Exploiting Twitter to fulfill information needs during incidents

Master Thesis, June 29, 2011

Richard Stronkman

Exploiting Twitter to fulfill information needs during incidents

MASTER THESIS

submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE
in
COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Richard Stronkman
born in Eindhoven, The Netherlands

Web Information Systems Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, The Netherlands
http://eemcs.tudelft.nl

Exploiting Twitter to fulfill information needs during incidents

Author: Richard Stronkman
Student id: 1150669
Email: richardstronkman@gmail.com

Abstract

How can Twitter be exploited to fulfill information needs of users on Twitter during incidents? In this Thesis this question wil be investigated and a strategy to locate tweets which fulfill information needs will be introduced. First, techniques are proposed to automatically detect incidents given an unstructured incident data source and subsequently tweets which report on these incidents are tracked. An interpretative analysis is carried out to derive and classify typical information needs during incidents. Furthermore, a metric is proposed to compute a tweet's relevance by measuring its informativeness and the user's trustworthiness. Finally, tweets are ranked according to this metric such that for each information need at a particular time the most relevant tweets can be found.

Keywords: Twitter, incident detection and tracking, information retrieval

Graduation Committee:

Prof. dr. ir. Geert-Jan Houben, Faculty EEMCS (WIS), TU Delft
Dr. Laura Hollink, Faculty EEMCS (WIS), TU Delft
Dr. M. Birna van Riemsdijk, Faculty EEMCS (MMI), TU Delft
Ir. Jeroen Broekhuijsen, TNO

Contents

1 Introduction **9**
- 1.1 Background and motivation . . . 10
- 1.2 Problem definition and goal . . . 10
  - 1.2.1 Research goal and questions . . . 10
- 1.3 Research methodology . . . 11
- 1.4 Document structure . . . 12

I Incident detection and tracking **13**

2 Building incident profiles **14**
- 2.1 Incident data sources . . . 14
  - 2.1.1 Selecting an incident data source . . . 14
  - 2.1.2 Characteristics of selected data source . . . 15
- 2.2 Incident profile builder . . . 19
  - 2.2.1 Designing an incident profile . . . 19
  - 2.2.2 Building incident profiles . . . 19
  - 2.2.3 Evaluating the profile builder . . . 21
- 2.3 Conclusion . . . 23

3 Filtering tweets based on incident profiles **24**
- 3.1 Twitter as data source . . . 24
  - 3.1.1 Characteristics of Twitter . . . 24
  - 3.1.2 Retrieving tweets from Twitter . . . 26
- 3.2 Incident tweet filter . . . 28
  - 3.2.1 Defining incident tweets . . . 28
  - 3.2.2 Filtering tweets based on incident profiles . . . 29
  - 3.2.3 Evaluating tweet filter . . . 30
  - 3.2.4 Optimization . . . 31
- 3.3 Conclusion . . . 33

II Fulfilling information needs during incidents **34**

4 Identifying information needs **35**
4.1 Data collection . . . 35
4.1.1 Archiving application . . . 35
4.1.2 Collected data statistics . . . 36
4.2 Tweeting activity . . . 37
4.2.1 Tweet distributions . . . 37
4.3 Tweet content interpretation . . . 39
4.3.1 Selecting an incident scenario . . . 39
4.3.2 Analyzing tweet content . . . 41
4.3.3 Analyzing re-tweets . . . 45
4.3.4 Deriving information needs . . . 47
4.4 Conclusion . . . 49

5 Ranking incident tweets based on relevance **50**
5.1 Information Retrieval . . . 50
5.1.1 Introduction . . . 50
5.1.2 Ranking documents . . . 51
5.2 Incident tweet ranker . . . 53
5.2.1 Assessing relevance . . . 53
5.2.2 Designing queries . . . 53
5.2.3 Ground truth rankings . . . 54
5.2.4 Ranking strategy 1: tf-idf . . . 55
5.2.5 Ranking strategy 2: tweet features . . . 56
5.2.6 Comparing ranking strategies . . . 58
5.3 Conclusion . . . 59

III Conclusions and future work **61**

6 Conclusions **62**
6.1 Summary . . . 62
6.2 Discussion . . . 64
6.3 Future work . . . 64

Bibliography **66**

A Parsing results **70**

B Query terms per incident type **72**

C Expanded search query **73**

D Most commonly used terms **74**

E Authority query terms **75**

List of Figures

1.1 System perpective showing three black boxes 11

2.1 Division of safety regions in The Netherlands 16

4.1 Incident types distributed by number of (a) incidents (b) tweets. . . . 38
4.2 Incident sizes distributed by number of (a) incidents (b) tweets. 38
4.3 GRIP values distributed by number of (a) incidents (b) tweets. 38
4.4 Tweet distribution at hour *x* after the incident. 39
4.5 Tweet distribution after correcting tweet gap during night 39
4.6 Tweet distribution at quarter *x* after incidents (500+ tweets) 40
4.7 Tweet distribution at hour *x* (no time-shift applied) 41
4.8 Tweets plotted on Google Maps at hour *x* after the incident. 44

6.1 System perpective showing three main components 62

List of Tables

2.1 Information in paging messages . . . 17
2.2 Incident size classifications. . . . 18
2.3 Incident scalings by GRIP value. . . . 18
2.4 Properties of incident profiles . . . 19
2.5 Parsing results for all/scaled paging messages . . . 21
2.6 Examples of paging messages sent to safety region Utrecht . . . 22

3.1 Example of incident profile . . . 28
3.2 Examples of incident tweets . . . 28
3.3 Queries built by the incident tweet filter . . . 30
3.4 Results of the incident tweet filter . . . 31
3.5 Results of the incident tweet filter given the Moerdijk incident . . . 31
3.6 Queries built by the incident tweet filter . . . 32
3.7 Applying search query expansion on a sample of Moerdijk tweets . . . 33

4.1 Ranking of most frequently used hashtags at hour x. . . . 42
4.2 Ranking of words from tweets mentioning `#dordrecht` . . . 42
4.3 Top 10 Internet domains mentioned in tweets . . . 43
4.4 Top 3 URLs mentioned in tweets (URLs are partly displayed) . . . 43
4.5 Classification of questions in tweets . . . 45
4.6 Top five re-tweets Moerdijk . . . 46
4.7 Top five re-tweets Moerdijk at $x = 3$ for topic Risk . . . 47
4.8 Matrix of information needs during incidents . . . 48
4.9 Sample answers for each classification . . . 48

5.1 Query for each typical information need. . . . 53
5.2 Statistics of ground truth rankings . . . 54
5.3 Average precision and recall for each information need at k=5,10,20 . . . 55
5.4 Sample top five tweets predicted by tf-idf based strategy . . . 56
5.5 Tweet feature per relevance dimension . . . 57
5.6 Average precision and recall for each information need at k=5,10,20 . . . 59
5.7 Sample top five tweets predicted by custom ranking strategy . . . 59

A.1 Results of parsing public safety paging messages . . . 71

B.1 Query terms per incident type 72

C.1 Additional query terms to find tweets of the Moerdijk incident 73

D.1 Most commonly used terms in tweets 74

E.1 Query terms per incident type 75

Chapter 1

Introduction

With the advent of Internet-enabled smartphones, more and more people are continuously connected to the web. Telling friends what you are doing and where you are seems easier than ever before. Popular services like Twitter[1] and Facebook[2] attract users who share with friends what is happening in their daily lives. On Twitter each hour millions of messages are broadcasted by users to tell the world publicly what is on their mind [7]. This type of content, also known as user-generated content, has become more present on the web and due to its public nature even appears in search results from search engines like Google[3] and Bing[4]. The paradigm shift, that users are not only consuming content from the web, but also producing it, is marked as one of the principles of web 2.0 [19].

