0	1	0.000	0	0.000
42644	13	85	81	13	1	0.160	13	0.160
165697	23	1	1	0	1	0.000	1	1.000
137530	24	9	4	0	1	0.000	1	0.250
517159	24	0	0	0	1	1.000	0	1.000
222241	36	204	180	33	0.97	0.183	72	0.400
234649	45	35	4	1	1	0.250	1	0.250
207358	54	68	4	4	1	1.000	4	1.000
429517	60	171	169	27	1	0.160	41	0.243
437747	65	144	142	8	1	0.056	13	0.092
117886	68	99	97	4	1	0.041	11	0.113
150390	71	16	16	1	1	0.063	1	0.063
350591	79	85	85	6	1	0.071	66	0.776
472733	93	500	478	8	1	0.017	18	0.038
176257	97	260	255	47	1	0.184	147	0.576
Avg	785	1766	1573	180	0.998	0.114	438	0.278

A.2 Découverte d' événements

Dans cette section, nous discutons de la facon de detecter des événements sociaux a l'aide de documents multimédia. Sur la base de l'analyse des données des médias sociaux, nous proposons deux approches pour résoudre le problème. Dans un premier temps, nous considérons que beaucoup de médias les données sont téléchargées lorsque des événements se produisent, et proposons une approche de détection des pic de téléchargement sur des zones géographiques ciblées. Deuxièmement, l'événement est considéré comme pertinent avec les thèmes cachés de la vie humaine, et nous intégrons le modèle sujet et faire la règle de décision à la validation des données pour identifier les événements.

A.2.1 Détection d' événements basée sur l'activité des médias sociaux

Dans un premier temps, nousproposons notre méthode pour découvrir les événements a partir du flux de médias sociaux en analysant les téléchargements en un lieu donné. L'heuristique est assez simple: Nous savons déjà que de nombreuses photos sont prises par des personnes différentes lors d'un évènement, donc si on remarque sur une platforme de partage en ligne que 1) de nombreuses photos sont uploadées 2) beaucoup de gens uploadent des photos, nous pouvons en déduire qu'un événement se produit. Comme le montre la figure A.4, il existe principalement 3 étapes dans l'ensemble de cette methode. Dans un premier temps, nous recueillons des données sur un lieu donné, puis utilisons la technique d'analyse de séries temporelles de trouver les points d'intérêt. Enfin, nous résumons et de présentons visuellement les événements trouvés avec les mêmes méthodes que celles proposées dans la section A.1.2.

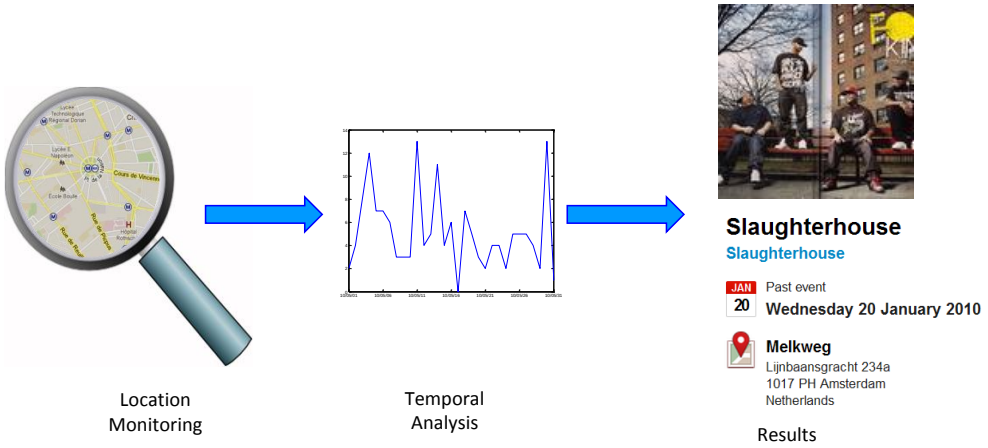
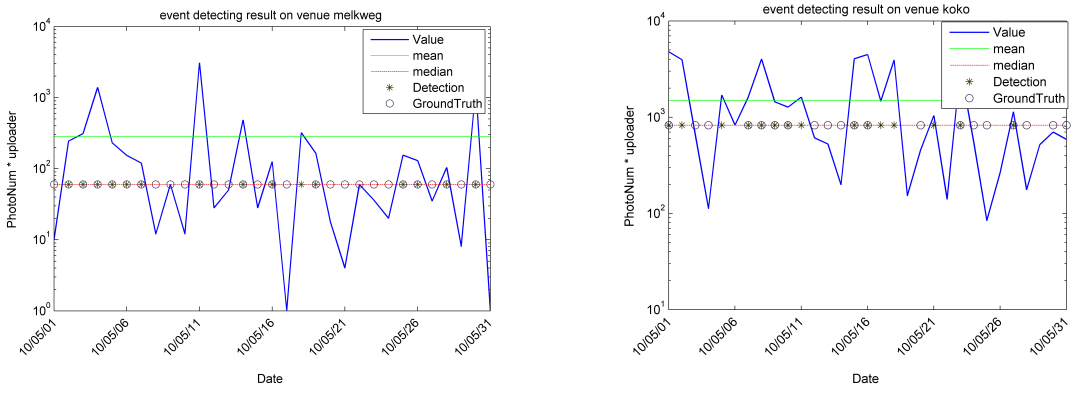


Figure A.4: L'ajout de découvrir l'analyse des événements

A.2.1.1 Détection d'évènement par analyse de mise en ligne

Nous sommes intéressés par la detection d'événements par la surveillance de l'activité de partage de médias sociaux à des endroits précis. Nous visons à miniers automatiquement les événements en fonction de l'activité de téléchargement de photos à des endroits particuliers. Notre objectif en termes de détection des événements est d'identifier la date et le titre de l'événement en raison de son lieu ou l'emplacement. Les figures A.5(a) et A.5(b) représentent l'activité d'ajout en ligne de media sur Flickr en mai 2010 pour deux sites, Melkweg et Koko. En regardant les deux courbes, on peut voir que le nombre de photos prises et téléchargées varie dans le temps en fonction des jours. Notre approche consiste à choisir avec soin les dates avec un nombre élevé d'ajouts et de les considérer comme dates candidates pour un événement. Plus formellement, considérons la série $\{d_i, i \in [1, T]\}$ qui représente l'évolution temporelle du téléchargement de photos caractéristique à un lieu donné v. L'événement e commençant au temps t est détecté lorsque le nombre de téléchargement de photos est supérieur à un seuil donné.



