NTNU

Norwegian University of
Science and Technology

Event Image Detection with Knowledge Base Lookup

Hedda Nonstad

Master of Science in Computer Science
Submission date: January 2016
Supervisor: Herindrasana Ramampiaro, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Abstract

The growth in interest of sharing images on social media have brought new opportunities, but also resulted in challenges and needs. New developments for photo-sharing applications and new websites with the same purpose, are resulting in easy access and endless of gigabytes for uploading their personal images. The technical development increases the amount of information gathered when a picture is taken and this information is often attached when uploading a picture. Some of the more interesting data attached to a image are descriptive annotations, time stamps and locational information.

This thesis explores the correlation between tags, time and locational information to detect events occurring amongst the images. By clustering the data and searching DBpedia for events based on tags, the events emerges from descriptive information from user-based DBpedia pages.

My experimental results and the user-based evaluation have shown that the method used in the thesis detects events in a highly distributed dataset. The methods used in this thesis has great potential for improving the detection of events from data collected from Flickr. The system detects events in a wide range and the user-based evaluation indicates perfect matches with both tags and descriptions. The DBpedia semantics provides information to a full and correct identification of events in various size.

Sammendrag

Den økende veksten i interessen av å dele bilder på sosiale medier har gitt nye muligheter, men også resultert i utfordringer og behov. Nye utviklinger for bildedelingstjenester og nye nettsteder med samme formål, resulterer i en enklere og mer tilgjengelig prosess for å dele store mengder personlige bilder. Den tekniske utviklingen øker mengden av informasjon som blir lagret når et bilde blir tatt, og denne informasjonen blir ofte med når man laster opp et bilde på nettet. Noen av de mer interessante dataene knyttet til et bilde er beskrivende ord, tidsstempler og stedsmessig informasjon.

Denne oppgaven utforsker sammenhengen mellom tags, tid- og stedsmessig informasjon til å oppdage hendelser som forekommer blant bildene. Ved å gjennomføre grupperinger av dataene og søke DBpedia for hendelser basert på tags, blir hendelsene funnet basert på semantiske data og returnet med beskriveler basert på brukeroppdaterte DBpedia sider.

Mine eksperimentelle resultater og den brukerbaserte evaluaringen viser at metoden som benyttes i masteravhandlingen detekterer hendelser i et svært distribuert datasett. Metoden identifiserer flest private og ukjente hendelser, men gir også en beskrivelse av de mer populære og velkjente hendelsene i datasettet. De semantiske etikettene på DBpedia er gjennomførte og kan bidra til å detektere hendelser i et bredt spekter.

Preface

This master thesis is submitted to the Norwegian University of Science and Technology, NTNU, with specialization in Data and Information Management, in Department of Computer and Information Science. The thesis is the last part for fulfilment of the requirement for the degree of Master of Science.

In the in-depth specialization project at NTNU in the previous semester, I and my then partner, Marte Nordrik Hallan, conducted a project on novelty detection based on a paper by Zhang et al. [30]. The relevant theoretical parts within novelty detection and information retrieval will be presented here, since the same methods are used within the scope of the master thesis. The thesis is conducted with relation to the doctoral thesis Geo-Temporal Mining and Searching of Events from Web-based Image Collection of Massimiliano Ruocco [25].

Signed: ____

Trondheim, January 2016

Acknowledgements

The thesis work have been conducted with guidance of Associate Professor Heri Rampiaro at the Department of Computer and Information Science(IDI). He has given me much helpful guidance and feedback during the process of completing my masters degree.

Martin Nordal has help me bring new ideas to the experimental part of the project and especially debugging. I would also thank others related to the work I have done and those who have help me read through the thesis close to delivery.

Contents

Part I

Introduction & Motivation

Chapter 1

Introduction

In this chapter the master thesis is introduced. Section 1.1 explains my motivation to work within the area of clustering and linking knowledge bases with data from social media sites. In Section 1.2 the background theory is briefly introduced. The third Section, 1.3, introduces the scope and the limitations for the thesis and in Section 1.4, I discuss the problems and research questions I have investigated during this master thesis. And finally in Section 1.5 I outline the structure of this thesis.

1.1 Motivation

Development in digital technology have made image and information sharing much easier and accessible for users all over the world over time. Today's mobile phones all have a camera and the increasing interest in image sharing makes the demand for sites and applications that offer countless gigabytes of on line storage enormous. Some of the most popular sites are Panoramio[1], Instagram[2], Imgur[3] and Flickr[4], and the latter offers every user 1 terabyte of free space for uploading their own photo collections and the possibilities for the community to like and comment on their pictures. With a user base of over 112 millions and approximately 1 million photo uploads a day, the amount of pictures is endless. Based on these facts and the availability of data on their site, both images and metadata, I decide to use Flickr as a basis for the experimental part of this thesis.

Upon uploading images to Flickr, the user often adds descriptive information and Flickr adds additional metadata to the image. User based information may include tittle,

[1] hht://www.panoramio.com

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

[3] http://imgur.com/

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

description and tags. Based on settings on the camera, a number of parameters of the photo is uploaded as metadata, and for a similar tagging base, Flickr uses an automatic system for tagging uploaded pictures. The metadata is presented as EXIF[5] data, and can contain time photo is taken, GPS location, camera model, ISO and much more. The massive production of affordable GPS-enabled equipment installed in camera gear, and the lack of knowledge about it, are making the metadata about every pictures more interesting. With location data presented with longitude and latitude, a photographer can be tracked based on their uploaded photo collection down to the meter.

The amount of uploaded images on line is increasing rapidly and in line with this increase, available information is also tremendous. To inform the users in a proper manner, continuity and coordination among the websites is important. The use of semantic web technology will improve how the web is build together and become a more credible source. In the beginning of Wikipedia[6], the users could edit and write anything, and the site was not a safe source. To become the leading site of information, Wikipedia had to keep control over their site with more restrictions and administration approval for all editing. With these additions, a new way to expand the web has taken form. With use of Media Wiki[7] one can easily store and query data within Wikipedia or Semantic Mediawiki[8] which publishes the data through the semantic web and by that allowing other systems to use this data seamlessly.

The motivation behind this thesis is a composition of the above mentioned facts and the desire to be able to use the ignorance of users uploading their photos and knowledge to others. Based on simple photo uploading, a person can easily be tracked on the web without the knowledge and by tagging they are connected to possible events. It is these events I am looking for and by crawling Flickr for images and connecting them with the knowledge base, I want to identify them.

1.2 Background

In the work done by Rattenbury et al. [24], they present a way of modifying searching and the disambiguation problem based on semantic data from Flickr. They use a method called scale-structured identification and by using the tags, location- and time data, they investigated how to use the data to identify further information relevant to that place at a given time. By finding relevant information, a new expanded search can be performed with more relevant results.

[5] Exchangeable image file format,

[6] https://en.wikipedia.org

[7] https://www.mediawiki.org/wiki/MediaWiki

[8] https://semantic-mediawiki.org

Another usage of Flickr data was presented by Sigurbjörnsson and Zwol [28], and they investigated how to assist the users in the tagging phase. By analysing characterisation of how users tag their photos and the available information in the tags, they further present an effective recommendation system for the users when they are tagging their new photos. By looking at the user groups on Flickr, the system was most effective for the group who used few tags themselves, rather than the group not adding any tags or those adding almost too many.

In 2009 Kobilarov et al. [18] published a paper of how the BBC are using semantic web to link between their domains and to add additional information to their sites. When they started the process of linking music and programs together, they did not care for the problem concerning disambiguation between multiple vocabularies. As the problem grew, they needed a solution for linking to the right information. Their solution was to link to DBpedia with a label look up and context-based result disambiguation, resulting in a system with a more meaningful navigation paths across the whole web.

The above shows the general need to continue research in the field of event detection and contributions for further developing new ways of exploring the available data on line.

1.3 Scope & Limitations

The available amount of images and information indicates that there are great opportunities and challenges that can be researched. The main focus of this thesis is based on image sharing and event detection based on data from image sharing applications, such as Flickr.

The work done in this thesis encounters general problems with media sharing. I have identified some interesting possibilities based on experiences on Flickr and some obtained from related work and literature review:

- User based tagging can lead to various problems based on different applications. When users do upload images, countless tags may be declared and each of these do not necessarily mean the same or have the same importance for the various types of users. Tags can have ambiguity both within a language, but also across languages and dialects, the search based on tags may lead to different pictures not related or not relevant for the intended search.

  For example when searching for "Glittertind" on Flickr the majority of pictures are resulting in Norway's second highest mountain, but there are also a number of images of a band called Glittertind and a doll[9]. But the problem may not stop there. There are many users and the way they describe images have large deviations and the spelling is also in a wide variation.

- Issue of disambiguation between multiple controlled vocabularies are often spotted when exploring Flickr. Disambiguation can be eliminated by users using entity ambiguity or entity variation and technically by adding additional metadata to records for unambiguously identification.

  To address these issues the BBC, Freie Universität Berlin and Rattle Research are investigating how to use DBpedia to provide a common ”controlled” vocabulary and equivalency service, which in turn is used to add ”topic badges” to existing, legacy web pages [18].

The main challenges is the user-generated data, including the view of both the descriptive vocabulary used for tagging and the foundation behind the phraseology which can lead to disambiguation. The problem lies with the diverse user group, where everyone has a different view of the world and it reflects how images on Flickr are being tagged. Some users do not tag their pictures at all, some post a few tags and some have exaggerated many, thus the variability of relevant tags are great and many of these contribute to noise in the data. Some users are simply lazy, while others are unaware of what they photographing and what the purpose of tagging is. It is understandable that the general user do not understand why tags are an essential part of shared images, although knowledge concerning tagging have become more commonplace terminology and used in several forms of social media.

To summarize the scope of the thesis; cluster the Flickr collection based on time and location for event detection, and based on the dominant tag and semantic data, use DBpedia for information gathering and event detection. With this in mind, the following section describes the main problem addressed in this thesis.

1.4 Problem Definition

The increasing availability of information and the increasing heterogeneity of information sources have resulted in challenges related to extraction and identification of wanted data. With the advances in technology and use of various digital resources, a wide set of new information are uploaded to the web. A large share of the uploaded images to Flickr consist of geographical location data, timestamps and have related annotations. These

[9] https://www.flickr.com/search/?text=glittertind

annotations have a diverse range and there are possibilities to find relations between them and possible patterns in the associated noise. Seeing the media as a collection, they can represent a real-time event. By exploring the geographical and temporal raw data with association to the available textual data, the collection of resources, with the objective of improving search, browsing and detection of events, the main research question addressed in this thesis is as follows:

[RQ] How can we improve event image detection using knowledge base look up?

To solve the main research question for the thesis, I have chosen to divide it into four sub-questions.

[RQ1] How can an event be detected based on metadata stored for images on Flickr?

[RQ2] What is the impact of cleaning the dataset?

[RQ3] How to identify the correct tags in a given cluster, for further event identification?

[RQ4] What contributions can DBpedia provide in identifying events?

1.5 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 - Background: An introduction to relevant theory within information retrieval, data mining and on line picture handling.

Chapter 3 - Related Work: A survey of related work in the field of entity linking, knowledge bases, event detection and more.

Chapter 4 - Methodology The methods used in the experimental part of the thesis are presented.

Chapter 5 - Experiments: Contains the evaluation methodology and result of the experimental work.

Chapter 6 - Discussion & Evaluation: The evaluation and discussion for the findings of the thesis is presented.

Chapter 7 - Conclusion: A final conclusion and suggestions for further work.

Chapter 2

Background

In this chapter I introduce standard theory for information retrieval, data mining and relevant theory for image handling. The chapter provides background information and the foundation of the work done in this thesis. The information is not necessarily directly used, but is intended as a comprehensive introduction to the field of study and to be used as a base layer for subsequent work.