The type of content shared on the web 2.0 ranges from meaningless chatter to informative news. One can imagine that during incidents (e.g. fires, accidents) local citizens are the first to know what is actually happening at the incident site. Particularly, with the easy-to-use mobile interfaces to Twitter and Facebook, their perception of incidents can be easily shared by means of text, photos or videos with friends or families, or can even reach a broader audience of web 2.0 users.

In earlier days journalists were responsible for obtaining, verifying and redistributing incident news. Nowadays, citizen journalism plays a key role in the journalistic process, as citizens' pictures and videos can be instantly distributed online [6]. During incidents both members of the public and emergency responders want to obtain a good overview of the incident, as it enables them to make better decisions.

This chapter will elaborate on the background and motivation of this research. Subsequently, the problem will be defined and a research goal and questions will be stated. Finally, the research methodology and document structure will be explained.

[1] http://www.twitter.com
[2] http://www.facebook.com
[3] http://www.google.com
[4] http://www.bing.com

1.1 Background and motivation

The micro-blogging service Twitter has become one the fastest growing trends on the Internet, with an exponentially-growing user base exceeding 190 million users in July 2010, producing more than 65 million messages (*tweets*) a day[5]. In previous research the topological characteristics of Twitter and its power as a new medium of information sharing are studied [12, 7]. Exploiting Twitter to spread scientific messages during scientific conferences has been studied by Julie Letierce [9]. Furthermore, studies are conducted to exploit Twitter sentiment during elections [28], or to analyze debates performance [4], or to construct user profiles by semantic enrichment of Twitter posts [1, 26]. Other studies show how user influence [3, 30] or malicious behaviour [13, 14] can be measured on Twitter.

Twitter has also been studied in relation to incidents. The earthquake warning system proposes measures to calculate impact area based on geo-annotated tweets [24]. A study on false rumor propagation shows how tweet contents can be analyzed [17]. Specifically the studies of Palen et al. [21, 22, 25, 29] contribute to a general understanding of micro-blogging behaviour during disasters. Yet, no studies are present that exploit Twitter to fulfill information needs during incidents, focus entirely on Dutch tweets, or study a variety of incident types.

1.2 Problem definition and goal

In this Thesis a novel approach will be presented to locate tweets that fulfill typical information needs during incidents. The need of such approach is motivated when examining disasters, for example the earthquakes in Chile on February 27, 2010. During this earthquake nearly 5 million tweets were broadcasted in five days, causing a continuous information overload problem [17]. A user trying to find relevant information in these enourmous data streams experiences overhead. The information retrieval system proposed by this research will solve that problem.

1.2.1 Research goal and questions

The research goal of this Thesis is to:

> Develop a strategy to locate tweets that fulfill information needs during incidents, by detecting incidents from unstructured data sources and tracking tweets that report on an incident.

In order to achieve this research goal, research questions are formulated. Each of the following chapters will answer a research question:

- Chapter 2: How can incidents be detected based on unstructured data sources?
- Chapter 3: How can tweets, that report on an incident, be tracked?
- Chapter 4: What information needs do users have on Twitter during incidents?

[5] http://techcrunch.com/2010/06/08/twitter-190-million-users

- Chapter 5: How can tweets, that fulfill information needs during incidents, be located?

Furthermore, subquestions are formulated to guide the process of answering each main research question. These subquestions are stated in the introduction sections of each chapter and will be answered in each subsequent section of that chapter. Finally, a general conclusion at the end of each chapter will answer each main research question.

1.3 Research methodology

It is essential to see that Chapter 2, 3 and 5 describe subsequent process steps. From a systems perpective Figure 1.1 displays these steps. The implementation of each step is depicted as a black box, and each box returns information that is input for a subsequent black box. For instance, the first research question denotes that incidents will be detected based on unstructured incident data sources (i.e. incident data), depicted as first black box. Here, we have implicitly decided that incident profiles are the output of this step, simply because it is a suitable means to store incident information. The second black box uses these incident profiles to track tweets for an incident. The third black box exploits the incident tweets of each incident profile to locate relevant tweets, that is, tweets that fulfill information needs. Chapter 4 first identifies these information needs.

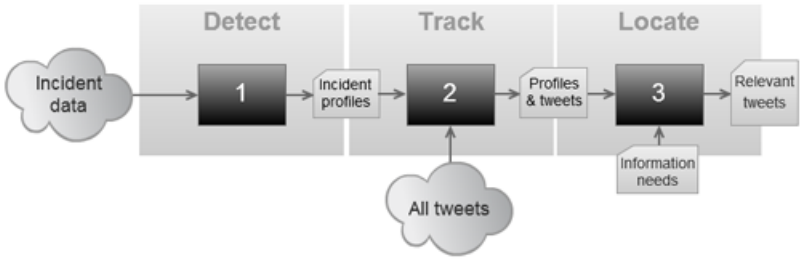

Figure 1.1: System perpective showing three black boxes

In this research a systematic approach will be applied to engineer each component:

1. The input and output of each component will be selected (incident data) or designed (incident profiles). Techniques will be described to retrieve external input data (incident data and tweets) and an input dataset will be constructed.

2. A black box component will be designed and engineered, based on techniques from (scientific) literature. A structured approach, involving domain experts, will be applied to (semi-)manually construct an output dataset (ground truth). An evaluation is done by comparing the manually constructed dataset with the component's generated dataset, using common effectiveness measures.

This approach is applied in Chapter 2, 3 and 5 to construct the black boxes.

1.4 Document structure

The document is divided into three parts. In Part I, starting with Chapter 2, a strategy will be devised to build incident profiles. In that chapter online data sources will be assessed, an incident profile will be designed and an incident profile builder will be engineered and evaluated. In Chapter 3, a strategy will be devised to filter incident tweets based on incident profiles. In that chapter Twitter characteristics will be explained, as well as techniques to retrieve tweets and finally an incident tweet filter will be engineered and evaluated.

In Part 2, starting with Chapter 4, typical information needs during incidents will be identified. In that chapter a data collection of incident profiles and incident tweets will be built, the tweeting activity and tweet contents will be analyzed and typical information needs will be derived by means of interpretative analyses. In Chapter 5, a strategy will be designed to locate tweets that fulfill information needs during incidents. In that chapter an introduction is given to Information Retrieval techniques, and an incident tweet ranker will be engineered and evaluated.

In Part 3, Chapter 6, the research will be concluded. In this final chapter the research goal and questions will discussed based on the findings of this research. Finally, future research directions will be proposed.

Part I

Incident detection and tracking

Chapter 2

Building incident profiles

The first step in this research is to find a strategy that can automatically detect incidents in real-time by means of an online incident data source. The strategy is responsible for building incident profiles which uniquely identify an incident and, as will be discussed in Chapter 3, are used to track tweets that report on the incident.

In this chapter the following research question will be answered: how can incidents be detected based on an online data source? To find an answer to this question, the following subquestions have been derived: what online data sources contain reliable and real-time incident information? What properties should an incident profile have? What strategy can be applied to build an incident profile? How can the effectiveness of this strategy be evaluated?

2.1 Incident data sources

Information about incident occurrences can be found in various online data sources. In this section these data sources will be explored and one such data source will be selected. Characteristics of the selected data source will be investigated to understand what incident information is published on it.

2.1.1 Selecting an incident data source

Real-time information about incident occurrences can be found in the following online media categories:

- *News media*. Among the news media many popular national online news papers can be found. Content in online newspapers is created by professional journalists and can therefore be considered as reliable.
- *Social media*. Among social media typically web logs, social networking sites, and others, can be found. A comprehensive list of social media is given by Kaplan and Haenlein [10]. Content in these sources can often be created by anyone using the service, making the content less reliable.

- *Emergency media*. Among the emergency media the incident monitoring services[1] can be found. These services tap directly into the public safety paging network used by professional emergency responders. Content in these services can be considered as reliable.