(a) Melkweg, mai 2010

(b) Koko, mai 2010

Figure A.5: Statistique de téléchargement a deux endroits

$$e_t = \arg_i(t_i > THRESHOLD) \tag{A.3}$$

A.2.1.2 Résultat

Pour évaluer l'approche de détection proposée, nous avons collecté un ensemble de photos pour 9 sites populaires. 9178 photos ont étés réunies en recherchant sur le site de Flickr pour le mois de mai 2010 en utilisant soit les boites englobantes géographiques ou le nom des lieux. Elles serviront de base pour les expériences de détection d'événements. Le tableau A.3 indique le nombre de photos obtenues pour chaque emplacement utilisant à la fois des requêtes basées sur les géo-tag et des requêtes sur le nom du lieu.

Table A.3: Nombre de photos prises dans les 9 sites sélectionnés en mai 2010

Name	Geotagged Photos	Venue Name tagged Photos	Duplicate	Total
Koko	372	2040	3	2409
Rotown	90	273	1	362
Melkweg	363	700	8	1055
HMV Forum	184	412	0	596
111 Minna Gallery	937	3	0	940
Ancienne Belgique	2206	288	2	2492
Circolo degli Artisti	70	553	1	622
Circolo Magnolia	95	236	0	331
Hammersmith Apollo	287	84	0	371
All	4604	4589	15	9178

Pour les 9 salles, un programme officiel est disponible sur leurs sites Web. Nous les collectons afin de créer une vérité terrain des évènement planifiés en ce lieu. Dans cette expérience, nous avons d'abord comparer trois caractéristiques différentes concernant la mise en ligne de photos . La premiere est le nombre de photos téléchargées $d_i = ||p_i||$. La seconde intégre la dimension sociale de l'événement et représente le nombre de photo uploaders différents u_i. La troisième caractéristique combine les deux précédentes en multipliant le nombre de photos téléchargées avec le nombre de personne différente qui ont téléchargés les photos de $d_i = ||p_i * u_i||$.

La méthode de détection est exécuté sur les neuf sites sélectionnés. En mai 2010, un total de 242 événements sont signalés dans les calendriers officiels des neuf sites en question. Afin d'évaluer la performance et la précision de notre approche, nous avons besoin d'aligner les événements détectés avec des manifestations officielles. Une fois que les heures de début des événements sont identifiés, nous utilisons les annotations des photos correspondantes pour en déduire le sujet de chaque événement. Tous les mots contenus dans les annotations et les titres des photos prises ce jour-là sont analysées et triées en fonction de leur fréquence d'apparition. Les 15 premiers mots-clés sont gardés pour decrire le sujet de chaque événement. Les événements détectés sont ensuite mis en correspondance manuellement avec des événements de la vérité terrain basés sur la date et le titre. Les événements correspondants sont ceux partageant une date de départ communne et pour lesquels au moins un mot peut être trouvé à la fois dans la liste des 15 mots clés les plus frequents des événements détectés et le titre des événements dans la vérité terrain.

Nous utilisons la même stratégie basée sur la"valeur médiane", qui se trouve être la meilleure méthode pour sélectionner un seuil statique, pour décider du seuil et l'utiliser pour decouvrir des événements. Les résultats sont présentés dans la dernière ligne du tableau A.4. Il est bon de constater qu'avec un seuil heuristique, 69 événements sont trouvés au total, ce qui conduit à la meilleure performance en termes de rappel parmi les approches testées. Cependant, avec un tel seuil, les mesures de précision et F1 se dégrade beaucoup comparé avec le seuil basé sur la "médiane", car trop de fausses detections sont produites. Enfin, nous avons choisi la "médiane" comme critère de seuillage pour la découverte d'evenement a partir du flux d'activités sur les plateforme de partage sociales de docuements multimédia.

Table A.4: Etude de différents critères pour la détection d'événement

Source	Threshold	True Predict	False Predict	F1
Image	mean	43	21	0.281
	median	64	51	0.359
Owner	mean	56	56	0.316
	median	58	62	0.320
Image*Owner	mean	34	**18**	0.231
	median	**67**	53	**0.370**
	hieratical	**69**	90	0.327

A.2.2 Détection d'événements par l'analyse latente des sujets

Dans cette section, nous étudions la détection d'événements dans les médias sociaux par l'analyse latente des sujets. Il est bien connu qu'il y a beaucoup de concepts dans le monde réel, et certains d'entre eux sont plus relatifs aux événements, tandis que d'autres ne le sont pas. Ici, on prend les événements comme distribution spéciale sur ces sujets, et essayons de découvrir les événements par leur distribution sur ces sujets. Comme le montre la figure A.6, les sujets sont tout d'abord extraits de grandes quantités de données capturées à un endroit donné. Ensuite, nous utilisons l'algorithme des moindres carrés pour estimer la distribution des événements sur un groupe d'échantillons de données validées. Nous détectons les événements, à partir d'un ensemble de données de test, si elles correspondent à la répartition sur des thèmes latents.

A.2.2.1 Apprentissage du sujet

Pour un lieu donné (une ville par exemple), l'ensemble des sujets liés à une période de temps peut être considéré comme stationnaire. La sémantique des événements peut être considérés comme une distribution spéciale sur ces sujets. Il y a déjà eu beaucoup d'approches pour inférer des sujets importants parmi les documents. La méthode utilisée dans cette étude est le modèle LDA qui est un modèle probabiliste de génération graphique de découvrir des sujets abordés dans les documents. Dans la pratique, nous recueillons des photos géo-localisées sur Flickr pour une ville donnée (ou un lieu donné). Nous choisissons la racine des mots du titre et des annotations de chaque photo comme la représentation de chaque médias a partir de laquelle l'apprentissage du modèle LDA est fait. Lorsque les modèles sont obtenus, ils sont utilisés pour déduire la distribution sur ces sujets sur les

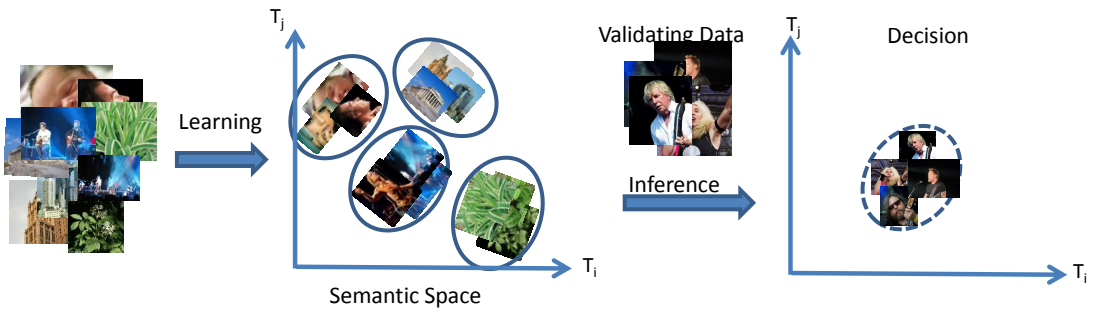


Figure A.6: L'approche proposée pour la détection d'évènements par analyse semantique latente

données validées et pour estimer la distribution des événements, comme détaillé dans la section suivante.