Section 2.1 introduces tags and tagging related to the thesis and in Section 2.2 I outline the basic concepts of information retrieval. Section 2.3 provides information to text mining and related subjects, Section 2.4 introduces the concept of DBpedia, and lastly Section 2.5 introduces crowd sourcing.

2.1 Tags & Tagging

The Oxford Dictionary[1] defines a *tag* as a label attached to an object for a purpose of identification or to give information, and *tagging* as adding characters to an object for information or categorization. Taxonomic classification is the original way to specify the contents that was published by the owner only, but now folksonomy is used, it is where users apply tags to on line items for re-finding and information.

There are different kinds of tags used in social media. The machine tag is used for better and more meaningful handling of tags, a special syntax is used to define extra semantic information of the given tag. A system who adds machine tags often use present information to add the new ones, based on available metadata. A type of machine tags is geotags, adding geographical identification to an object. Another name for machine tag, is triple tag, consisting of three parts: namespace, predicate and a value. E.g.

[1] http://www.oxforddictionaries.com/

"$geo : long = 50.124235$", a term added to Flickr to formalize how tags are treated. Another form of tagging is user based and called hashtags, because the word is initialized by the character "#". In some social media can users add additional tags to other posts added by other user for a more comprehensive informational purpose.

There are many positive improvements done by tagging; better search, spam detection, reputation systems and personal organization. On the other hand are there some challenges operating with user based tags. The lack of knowledge to association to unstructured ontology's, linguistic and grammatical variations and typing errors to mention some. Tags are a reflection of how the users sees and observes the world from a personal perspective, and not normalised [16, 20].

2.2 Information Retrieval

Tagging is a central part of the thesis and to use it in a orderly manner, Information Retrieval is an important foundation for the extraction, usage and evaluating the results. Information Retrieval, IR, is the field based on collections of either documents or other information giving resources and the task of retrieving relevant information based on queries. Most IR systems share a common architecture, where search is a central task, but the field covers a wide range of processes, including storage, manipulation and representation [5].

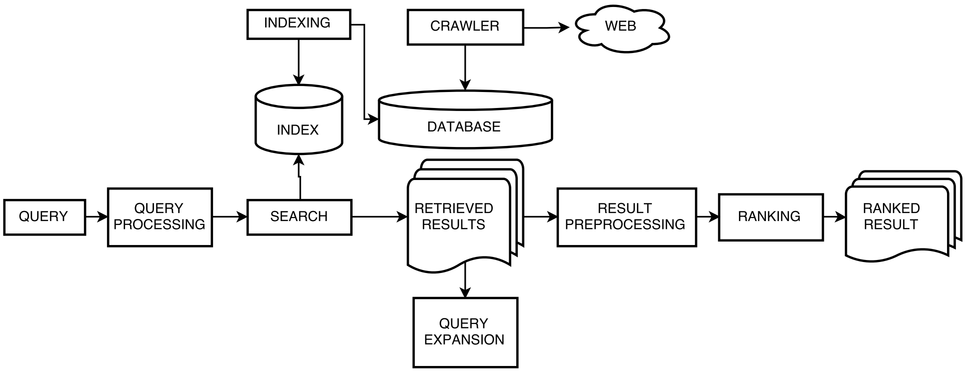


FIGURE 2.1: Architecture of a General Information Retrieval System

An IR system is a complex process consisting of many components, that can be modified to specify different scopes. A general architecture of an IR system can be viewed in Figure 2.1. The system presented for the user must be efficient and effective, giving the user relevant information in no time.

In the following subsections some of the parts of the system presented in Figure 2.1 is described more, including issues and challenges.

2.2.1 Web Crawling

A web crawler crawls for documents or other wanted information belonging to a wide range of topics. To specify the crawler, a set of limitations can be given. That can be available pages, specific topic, publishing time or other data. The crawler usually starts out with an URL, then collect all URLs on that page, adding them into the queue for visitation and repeats the process.

Occurring problems with crawlers are freshness, efficient resource usage and duplicate result objects. Freshness represents the fraction of documents or pages that is up to date in the collection. Efficient use of available resources is related to response time and use of minimal resources as CPU time, I/O time, network bandwidth, memory and disk space. The issue with duplicate documents is the increasing size of databases as the search engines index every single web-page and crawlers visit billions of pages every day and much of the time used on irrelevant duplicates [2, 9, 19].

2.2.2 Indexing

An index is a data structure mapping search concepts after occurrence, to speed up the search. To evaluate the efficiency of an index, five measures can be used. The time used to build the index, spaced used during generating, space used for storage, time spent producing result for an arriving query, and average queries processed every second.

Inverted file structure is the most efficient index structure used on textual queries. An inverted file has two major part; (1) a search structure, containing all of the distinct values being indexed, and (2) for each distinct value an inverted list, storing the identifiers of the record containing the value [31].

For representations of strings, the suffix tree structure is widely used. Every suffix in the string is presented as nodes, suitable for pattern matching and longest common substring problem. The tree consists of n leaves and every internal node has at least two children, giving the total size of the tree $O(n)$.

The main focus for a search engines is crawling, indexing and sorting. To be the leading search engine on the marked, the hours spent optimizing the processes is crucial. Google crawlers crawl 48,5 pages a second and manages to index the pages slightly faster. A goal for the developers was to create an index fast enough to not create a bottleneck in the process [4].

2.2.3 Preprocessing

Within the information retrieval domain, the preprocessing step is often used to make a common ground of similarity within the textual documents. Modern Information Retrieval [2] have split preprocessing into five actions. The first one is a lexical analysis of the text, meaning treating digits, hyphens, punctuation marks, and the case of letters. The second step is elimination of stop words, meaning filtering out words with very low discrimination value. Then there is stemming, removing affixes and allowing the retrieval of documents containing syntactic variations of query terms. The forth action is to select the index term to determine which words will be used as indexing elements. The last step is the construction of term categorization structure, thesaurus, or extractions of structure directly represented in the text.

2.2.4 Retrieval Evaluation

In related work in the field of information retrieval there is a various number of possible methods relevant for the numerous tasks for extracting information and ways to evaluate the given results. The most common methods for measuring the retrieval is precision, recall and accuracy.

The confusion matrix in Table 2.1 reports the number of true positives, false positives, false negatives and true negatives, after finished classification. The distribution of the results within these categories allow a more detailed analysis.

	Relevant	Non-relevant
Retrieved	true positive(tp)	false positive(fp)
Not Retrieved	false negative(fn)	true negative(tn)

TABLE 2.1: Confusion matrix

Precision and Recall

Precision is the fraction of relevant retrieved pictures, or the probability that a new picture processed is relevant. Recall is the fraction of relevant instances that are retrieved, or the probability that after a search, a relevant image is retrieved.

Both precision and recall is a combination of the same components, and they are defined as followed:

$$Precision = \frac{|R \cap A|}{|A|} = \frac{tp}{tp + fp} \tag{2.1}$$

$$Recall = \frac{|R \cap A|}{|R|} = \frac{tp}{tp + fn} \tag{2.2}$$

Where $/R/$ is the total number of relevant images, $/A/$ is the number of retrieved images, and $|R \cap A|$ is the intersection between the to sets of pictures.

Accuracy
Accuracy gives the fraction of the classified images that are correct. In other words, the number of true results (true positive and true negative) of all of the examined cases.

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{2.3}$$

2.3 Text Mining

Text mining can be seen as a form of machine learning and based on given input data, the processes is to design and develop algorithms that learn patterns present in the data. The machine learning area is mostly dependent on learning for further processing to evolve methods who encode patterns. The process of evolving these methods are split into two main areas; supervised and unsupervised learning. Supervised learning algorithms needs a set of training data as input and requires to learn a function from it, to succeed and on the other side there are unsupervised methods who do not need any training.

With methods related to textual evaluation, the problem with ambiguous word will occur and all of these topics will be further introduced in the following subsections and a subsection with usage of extracting information will end the section.

2.3.1 Supervised Learning

Supervised learning is the form of text mining called classification and supervised learning algorithms work to classify every available document, based on a preclassified user-based training set. For the algorithms to work perfectly it is important that the training data are diverse, covering all cases, and the desired solutions are provided. The methods are later used to predict outcomes from new and unseen data, similar to the training set.

Based on supervised learning principles Carneiro et al. [6] have proposed a probabilistic formulation for semantic image annotation and retrieval. With a database of images labeled with a shared semantic label defying the classes, the formulation is posed as a classification problem. The images are represented as bags of localized feature vectors, a mixture density estimated for each image, and the mixtures associated with all images annotated with a common semantic label pooled into a density estimate for the corresponding semantic class. The result of the research outperformed previous methods and had an excellent performance when semantic querying was conducted.

Another study in this field was presented by Martin et al. [21] and their goal was to accurately detect and localize boundaries in natural scenes using local image measurements. They formulated features to respond to a set of characteristics associated with natural boundaries; change in color, brightness, and texture. To train the classifier they used human labeled images, to combine the information in the features in an optimal manner. Based on the changes in image features they managed to outperform other similar methods and concluded the paper with saying that proper treatment of texture is an essential part of detecting boundaries in natural images.

There is a wide set of available classification algorithms and most of them are usable for specific cases of input data and with a desired solution of a given problem. Some examples of classification methods are decision tree, support vector machine classifier, and k-nearest neighbor classifier. The latter classifies document based on a distance function in a predefined metric space, by classifying the k-nearest documents to a centroid as the same class.

2.3.2 Unsupervised Learning

Unsupervised learning algorithms for classifying text are either clustering or naive classification, and are doing so without any information regard to training examples. According to the book *Modern Information Retrieval* by Baeza-Yates [2] a cluster is defined as "A method used do dived a given set of documents D into a wanted set of K clusters according to some predefined criteria". Results after clustering are often difficult to interpret, but a result allows insight into the data and natural properties within the set.

A proposed unsupervised leaning technique, Probabilistic Latent Semantic Analysis, by Hofmann [14] is developed to identify and distinguish between different contexts of word usage without redirecting to a dictionary or a thesaurus. By doing so the method can differentiate disambiguous words and reveal topical similarities by grouping together words that are part of a common context.

Over the years multiple clustering algorithms have been developed in a variety of domains for different types of applications and data. None of the algorithms are suitable for all types of data, clusters or applications, and often another algorithm can be more efficient, but there is always compromises amongst them. In fact, many clustering algorithms have time or space complexity of $O(m^2)$, where m is the total number of objects, resulting in pore performance when the dataset grows [22].

2.3.3 Disambiguation

The problem with disambiguation is briefly presented in Section 1.3, and can further be described as determination of the sense of a word. Word ambiguity is not something we normally experience, when ambiguous word occur in sentences the brain will choose the right meaning, but the process for computers are more complicated. Yarowsky [29] presented an algorithm determining the sense of a word by looking at the nearby words which provide a strong and consisting clue of the right meaning. The clue includes the relative distance, order and syntactic relationship between words within a specific document the word has often one meaning.

When searching the web, the problem with disambiguation often occurs. Another approach to solve the disambiguation problem is presented by Cucerzan [10]. An example of the disambiguation problem from his paper is presented here; *For queries such as "Bush", search engines return a various set of possible persons and other relevant objects with the name. Bush could be the two former U.S presidents, Jeb Bush (candidate for the presidential election in 2016), Reggie Bush(an American football player) or a rock band balled Bush.* He present a large-scale system for recognition and semantic disambiguation of name entities based on information extracted from Wikipedia. In his solution to solve disambiguation he linked entities together either if the surface form was an exact match, or else he chose the entity which was most frequently mentioned in Wikipedia and used its surface form.