Reliability and timeliness of incident information are the criteria we apply to choose one of the mentioned media categories. Since social media are less reliable than the other two categories, we will not select this category. Furthermore, we notice that journalists working at news media organizations require a considerable amount of time to write a news item. The incident monitoring services from the other media category on the other hand, provide real-time information updates about incidents, making them preferable over the news media category. Therefore, we select these incident monitoring services as incident data source.

2.1.2 Characteristics of selected data source

As the reader's knowledge about incident monitoring services is still limited, this section will go in detail about this data source. It was previously mentioned that incident monitoring services tap into the public safety paging network used by professional emergency responders. This section will elaborate on this public safety communication network and paging network and describe what kind of incident information is paged over it.

Public safety communication network The public safety communication network in the Netherlands is called C2000 and was introduced in June 2004 [2]. All emergency disciplines (i.e. the police, fire, ambulance and rescue brigades, coastguard, border patrol and Royal Marechaussee) are connected to this network. Marnix Heskamp [16] describes its three components:

- *T2000*. This network component handles all voice and data traffic between public safety responders. The component is built upon the TETRA standard – a worldwide adopted protocol, like GSM, with additional features such as group calling, advanced security and direct mode operation between individual radios. The latter is particularly useful as it allows the mobile stations to communicate with each other, even when they are outside the range of the base station.

- *P2000*. Paging is an important communication application in public safety where short predetermined text messages are transmitted and displayed on pager devices. For instance, they are used to alarm fire fighters. The component uses the FLEX protocol in the 169.650 MHz band – an unidirectional communication protocol, similar as used with SMS text messaging.

- *M2000*. This software system is used in the public safety answering point (PSAP). A PSAP is a call center responsible for answering calls to an emergency telephone number (i.e. "112" in the Netherlands) for police, fire fighting, and ambu-

[1] These services provide real-time updates about incident occurrences.

lance services. The system helps the employees in the PSAP to identify the scale of the emergency and to allocated resources to the emergency.

P2000 paging network Data transmission over the P2000 network is unencrypted and insecure, in contrast to T2000. As a consequence, the network is not used by the police during critical crime fighting operations, as criminals may intentionally intercept messages and jeopardize police operations. Fire fighters and ambulance services do use the network during emergencies. The insecurity of the paging network has enabled third parties to intercept the paging message stream entirely and republish messages online. Examples of such services are LiveP2000[2] and P2000Alarm[3]. The Dutch government has taken no action to bring these services offline.

The Netherlands is divided into twenty-five safety regions, as depicted in Figure 2.1[4]. Division into regions enables a well-organized coordination and management of public safety disciplines during emergencies. Depending on the incident scaling, responsiblity can be at held at various levels in the public safety organization. In small scale incidents the officers from the involved disciplines are often in charge of operations. As soon as an incident scales up, local government agencies are involved and the mayor is responsible for the crisis management. If the incident spans multiple municipalities, then more than one mayor is responsible. Whenever incidents cover multiple safety regions, the National Operational Coordination Center and the Dutch ministries are responsible for disaster management.

Figure 2.1: Division of safety regions in The Netherlands

[2]http://monitor.livep2000.nl
[3]http://monitor.p2000alarm.nl
[4]http://nl.wikipedia.org/wiki/Veiligheidsregio

Incident information Since this chapter is about building incident profiles, it is important to understand what incident information is contained in a paging message. Table 2.1 shows what information can be contained in a paging message[5].

Item	Description
Priority	The priority indicates the urgency of the required emergency response. With priority '1' the emergency responders may behave as a priority vehicle in the traffic characterized by its flashing lights and sirens, e.g. in case of life-threading incidents. With priority '2' or '3' less urgency is required.
Classification	An incident can be classified as fire, accident, power failure, etc. In some cases a more precise classification is given, such as 'container fire', 'water accident' or 'storm damage'. Based on this classification response units can be assigned.
Address	The address can be a reference to a street address, road or highway number; sometimes a house number, zip code, and indication or hectometer number is given.
City	The city refers to a Dutch city name. In some safety regions acronyms are used to refer to a city, e.g. ASD stands for Amsterdam.
Descriptive location	Most public buildings have detectors for smoke, fires, intoxication, etc. which are connected to the PSAP. This public incident detection system contains a register of these buildings, including a descriptive location name, e.g. City Hall.
Vehicles	Each brigade has different vehicles at his disposal, which can be used to move to the incident site. Vehicles in a safety region are often identified by a three or four digit number and optional letters.
Scaling	Whenever the allocated emergency responders have insufficient manpower or material to master the emergency, an additional call can be made to the PSAP to assign more emergency responders. The incident scaling increases accordingly.

Table 2.1: Information in paging messages

Formal procedures are established to denote the scaling of an incident. Generally, the scaling influences the number of responding resources to be allocated and the level of coordination applied. Based on the scaling all emergency responders know who is in charge of operations. Table 2.2 displays a classification of incident sizes.

The Coordinated Regional Incident-Management Procedure (abbreviated to GRIP in Dutch) is a nationwide emergency management procedure in the Netherlands. The procedure is used by all emergency services, different layers of government and government agencies. The affected area of an emergency determines the level of scaling. Table 2.3 displays the different GRIP values that can be assigned to an incident.

[5]http://www.112kampen.nl/alarmeringen/uitleg

Size	Description
Small	One vehicle is sufficient to master the incident.
Medium	Two vehicles and an officer are required to master the incident.
Large	Three vehicles, an officer and a command support unit are required.
Very large	Four vehicles, an officer and various other units are required, depending on the type of incident.

Table 2.2: Incident size classifications.

GRIP	Description
0	Indicates an incident of limited proportions and requires no special coordination.
1	Indicates an incident of limited proportions and requires coordination between the various emergency services.
2	Indicates an incident with a definite effect on the surrounding area.
3	Indicates a threatening on the well-being of the population within a single municipality.
4	Indicates a threatening on the well-being of the population spanning more than one municipality, province or country.

Table 2.3: Incident scalings by GRIP value.

Paging message example In order to understand in what format information is disseminated over the public safety paging network, an example is given in this paragraph. The following paging message[6] was sent to five paging devices in safety region IJsselland on January 5, 2011.

```
   Prio 1 Woningbrand :  :  MONUMENTSTRAAT : 37 8102AJ RAALTE
2690 2631 (Opsch.Brw:  middel brand)
```

In this message a priority is given (i.e. `PRIO 1`), a descriptive location is missing, a street address (i.e. `MONUMENTSTRAAT 37`) is provided, as well as a zip code (i.e. `8102AJ`), a city name (i.e. `RAALTE`) and a classification (i.e. `Woningbrand`). The classification indicates that there is a house fire occurring. The fire is expected to be mastered by two fire fighting vehicles (i.e. vehicle numbers `2690` and `2631`) and an incident scaling is given (i.e. `Middel brand`), indicating that the fire fighters are dealing with a medium sized fire.

[6] This message was extracted from P2000Monitor.

2.2 Incident profile builder

In the previous section a incident data source was selected and the contents of paging messages were investigated. In this section the properties of an incident profile will be defined and a strategy is devised to build these profiles and evaluate the effectiveness of this strategy.

2.2.1 Designing an incident profile

The purpose of an incident profile is to store information about an incident that can be used to track incident tweets in Chapter 3. The following question will be answered in this section: what properties constitute a well-designed incident profile? To answer this question we will investigate what incident properties can distinguish one incident from another.

In the previous section various incident properties have already been mentioned in relation to the public safety paging network. A paging message is designed such that the paged emergency responder knows what to do, where to go, what vehicle to take, etc. There is no ambiguity in the content of a paging message. We decide to use some of the properties of the paging message. Table 2.4 shows the design for an incident profile, by means of six profile properties.