A.2.2.2 Estimation des événements

A partir du modèle LDA, la distribution d'un document d sur des sujets latents peut être déduite par l'équation A.4.

$$P(\theta_d|\alpha,\beta,d) = \int\int p(w,z,\theta|\alpha,\beta)dwdz \tag{A.4}$$

avec w représentant les mots dans les documents, z le sujet des mots dans un document, et θ la distribution des sujets.

Ensuite, nous estimons la distribution d'événements sur les thèmes inférées à partir des données de validation, qui sont les documents multimédias provenant d'un événement. Supposons que D est l'inférence des données de validation sur les thèmes latents, la distribution de l'événement $\mathbf{e}$ peut être estimée grace a la théorie d'optimisation des moindres carrés. L'objectif est de minimiser l'équation A.5.

$$\mathbf{e} = \underset{e\in R^N}{\operatorname{argmin}} \sum_i Dist(D_i, e) \tag{A.5}$$

La fonction $Dist$ mesure la distance entre une instance de validation D_i et l'estimation des événements $\mathbf{e}$. Nous utilisons la version standard symétrique comme mesure de distance

$$\begin{aligned} Dist(p,q) &= D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p) \\ &= \sum_i p(i)\log\frac{p(i)}{q(i)} + \sum_i q(i)\log\frac{q(i)}{p(i)} \end{aligned} \tag{A.6}$$

La distribution de l'événement est estimée à partir de l'équation A.5, qui peut être utilisé pour vérifier si un nouveau document d est un événement lié ou non, selon la règle définie dans l'équation A.7.

$$d \;\; is \begin{cases} event, & \text{if } Dist(d,\mathbf{e}) \le T \\ noevent, & otherwise \end{cases} \tag{A.7}$$

Ou T est le seuil de la fonction de décision, qui est utilisé pour décider si un document est pertinent ou non dans le processus de détection. Dans la pratique, la valeur de T peut être déduit de l'ensemble de données de validation D, comme suit:

$$T = k\max_i\{Dist(D_i, \mathbf{e})\} \tag{A.8}$$

où k est utilisé pour supprimer l'influence du bruit contenu dans l'ensemble de données de validation. La valeur de k est définie à 0,3 expérimentalement.

A.2.2.3 Résultats

Afin de valider l'approche proposée, nous avons recueilli un large ensemble de données à partir d'Internet, qui comprends les photos prises près de 7 salles en mai 2010. Nous utilisons les photos recueillies dans les villes où les salles de concerts sont situées comme données d'apprentissage pour apprendre le modèle LDA. Les photos prises pendant les événements avant mai 2010 dans les même salles de concerts serviront de données de validation, alors que les photos prises pendant le mois de mai 2010 constituent les données de test dans lesquel la fouille d' événements est realisée. Comme nous utilisons l'approache ń bag of word żpour représenter les documents multimédias, certains pré-traitement sont effectués pour éliminer les stop-words et les mots de géolocalisation. Le detail des photos de nos trois jeux de données peut être trouvé dans le tableauA.5. On calcul le modèle

Table A.5: Collections de photos par Lieux.

Venue	**City**	**Training Set**	**Validating Set**	**Testing Set**
Melkweg	Amsterdam	3786	179	355
Koko	London	23384	194	724
111 Minna Gallery	Chicago	11725	175	313
Ancienne Belgique	Brussel	2120	321	496
Rotown	Rotterdam	1575	71	118
Circolo degli Artisti	Rome	6551	107	167
HMV Forum	London	23384	189	97

LDA sur les données d'apprentissage et d'utilisons les modèles formés pour en déduire la distribution des sujets sur les données de validation et sur les données de test. Ensuite, la règle de décision peut être apprise après le processus d'inférence sur les données de validation.

La dernière étape de notre processus de détection d'événement est effectuée sur les données de test et les résultats sont contrôlés manuellement et individuellement, sur la base de la correspondance entre les descriptions textuelles des images d'événements et la réalité de terrain. Le tableau A.6 rapporte les statistiques de données multimédia sur la détection d'événements pour les 7 sites. Dans ce tableau, le nombre de documents qui sont détectés comme étant liés à un événement est représenté. Au total, 265 sur les 2270 photos sont identifiés comme liés à un événement et 160 photos (sur la 265) sont affectés correctement à l'événement en question, ce qui conduit à une précision moyenne de 0,60.

Sur les 203 événements disponibles dans le jeu de données (selon les méta-données), 63 d'entre eux sont détectés par notre approche. Cela corresponds à un rappel de 0,31.

Table A.6: Performance de la Détection d'événements sociaux

Venue	GroundTruth	Total Post	Detection	Postive	Events	Precision	Recall
Melkweg	69	355	42	32	14	0.76	0.52
Koko	21	724	95	44	12	0.46	0.80
111 Minna Gallery	23	313	26	10	4	0.38	1.00
Ancienne Belgique	38	496	32	19	10	0.59	0.53
Rotown	16	118	6	4	2	0.67	0.29
Circolo degli Artisti	22	167	46	36	15	0.78	0.88
HMV Forum	14	97	18	15	6	0.83	0.60
total	203	2270	265	160	63	0.60	0.64

A.3 Creation automatique d'ensembles de donnes d'entrainement pour la modélisation d'événements sociaux

Dans cette section, nous proposons un nouveau cadre pour la collecte automatique de données de grande qualité pour l'apprentissage de modèles visuels d'évènements formation. Les données sont acquises sur la base de l'analyse de contexte d'événements. Les échantillons positifs sont obtenus à partir de balises spécifiques qui identifient les événements avec précision, sous la forme de balises ń machines ż et d'abréviations de titre de l'événement. Les échantillons négatifs sont sélectionnés en classant les photos localisées candidates. Enfin, les deux ensembles d'échantillons (positifs et négatifs) sont utilisés pour apprendre des modèles visuel d'événements, qui sont vérifiés avec l'ensemble de vérité terrain manuellement étiquetée. Les resultats obtenus dans nos expériences atteignent une précision honorable.