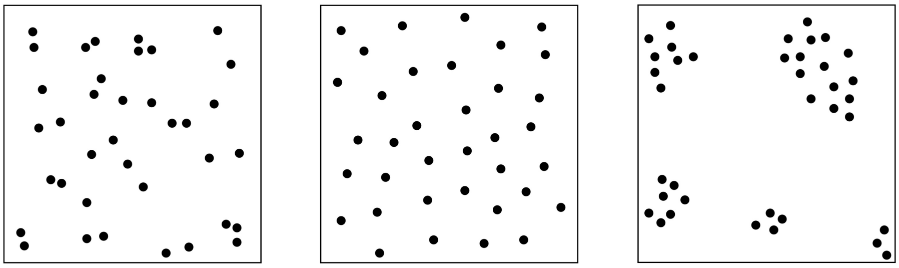
2.3.4 Information Extraction

Information is distributed and collected in a range of forms, from text documents, images, videos, sound etc. A part of the mining field is to extract information from a source, not seen when only one element is evaluated. The information available when seeing multiple objects as a whole, and finding patterns in these objects, can be of great importance and bring new relevant information. Related to this thesis is pattern extraction, based on spatial positioning of the gathered images.

Spatial Point Pattern

Analysis of spatial points aims to explore structures and patterns present in points distributed in two- or multidimensional spaces. The first methods for analysis of spatial points can be categorized into two groups, distance-based, using mean distance to nearest neighbour and area-based, relied on frequency distribution characteristics on sub-parts of the total area [12].

With a set of points in space, the different structures and correlations may happen between points. In Figure 2.2 three categories for this phenomenon is shown. Complete spatial randomness is where points are distributed independently in space, or they are regularly distributed or clustered. Regular positioning is where the distribution is equally spread in space, and clustering is where points are grouping together.



(a) Random (b) Regular (c) Clustered

Figure 2.2: Different spatial structures

2.4 DBpedia

The DBpedia[2] site is a crowd source community and is a structured information extraction from Wikipedia. The community is trying to link different data sets to Wikipedia by sophisticated queries, and to find new ways to use the huge amount of information available on Wikipedia in better ways. In 2014 DBpedia released a set of 3 billion Resource Description Framework(RDF) triples, where 580 million was extracted from the English Wikipedia. RDF, is a conceptual description or modeling method for information, with use of a various syntax notations and data serialization. And in the English DBpedia set there are 45 377 RDF triples in the event ontology. An ontology is a formal naming and definition of types, properties and interrelationships of entities.

2.5 Crowd Sourcing

User studies is an important factor for any success and input in every stage can substantially improve the product. Feedback from the crowd can come in different forms; surveys, usability test, rapid prototyping, cognitive walkthroughs, quantitative ratings,

[2] http://dbpedia.org/about

and other performance measures. Crowd sourcing is not always free, but the cost is increasingly less than having employees doing the same job. Crowd sourcing is a new way to outsource tasks too expensive for companies to do in house or in low-cost countries [17].

Amazon have developed a crowd source market, Mechanical Turk[3], where users can post different task they want done and a price for doing them. The system encourages the crowd to fulfil tasks that would be too difficult for computers to perform. Some of the tasks a human can conduct more optimally than a computer are identification of objects in images or find relevant information in document collections.

[3] https://www.mturk.com/mturk/welcome

Chapter 3

Related Work

In this chapter a survey of related work is presented and related work to this thesis is introduced. In Section 3.1 a general presentation of entity linking is described and in Section 3.2 the challenges of how to populate knowledge bases are presented. The Section 3.4 introduces the knowledge base acceleration research and Section 3.5 describes principles with automatically tag extension and generation.

3.1 Entity Linking

The size of the web is producing challenges for all parties working on the web. A solution for combining knowledge from a wide set of different sources, is the use of entity linking. Entity linking is the reference between an entity mentioned in textual data, to their representation in a structural knowledge base. These knowledge bases harbor a rich knowledge of entities, their semantic properties, and the semantic relationships between one another.

The main problem with entity linking is the ambiguation of words, because an entity can have multiple mentions in the knowledge base and a mention can be shared among different entities. The crucial element of entity linking is the process of disambiguation, which is briefly explored in Subsection 2.3.3 [13, 23].

3.2 Knowledge Base Population

The ability to recognize entities, extract their attributes and identify entity relations is the definition of knowledge base population. The process is often seen as two separates sub-tasks; (1) entity linking and (2) slot filling.

In the work presented in [11], Dredze et al. see three challenges for populating the knowledge base. (1) Name variations, (2) solving the disambiguation problem, and (3) learning when there is no result(NIL). To solve the first challenge, they introduced five rules for linking entity names and name variations, from exact match to strong string similarity score. Disambiguation is handled by a learning machine and several possible features for name variations. To learn the NIL predictions, they use a support vector machine ranker by including NIL in the argumentation.

The National Institute of Standards and Technology(NIST)[1] community and their yearly Text Analysis conference(TAC)[2] have in the later years focused on temporal slot filling for the knowledge base population track. Temporal slot filling requires the developed system to discover temporally bound facts about given entities and extract the values of their attributes/slots to fill the structured knowledge base, the temporal part is to extract the start and end dates for the slots. The research done in the NIST community increases the development done in the field, resulting in new and creative ways of populating the knowledge bases.

3.3 Event Detection

The topic of event detection came originally from the field of topic detection and tracking(TDT). Event detection can be performed on various medias, and the use of images for event detection can be seen as a recent form of TDT [1].

Extracting information from textual document streams is a popular field of study. The microblogging service, Twitter[3], is often used in the later years after the increasing usage and posting rates. Sakaki et al. [26] present an algorithm detecting events with use of a classifier and a probabilistic spatio temporal model to detect events related to earthquakes based on Twitter messages.

With the wide range of available media sharing sites, the event detection field have started using other media forms than textual documents. Based on collections of crowd videos Ke et al. [15] have used real time actions sequences to detected events. They filmed training sequences and extracted event models, and used the models to detect the same pattern in crowd videos.

In [7], Chen and Roy look at the user tags, temporal and locational distributions to analyse pictures on Flickr to determine relevance to occurring events. They analysed

[1] http://www.nist.gov/

[2] http://www.nist.gov/tac/

[3] https://twitter.com/twitter

temporal and locational metadata distribution of tags to discover event-related tags with a significant distribution pattern in both temporal and locational dimension.

3.4 Knowledge Base Acceleration

Knowledge Base Acceleration is the way of recognizing new information in a stream of information related to an entity. The new information retrieved in the stream should be queried and constantly updated its entity in a knowledge base. In regard to the knowledge base population, is this a real time updating problem.

One of the most used web pages on the world wide web, is Wikipedia.com and it is an example of a knowledge base. A companion to Wikipedia is DBpedia[4], and it has advanced queries against Wikipedia and links huge amount of data on the web to Wikipedia for common structure. The goal of DBpedia is to ease the way of sharing the available data through a common semantic form.

When people are updating sites or adding new information to the web, knowledge base acceleration wants to update the knowledge base in zero time with this information. And by conducting this update, the other pages linking to the knowledge base will be updated in the same time.

3.5 Automatic Tag Extension & Generation

Tags and the purpose of tagging web pages and social media post is described in Section 2.1, but user generated tags may cause some discrepancies. Not every user have the same views and the same vocabulary to make the tags a shared ground, therefor automatic tagging may be a solution. Based on different parameters or techniques, the software can produce new tags with a shared basic perception.

Shen et al. [27] have developed a method for automatically tagging videos based on geographical properties. The increasing amount of sensors available on today's cameras are making the location and orientation of a camera available for continuously acquiring. Their technique utilizes these sensors and their appurtenant metadata to automatically tag outdoor videos, firstly based on viewable geographical objects by querying geo-information through viewable scene descriptions, then extracting tags from databases based on the data. Secondly they score the new tags against the obtained tags based on

[4] http://wiki.dbpedia.org/

relevance from the given segment of the video. Based on the result from a conducted user study, they developed a prototype tag generator for outdoor videos from the research.

Another way to generate tags was introduced by Chirita et al. [8] and their technology generated personalized tags, P-TAG. When a user is browsing a web page, the P-TAG generates keywords relevant both from textual content and data residing from the users desktop, giving the personalized viewpoint. They believe these tags with large scale metadata annotations for web pages can provide for an important step towards the realizing of the semantic web.

Part II

Approach

Chapter 4

Methodology

This chapter presents the methods and basic principles used for the basis for the experimental work done on this master thesis. Section 4.1 the used Flickr crawler is presented and the following Section 4.2 presents how the clustering process is implemented. In Section 4.3 the used data, how the information available is used, and how the information from DBpedia was used to identified an event is presented.

4.1 Flickr Crawling

To gather the wanted images and corresponding data, a Flickr crawler was developed and used. In the first experiments with the crawler, an argument was used and the result was pictures containing the given argument as a tag with wanted metadata. After further testing, the crawler was changed to find all images containing geographical information, so the complete dataset used is a collection of images from all around the world. The restraints for images in the collection is that they contain geographical location and a time stamp, but there was no requirement for a photo to have a collection of tags.

To avoid redundant images in the directory,the crawler is checking if the image is already in directory, and if it is the image is skipped. The importance here is to save all downloaded images in the same directory. All images is saved in locally within the workspace and metadata is printed to a file after wanted design. The metadata is further processed before the strings are used for clustering.

4.2 Clustering

In Subsection 2.2.2 an introduction to indexing and an example of a good method was presented. In the experimental part of the thesis, I did not implement the desired solution mention in the Subsection 2.2.2 for string handling, but rather a more simple variant. Based on the data and the spread in data points, the need for a complex indexing system was not present.

4.2.1 Clustering Process

To cluster the data strings a logical method for getting the top most frequent elements was implemented. The implementation consists of a class with map storing the strings and their representative count, an interface for that class to sort the string by count and alphabetically, a new list of the previous map where the entities are added and sorted, returning the top n items by using a sublist.

4.2.2 Cluster Scope

For detecting events with a variety in time and location data, there are some possibilities to how the limitations is set for these measurements.

Time
For detecting events the time is an important factor. It is possible to put every image with metadata-time down to the second for further processing, but the possibility to finding an event for that second is minimal. After looking through the data and the wide spread of it, I decided to use days as a limitation when setting restriction regarding time. Most of the pictures relevant to an event, are taken within a day. The day is further presented as the date the image was taken.

Geographical
Location is a central part of detecting an event. If the range is too wide, the possibility of disambiguating events is increasing, but set the range too small, nothing will occur. In Table 4.1 the range for the geographical coordinates is displayed. Most of the data collected from Flickr are presented with five to six decimals, not so relevant for further detections. For the event detection system presented here is two, three and four decimals used.

Decimals	Range	Qualitative scale
0	111km	country/large region
1	11 km	large city
2	1km	town
3	100m	neighborhood
4	10m	street
5	1m	human

Table 4.1: Decimal range of geographical coordinates

4.3 Event Extraction

To handle image data, a set of processes is needed for further identification of any present events. The data must be rewritten to a desired output and cleaned for all unnecessary noises. The most frequent of the new output strings must be detected and further looked up on DBpedia for a clarification if there is a present event.