Profile properties	Description
Classification	Type of incident (e.g. fire, gas leak, accident, power failure)
Address	Street name, number and zip code, or alternatively a highway number.
City	City in which the incident originated
Descriptive location	Name of a building or place
Scaling	Size and/or GRIP value
Time of occurrence	Timestamp

Table 2.4: Properties of incident profiles

Compared to the paging message content, we have intentionally left out vehicle number and priority, and added time of occurrence. Since an incident profile serves to track tweets reporting on an incident, we see no value in storing a priority and vehicle number, yet time of occurrence is trivial.

2.2.2 Building incident profiles

Previously, an incident data source was selected and an incident profile was designed. In this section a strategy will be chosen to use the former to obtain the latter. Hence, the objective is to build incident profiles based on data from an incident data source.

Retrieving paging messages Incident monitoring services republish the paging message stream through an Internet site or RSS feed. In a RSS feed each RSS item com-

prises a (1) title, containing the paging message, (2) description, containing the paging device number, and (3) timestamp, containing the time of occurrence. We will use the title and timestamp data field of a RSS item, as it contains the incident information required to build our incident profiles.

Information extraction strategy In this paragraph a strategy is devised to map the information contained in a paging message to the properties of an incident profile. The following message example is used to elaborate on information extraction techniques.

```
   Prio 1 Woningbrand :  :  MONUMENTSTRAAT : 37 8102AJ RAALTE 2690
2631 (Opsch.Brw:  middel brand)
```

A straightforward strategy is to derive the paging message syntax. Generally, a paging message is created conform some predefined message format. A string pattern matching strategy can be applied to match strings of text within such message, or split the string of text into multiple information fragments.

Regular expressions provide a concise and flexible means for matching strings of text, such as characters, words, or patterns of characters. It is written in a formal language that can be interpreted by a regular expression processor, a program that either serves as a parser generator or examines text and identifies fragments that match the provided specification. Given the previous example, we have designed the following regular expression[7]:

```
   ^Prio (\d) (.*) :(.*) :(.*) :(.*) (\d{4}\w{2}) ([^\d]+) ([\d
]+) (.*)
```

The subsequent pairs of parentheses match from left to right: priority number, classification, descriptive location, street address, house number, zip code, city name, vehicle numbers and incident scaling. In the example message the descriptive location is omitted as the incident is classified as a house fire; so, no building number exists. The set of classification values in a paging message can be mapped to the classification values of the incident profile.

Each safety region exploits a different message format per discipline. It is time-consuming to create message formats for all these cases. Therefore, since the police does not extensively use the public safety paging network, we will omit these paging messages. Ambulance services on the other hand are often assisted by fire fighters as soon as incidents scale up. For instance, when an incident scales to GRIP 1 or higher, a fire fighter officer must formally be present. Therefore, we choose to create regular expressions only for paging messages sent to fire fighters; that is, one expression for each of the twenty-five safety regions.

Limitations Not all paging messages can be parsed as assumed in the previous paragraph. Issues rise when delimiter characters are missing in the text. Without these characters, which enable text splitting, it is in some cases ambiguous whether a string of words is a part of street address or a descriptive location. This problem could be

[7]This regular expression has been simplified for readability reasons. Deviations from the example paging message are present, which require more complex regular expression patterns.

tackled by adding semantics, e.g. lookup and match street names through Google Maps API[8]. Unfortunately, such services have usage limitations, which are quickly reached.

2.2.3 Evaluating the profile builder

The final step of this chapter is to evaluate the effectiveness of the profile building strategy. As incident profiles are used in the subsequent chapter to track tweets, it must be clear to what extent incident profiles are correctly build. A straightforward strategy for evaluation is to supply a set of paging messages and manually analyze whether or not an incident profile is correctly created. The author can be the assessor, as the task of identifying a correct or incorrect incident profile is rather simple.

For the evaluation step 6,521 paging messages are retrieved which were sent to members of the fire fighting brigade in 25 safety regions in a period of 2 months. Informal and test messages (e.g. stating "good luck on your duty today" or "please contact the PSAP") are excluded as they do not provide any incident profile information. Table 2.5 shows that 88% of all paging messages is correctly parsed. This number does not imply that the remaining 12% are incorrectly parsed, but they simply do not match the proposed regular expression. Appendix A displays the parsing results for each safety region. Here, one can observe that some regions achieve 100% parsability, while other regions only reach 70%. Further analysis shows that 88% of the paging messages mentioning a scaled incident could be parsed as well. Due to one safety region this number was negatively influenced: Limburg Noord. Here, it was impossible to match paging messages mentioning a scaled incident (resulting in a parsing score of 22%), since they have a whole different syntax. Finally, in four safety regions (i.e. Friesland, Drenthe, Zaanstreek-Waterland, Brabant Noord, Limburg Noord) it was impossible to distinguish the incident type from the place name. To overcome this problem an index could be created that stores all incident type text combinations. Here, we have only extracted an incident type (based on a list of most frequently recurring incident type terms).

Safety region	#Messages	%Parsed	
		Yes	No
All regions	6,521	88%	12%

Table 2.5: Parsing results for all/scaled paging messages

Table 2.6 shows four paging messages sent to the fire fighting brigade in safety region Utrecht. The regular expression applied to parse the messages is as follows:

```
^P \d ([^a-z ]+) (.*) \(?[A-Z]{3}\)?  (.*)Eenh:.*$
```

The first pair of parentheses extracts the incident type, the second pair extracts the address and city, and the third pair extracts the descriptive location with optionally a classification. Distinguishing an address from a city requires an additional operation, which is straightforward considering the general syntax of an address (containing a street name, house number and optional suffix). The first two messages in Table 2.6

[8] http://code.google.com/apis/maps

are successfully parsed, while the last two are not. Here, the third message fails to be parsed as it misses a reference to the vehicle number. Although this information is not required to build an incident profile, we notice that a vehicle is required to actually move to an incident site. Indeed, after interpreting this message more carefully one can read that the intention of the message is to request contact to the PSAP. The fourth message also fails to be parsed, as it applies a different syntax, i.e. the street address is provided in a lowercase sentence.

Paging message	Parsed
`P 1 OMS-BRANDALARM JAN VAN SCORELSTRAAT 34 UTRECHT (UTR) De Rading Meidenhuis Eenh: UTR922`	Yes
`P 1 BEDRIJFSBRAND LEIDSESTRAATWEG 235 a WOERDEN (WRD) Baderie Blankestijn (Classificatie: grote brand) Eenh: REG799 COH591 IJS916 ABH521 REG646`	Yes
`P 2 HULPVERLENING ODIJKERWEG 25 DRIEBERGEN (DRB) GAARNE TEL. KONTAKT RAC`	No
`P 1 SCHOORSTEENBRAND Jasmijnstraat 37 EDER Eenh: VND895`	No

Table 2.6: Examples of paging messages sent to safety region Utrecht

Non-parsable messages For each of the regular expressions we have analyzed why some paging messages cannot be matched. In general the main reasons are:

1. *Missing information*. Sometimes information is missing, such as a house number or vehicle number. When for example a house number is missing in a street address, it can be impossible to distinguish a streetname from another property, without applying semantic analysis.

2. *Different syntax*. Some messages have a divergent syntax, requiring a case-based solution. For example, when locations are given by street crossings or by highway, hectometer number and left/right lane indication. Additional regular expressions are required to match these cases.

By devising a case-based expression we could increase the parsing score for safety region Limburg Noord, from 22% to 67%, taking only paging messages with a scaling, which would increase the overall score as well. Similar expressions could be devised for other cases, but given the time-constraints of this research we will omit that step.