La figure A.7 représente les étapes automatisées menant à la création du jeu de données d'apprentissage des modèles d'événements individuels. Grace aux balises machine, qui fournissent des liens explicites et précis entre les événements et les médias sociaux, il est possible d'obtenir automatiquement l'ensemble des documents multimédia qui font référence à des événements spécifiques et peuvent être utilisés comme échantillons positifs lors de l'apprentissage d'un modèle visuel d'événement.

Toutefois, les échantillons positifs ne sont pas suffisants pour construire des modèles précis. Nous proposons une approche pour recueillir les échantillons négatifs à partir de données de médias sociaux en ligne inspirée de l'algorithme ń apprendre à classer ż.

A.3.1 Collecte d'échantillons positifs

Nous recueillons les échantillons visual positifs pour différents évènements sociaux en interrogeant les plateformes de médias sociaux avec des balises d'identification événement. Une balise d'identification d'un événement, est une balise qui se réfère précisément et uniquement à un événement. Il existe différents types d'étiquettes pour identifier les événements dans les données des médias sociaux. La balise machine est une métadonnées additionnelle qui est disponible à partir de certains sites de référencement d'événements (comme

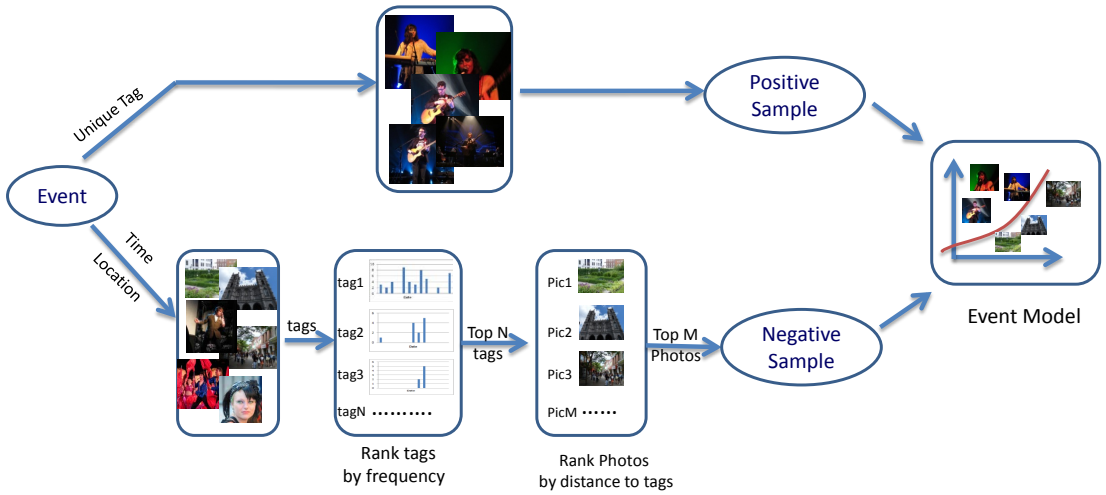


Figure A.7: la cadre de modélisation pour les événements sémantiques

Last.FM, Upcoming, Facebook). Cette balise est utilisée pour désigner l'événement lorsque les utilisateurs téléchargent des données multimédia prises lors de l'événement. Elle est assez couramment employée pour mettre en correspondance des événements et des photos / vidéos sur les plateformes de partage de médias comme Flickr. Lorsque les utilisateurs prennent des photos pendant un événement, le système leurs proposes d'ajouter une étiquette qui identifie les photos déposer commeprovenant d'un événement particulier. Les documents multimédias contenant la balise machine appropriée sont utilisés comme échantillons positifs pour l'événement correspondant.

La balise machine est fréquemment et couramment utilisée de nos jours. Cependant, de nombreux événements du monde réel ne sont toujours pas dotées de telles métadonnées. Pour surmonter ce problème, nous utilisons le nom abrégé de l'événement pour identifier de tels événement. Ces abréviations sont bien connu et populaire parmi les participants a l'événements. Par exemple, " ACMMM10" est l'abréviation de conférence ACM Multimédia 2010, sans aucune ambiguïté. Toutes les photos avec le label abrégé de l'événement sont supposés être des échantillons positifs de cet événement.

A.3.2 Collecte d'échantillons négatifs

Puisque les événements sociaux sont caractérisés par un groupement de personnes à un moment et un lieu donné, les échantillons négatifs les plus pertinents sont des images prises autour de la même période et autour du même emplacement que l'événement mais qui ne proviennent pas de l'événement lui même. Voici un exemple pour motiver notre hypothèse. Sur la base de cette hypothèse, nous prélevons des échantillons négatifs (photos) avec des étiquettes faisant référence à des concepts parmi les plus courants pour l'emplacement en question. Nous mesurons la banalité d'une étiquette (label) de par sa fréquence d'apparition au cours d'une période donnée. Notre approche vise à collecter des échantillons négatifs à partir de données localisées. Les étiquettes peuvent être intégralement considérées comme portant un concept latent. Soit C representant un concept cible et $T = T_i$ la liste des etiquettes d'une image I contenant n labels. La probabilité que C soit present dans I est défini comme suit:

$$P(C|I) = \frac{P(T|C) * P(C)}{P(T)} \tag{A.9}$$

où la probabilité a priori $P(C)$ et $P(T)$ peut être considéré comme une constante dans le but de classer les images. Nous supposons que le concept C est dominant pour le lieu en question mais pas en relation avec l'événement. Notre solution pour estimer C et calculer $P(C|I)$ comporte 3 étapes.

La première étape consiste à rassembler les photos candidates. Pour chaque événement, les services en ligne sont utilisés pour identifier un lieu et une date unique. Ces paramètres sont ensuite utilisés pour interroger l'API Flickr pour une série de photos (P). Le lieu geographique est défini par un cercle dont le centre est déterminé par les coordonnées GPS du lieu de l'événement et la valeur du rayon (R). L'intervalle de temps est la période de jours D, avant et après la date de l'événement. Afin d'obtenir un grand nombre de photos, les valeurs appropriées doivent être déterminées à la fois pour D (jours) et R (kms).