4.3.1 Input Data

The collection of images downloaded from Flickr represent my dataset and can be presented as $\mathcal{F}$. We can assume that every picture in the dataset contains different kinds of textual information related to the picture presented as metadata. In addition, the temporal information for *when* and geographical information for *where* the picture is taken. Hence each image I $\in \mathcal{F}$ is represented as

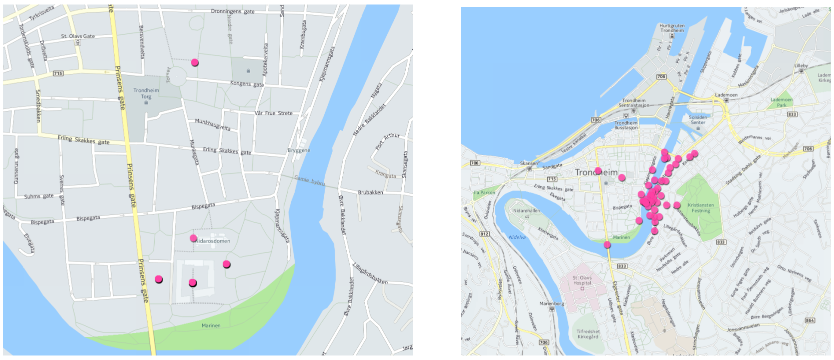
$$I = \{d_t, g, \mathcal{T}\} \tag{4.1}$$

where d_t denotes the date and time the picture is taken, $g = \{lat, long\}$ is the pair of numbers representing latitude and longitude, and $\mathcal{T}$ representing the set of annotations for I. It can be assumed that every image downloaded from Flickr included in the dataset contains all this information.

4.3.2 Image Occurrences

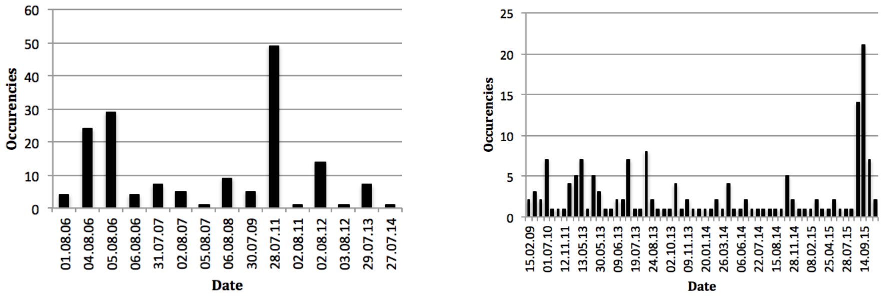
Image uploading to Flickr occurrence every millisecond with attached information can the image be presented by both temporal and spatial distributions. In the desire of clustering images with relation to a shared happening or event, these distributions are interesting. An event can be characterized by time and geographical information and relevant tags as parameters.

An example is presented in Figure 4.1 and 4.2 over the tags `Olavsfestdagene` and `Bakklandet`. The same amount of images is used for both tags and illustrate the different possibilities when looking at an event tag and a location tag. `Olavsfestdagene` present an reoccurring event in Trondheim and by the Figure 4.1a the spatial distribution is only in five places in the city, compared to `Bakklandet` in Figure 4.1b where almost every picture is in a new place. By the temporal distribution of the tag, the picture posted with the tag `Olavsfestdagene` only occur three days every summer seen in Figure 4.2a, in comparison to the `Bakklandet` tag which can be uploaded all year long seen in Figure 4.2b.



(a) `Olavsfestdagene` (b) `Bakklandet`

Figure 4.1: Spatial image distribution for the tags `Olavsfestdagene` and `Bakklandet`



(a) `Olavsfestdagene` (b) `Bakklandet`

Figure 4.2: Temporal distribution for the tags `Olavsfestdagene` and `Bakklandet`

In Subsection 2.3.4 a set of possible spatial point pattern where shown as an example, and in the example shown in the Figure 4.1 the first mentioned example is shown in a real example. Where there is an event occurring the points are aligned as a an example of Figure 2.2c, while there are random uploaded images there is a spatial clutter as in Figure 2.2a. The `Bakklandet` tag can be seen as a cluster when only looking at it as an geographical group, not surprisingly when the search is a geographical place, while `Olavsfestdagene` is an event.

4.3.3 Cleaning

When user based tags are the base for a similarity comparison, some textual cleaning is necessary. A lexical analysis is done and tags are presented as individual terms for further processing. The second step in preprocessing is stopword removal, when tagging a picture the use of stopwords is low, but can be relevant if the occurrence of words e.g. "is", "are" and "I" is present with a high percentage. A step that may cause some difference is stemming, removing of affixes and allowing the retrieval of tags containing syntactic variations of same terms.

The first step after acquiring the dataset is a preprocessing step, mainly consisting of cleaning the image annotations. The image represented as $I = \{d_t, g, \mathcal{T}\}$ is now transformed into $I' = \{d_t, g, \mathcal{T}'\}$, where all terms are cleaned. To start the cleaning every annotation is converted to lowercase letters and all words containing a diacritic sign are removed for easier handling. Further are all digits removed from the annotations, even though they can contain temporal information, they are not useful in the clustering process. Next, many annotations from Flickr images contain semantically irrelevant terms, or they are so common that they do not contribute to a real value for later clustering. For example camera names like `Canon`, `Nikon`, or other terms like `photography`, `instagram` and `geotag` are very common and may safely be removed.

The additional terms are considered as non-relevant and not contributing for later knowledge base lookup. The set of additional words was divided into four classes: (1) temporal terms, (2) camera related terms (3) top non-relevant terms and (4) general noise. The terms came from the most frequent used tags from Flickr's own overview[1] and personal observations of images and attached annotation in the database. The top k non-relevant terms list was an subjective evaluation of the given terms presented, and they was considered not relevant for further event recognition. The list of terms removed from the annotations are listed in Figure 4.2.

Within the temporal terms removed from the set, I did not include the month May, June and July, when I thought they may lead to relevant events.

The presented list of word in Table 4.2 are removed with three different methods. The first one removes all words starting with a given set of characters, e.g. `geo` are also removing `geotag` and `geotagged`, second method removes all tags containing a given set of characters, the third method only removes words if there is an exact match between the given word in the table and the tag. All of the camera related terms are either removed with the first or second method, when many of the tags contain a sequence of characters

[1] https://www.flickr.com/photos/tags/

Temporal terms	january, february, march, april, august, september, october, november, december, summer, winter, autumn, fall, spring, week
Camera related terms	instagram, square, squareformat, nikon, sony, fuji, pentax, canon, olympos, fisheye, exposure, longexposure, photo, camera, macro, eos, portrait, filter, iphone
Top non-relevant terms	unitedstates, france, nature, landscape, england, travel, italy, sky, taiwan, germany, usa, japan
General Noise	geo, uploaded, flickr, foursqaure, outdoor, indoor, exif, follow, test, ???

Table 4.2: Extended stopword list

after those given terms, e.g. `canonfullframestreetphotography`, `mynikonlife` and `photoshop`

The crawler is not restricted by any locations, so there is a wide set of languages among the tags. The majority is in English, so there is no further procedures performed to rectify tags with diacritics, since the lookup on DBpedia will only happen in English.

4.3.4 Event Detection from DBpedia

To determine if an event is among the most frequent clusters, based on the metadata from the dataset and the representation of images, $I = \{d_t, g, \mathcal{T}\}$, a lookup on DBpedia is performed. The tag is extracted from the string and is searched for among the semantic labels present in the RDF information stored in every DBpedia cite. The process of finding events among the clusters is shown in Figure 4.3.

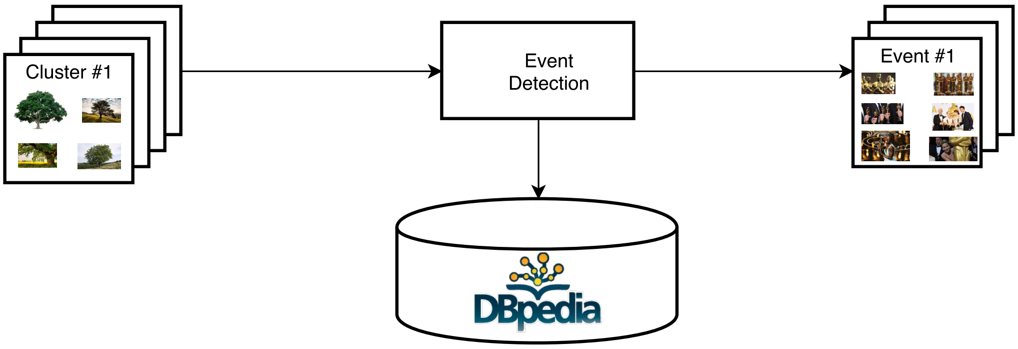


Figure 4.3: The process of detecting events with DBpedia

To differentiate between events and non-relevant DBpedia pages, the tags must be labeled by DBpedia as a relevant event definition. The list used for this differentiation is presented in Table 4.3

Events	event, award, festival, concert, conferences, ceremonies, convention, happening, holiday

Table 4.3: Labels from DBpedia defined as an event

The semantic labeling on DBpedia is evaluating every tag and based on the given label can determine if there is an event for the given tag. This will solve the problem with disambiguation among words where one of them are an event while another may be something totally different, e.g. *OSCAR*, disambiguate with approximately 60 pages on Wikipedia, where the academy awards are an event and the orbital satellite carrying amateur radio also is referred as an Oscar. Following is the first result presented on the DBpedia site presented for the term *OSCAR*[2]:

```
<Result>
<Label>Academy Award</Label>
<URI>http://dbpedia.org/resource/Academy_Award</URI>
<Description>
An Academy Award is an award bestowed by the American Academy of
Motion Picture Arts and Sciences (AMPAS) to recognize excellence of
professionals in the film industry, including directors, actors and
writers. The Oscar statuette is officially named the Academy Award of
Merit and is one of nine types of Academy Awards.
</Description>
<Classes>
<Class>
<Label>award</Label>
<URI>http://dbpedia.org/ontology/Award</URI>
</Class>
<Class>
<Label>owl#Thing</Label>
<URI>http://www.w3.org/2002/07/owl#Thing</URI>
</Class>
</Classes>
<Categories>
<Category>
<Label>Awards established in 1929</Label>
<URI>http://dbpedia.org/resource/Category:Awards_established_in_1929
</URI>
```

[2] http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryString=oscar

```
</Category>
<Category>
<Label>Academy Awards</Label>
<URI>http://dbpedia.org/resource/Category:Academy_Awards</URI>
</Category>
<Category>
<Label>American film awards</Label>
<URI>http://dbpedia.org/resource/Category:American_film_awards</URI>
</Category>
</Categories>
<Refcount>12260</Refcount>
</Result>
```

In the presented example above there are multiple labels present for the first result and in this example all of them are relevant. In other results there are some label not so relevant for the result and often are non relevant result presented for the result as well. Some examples of other results in the same result page in DBpedia as the example above is: Rey Mysterio(Mexican American professional wrestler), Un ballo in maschera(an opera) and Arthur Honegger(Swiss composer). What they all three have in common is that Oscar is mentioned in the representative articles so that it can be a indirect relevance to the keyword, but not in the direct sense for a relevant result.

Chapter 5

Experiments

In this chapter the experimental setup for this thesis is explained and the following result from the experimental part are presented. In Section 5.1 the evaluation methodology is presented and with that the experimental setup, information regarding the used dataset, results from cleaning and how the data was transformed from raw data to clusters. Section 5.2 contains all result collected from the experiments done.

5.1 Evaluation Methodology

The system is presented in Figure 5.1, the figure is an overview of the architecture. The first steps include gathering and cleaning of data, followed by cluster construction and event identification, and lastly refinement and merging before a number of clusters can be returned.

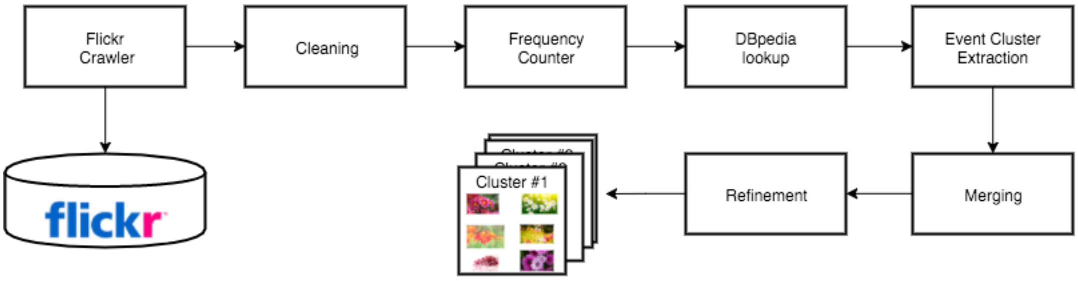


FIGURE 5.1: Overview of the system

In this section I present how the experiments was conducted and basics for the conduction, secondly the facts regarding the dataset and thirdly the effect of the cleaning process. Lastly the use of the metadata is presented and how they were transformed to usable strings.