Handling duplicates Multiple paging messages may refer to the same incident. Therefore, a collection of created profiles is maintained in order to check for duplicates. A profile is considered a duplicate from another profile if the incident type, address and city are equal and the creation time differs no more than 24 hours. If an incident profile already exists, the scaling property will be updated accordingly.

2.3 Conclusion

The first step in this research was to find a strategy that can automatically detect incidents in real-time by means of an online incident data source. This strategy is then responsible for building incident profiles which uniquely identify an incident and, as will be discussed in Chapter 3, are used to track tweets that report on the incident.

First we investigated which online data source contains reliable and timely information about incidents. We discovered that the Dutch public safety paging network was the best option as it meets both criteria. We then analyzed the paging message content and identified what incident information is contained. Subsequently, we designed incident profiles, including the following profile properties: classification, address, city, descriptive location, scaling, time of occurrence. The challenge then was to find a strategy that extracts information from the paging messages and maps it to the incident profile. We chose to use regular expressions as they provide a concise and flexible means for matching strings of text, for example fragments with incident information. To measure the effectiveness of this strategy an evaluation step was carried out. Here, it was discovered that the regular expressions matched 88% of 6,521 paging messages. The remaining 12% could not be matched, because of missing information or because a distinct syntax is applied. Furthermore, in four regions it was impossible to distinguish an incident type from a place name, requiring another strategy (e.g. building an index of incident type strings of text) to distinguish one from the other. Finally, we showed a case-based regular expression could be applied to further increase the effectiveness of the incident profile builder.

Chapter 3

Filtering tweets based on incident profiles

The second step in this research is to find a strategy that can track tweets that report on an incident. The strategy is responsible for filtering tweets based on incident profiles and some of these tweets, as will be discussed in Chapter 4, are used to fulfill particular information needs during incidents.

In this chapter the following research question will be answered: how can tweets reporting on an incident be tracked? To find an answer to this question, the following subquestions have been derived: what is Twitter and how can tweets be retrieved from Twitter? What strategy can be applied to filter tweets per incident? How can the effectiveness of an incident tweet filter be evaluated?

3.1 Twitter as data source

This section serves as an introduction to Twitter. Any social networking service has its own characteristics, which need to be understood before one can successfully exploit the service. After this introduction to Twitter, tweet retrieval techniques will be discussed.

3.1.1 Characteristics of Twitter

Twitter, a popular micro-blogging service, has become one the fastest growing trends on the Internet, with an exponentially-growing user base exceeding 190 million users in July 2010[1]. On Twitter every user can publish short messages with a maximum of 140 characters, so-called tweets, which are visible publicly or semi-publicly (e.g. restricted to the user's designated contacts) on a message board of the website or through third-party applications. Its founders' original idea was to provide a service that enables personal status updates. Due to its popularity, tweets cover every imaginable topic,

[1] http://techcrunch.com/2010/06/08/twitter-190-million-users

ranging from political news to product information. The public timeline conveying the tweets of all users worldwide is an extensive real-time information stream of 65 million of messages per day [7].

Following someone on Twitter simply means receiving all their Twitter updates. Every time someone posts a new message, it will appear on a user's personal profile webpage on Twitter. New messages are added to this webpage as people post them, so the user always gets the latest updates in real time. The relationship of following and being followed requires no reciprocation: a user can follow any other user, and the user being followed is not required to follow back; this differs from other popular social web services like Facebook and MySpace[2].

In Twitter the following types of tweets are distinguished:

- *Singleton*. A singleton is a normal update, which contains no references to other tweets, users or tags. The message is undirected.
- *Reply*. A reply is directed to another Twitter user, and is distinguished from normal updates by the username prefix. If a message begins with `@username`, it is indicated as a reply.
- *Mention*. A mention is when a Twitter user refers to another person on Twitter. If a message contains a `@username`, but is not a reply, it is called a mention.
- *Direct message*. A direct message is a private form of communication between two Twitter users. Users can start their tweet with `D username` to send such message. These messages can only be sent to users who follow them.
- *Re-tweet*. RT is short for re-tweet, and indicates a re-posting of someone else's tweet. Although it is not an official Twitter command or feature, people add `RT` somewhere in a tweet to indicate that part of their tweet includes something they are re-posting from another person's tweet, sometimes with a comment of their own.

Each user on Twitter maintains a user profile, which can include a full name, the location, a web page, biography and number of tweets of the user. Furthermore, the people who follow the user and those that the user follows are listed.

Since Twitter provided no easy way to group tweets, the Twitter community came up with their own way: hashtags. A hashtag is a convention among Twitter users to create and follow a thread of discussion by prefixing a word with a `#` character. Popular words and hashtags are captured in trending topics. Twitter has built an algorithm, to select those words and hashtags that are used most at the moment. A list of the top ten trending topics according to Twitter is presented on Twitter's homepage.

Next, we will present a Twitter update from the Southern Arizona Red Cross in America (a specific division of the American Red Cross), officially published on March 30, 2011:

```
RT @RedCross In the next few days, we will increase our
assistance for #Japan by $50 million http://rdcrss.org/hJXyPq
```

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

The message is a re-tweet, originally posted by Twitter user `RedCross`. The message contains one hashtag (e.g. `#Japan`) and one link to an online article (e.g. `http://rdcrss.org/hJXyPq`). Often, Twitter messages contain links to external sources, such as online videos, photos, blog posts, news articles, etc. Due to the limited number of characters that can be used in Twitter updates, several initiatives have been founded to shorten hyperlinks, such as Bit.ly[3] and TinyURL[4], and the American Red Cross applies its own shortening service[5].

Tweets can be annotated with geo-location information. When enabled, the user can select a location to be assigned with the tweet or let the browser or mobile device determine the exact location.

User intentions Why do people use Twitter? Research shows that people use Twitter for various purposes [7, 18, 23]. Java et al. [7] analyzed user intentions of 94,000 Twitter users given a data collection of more than 1.3 million tweets. Four main intentions have been distinguished:

- *Daily chatter*. Users discuss their daily routine and what they are currently doing. This is the largest and most common user of Twitter.
- *Conversations*. About 13% of all posts in the data collection are replies, which indicate that Twitter users have a conversation.
- *Sharing information/URLs*. About 13% of all the posts in the collection contain URLs.
- *Reporting news*. Many users report latest news or comments on current events via Twitter. Due to Twitter's APIs various automated users have been subscribed to the service that post updates, like news stories or weather updates from RSS feeds.

Depending on these intentions, users will spread different types of content. Particularly during incidents it is valuable to know what content is being disseminated, and how this information can fulfill information needs during incidents. In Chapter 5 we will investigate these research questions.

3.1.2 Retrieving tweets from Twitter

In this section it will be investigated how tweets, with or without incident information, can be retrieved from the Twitter service. The developer usage limitations will be discussed too, as these limitations may influence how many tweets may be retrieved from the micro-blogging service, at which rate.

Twitter APIs Twitter offers third party application developers access to its data store. The Twitter API consists of three parts: two REST[6] APIs and a Streaming API. The

[3] http://bit.ly

[4] http://tinyurl.com

[5] http://rdcrss.org

[6] REpresentational State Transfer is a style of software architecture for distributed hypermedia systems such as the World Wide Web [5]

two distinct REST APIs are entirely due to history, as a company called Summize[7] was acquired by Twitter for search functionality and was rebranded as Twitter Search. On its website[8] Twitter notes that fully integrating Twitter Search and its own REST API into the Twitter codebase is more difficult. Until resources allow Twitter to unify these two APIs they will likely remain as separate entities. The Streaming API is distinct from the REST APIs as Streaming supports long-lived connections on a different architecture. Functionality of the different APIs are described as follows:

- *Original REST API*. These API methods allow developers to access core Twitter data, allowing developers to modify timelines, tweets, and user information.
- *Search API*. These API methods give developers ways to interact with Twitter Search[9] and trends. It is read-only and when developers access it, they are looking up tweets in the search database from the past.
- *Streaming API*. The Streaming API is the newest API, and provides near real-time high-volume access to tweets in sampled and filtered form. A sampling process serves to limit the processing and bandwidth capacity Twitter must provide to publish the stream, and also the processing and bandwidth capacity a client must provide to accept the stream. The stream can be filtered by keywords, lists of users that created the tweets, or locations of geo-tagged tweets.