La deuxième étape consiste à représenter concept C avec les labels les plus "communs". Ici, nous définissons les "label communs" comme des etiquettes qui représentent les concepts les plus généraux liés aux photos prises dans un endroit particulier. Ils sont communément et fréquemment associée à l'ensemble des photos prisent dans cette zone geographique, mais ils sont différents de ceux qui definisse l'événement. La banalité d'une etiquette peut être influencée par le nombre de jours qu'elle apparait dans un délai donné. Plus formellement, la banalité du label t peut être calculé comme suit:

$$Score_1(t) = \sum_{i=1}^{D} SD(t,i)/D \tag{A.10}$$

où la valeur de $SD(t,i)$ est 1 si le label t apparait le jour i, et 0 sinon.

Nous classons les étiquettes selon les deux types de critères de banalité de manière décroissante. Les N étiquettes les plus frequentes sont conservées comme le groupe de labels communs $CTags$.

La dernière étape consiste à calculer la probabilité de C dans I de sorte que les échantillons négatifs photo pourrait être choisies en fonction du classement banalité. Pour chaque photo p de P, on extrait le titre et les labels de leur description textuel $Texte(p)$, et de calculer la similarité entre les mots et les labels communs obtenus précédemment. La mesure utilisée ici est la distance cosinus.

$$P(C|I) = \frac{CTags \cdot Text(p)}{\|CTags\|\|Text(p)\|}$$

Tous les candidats négatifs sont classés selon leur similarité textuelle avec l'ensemble des labels les plus communs ($CTags$) et les premieres M photos sont conservées comme échantillons négatifs pour l'apprentissage du modèle visuel.

Après avoir recueilli des exemples visuels à la fois positifs et négatifs d'un événement particulier, les approches d'apprentissage automatique peuvent être utilisées pour apprendre le modèle visuel. La méthodologie utilisée pour entrainer les machines a support de vecteurs utilisées dans ce travail est détaillé dans A.3.3.

A.3.3 Entrainement des modèles visuels

Le modèle visuel de chaque événement est appris comme suit. Nous représentons chaque photo dans les données d'apprentissage comme dans l'approche *Bag of Words* (BoW). Pour

chaque image, les caractéristiques visuel de 400-dimension formant le ń bag of words ż sont extraites en trois étapes. Tout d'abord, un filtrage de type différence de gaussiennes (DoG) est effectuée sur les images en niveaux de gris pour trouver respectivement les points clés et les échelles. Ensuite, la caractéristique SIFT de 128D est calculée sur la région définie par les points-clés et les échelles. Enfin, nous regroupons les caractéristiques visuelle avec K-means pour chaque événement, et les representations SIFT sont quantifiées pour créer le ń bag of words ż de 400 dimensions.

Le model d'événement visuel est appris par *Support Vector Machine* avec une fonction à base radiale (RBF) comme noyau. Nous utilisons la dernière LIBSVM dans notre implemantation. Le principe de validation croisée est utilisée pour optimiser les paramètres du modèle SVM.

A.3.4 Résultats

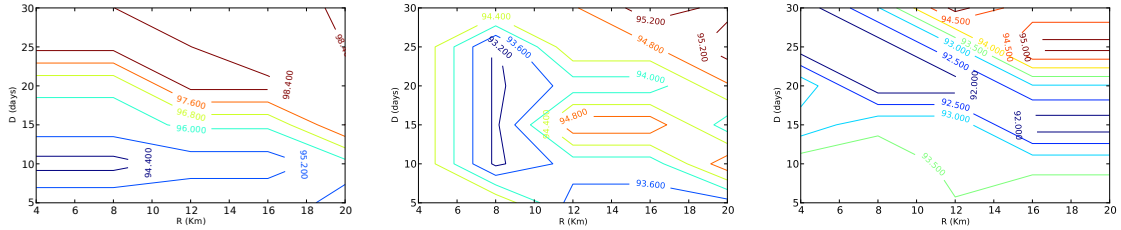
Notre nouvel algorithme est évalué sur différents types d'événements, 10 concerts provenant du site Last.fm, 3 conférence internationale et 1 carnaval. Les photos sont collectés sur le site de partage de photo Flickr et sont utilisées pour l'apprentissage automatique des modèles visuel comme dans l'approche décrite précédemment.

Pour nos expériences, trois séries de photos sont créés. Le premier ensemble contient toutes les photos Flickr qui font référence à la balise identifiée de l'un des 14 événement sélectionné. Nous divisons aléatoirement les photo positives provenant de chaque événement en trois parties selon les usages suivants: 50 % pour l'entrainement du model, 25 % pour la vérification et 25Le nombre de photos pour chaque événement de la collection peut être trouvés dans le tableau A.7.

Table A.7: Nombre de données utilisées pour la modélisation visuelle d'événements

EventID	Positive Samples	Negative Candidate	Testing	
			Pos	Neg
lastfm:804783	441	1063	466	64
lastfm:1830095	716	748	398	134
lastfm:1858887	408	745	431	266
lastfm:1499065	348	712	16	153
lastfm:1787326	446	913	0	313
lastfm:1351984	307	584	498	19
lastfm:1842684	602	1125	535	78
lastfm:2020655	538	745	750	6
lastfm:1301748	944	541	1157	80
lastfm:1370837	592	1025	592	115
SIGIR2010	100	557	178	23
ACMMM07	30	525	0	201
ACMMM10	118	64	15	44
NICECarnival2011	52	848	60	209
Total	5642	10195	5096	1705

Dans cette expérience, nous avons d'abord étudier l'impact des paramètres R et D, la taille de la zone géographique et l'intervalle de temps entre la capture de la photo et l'événement organisé, sur la qualité du modèle visuel d'évènements. Plus précisément, R est choisi entre 4 à 20 à l'increment de 4 km, et D varie de 5 à 30 par étape 5 jours. La validation croisée sur les deux paramètres est effectuée pour apprendre le modèle. La figure A.8 montre des exemples de la classification, en terme de precisionmoyenne différentes tailles de vocabulaire de mot commun. Les résultats montrent que pour tous les événements sélectionnés il est interressant de favoriser l'utilisation de paramètres assez large pour



(a) Event 804783 (b) Event 1351984 (c) Event 1351984

Figure A.8: Cross Validation on R and D for 3 Events

intervalle de temps et la la taille de la zone geographique. Dans nos expériences, nous avons également évaluer l'influence du nombre de label couramment utilisées sur la précision du modèle visuel d'événement. Les résultats, présentés dans la figure A.9, indiquent clairement que la meilleure performance est obtenue avec un vocabulaire de 10 mots.