5.1.1 Experimental Setup

To conduct the experimental work in this thesis, I have used a Dell computer with Windows 7 from the University and a private mac book pro. For the development of the presented method in Chapter 4 Eclipse Mars was used on both computers.

When Eclipse is sorting and counting the frequency of the approximately 1 million strings the amount of time used was from 11 seconds to 14 seconds based on the number of decimals used for the geographical coordinate data. When an additional lookup to DBpedia was added to the task, a predestined number of strings decide how many of the top clusters are looked up on DBpedia. For identifying approximately 60 events on DBpedia, the predestined number is set to 1000 and the time used for this operation is 151 seconds approximately 2.5 minutes, when the number of decimals are four. The results for this example is shown in Figure 5.2, where every detected event, returned by a DBpedia description for a given tag, is displayed by a green line. The chart displays the top 1000 clusters and which of the top thousand clusters an event is detected. The Figure shows there is no correlation in where in the top thousand clusters there is an event and the possibility of finding of events can be from cluster #1 to cluster #million, it is all up to the tag.

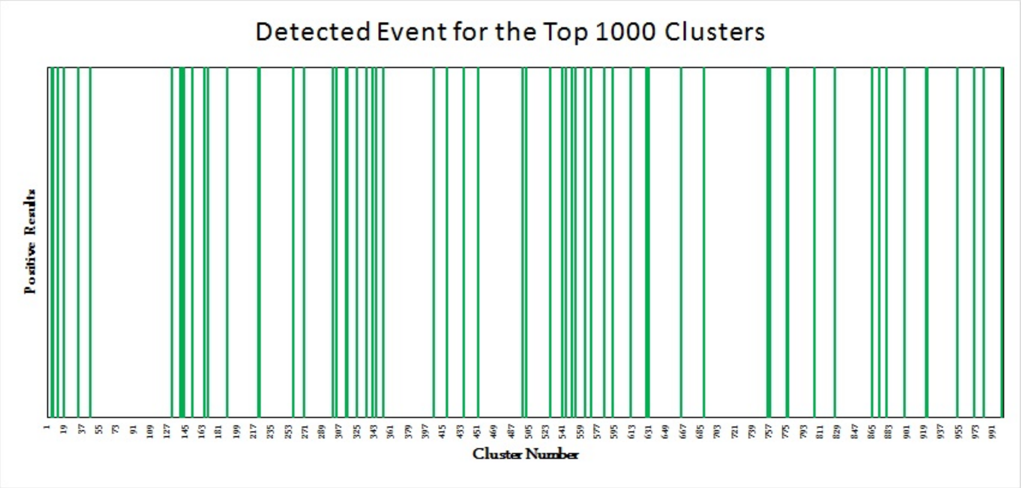


FIGURE 5.2: Overview of where events are detected among the thousand most frequent clusters

The system consist of three separated parts. The beginning of the system is the Flickr crawler gathering all the images and information for further examination. The second part are preprocessing the metadata for the dataset and transforming the data for further processing. The third part consist of counting the string frequencies in the set of strings from the data and do a DBpedia lookup. The lookup searches for the tag on DBpedia and stores a description for the given tag if the site is stored with one of the labels presented in Table 4.3. The last to steps of the system is presented in a sequence diagram in Figure 5.3, where the last two parts are displayed and their internal functions.

The first part of the system displayed is a reader and cleaning process and is essential for further processing of the data. The tag transformation method is transforming the string into new string based on desired output for the number of decimals for the radius measurements. To further connect a string to a given image, a file with image ID is saved.

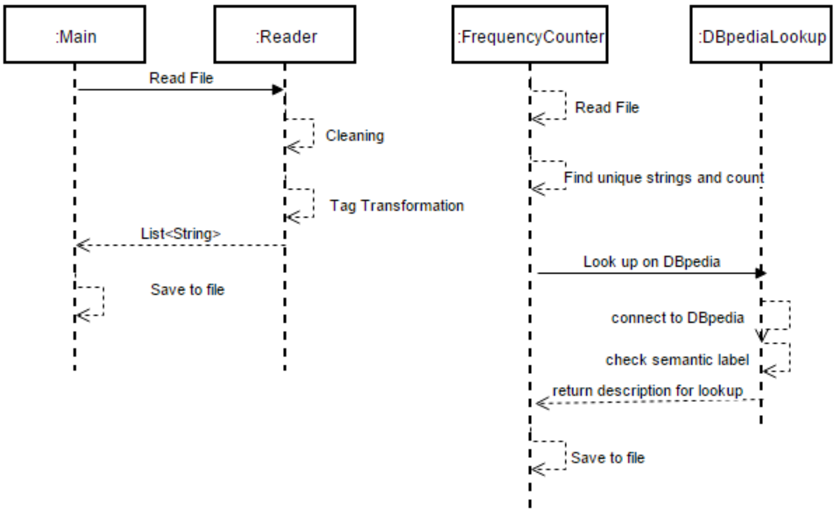


FIGURE 5.3: Sequence diagram of the system handling the data

The frequency counter system consists of a comparable class storing the string values of the metadata strings and its counter of the number of instances in the set. Then an interface which sort the occurrences firstly and also sort them alphabetically when the occurrences is matched, by creating a new list from an original map and add the strings. Then sort this list, extract the k first items in the list, using sublist, to return those for further lookup on DBpedia.

The DBpedia lookup checks for matches for a tag search and the labels available amongst the results. The labels are then returned and a new call for the given description for the wanted result. The description is then returned to the frequency counter. If there is no result for the tag search, the search goes to the next tag.

5.1.2 The Dataset

The Flickr crawler was downloading images and metadata from Flickr in the period 26th of October 2015 to 27th of November, so in a moth the crawler downloaded approximately 250 000 images and units of metadata. After reviewing the data and the timestamps

I found that most of the images in the dataset are taken in 2015, nearly 90% to be more correct. The crawler downloaded 174 gigabytes of images for the conduction of the experiments done.

The dataset consists of approximately 250 000 images, but a huge amount of these images do not have any tags, so approximately 90 000 images are removed from further processing. For further work the 160 000 images are transformed to 1 396 229 strings which are being used for event detection. A summery of relevant statistics is presented in the Table 5.1.

Image Statistics	
Images	253 873
Tags	1 396 229
Images without tags	91 635
Unique Tags	102 600
Avg. tags/Image	8,6
Usable Images	162 238

Table 5.1: Summary of related information to the dataset

When looking at the unique tags in the dataset, an amount on irrelevant tags are among the top most frequent once. In Table 5.2 a list of the top 20 most frequent tags is presented with associated number of frequency. It is from this list some of the tags in Table 4.2 is determined. None of the top 20 tags are related to any events, but rather social media tags, camera related tags and countries.

After the cleaning process are almost all of the tags in Table 5.2 removed from further processing and it contributes to a better performance for later use. The only tags left are `Newyork` and `us`, which can be used for later identification and evaluation of the events.

5.1.3 Dataset Tag Cleaning

To increase the performance and remove some noise, the dataset is cleaned. The cleaning process is describes in Subsection 4.3.3 and in Table 5.3 the result of the process is shown with relation to the Table 5.1 which contain related information before the cleaning is done.

Tag	Frequency
square	23 135
iphoneography	22 952
squareformat	22 770
instagramapp	22 743
autumn	5991
unitedstates	5624
france	5569
nature	4880
geotagged	4415
landscape	4306
england	4215
nikon	4162
travel	3885
us	38868
italy	3748
sky	3730
newyork	3690
outdoor	3601
canon	3585
taiwan	3367

Table 5.2: Top 20 frequent tags in the dataset

Image Statistics After Cleaning	
Total Images	253 873
Tags	1 193 454
Images without tags	106 487
Unique Tags	100 961
Nr removed tags	202 775
Avg. tags/Image	8,1
Usable Images	147 386

Table 5.3: Summary of related information to the dataset after cleaning

In Figure 5.4 the effect of cleaning the dataset is shown with a comparison to the uncleaned set. The effect results in 15% less tags, but only 1,5% of unique tags are removed, resulting in a better distribution of the more relevant tags in the set. The cleaning increases the set of images with empty tags collections from 36% to 42%.

In Table 5.4 a new list of the top 10 most frequent tags in the dataset after cleaning is presented. Some of the locational tags are the same and the list consist mostly of geographical locations. I decided to keep those, because they can help to further evaluate if there is an event and if it is identified correctly.

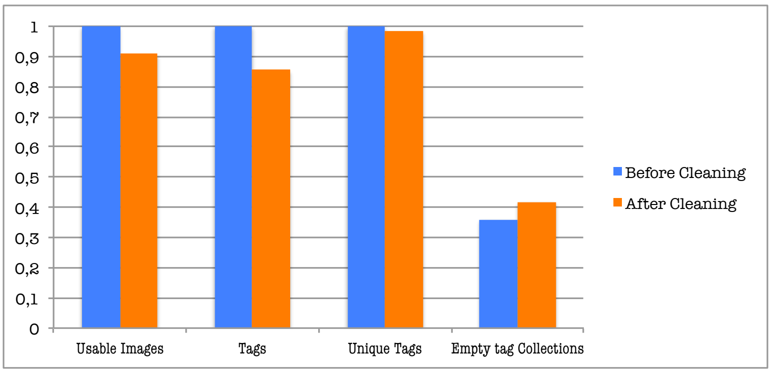


FIGURE 5.4: Effect of the cleaning on the tags of the dataset.

Cleaned Tag	Frequency
us	3868
newyork	3690
de	2974
australia	2973
uk	2808
london	2773
deutchland	2750
city	2749
europe	2709
water	2705

TABLE 5.4: Top 10 frequent tags in the dataset after cleaning

5.1.4 Dataset Transformation

To present how the data from Flickr was changed for event detection there is an example of downloaded metadata from the crawler displayed below.

```
ID:22353179728.jpg GeoData[longitude=24.108929 latitude=56.947533
accuracy=16]Date: 20150913[Tag [value=riga, count=0], Tag [value=
latvia, count=0], Tag [value=restaurants, count=0], Tag [value=
church, count=0], Tag [value=cathedral, count=0], Tag [value=stpeters
, count=0], Tag [value=rigacathedral, count=0]]
```

After a round of textual preprocessing and cleaning, the metadata is transformed into strings containing the same geographical and temporal data, but the textual annotations are listed separately. The new set of strings produced from the example of metadata over is presented here:

```
2015091324.1056.94riga
2015091324.1056.94latvia
2015091324.1056.94restaurants
2015091324.1056.94church
2015091324.1056.94cathedral
2015091324.1056.94stpeters
2015091324.1056.94rigachathedral
```

The data is transformed into a string where the date comes first, *yyyymmdd*, followed by longitude and latitude with wanted number of decimals, and last the tags. The raw data from Flickr are presented with 6 decimals in the longitude and latitude, and from Table 4.1 that indicates a radius of under 1 meter. When wanting to find clusters of similar tags, the number of decimals is cut short to only 2, so the radius is now 1000 meters. This change will increase the possibility for more images containing the same tags at the same location.