Developer usage limitations Mendoza et al. [17] showed that the amount of tweets published during incidents can be enormous, e.g. during the earthquake in Chile on February 27, 2010 nearly 5 million tweets were published in only five days. Obtaining such number of tweets is not straightforward, since Twitter applies usage limitations on its APIs.

The original REST API Rate Limit is 800 tweets per request with a maximum of 150 or 350 requests per hour, depending on the authorization type. The Search Rate Limit is not made public to discourage unnecessary search usage and abuse, but it is higher than the REST Rate Limit. Search queries return up to 1500 tweets from up to six days back in time. Unfortunately, this time frame is getting smaller as Internet traffic to Twitter increases and has impact on the service's performance.

The Streaming API distinguishes five methods to access public statuses in near real-time: filter, sample, firehose, links, re-tweet and sample. The filter method returns public statuses that match one or more filter criteria, such as track keywords or geo-location bounding boxes. It allows up to 400 track keywords, 5,000 follow userids and 25 location boxes, given a default access level. The sample method returns a random sample of 1 to 10% of all public statuses, depending on the client's access level. The other three methods are not a generally available resource, as only few applications require this level of access. The firehose method returns all public statuses; the links method returns all statuses containing "http:" and "https:"; and the re-tweet method returns all re-tweets.

[7] http://blog.twitter.com/2008/07/finding-perfect-match.html
[8] http://dev.twitter.com/pages/api_overview
[9] http://search.twitter.com

3.2 Incident tweet filter

In the previous chapter incident profiles were designed and in the previous section techniques to retrieve tweets from Twitter were explored. In this section a definition of incident tweet will be given, a strategy to filter these tweets is devised and the effectiveness of this strategy will be evaluated.

3.2.1 Defining incident tweets

An incident tweet is a tweet that mentions or reports an incident. In this research a tweet is considered relevant to a particular incident when the tweet's content refers to or reports on the incident in question. On the contrary, a tweet is considered not relevant when a different incident or no incident at all is reported. Consider the incident profile as displayed in Table 3.1.

Profile properties	Values
Classification	Fire
Address	Vlasweg 4
City	Moerdijk
Descriptive location	Chemie Pack
Scaling	{Very large, GRIP4}
Time of occurrence	2011-01-05 at 14:28h

Table 3.1: Example of incident profile

Subsequently, consider the tweet examples of Table 3.2.

Tweet's content[10]
Big fire at Chemie Pack in Moerdijk: http://bit.ly/hElW67
Big fire at the port in Amsterdam http://tinyurl.com/4ddkv8b
I am on my way to Moerdijk for another day of work.
Citizens in Moerdijk report health problems.

Table 3.2: Examples of incident tweets

One can judge that the first tweet actually reports on the incident as captured by the incident profile. The second tweet reports on a different incident, the third example refers to no incident at all and the final tweet might refer to the incident, but one cannot judge without knowning the context. Actually, the fourth example does refer to the Moerdijk fire incident, since it was published shortly after the incident, and one could observe that more users started tweeting about health problems in relation to the Moerdijk fire incident. These cases will be cited again in a subsequent section.

3.2.2 Filtering tweets based on incident profiles

Previously, an incident profile was designed and incident tweets were defined. In this section a strategy will be chosen to use the former to obtain the latter. Hence, the objective is to filter incident tweets from the entire stream of Twitter messages based on incident profiles.

Selecting a Twitter API The first step in the process of filtering tweets is to choose an API to retrieve tweets. Both the Search API and Streaming API have features that enable tweet filtering based on various parameters, such as search terms, geo-location. However, their differences should be considered when selecting one or the other.

The Streaming API operates in nearly real-time and will therefore return the most recent tweets. The filter method allows developers to enter query terms, also known as track terms, or geo-location bounding boxes and the method returns tweets that match these parameters. By design, this API does not allow developers to retrieve tweets from the past. Furthermore, whenever the developer exceeds a rate limit, the filter method will return a sample stream, just like the sample method does.

The Search API has similar functionality, allowing developers to specify a search query, based on keywords, geo-location, language and time frame. The Search API will in turn return tweets that match the query. Although similar rate limits are applicable, one can delay the retrieval over time since the Search API can retrieve tweets from the past. Furthermore, by using multiple Twitter accounts (and IP addresses) the majority of tweets matching the query can be retrieved.

For the purpose of building and evaluating the incident tweet filter, we want to ensure that the complete collection of tweets is retrieved per incident. As Sakaki et al. [24] mention that large scale incidents yield many tweets, we consider the Search API to be more useful[11].

Filtering incident tweets The next step is to devise a strategy for filtering incident tweets. This filter should return all relevant incident tweets for a given incident profile. Basically, we could use the incident profile to build a search query which in turn will be used to filter tweets via the Twitter API.

A straightforward strategy is to use the information from an incident profile as input for the search query. An incident profile has the following properties: classification, address, city, descriptive location, scaling and time of occurrence. We assume that Twitter users will not likely mention an address in a tweet, and neither will they refer to an official scaling of an incident as given by emergency services. We assume that a classification is useful since we could guess the words that Twitter users will use when referring to such incident. For example, given a fire incident we could guess that users will mention terms like “fire” or “smoke” in their tweet to report on the incident. We derived a complete list of query terms per incident, which can be found in Appendix B.1.

[11] We do note that in a real-time situations where users want to be updated immediately the Streaming API is a better choice, optionally in combination with the Search API.

Time of occurrence is a useful property to obtain tweets after some point in time. Furthermore, we consider either the descriptive location or city to be useful in the search query, since we assume a Twitter user will not mention both in one tweet. Then basically our search strategy matches tweets that (1) refer to the incident type, (2) mention the city or descriptive location, and (3) are published after the time of occurrence. Table 3.3 displays two queries built by the incident tweet filter, which can be applied to search for tweets with the Twitter Search API. Since the API does not allow to filter on time of occurrence (only date), the application utilizing the API should take care of that.

Incident	Queries
Moerdijk fire	`(Moerdijk OR Chemie-Pack) AND (fire OR smoke OR flame OR cloud) SINCE:2011-01-05`
Goes fire	`(Goes OR Houtkade) AND (fire OR smoke OR flame OR cloud) SINCE:2011-02-15`

Table 3.3: Queries built by the incident tweet filter

3.2.3 Evaluating tweet filter

The next step of this chapter is to evaluate the effectiveness of the incident tweet filter. As incident tweets are used in the subsequent chapter to fulfill information needs, one should measure to what extent incident tweets are correctly filtered. A straightforward strategy to evaluate the incident tweet filter is to let a domain expert carry out the searching/filtering manually and then compare the outcome with the outcome of the incident tweet filter. Finding an external domain expert and committing this expert to frequently perform searches over a period of two months is far from being practical. Therefore, we choose the author of this research as executive domain expert.

In order to obtain a ground truth dataset of incident tweets the author applied an heuristic search approach. Heuristic methods are exploratory problem-solving techniques that utilize self-educating techniques (as the evaluation of feedback) to improve performance[12]. By trial and error the author attempted to devise an optimal search query, where optimal refers to maximum coverage of relevant tweets (tweets that report on the incident) and minimal coverage of non-relevant tweets (tweets that do not report on the incident). By analyzing and evaluating search results of each query, the author could identify the terms that contribute to an optimal search query.