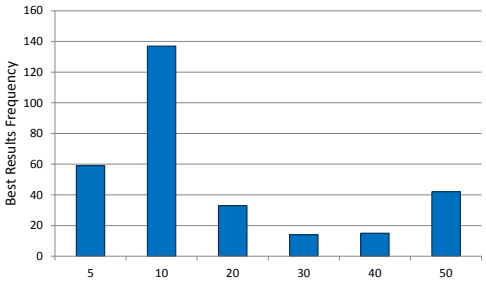


Figure A.9: Performance de la taille du vocabulaire commun tag

Après avoir soigneusement choisi les paramètres (R, D et $CTags$), nous évaluons les modèles visuels optimisés sur la collection de photos ń données réelles ż manuellement étiquetés (**RO**). Les résultats des essais d'évaluation sont mesurées en termes de précision de la classification (Acc) Manning et al. [2008] et présentés dans le tableau A.8. Nos modèles d'événements visuels appris automatiquement sont comparés avec quatre autres approches. La première consiste simplement à lancer une requête sur Flickr (comme celle utilisée pour créer les données réelles (**RO**)) et en supposant que tous les médias retournés sont positifs. En d'autres termes, la valeur de précision indiqué dans la colonne **Query**, indique la proportion de bonnes associations entre événements et photos dans (**RO**). La seconde approche signalés aux fins de comparaison est celle où le modèle SVM est remplacé par un classificateur de type k plus proche voisins (**K-NN**). En outre, nous avons également comparer le modèle avec différentes méthodes de prélèvement d'échantillons négatifs. Dans la troisième approche de comparaison (colonne **Localization Aware**), nous formons modèle SVM avec des données négatives choisis au hasard parmi les candidats négatifs. Pour l'évaluation de l'influence de la localisation, nous avons également sélectionner au hasard des échantillons négatifs générés par notre approche et le modèle SVM de chaque evenement est entrainé avec le même ensemble négatif (colonne **Localization UnAware**). D'après les résultats du tableau A.8, il est intéressant de noter que l'approche utilisée dans la section A.1.2.3 pour l'analyse du contenu visuel réalise que des résultats similaires par rapport à **Query** seule approche. En comparaison avec l'approche adoptée

Table A.8: évaluation de la performance (précision)

EventID	Query	Our Algorithm	Pruning in Section A.1.2.3	Localization Aware	Localization UnAware
lastfm:804783	87.92	88.68	46.98	50.00	75.85
lastfm:1830095	74.81	78.38	80.26	96.62	84.96
lastfm:1858887	61.84	63.41	63.56	76.47	73.89
lastfm:1499065	9.47	90.53	89.94	92.90	89.35
lastfm:1787326	0.00	98.40	92.65	97.12	42.49
lastfm:1351984	96.32	96.32	55.32	86.65	93.81
lastfm:1842684	87.28	87.93	67.86	79.28	87.11
lastfm:2020655	99.21	91.80	71.69	75.00	94.58
lastfm:1301748	93.53	93.53	73.73	64.83	93.21
lastfm:1370837	83.73	85.15	73.83	60.25	80.62
SIGIR2010	0.00	60.19	42.28	16.41	22.38
ACMMM07	25.01	57.62	46.61	28.81	27.18
ACMMM10	85.83	91.04	87.56	86.57	89.05
NICECarnival2011	22.30	76.58	59.10	55.39	56.51
Average	69.41	83.31	68.64	70.07	73.42

dans la section A.1.2.3, le notre modèle d'apprentissage visuel offre des résultats nettement et systématiquement meilleurs (83,3 % vs 68,6%). Les résultats montrent l'importance de la modélisation du contenu visuel.

En outre, par rapport à notre approche, les modèles formés des échantillons négatifs tirés par hasard dégradent beaucoup les resultants(de 83,3 % à 70,1 %). La performance des modèles formés avec ensemble de données uniformes négative est meilleure que l'échantillon aléatoire, mais quelque peu en dessous des resultats de notre nouvelle approches. On peux en conclure que notre approche est efficace pour collecter des échantillons négatifs.

Dans l'ensemble, les expériences ont clairement montré l'intérêt d'utiliser l'analyse visuelle pour modéliser le contenu des événements. De plus, nous avons démontré que la construction du modèle visuel d'événement peut être automatisé sans compromettre les performances qui en résulte.

A.4 Conclusion

Dans cette section, nous allonsconclure sur le contenu et les contributions de cette thèse et de discuter des perspectives d'avenir.

A.4.1 Réalisations

Dans cette thèse, nous avons étudié la sémantique intérieure et les connexions entre les événements et les médias sociaux dans de nombreuses directions différentes. Il existe principalement trois parties soigneusement étudiés dans ce travail .

1. **L'illustration d'événements**
   Nous avons développé un système de recherche de media basé sur les événements afin d'étendre le nombre d'illustrations pour chaque événement. Dans le système d'illustration d'événement, nous recherchons sur les médias sociaux des documents pour enrichir ensemble des illustrations d'un événement . Nous avons également conçu un flitre visuelle efficace pour éliminer les données bruitées dans les résultats

de la requête en ligne. Nous employons également l'approche "raffinage propriétaire" permettant de retrouver des échantillons positifs visuellement different mais provenant assurément de l'évènements pour améliorer la performance finale. Enfin, une interface conviviale pour démontrer les résultats est proposée.

2. **La Découverte d'événements**
   Nous avons étudié deux approches pour découvrir les événements à partir de données de médias sociaux. La premiere approche de détection est basé sur des faits que beaucoup de document multimédias sont téléchargés pendant et après les événements. Nous avons étudié les facteurs de seuil afin d'obtenir une meilleure performance. Dans l'autre approche, basée sur la methode statistique de modèle latent du sujet, l'hypothèse est faite qu'un événement est l'un des concepts communs de la vie quotidienne. Nous recueillons des données prises dans une region geographique et d'employons l'analyse LDA pour extraire les sujets d'actualité. Nous utilisons les données de validation pour mettre en place une règle de décision efficace.

3. **Modélisation d'événement**
   Nous avons proposé un cadre pour recueillir les échantillons d'apprentissage automatique à l'événement modèle dans l'aspect visuel. La collecte est effectuée sur les médias sociaux grâce a l'analyse du contexte des événements. Les échantillons positifs sont collectées si ils contiennent des balises machine identifiant unévénement ou si les annotations contiennent le nom de l'événement abrégé. Les échantillons négatifs sont obtenusgrace a une approche derivée de la methode ń learning to rank ż. Dans nos travaux de recherche, les paramètres qui peuvent influer sur le classificateur final, comme le nombre de balises communes, le laps de temps et dans l'espace bounding box pour collecter des échantillons négatifs sont bien étudiés.