The metadata from Flickr contains a measurement called accuracy and it is an indicator for the location information level.The current range used by Flickr today is 1-16. World level is 1, Country is 3, Region 6, City 11, Street 16. Defaults setting is put to 16, if there is no specification. This is not currently used further in this thesis, but may be an opportunity in further work.

After reading over the set of tags presented for this given example, it can be said that no relevant events has occurred at this given time and place. If one or more of the tags from the example is occurring so often that a DBpedia lookup is performed, it is the DBpedia search which determine if there is an event present or not.

5.2 Results

To evaluate the method used in the thesis, a set of results is presented in the following subsections. In Subsection 5.2.1 the top clusters for the dataset are presented as well as other relevant information. A survey for user-based evaluation is presented in Subsection 5.2.2, where a set of bystanders are used to evaluate the results by judging a set of images against Flickr tags and DBpedia descriptions.

5.2.1 Top Clusters

In the previous Section the experimental setup for this master thesis was described. By clustering the data string after occurrence, a list of most frequent tags are presented in Table 5.5. The three columns present the top 10 clusters for the three geographical measurements presented in Subsection 4.2.2.

Top 10 Most Frequent Clusters					
Radius 10m		**Radius 100m**		**Radius 1000m**	
Tag	Count	Tag	Count	Tag	Count
antwerpen	470	antwerpen	470	antwerpen	470
anvers	470	anvers	470	anvers	470
conference	470	conference	470	conference	470
congres	470	congres	470	congres	470
devoxx	470	devoxx	470	devoxx	470
discourse	470	discourse	470	discourse	470
disquisition	470	disquisition	470	disquisition	470
geek	470	geek	470	geek	470
huddle	470	huddle	470	huddle	470
java	470	java	470	java	470

Table 5.5: Top 10 most frequent clusters based on three set of geographical radius distances

In Table 5.5 the three radius's are presenting the same result for the top 10 clusters and with the same frequency count. In comparison to Table 5.4 the top ten cleaned tags in the dataset represent the most frequent tags used, while these tags are the most commonly used in a given day and place, by geographical coordinates.

When comparing the three columns of results with different geographical constraints, the results are all similar between the three. The top ten tags for these clusters can be identified as one event, when there can be context and correlations among the tags. The image set for these tags is the majority of the cluster set, and further down the list of image cluster the count for each cluster can indicate a possible relation between the tag(s).

For further identification of the clusters a lookup on DBpedia is conducted and the result for the top ten events with additional image count is presented in Table 5.6.

Top 10 Events Identified

Radius10m		**Radius100m**		**Radius1000m**	
Event	Count	Event	Count	Event	Count
congres	470	congres	470	congres	470
devoxx	470	devoxx	470	devoxx	470
lecture	470	lecture	470	lecture	470
taekwondo	361	taekwondo	361	taekwondo	361
sydney	274	sydney	274	sydney	274
betrayal	240	betrayal	240	betrayal	240
syria	192	car	195	meer	217
us	186	veteran	195	car	195
deutschland	183	syria	192	veteran	195
stuttgart	183	meer	191	syria	192

Table 5.6: Top 10 events for three set of geographical radius distances

The identified events in Table 5.6 displays some tags with the similar count and with an eye on Table 5.7 they are the same image. The latter table presents the identified events with the identical metadata string, containing the same date and location for the given tag(s). With that in mind the top ten identified events are not unique, but some are overlapping.

Comparing Table 5.5 to Table 5.6, all of the top ten clusters is probably the same event and consisting of the same images. This assertion is proven after further investigation of the raw data. For the top ten identified events by DBpedia, the result of top ten unique events is shown in the following table, but there is no answer to the question what is the right tag and rightly identified tag for the image set. To answer this question, I have conducted a survey for a objective evaluation, see the following subsection.

The DBpedia description identifying some of the top ten events is shown in Appendix A.

Top 10 Image sets

Metadata	Tags
201511114,418251,2461	congres, devoxx, leture
201510318,606244,9161	taekwondo
20151124151,2347-33,9177	sydney
20151103-79,394243,6630	betrayal
20151101-0,18551,040	car, veteran
2015060143,957336,1698	syria
2015110728,27436,851	meer
20151123-74,339040,5632	us
201511219,233348,7915	deutchland, stuttgart, volleyball
20151031-0,131051,5945	classic, retro

Table 5.7: Top ten image set with their associated metadata and tag collection

When looking at Table 5.7, there are two tags present there and in the top ten most frequent tags in Table 5.4. `deutchland` and `europe` is detected by DBpedia, but the count for the present event is far less then the overall count for the two given tags.

5.2.2 User-based Evaluation

To evaluate the results, I decided to perform a user-based evaluation through a survey to get an independent and objective evaluation. The setup for the evaluating survey was conducted with a similar setup as [25]. In Section 2.5 the principles of crowed sourcing is introduced and here taken into action. The implementation sorts the string based on frequency of occurrence, based on the time and geographical location and lastly the given tag. The top 10 groups are presented in Table 5.5, but not necessarily is there an event amongst them. The survey presents the detected events with a set of 10 random images from the crawled dataset with the right corresponding data. If two or more tags are presented for the same time and location, the set of tags and descriptions is presented in the survey as possible matched. In the survey it is given that multiple tag and description can be suitable for the given image set, so there is overlapping amongst the answers. With thees two question for each event detected, a view of possible disambiguation between events are detectable.

To carry out the user-based evaluation, a set of 17 highly educated humans assessors[1], aged between 20 and 30 with a base at the Norwegian University of Science and Technology. They were provided with the top-50 detected events detected and extracted by the used algorithm and the survey was divided into two parts because of its length and the time the survey took in carrying out. This will give some variation of what lies in the foundation behind each decision, but will be negligible when everyone contributes their own unbiased opinions. Based on the images the crowd is asked if the tag and a given description from DBpedia are a fit for the image set, the questions can either be answered as “suited”, “possible” or “unsuitable”. This means that for each group of images the possible answers were *yes* - i.e. the group of images surely represent the tag, *probably* - i.e., the group may be an event presented by the tag, and *no* - i.e. the cluster was clearly not an event presented by the tag. The same examples can be used for questions related to the given description and if it is fitting.

Each of the answers was given a score depending on the given alternative: 1 for *yes*, 0.5 for *maybe*, and 0 for *no*. Further, the final relevance score for each cluster of images was calculated as the average score given by the assessors. The values gathered in the score is used as an basis for the precision-oriented evaluation metrics, introduced in Subsection

[1] The number of persons that participated in the two surveys for evaluation varied between 7 and 10

2.2.4. For example, suppose a cluster obtained a set of scores 1, 0.5, 0, 1, 1 from five different assessors and resulting in an average score of 0.7 for the given cluster. And further will a set of cluster with average scores of 1, 0.8, 0.6, 1, 0,9 result in a precision score for the algorithm to 0.86.

For a more formal definition, let c_{ij} be the score for the cluster c_i given by the assessor a_j. Giving the c_{ij} a specification as follows:

$$c_{ij} = \begin{cases} 1 & \text{if } c_i \text{ is surely an event} \\ 0.5 & \text{if } c_i \text{ is probably an event} \\ 0 & \text{if } c_i \text{ is surely not an event} \end{cases} \tag{5.1}$$

Then, let α_i represent the final score for c_i, and the total number of assessors evaluating c_i is n_i. Further, the score for c_i is:

$$\alpha_i = \frac{1}{n_i} \sum_{j=1}^{n_i} c_{ij} \tag{5.2}$$

Resulting in the precision value for the set of image clusters for top k detected events by:

$$\mathrm{P}(k) = \frac{1}{k} \sum_{i=1}^{k} \alpha_i \tag{5.3}$$

Using the information presented above, a set of precision values is computed with four different k values. The k values is 5, 10, 25 and 50 for the above relevance-score values. The result from this assessor evaluation survey is presented in Table 5.8 and with the three geographical measurements.

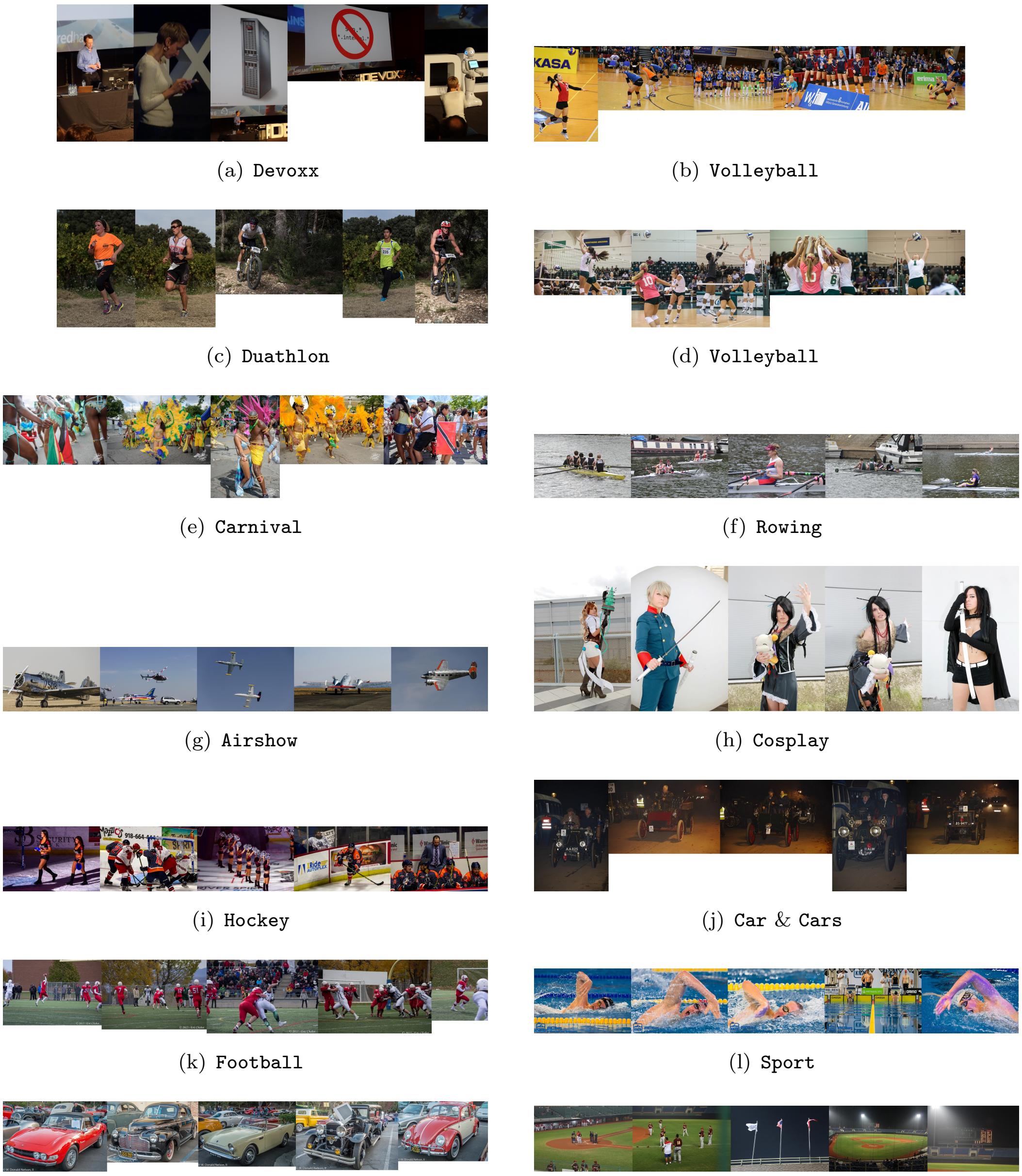
	Radius 1000m		Radius 100m		Radius 10m	
k value	Tag	Description	Tag	Description	Tag	Description
5	0.68	0.47	0.68	0.47	0.64	0.46
10	0.65	0.35	0.65	0.35	0.64	0.36
25	0.61	0.33	0.61	0.33	0.59	0.32
50	0.61	0.28	0.61	0.29	0.62	0.30

TABLE 5.8: Precision for event clusters extracted based on DBpedia label matching, with four cluster sizes

The top ten identified events and unique events shown in Tables 5.6 and 5.7 indicate the similarity among the result and when looking at the result for the top 50 returned unique events for the three radius's is the similarity present there as well. The list of top

50 events for all three measurement is somewhat similar, so the result in Table 5.8 for the different precision measurement is not surprising. The assessors think there are more in common between the image set and the given tags, rather than the given description. The results from the survey indicates that the assessors more often uses the term *probably* when ruling over a tag, rather when the description is evaluated.