For the evaluation step 72,929 tweets are retrieved given 4 incident profiles (excluding re-tweets). We have included only those profiles that covered at least 500 tweets. Table 3.4 displays the properties of the four incidents, the number of tweets and percentage of tweets that was found by the incident tweet filter. One can observe that all incidents are fires, and also a clear pattern is visible: the larger the incident scaling, the larger the number of tweets and the smaller the percentage of tweets found. The percentage of non-relevant tweets filtered is not provided, since that problem will be tackled by means of the ranking algorithms proposed in Chapter 5.

[12]http://www.merriam-webster.com/dictionary/heuristic

City	Type	Date	Size	#Tweets (manual search)	%Found
Moerdijk	Fire	Jan 5, 2011	Very Large	66,715	44%
Goes	Fire	Feb 15, 2011	Large	630	94%
Utrecht	Fire	Feb 18, 2011	Medium	605	92%
Amsterdam	Fire	Feb 22, 2011	Very Large	4,979	79%

Table 3.4: Results of the incident tweet filter

Table 3.5 displays the percentages at different time slots of the Moerdijk incident. Here, one can clearly observe that the effectiveness of the search query degrades over time. A suitable explanation is that the dynamics of the incident change the subject of discussion. For example, at $x = 0$ Twitter users mention "Fire in Moerdijk!", but at $x = 2$ the message is "Civic air alarm activated in Moerdijk!" [13].

Timeslot x (in hours)	#Tweets	%Found
0	591	79%
1	2,031	64%
2	6,440	53%
3	15,312	42%
4	8,914	39%

Table 3.5: Results of the incident tweet filter given the Moerdijk incident

3.2.4 Optimization

The previous section showed that the incident tweet filter does not perform well over time. Therefore, two approaches are considered to enhance the incident tweet filter:

- *Query reduction*. The easiest way is to omit one of the previously specified search parts. By omitting for example the guessed terms derived from the incident type, all tweets mentioning the incident profile's city or descriptive location would be matched. The side effect is that tweets not reporting on the incident can be falsely included. This number can be large when the incident occurred on a location which is already extensively mentioned in other contexts.
- *Query expansion*. A more comprehensive way is to analyze the contents of tweets matching the search query, and find relevant keywords in these contents. For example, when a few users start mentioning "health problems", this could be noticed and these keywords could be added to the search query.

We decide to apply query expansion to optimize the incident tweet filter, as query reduction has negative side effects.

[13] http://www.nrc.nl/nieuws/2011/01/05/grote-brand-chemisch-bedrijf-moerdijk/

Expanding search query An issue with query expansion is that most terms are stop words or other commonly used terms, and not the descriptive terms one would apply for query expansion. These commonly used terms should be disregarded when applying query expansion, so the first step is to devise a list of these terms.

The obtain a suitable list of commonly used terms, the Streaming API is utilized to retrieve a random sample (1-10%) of all Dutch tweets. Since this API has no feature to filter based on language, the hundred most frequently used Dutch terms[14] are tracked and only tweets from users with timezone "Amsterdam" are included. In a period of two weeks over 2 million tweets are retrieved. All terms are extracted from these tweets, indexed and counted. Subsequently, we decided to select the 10.000 most commonly used terms and manually remove all terms that relate to incidents. Appendix D shows a part of this list.

For each timeslot the incident tweet filter is applied to find incident tweets, as explained in the previous section. Within these filtered tweets all terms are extracted, indexed and counted. Subsequently, all commonly used terms are removed given the list we devised, so only terms remain that are specificly used during the incident. Finally, only the terms mentioned five times or more are selected to expand the search query. Table 3.6 shows the resulted expanded queries[15]. An overview of the discovered query terms for additional timeslots can be found in Appendix C. In Table 3.3 one can observe that the query expands at each timeslot. In Chapter 4 the content of tweets will be analyzed to better understand why Twitter users mention such terms.

Timeslot *x* (in hours)	Expanded queries
0	`(Moerdijk OR Chemie-Pack) AND ((fire OR smoke OR flame OR cloud) OR (company OR "industrial area" OR chemical OR grip2 OR vlasweg)) SINCE:2011-01-05`
1	`(Moerdijk OR Chemie-Pack) AND ((fire OR smoke OR flame OR cloud) OR (company OR "industrial area" OR chemical OR grip2 OR vlasweg) OR (strijen OR dordrecht OR air-alarm OR "crisis control" OR "emergency channel" OR grip4 OR "smoke clouds" OR crashtender)) SINCE:2011-01-05`

Table 3.6: Queries built by the incident tweet filter

Table 3.7 displays the results of the query expansion approach for the first five timeslots of the Moerdijk incident. One can observe that the number of tweets found is higher than without query expansion, but it still decreases over time. An explanation is that some tweet contents, such as "The crisis.nl website is offline!" cannot be found, since the tweet does not contain the profile's city name or location description. This is

[14]Dutch terms are derived from: http://wortschatz.uni-leipzig.de/Papers/top100nl.txt

[15]All terms are translated from Dutch to English for readability reasons

a result of our desicion: we argue that at least the city name or location reference, as stored by the incident profile, should be mentioned in a tweet, since otherwise tweets from other incidents might be unintentionally tracked.

Timeslot x (in hours)	#Dutch tweets	#Tweets (ground truth)	%Found (original strategy)	%Found (query expansion)
0	9,275	65	79%	80%
1	9,299	206	64%	68%
2	10,317	506	53%	66%
3	11,594	1054	42%	58%
4	11,040	705	39%	49%

Table 3.7: Applying search query expansion on a sample of Moerdijk tweets

Limitations Whenever two scaled incidents occur at the same time in the same city, the incident tweet filter might not be able to correctly distinguish one incident from another, tracking all tweets for both profiles. Likely, a portion of tweets will actually report on both incidents, making it even a more complicated task allocate tweets to one incident or another. Here, we argue that a system does not necessarily have to make this distinction, as long as the users of the system are aware of this situation.

3.3 Conclusion

The second step in this research was to find a strategy that can track tweets that report on an incident. This strategy is responsible for filtering tweets based on incident profiles and these incident tweets are then used, as will be discussed in Chapter 4, to fulfill particular information needs during incidents.

First we investigated what characterizes Twitter and how tweets can be retrieved from the micro-blogging service. We showed that both the Search and Streaming API have functionality to filter tweets, but one should consider the developer usage limitations when choosing one or the other. Subsequently, we provided a definition of incident tweets and presented tweet examples to clarify the difference between a relevant tweet (one that is actually reporting on the incident as represented by the incident profile) versus a non-relevant tweet. The challenge then was to build search queries based on incident profiles that can filter incident tweets. First, we chose to use the Search API as its developer usage limitations can be circumvented by retrieving tweets from the past with a delay. Next, we used the properties of the incident profile to build a search query. For instance, the classification property was used to guess the words Twitter users will use when referring to such incident. Finally, the filtering strategy was evaluated based on effectiveness measures. It was shown that over 90% of all medium sized incident tweets were correctly filtered, but at large sized incidents this percentage dropped over time. Expanding the search query based on frequently mentioned terms (excluding stop words) improves the incident tweet filter.

Part II

Fulfilling information needs during incidents

Chapter 4

Identifying information needs

In the previous chapters strategies were devised to build incident profiles and filter incident tweets based on incident profiles. The objective of this chapter is to identify information needs of Twitter users during incidents by analyzing the tweeting activity and tweet contents during incidents. In Chapter 5 a strategy will then be presented to fulfill these identified information needs.

In this chapter the following research question will be answered: what information needs do users have on Twitter during incidents? To find an answer to this question, the following subquestions have been derived: how can a data collection of incident profiles and incident tweets be built? What analysis can be conducted to investigate tweeting activity and tweet contents? What information needs can be derived from these analyses?