Pour résumer, dans cette thèse, nous avons présenté des methodes d'analyse de document multimedia pour l'illustration, la découverte, et la modélisation visuelle d'évènements. Toutes les approches proposées contribuent à extraire les interconnections entre les événements et leurs documents multimédias associés. La relation entre les médias sociaux et les événements est étudié avec rigueur et détails.

A.5 Perspectives

Ces dernières années, de nombreuse recherches ont été effectuées afin d'étudier la relation entre les événements et les médias disponible sur les réseaux sociaux et les plateforme sociales. Bien que ces recherches proposent de nombreux resultat interressant, il demeure encore des travaux futurs. Certaines directions et les améliorations possibles sont les suivantes:

1. **Les événements personnels/privés**
   Dans les plateforme web de medias partagés, la plupart des documents multimedia sont capturer lors d'événements personnels avec les gens proches, tels que anniversaire, mariage. L'exploitation et la gestion de ces événements personnels pourraient faire progresser la recherche basée sur l'événement de facons importante. Comment étudier le problème des événements personnels, sans compromettre la vie privée est toujours un défi ouvert.

2. **L'analyse de base de donnée de tres grande taille**
   Les problèmes d'analyse de très grande quantités de données sont une nouvelle tendance due a la croissance rapide des plateformes de partages de media.. Il y a une demande importante de solutions logicielles pour extraire l'information pertinente automatiquement dans les media sociaux. Comment appliquer l'apprentissage machine traditionnel et l'analyse statistique de problèmes à grande échelle est un nouveau défi à la fois dans la recherche et l'industrie.

3. **Vérité terrain**
   Lacréation de vérité terrain pour développer et évaluer les différents systèmes de recherche d'information est un processus essentiel. Cependant, dans la recherche basée sur l'événement, il n'y a toujours pas de données publiques fournissant des relations entre des événements et les données multimédias les caractérisants. Pour obtenir une vérité terrain de qualité et de taille suffisament important, , il semble interressant de s'appuyer sur une plate-forme comme Amazon's Mechanical Turk [2].

[2] `https://www.mturk.com`

Bibliography

J. K. Aggarwal and Q. Cai. Human motion analysis: a review. In *IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102, 1997.

P. Andrews, V. Murdock, A. Rae, F. De Natale, S. Buschbeck, A. Jameson, K. Bischoff, C. S. Firan, C. Niederée, V. Mezaris, and S. Nikolopoulos. GLOCAL: Event-based Retrieval of Networked Media. In *Proceedings of the 21st international conference companion on World Wide Web*, page 219, New York, USA, Apr. 2012.

K. Aouiche, D. Lemire, and R. Godin. Web 2.0 OLAP: From Data Cubes to Tag Clouds. *Lecture Notes in Business Information Processing*, 2009.

Y. Arase, X. Xie, T. Hara, and S. Nishio. Mining people's trips from large scale geo-tagged photos. In *18th ACM International Conference on Multimedia*, pages 133–142, Firenze, Italy, 2010.

T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis, and Y. Kompatsiaris. Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. In *Proceedgins of 5th International Workshop on Knowledge Markup and Semantic Annotation*, pages 1–10, 2005.

L. Ballan, M. Bertini, A. Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, Nov. 2010.

H. Becker, M. Naaman, and L. Gravano. Event Identification in Social Media. In *12th International Workshop on the Web and Databases*, Providence, USA, 2009.

H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Web Search and Data Mining*, New York, NY, USA, 2010.

H. Becker, D. Iter, M. Naaman, and L. Gravano. Identifying content for planned events across social media sites. In *ACM conference on WSDM*, 2012.

T. Berg and D. Forsyth. Animals on the Web. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1463–1470, June 2006.

M. Bertini, G. D'Amico, A. Ferracani, M. Meoni, and G. Serra. Web-based semantic browsing of video collections using multimedia ontologies. In *Proceedings of the international conference on Multimedia*, page 1629, New York, USA, Oct. 2010.

D. Billsus and M. J. Pazzani. A hybrid user model for news story classification. In *Proceedings of the seventh international conference on User modeling*, pages 99–108, June 1999.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, May 2003.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D Haussler, editor, *Proceedings of the fifth annual workshop on Computational learning theory*, number 8 in COLT '92, pages 144–152, 1992.

R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, New York, USA, June 2006.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—-27:27, 2011.

L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *ACM conference on CIKM*, 2009.

S. Chengjie and G. Yi. A Statistical Approach for Content Extraction from Web Page. *Journal of Chinese Information Processing*, 18(5):17–22, 2004.

T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece, 2009.

R. Datta, D. Joshi, J. Li, James, and Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2), 2008.

D. Delgado, J. a. Magalhães, and N. Correia. Assisted News Reading with Automated Illustrations. In *ACM conference on Multimedia*, pages 1647–1650, 2010.

N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122, Oct. 2010.

L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1959–1966, June 2010.

M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2009.

R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Tenth IEEE International Conference on Computer Vision*, pages 1816–1823, 2005.

A. Fialho, R. Troncy, L. Hardman, C. Saathoff, and A. Scherp. What's on this evening? Designing User Support for Event-based Annotation and Exploration of Media. In *1st International Workshop on EVENTS - Recognising and tracking events on the Web and in real life*, pages 40–54, Athens, Greece, 2010.

C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing order to your photos: Event-Driven ClassiïňĄcation of Flickr Images Based on Social Knowledge. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, page 189, New York, USA, Oct. 2010.

M. Gao, X.-S. Hua, and R. Jain. WonderWhat: Real-time Event Determination from Photos. In *20th World Wide Web Conference*, Hyderabad, India, 2011.

Y. Gao, M. Wang, Z. J. Zha, J. Shen, X. Li, and X. Wu. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. *IEEE transactions on Image Processing*, June 2012.

H. Glotin, Z.-Q. Zhao, J. Gao, and X. Wu. A matrix modular SVM robust to imbalanced data for efficient visual concept detection. In *Proceedings of the international conference on Multimedia information retrieval*, page 333, New York, USA, Mar. 2010.

J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

J. Hobbs and F. Pan. Time Ontology in OWL. W3C Working Draft, 2006.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, USA, Aug. 1999.