The image sets with a precision of k=1 is shown in Figure 5.5. The assessor was given ten images for each event, and only five of these are displayed here.



(a) `Devoxx`

(b) `Volleyball`

(c) `Duathlon`

(d) `Volleyball`

(e) `Carnival`

(f) `Rowing`

(g) `Airshow`

(h) `Cosplay`

(i) `Hockey`

(j) `Car & Cars`

(k) `Football`

(l) `Sport`

(m) `Cars`

(n) `Baseball`

(o) `Car`

Figure 5.5: Random images for cluster with precision of 1, when only the tag are evaluated

The fifteen image set display all real events, but their tags are general tags which can identify multiple events. All of these events are unique, but the tags identifying these events are general term for a commonly and some repeatedly occurring events. Three of the detected events and accurately described by `car` and/or `cars`, are an indication that DBpedia relates the tags with predefined events with ties to other events where these tags were an essential part. The example follows on other tags and are also shown in the precision measure for the description related to the tags in Table 5.8.

In Figure 5.6 is a set of images presented with the associated description given by DBpedia. These images and descriptions have been rated by the assessor to a precision like k=1 for either both the tag and description or only the description.

(a) Devoxx: *Devoxx (formerly named JavaPolis) is an annual European Java conference organized by the Belgian Java User Group. The conference takes place every year around November. With over 2,800 attendees in 2006, JavaPolis became the biggest vendor-independent Java conference in the world. In 2008, the conference was renamed Devoxx. With over 3300 attendees, Devoxx 2011 was sold out 6 weeks before the event. In 2012, the first edition of Devoxx France took place from 18/4 until 20/4 in Paris.*

(b) Duathlon: *Off-road duathlon is a form of duathlon, where the competitors have to go through a trail-running stage and a mountain-biking stage. Off-road duathlons are distinguished from conventional duathlons in that the terrain for the cycling and running stages are generally unpaved, rough, and very steep and hilly. They require different techniques than conventional duathlon races, and the athletes employ mountain bikes rather than road bikes.*

(c) Hadrian: *The Mausoleum of Hadrian, usually known as the Castel Sant'Angelo, is a towering cylindrical building in Parco Adriano, Rome, Italy. It was initially commissioned by the Roman Emperor Hadrian as a mausoleum for himself and his family. The building was later used by the popes as a fortress and castle, and is now a museum.*

(d) Football: *College football refers to American football played by teams of student athletes fielded by American universities, colleges, and military academies, or Canadian football played by teams of student athletes fielded by Canadian universities. It was through college football play that American football rules first gained popularity in the United States.*

Figure 5.6: Random images for cluster with Precision of 1, when only the description are evaluated

The Figure 5.6 displays the reason the tags and images are identified as events. The text in the description by each image set is a description for an event detected by DBpedia. Figure 5.6a is a correctly by my objective opinion a rightly classified tag and description, they are all three matching together. Figure 5.6b displays a set of images and the tag fits nice, but the description is a general term for the event and not a specific description for the given event occurring in the images. The third Figure 5.6c displays a family in a location and the description is not necessarily an event description, but it reflect that DBpedia can label a page as an event based on previously occurred happenings(historical events). This figure is the only one evaluated to a 100% precision for the description and not for the tag. The last Figure 5.6d is another example of the general description of an event, and not a result for the rightful event occurring in the images.

Chapter 6

Discussion & Evaluation

In this chapter I take a closer look at the findings in the previously presented chapter and I evaluate the completed work. In Section 6.1 I discuss the results and important facts for the thesis. Section 6.2 contains the evaluation of the research and the research questions, to give a better understanding of the research and the work done.

6.1 Discussion

In this section I discuss the findings of the thesis and possible factors impacting the result. The discussed parts include the dataset, the cleaning process, the knowledge base usage, use of tags, the clustering process and social contributions.

6.1.1 The Dataset

An important factor for success in a study like this, is a good dataset for processing. The dataset used in the thesis consists of approximately 250 000 images, but only 150 000 of them are usable. For a better result the dataset needs to be bigger and more complex. After cleaning a huge percentage of the images I ended up with empty tag collections, giving no results and consuming a lot of storage space. If the percentage is representative for users on Flickr, the dataset should take that into account and always be twice the size of desired amount of images.

Another point with the amount of images and the used dataset, is the distribution of the images. In Figure 6.1 the geographical coordinates is used to display each image on a world map and the color indicates the amount of images from few in yellow and many in dark red. With the amount of images approximately 150 000, the distribution of images

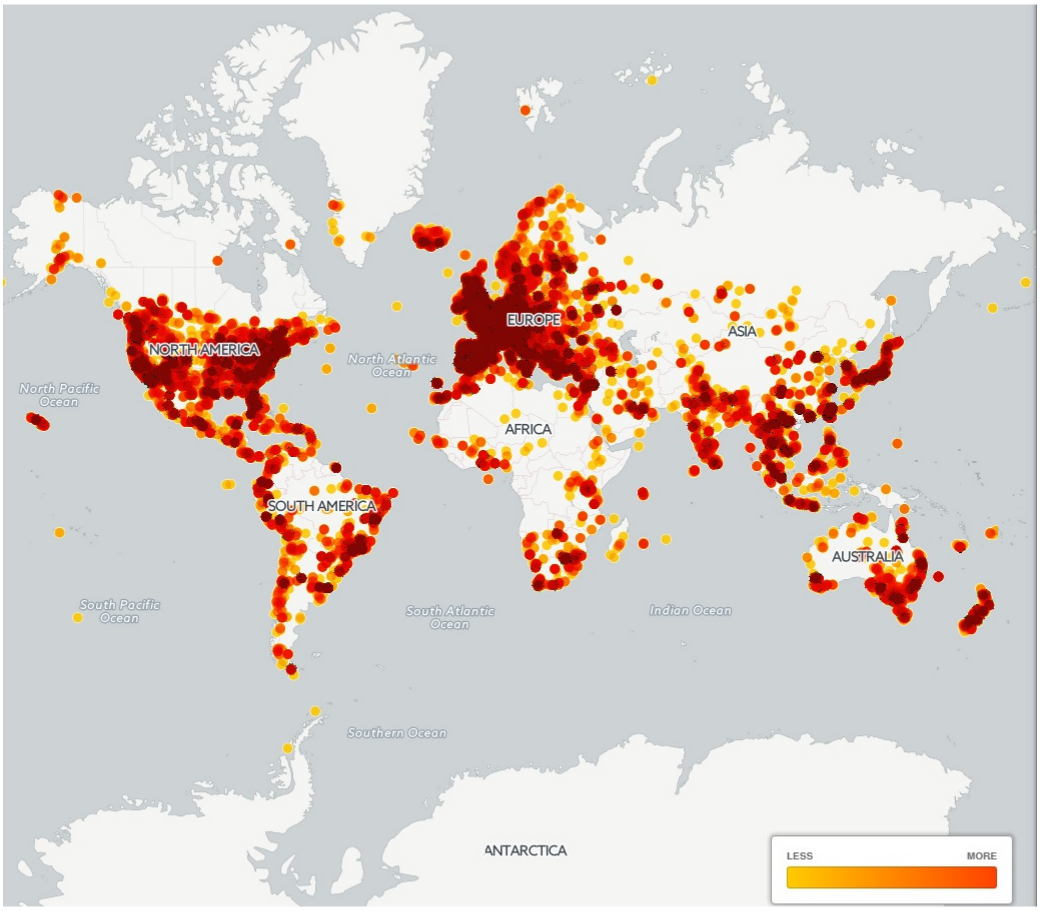


FIGURE 6.1: Distribution of the used images

are making the amount in one place considerably low. The high distribution is also seen in the #1 cluster with an occurrence of 470, considered low when the total amount of tags is over one million. This can be compared to the use of datasets with limitations to few cities [3, 25].

6.1.2 Cleaning Process

The cleaning process done in the thesis consists of tag removal. Based on exact or partial match tags are removed from their appurtenant collection and thereby cleaned from the dataset. When looking at the results, the amount of removed tags can be increased, since there is a majority with none identified tags.

Alternative method usable for the cleaning of the collections is stemming and affix removal. Stemming is a method converting words to a shared root form - the affixes is removed from the words, the disadvantage with this method is that words often can loose their meaning and the matching to DBpedia pages may be effected.

The amount of images is scattered around the globe and it included a variety of languages used in the tagging. The used method and DBpedia pages are only in English. To exploit the variety in image location and language, it would be wise to use a knowledge base that can handle variations and possible switch between languages themselves. Another possibility is to use a dataset with only one language present, preferable English.

6.1.3 The Knowledge base

The use of DBpedia as the source for identifying events in the dataset, may lead to refinements. The amount of pages in the DBpedia database and how they are labeled may affect the result. An alternative to DBpedia is Freebase[1] and Wikidata[2], they are both collections of knowledge based on user input and the latter is central storage for structured data from related projects including Wikipedia, Wikivoyage[3] and others. Freebase is a temporary knowledge base with 57 million topics, 3144 million facts and 600 000 domains for events, but is now only read-only and will be shut-down.

The use of an alternative knowledge base may lead to positive possibilities in identifying new and other kinds of events. To alter the method to be functional with another knowledge base can be seamless, but the building blocks may probably lead to difficulties. There is no certainties for the reachable usable information and the way of labeling their events is effective. The Freebase knowledge base is no longer maintained and the new events are not added to the database, since the used dataset used in the thesis consist almost only of images from 2015, the result may lack the same accuracy.

6.1.4 Tag Usages

To handle the tag collections attached to images there are multiple ways to use the information. In this thesis every tag is handled individually. This is done to avoid noise from irrelevant tags and to see every tag as an independent part of the annotations. The varying way of how users tag their images I did evaluate the possibility of the users to tag a photo with one individual and important tag high, rather than seeing the collection of tags as correlation.

In the work done by Massimiliano Ruocco [25] was the search for events was done in a different fashion. He assembled all the tags for each image and used these strings to look for events, resulting in the images need to have a similar collection of tags to get the count of similar images.

[1] http://www.freebase.com/

[2] https://www.wikidata.org

[3] https://www.wikivoyage.org/

Another solution can be to handle the tags individually, but with regard to the rest of the collection. Multiple tags can be identified as an event, but to identify the correct one a second or third tag can be used as a guidance for the right identification. The way of handling tags can effect the result and exploring the different ways of handling tags can be done in future work.

6.1.5 The Clustering Process

There are many was of handling the strings present in a task like this. I decided to use a simple and concrete comparison method, with an individual counter for occurrence throughout the dataset. Another approach to a similar problem was used in the Ruocco [25] thesis and briefly introduced in Chapter 2.

In the beginning of the thesis an indexing method for handling the metadata was introduced, suffix tree clustering. A suffix tree is a compressed tree containing all the suffixes of the given text as their keys and positions in the text as their values. The tree has n leaves and all internal nodes has at least two children, is resulting in the trees total size to be $O(n)$ [2].

The effect of using different kinds of clustering method is a task itself, but I will say my decision did not affect the result, both in accuracy and performance.

6.1.6 Social Contributions

The concept of using data from social media entails a series of challenges. There is no shared understanding of how to share images and what it is expected to accompany the image. The one thing effecting the performance of the result is the formulation of tags. Some users are sparse with adding tags, some adds none and some add almost to many, other problems are spelling, synonyms and disambiguation. All these problems may effect the result, but will always be a part of studies with user based data.

The evaluation of the result done in Subsection 5.2.2 are all subjective opinions done by assessors. They reflect over and evaluate the different question throughout the survey, they can develop an understanding and bias attitude. Making bystanders a part of the evaluation has two sides, an objective evaluation of the findings and an possible lack of knowledge and discernment on the other side.

6.2 Evaluation

Before moving on to the final conclusion, a brief evaluation of the research I performed. This section contains an evaluation of the thesis and the study done in the past months. The evaluation is done to give a better understanding of the research at hand, and to build confidence and trustworthiness in both the research and the thesis. Lastly is a review of the problem definition.

6.2.1 Evaluations

The master thesis have proven to identify new ways of seeing connections between images collected on Flickr and information shared on DBpedia. The process of collecting the images on Flickr, shows that the crawler mostly downloaded images in a certain time and the images was distributed all over the world. For a better dataset, the crawler can be more specific to a location or to a time period, but the former would be preferable.

For a better evaluation of the results, a bigger group of assessors will be helpful. The seventeen people answering the surveys can be seen as too narrow and like sighted, and by bringing in a group from different settings the result may be enhanced.

The method developed in this thesis connects Flickr and DBpedia in a special way. The used method was not accurate in determining the right events when tying images with descriptions together. The method detected 15 out of 50 events with a precision of 100% and an overall of 61%, indicating a good method for this purpose, but not perfect.

6.2.2 Research Questions

To end the work done in the thesis I will answer the questions asked in Section 1.4. I presented one main problem for the thesis and four underlying questions. Here I will begin answering the four sub-questions before ending up to answer the main problem.

[RQ1]*How can an event be detected based on metadata stored for images on Flickr?*
The work done is based on a tie between occurrence of images in one place at a time and based on the occurrence of tags, an image collection can be identified as an event. The metadata gathered with the crawler is not much, but with a large set of small information strings, an event can be identified with shared information among the set. To detect an event the used metadata was time stamp for photo taken, geographical coordinations for photo taken and a collections of tags added by the photographer. So yes, an event can be detected by information added to Flickr by ignorant users who do not hide important information.

[RQ2] *What is the impact of cleaning the dataset?*
When seeing through the list of removed tags from Table 4.2 the need for this cleaning may not impact the end result. Some of the removed *temporal terms* listen in the table, may be detected as an event on the DBpedia interlinked network, but both *camera related terms* and *general detected noise* is far less possible to be identified as an event. The top *non-relevant terms* removed includes in general countries and cities. To avoid identification of events happening specially for that country or city these are removed. E.g. `deutchland` and `stuttgart` was identified as events with `volleyball`, which was the correct tag for the image set, seen in Table 5.7. The amount of tags in the whole data set was approximately 1.4 millions and the cleaning removed 200 000 tags, resulting in better performance.

To evaluate the effect of the cleaning process, it is needed to look at the two sides. The one side will improve performance by removing tags, and the other side will keep the opportunity to identify all kinds of events. By removing a long list of tags, the more relevant tags will be considered and the returned precision score will be affected in a positive way.

[RQ3]*How to identify the correct tags in a given cluster, for further event identification?*
With use of a method where DBpedia is searched for the corrected labels for identification for an event, the right tags can be found based on the DBpedia labeling. The set of labels on the DBpedia pages is in a wide range and with more information regarding how the labeling is done, a more accurate matching and elimination of the search can be conducted. The semantic information in DBpedia are presented in labels and these labels identify pages defined as event. By using these labels the tags with ambiguity will return results only when the meaning contains an event result.

[RQ4] *What contributions can DBpedia provide in identifying events?*
The fourth research question may overlap with the previously answered question. DBpedia can be used to identify events by using a predefined label for a given event and with the right connections between a searched tag and a given event page. The pages on DBpedia are mainly big and well known events, e.g. wars, awards, sports event and historical important dates. Most of the images from Flickr consist of small and personal events, not with DBpedia pages, but perhaps a general description for the event.

The four research questions answered above set the lines for the main problem defined for the thesis.

[RQ]*How can we improve event image detection using knowledge base look up?*
Based on the work done in this thesis a way of identifying events of a certain size and popularity can be conducted. The implemented solution for the experimental work is a

good solution for identifying events with their own DBpedia page, but when identifying little well known or private events the result is with less accuracy. The tags presented with their image sets in Figure 5.5 are all general terms for specific events present in the images. All of the detected events is indeed events occurring, and they are detected by the DBpedia lookup, but the description given is not needed. When the amount of images at a given location and within a time range, an event can be identified by the majority of tags in common among the images and be checked with DBpedia for an confirmation.

Part III

Conclusion & Further Work

Chapter 7

Conclusion

This is the final chapter for the thesis and contains a final conclusion as well as proposal for further work within the field. Section 7.1 presents a conclusion for the work done and the experimental results. Finally, in Section 7.2 the suggestions for further work is the ending of the thesis.

7.1 Conclusion

The main focus of this thesis has been to detect events based in semantic data on DBpedia. In the presented method I used metadata from Flickr, such as timestamp, geographical information and textual annotations to search and detect events based on information in the DBpedia knowledge base.

The crawler collecting images from Flickr gathers images from the whole world, which were uploaded in 2015. When clustering the data collected the tags were sorted based on frequency. The method was a good fit for the desired result and the use of the data, but the images were highly distributed, resulting in small clusters.

With the dataset used, the DBpedia lookup method often only detects small local events. The result from the user-based evaluation indicates that the methods can indeed be used for detecting events with different backgrounds, both unknown and well known. My study has shown that the semantic labeling in DBpedia pages has been complete and well-conducted. Overall, it is well-suited for categorizing data and for matching other information sources, such as Flickr data.

The experiments and the user evaluation have also shown that the methods used in this thesis has great potential for improving the detection of events from data collected from Flickr. The system detects events in a wide range and the user-based evaluation indicates perfect matches with both tags and descriptions. The DBpedia semantics provides information to a full and correct identification of events in various size.

7.2 Further Work

For a better view of the used method a new set of data is to be used. The used data in the thesis has a too high distribution and the detected events are only small events with the maximum of 470 images for the biggest cluster. With a dataset with limitations to one location or a bigger set with more centered points, the results can be different. To check for validation of the presented work, the first step should be replacement of the used data.

To further validate the detected events, a new step can be added to the process. Multiple tags can be checked up against the matching DBpedia sites, for a thorough identification. Using more than one tag to identify an event, can be proven effective by implementing resourceful methods.

With the use of DBpedia and their database, a way of linking events and pages together is an interesting field. DBpedia provides an enormous network between their pages, and by using this, the collection of a page identified by tag search, can be used to identify the final event. This collaboration between multiple pages can produce a wider perspective of what happens in the image set and for further determination of the right event identification.

Another possible future work is to explore the correlation between tweets from Twitter and the images being posted on shared image platforms, such as Flickr. Twitter also attach tags, timestamp and location to their tweets, but only 3% of the tweets contains location information. The work to be done can tie the information they share together, and by that identify images through a set of tweets to a shared event.

Appendix A

Information Regarding Detected Events

Top 7 events with information for all three radius measures

Tag	Description
Congress	The Belgian National Congress was a temporary legislative assembly in 1830, established shortly after the Provisional Government of Belgium had proclaimed Belgian independence on October 4 of that year. Its primary task was to create a constitution for the newly formed state. The National Congress was elected by approximately 30,000 voters on November 3, 1830 and consisted of 200 members. Its president was Baron Erasme Louis Surlet de Chokier.
Devoxx	Devoxx (formerly named JavaPolis) is an annual European Java conference organized by the Belgian Java User Group. The conference takes place every year around November. With over 2,800 attendees in 2006, JavaPolis became the biggest vendor-independent Java conference in the world. In 2008, the conference was renamed Devoxx. With over 3300 attendees, Devoxx 2011 was sold out 6 weeks before the event. In 2012, the first edition of Devoxx France took place from 18/4 until 20/4 in Paris.

Lecture	A lecture is an oral presentation intended to present information or teach people about a particular subject, for example by a university or college teacher. Lectures are used to convey critical information, history, background, theories and equations. A politician's speech, a minister's sermon, or even a businessman's sales presentation may be similar in form to a lecture. Usually the lecturer will stand at the front of the room and recite information relevant to the lecture's content.
Taekwondo	The 2011 World Taekwondo Championships is the 20th edition of the World Taekwondo Championships, and was held at Gyeongju Indoor Stadium in Gyeongju, South Korea from May 1 to May 6, 2011.
Sydney	The Australian census is administered once every five years by the Australian Bureau of Statistics. The most recent census was conducted on 9 August 2011; the next will be conducted in 2016. Prior to the introduction of regular censuses in 1961, they had also been run in 1901, 1911, 1921, 1933, 1947, and 1954. Participating in the census is compulsory.
Betrayal	RuneScape is a fantasy massively multiplayer online role-playing game (MMORPG) released in January 2001 by Andrew and Paul Gower, and developed and published by Jagex Games Studio. It is a graphical browser game implemented on the client-side in Java, and incorporates 3D rendering. The game has approximately 10 million active accounts per month, over 156 million registered accounts, and is recognised by the Guinness World Records as the world's most popular free MMORPG.
Meer	Nitro compounds are organic compounds that contain one or more nitro functional groups. They are often highly explosive, especially when the compound contains more than one nitro group and is impure. The nitro group is one of the most common explosophores (functional group that makes a compound explosive) used globally.

Table A.1: Description from DBpedia and identification of top 7 events

Bibliography

[1] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. 1998.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. Modern information retrieval. 2011.

[3] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 291–300, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-889-6.

[4] S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

[5] S. Büttcher, C. L. Clarke, and G. V. Cormack. *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2010.

[6] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):394–410, 2007.

[7] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[8] P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: Large scale automatic generation of personalized annotation tags for the web. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 845–854, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242686. URL http://doi.acm.org/10.1145/1242572.1242686.

[9] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. 1999.

[10] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[11] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics, 2010.

[12] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274, 1996.

[13] X. Han and L. Sun. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 945–954, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://dl.acm.org/citation.cfm?id=2002472.2002592`.

[14] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[15] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[16] H. L. Kim, S. Scerri, J. G. Breslin, S. Decker, and H. G. Kim. The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In *International Conference on Dublin Core and Metadata Applications*, pages 128–137, 2008.

[17] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[18] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer, 2009.

[19] G. S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM, 2007.

[20] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40. ACM, 2006.

[21] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5):530–549, 2004.

[22] T. Pang-Ning, M. Steinbach, V. Kumar, et al. Introduction to data mining. In *Library of Congress*, 2006.

[23] D. Rao, P. McNamee, and M. Dredze. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer, 2013.

[24] 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*, pages 103–110. ACM, 2007.

[25] M. Ruocco. Geo-temporal mining and searching of events from web-based image collections. In *ACM SIGIR Forum*, volume 48, pages 119–120. ACM, 2014.

[26] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[27] Z. Shen, S. Arslan Ay, S. H. Kim, and R. Zimmermann. Automatic tag generation and ranking for sensor-rich outdoor videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 93–102, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0616-4. doi: 10.1145/2072298.2072312. URL `http://doi.acm.org/10.1145/2072298.2072312`.

[28] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.

[29] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[30] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88. ACM, 2002.

[31] J. Zobel, A. Moffat, and K. Ramamohanarao. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23(4):453–490, 1998.