L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48, July 2010.

R. Hong, G. Li, L. Nie, J. Tang, and T.-S. Chua. Explore Large Scale Data for Multimedia QA. In *ACM conference on Image and Video Retrieval*, Xi'an, China, 2010.

M. J. Huiskes and M. S. Lew. The MIR flickr retrieval evaluation. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, page 39, November 2008.

J. Jiang, Yu-Gang And Ngo, Chong-Wah And Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501, 2007.

D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, page 37, New York, USA, July 2008.

D. Joshi, J. Z. Wang, and J. Li. The Story Picturing Engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing Communications and Applications*, 2(1):68–89, 2006.

L. Kennedy and M. Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *18th ACM International Conference on World Wide Web*, pages 311–320, Madrid, Spain, 2009.

L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *15th ACM International Conference on Multimedia*, pages 631–640, Augsburg, Germany, 2007.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems*, page 453, New York, USA, Apr. 2008.

J. H. Lee and X. Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, page 129, New York, USA, June 2012.

M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval. *ACM Transactions on Multimedia Computing Communications and Applications*, 2(1):1–19, 2006.

J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

L.-J. Li and G. Wang. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 88(2):1–8, 2007.

X. Li, C. G. M. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 10, 2010.

X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Social negative bootstrapping for visual categorization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, New York, USA, Apr. 2011.

C. Liangliang, L. Jiebo, H. Kautz, and T. S. Huang. Image Annotation Within the Context of Personal Photo Collections Using Hierarchical Event and Scene Models. *IEEE Transactions on Multimedia*, 11(2):208–219, 2009.

D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *18th ACM International Conference on Multimedia*, pages 491–500, Firenze, Italy, 2010.

T.-Y. Liu. *Learning to Rank for Information Retrieval.* 2011.

X. Liu, B. Cheng, S. Yan, J. Tang, T. S. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *Proceedings of the seventeen ACM international conference on Multimedia*, page 115, New York, USA, Oct. 2009.

X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *1st ACM International Conference on Multimedia Retrieval*, Trento, Italy, 2011.

D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, 1999.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. 1 edition, July 2008.

R. Mattivi, G. Boato, and F. De Natale. Event- based media organization and indexing. *Infocommunication Journal*, 3:9–18, 2011.

R. Mattivi, J. Uijlings, F. De Natale, and N. Sebe. Categorization of a collection of pictures into structured events. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 1, New York, USA, June 2012.

T. Mei, B. Yang, S.-Q. Yang, and X.-S. Hua. Video Collage: Presenting a Video Sequence Using a Single Image. *The Visual Computer*, 25(1):39–51, Aug. 2009.

P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, and A. F. Smeaton. TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID*. NIST, USA, 2011.

C.-C. Pan and P. Mitra. Event detection with spatial latent Dirichlet allocation. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, page 349, New York, USA, June 2011.

S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, and I. Kompatsiaris. Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation. In *MediaEval 2011 Workshop*, Pisa, Italy, Sept. 2011a.

S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-Based Landmark and Event Detection for Tagged Photo Collections. *IEEE Multimedia*, 18(1):52–63, 2011b.

T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, page 47, New York, USA, July 2008.

Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The Music Ontology. In *8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 103, New York, USA, July 2007.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77 (1-3):157–173, 2007.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *International conference on WWW*, pages 851–860, Apr. 2010.

F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the Web. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):754–66, Apr. 2011.

K. Schwarz, P. Rojtberg, J. Caspar, I. Gurevych, M. Goesele, and H. P. A. Lensch. Text-to-Video: Story Illustration from Online Photo Collections. In *Knowledge Based and Intelligent Information and Engineering Systems*, volume 6279, pages 402–409. 2010.

R. Shaw, R. Troncy, and L. Hardman. LODE: Linking Open Descriptions Of Events. In *4th Asian Semantic Web Conference*, 2009.

A. Singhal. Modern Information Retrieval : A Brief Overview. *IEEE Data Engineering Bulletin*, 24(4):1–9, 2001.

J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the seventeen ACM international conference on Multimedia*, page 223, New York, USA, Oct. 2009.

J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by k NN-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology*, 2(2):14, Feb. 2011.

J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua. Semantic-gap-oriented active learning for multilabel image annotation. *IEEE transactions on Image Processing*, 21:2354–60, Apr. 2012.

H. Toda and R. Kataoka. A clustering method for news articles retrieval system. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, page 988, New York, USA, May 2005.

M. R. Trad, A. Joly, and N. Boujemaa. Large scale visual-based event matching. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–7, New York, USA, Apr. 2011.

R. Troncy, A. Fialho, L. Hardman, and C. Saathoff. Experiencing Events through User-Generated Media. In *1st International Workshop on Consuming Linked Data*, Shanghai, China, 2010a.

R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *6th International Conference on Semantic Systems*, Graz, Austria, 2010b.

E. Tulving. Precis of Elements of Episodic Memory. *Behavioral and Brain Sciences*, 7(2): 223–238, 1984.

A. Vailaya, M. A. Figueiredo, A. K. Jain, and H. J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.

W. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Combining Ship Trajectories and Semantics with the Simple Event Model. In *1st ACM International Workshop on Events in Multimedia*, Beijing, China, 2009.

T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou. Video Collage: A Novel Presentation of Video Sequence. In *IEEE International Conference on Multimedia and Expo*, pages 1479–1482, 2007.

J. Weng and B.-s. Lee. Event Detection in Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, pages 401–408. HP Laboratories, 2011.

U. Westermann and R. Jain. Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, 14(1):19–29, 2007.

H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized LDA. In *Proceedings of the seventeen ACM international conference on Multimedia*, page 573, New York, USA, Oct. 2009.

Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. GRAPH-BASED SEMI-SUPERVISED LEARNING WITH MULTI-LABEL. *ACM Trans. Program. Lang. Syst.*, 20(5):97–103, 2009.

L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. *International Conference on Image processing*, 2(x):721–724, 2001.

Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the World: building a web-scale landmark recognition engine. In *22nd International Conference on Computer Vision and Pattern Recognition*, Miami, Florida, USA, 2009.

G. Zhu, S. Yan, and Y. Ma. Image Tag Refinement Towards Low-Rank , Content-Tag Prior and Error Sparsity. In *ACM Multimedia*, pages 461–470, 2010.

X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock. A Text-to-Picture Synthesis System for Augmenting Communication. In *Proceedings of the 22nd national conference on Artificial intelligence*, number 2, pages 1590–1595, 2007.