4.1 Data collection

Before tweets can be analyzed, a data collection needs to be built. In this section a strategy is provided to build a data collection of incident profiles and incident tweets. Next, statistics are presented to better understand the characteristics of the collected data.

4.1.1 Archiving application

Both the incident monitoring service and the Twitter Search API have limitations regarding historical data retrieval, as described in Section 2.1.2 and Section 3.1.2 respectively. Both services allow retrieval of messages (paging messages and tweets) up to a limited number of days in the past. Hence, obtaining a large data collection with incident profiles and tweets requires an archiving solution.

Archiving incident profiles We have not discovered any third-party tool that features archiving of paging messages. Therefore, we have created a custom application. The application pulls the RSS feed of the incident monitoring service on a regular time

interval, parses each of the RSS items and stores the field's content of each RSS item in the database. The following fields are considered: paging message, pager number and timestamp. Measures are taken to avoid duplicates. Subsequently, another process is implemented which builds the incident profiles, as described in Section 2.2.2, and stores all profile property values in a database.

Archiving incident tweets Few third-party archiving tools exist that store tweets; examples are TwapperKeeper[1] and 140Kit[2]. These online tools allow users to enter keywords, whereas the service starts tracking these keywords in real-time for a period of time. Until recently, the collected tweets could be exported, but unfortunately Twitter's API Terms of Service[3] have changed regarding redistribution and syndication of content. As a consequence, all such tools have been forced to disable exporting features. Hence, without features to extract the data collection, it seems impossible to do extensive analyses.

We have chosen to create a custom application that can store tweets into a database (including a tweet's content, geo-location, user, etc.). An initial process builds a search query whenever a new incident profile is encountered. Another process establishes a connection with Twitter Search API, executes all search queries from the queue on a regular time interval. Whenever a search query returns no results in an interval of 24 hours it will be removed from the queue. Whenever the maximum number of requests is reached or when a user account is rate limited, an intentional delay is scheduled to avoid blacklisting by Twitter. As soon as the rate limit is reset by Twitter, the application continues to execute queries and retrieves new tweets.

Supervising the process In order to conduct a proper analysis it is important to obtain complete and accurate data. Although both the incident profile builder and tweet filter perform generally well, we choose to manually review and correct incident profiles and search queries to ensure that a complete and accurate dataset is obtained.

4.1.2 Collected data statistics

When collecting incident profiles and incident tweets we observe that most profiles do not return any tweets. Generally, all small sized and GRIP0 scaled incidents return zero or just a few tweets. The majority of incidents fall within this group. Hence, we have chosen to exclude these profiles, since these incidents do not have a substantial number of tweets for (statistical) analysis.

In a period of two months we collected 79 incident profiles and 110.976 tweets. Among these incidents we encounter fires, accidents, gas leaks and power failures. In Figure 4.1 the distribution of (a) incidents and (b) tweets are displayed for these types. As the illustration shows, mostly fire incidents occurred, covering 99% of the incident tweets in the dataset of tweets. Figure 4.2 and Figure 4.3 show the distribution of incident profiles and number of tweets by respectively size and GRIP value. The portion of

[1] http://www.twapperkeeper.com
[2] http://www.140kit.com
[3] http://dev.twitter.com/pages/api_terms

small sized and GRIP0 scaled incidents is not representative, as most of these incident profiles are excluded. One can observe that no GRIP3 incidents are reported during the collecting period. Furthermore, one can observe that high sized/scaled incidents occur less often compared to low sized/scaled incidents, but the number of tweets related to these few high scaled incidents is much higher.

4.2 Tweeting activity

In the previous section a data collection of incident profiles and tweets was build. In this section an analysis will be carried out to get an understanding of how Twitter is used during incidents, by analyzing the tweeting activity. A number of tweet distributions will be displayed and clarified.

4.2.1 Tweet distributions

The tweet activity x hours after each incident is displayed in Figure 4.4. Basically, at each hour the portion of tweets per incident, averaged over all incidents, is shown. One can see that most tweeting occurs in the first hour, but tweeting activity decreases just a few hours after. The solid line fits to some extent a power law distribution[4] including a long tail. Furthermore, one can see noticeable glitches in the solid line. These can be explained by interpreting the tweet contents at these timeslots. For example, at $x = 3$ one can read in the tweets that emergency services announced that soot in the smoke could potentially harm citizens, catching the attention of many users. At $x = 8$ a nightly car accident in Nederhorst on February 22, 2011 caused enormous disbelief among users the morning after. At $x = 13$ an online news story published about the parking garage fire in Amsterdam on February 18, 2011, hitting the attention of many online readers which dessiminated the news via Twitter.

We assume that the number of tweets published during night is lower than during daytime. Therefore, a modified tweet distribution is shown in Figure 4.4 which omits all tweets published at night (between 12AM and 6AM) and applies a time shift of six hours to bridge the gap. One can see that the number of glitches decreases. Furthermore, the glitch from $x = 13$ in Figure 4.4 shifts to $x = 7$ as result of the time shift. Indeed, the fire in the Amsterdam parking garage started at 10PM, while the majority of tweets were published the next day at 11AM.

Finally, a tweet distribution of incident profiles with more than five-hundred tweets is presented for eight quarters after the incident occurrence. In Figure 4.6 one can observe that the number of tweets increases between $x = 0$ and $x = 2$, and decreases between $x = 6$ and $x = 7$. The difference with the previous distribution figures suggests that each incident causes a unique tweeting activity. Therefore, an interpretative analysis will be done next to clarify such patterns.

[4]The equation is $y = 0.4003x^{-1,438}$

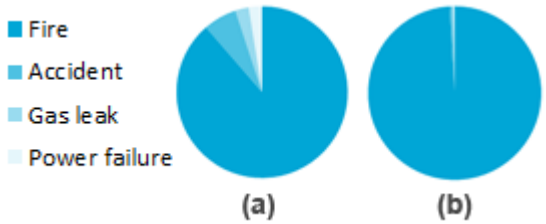

Figure 4.1: Incident types distributed by number of (a) incidents (b) tweets.

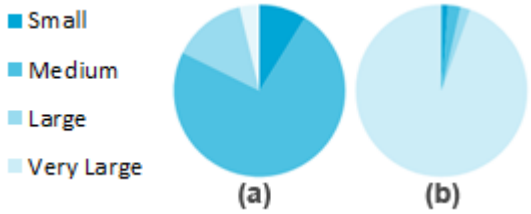

Figure 4.2: Incident sizes distributed by number of (a) incidents (b) tweets.

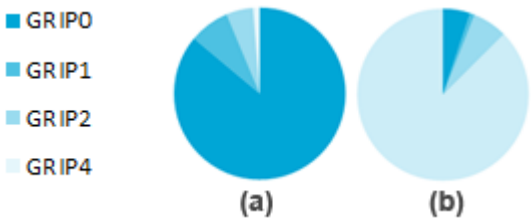

Figure 4.3: GRIP values distributed by number of (a) incidents (b) tweets.

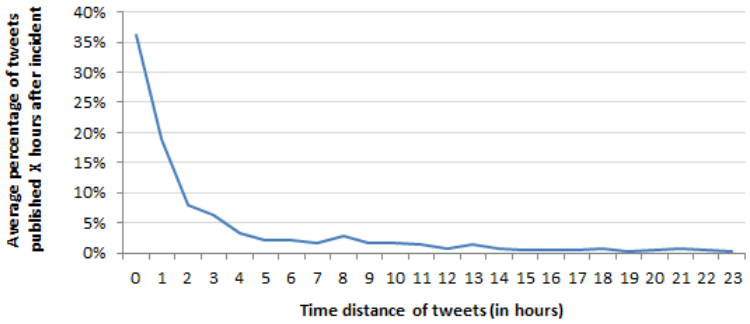

Figure 4.4: Tweet distribution at hour x after the incident.

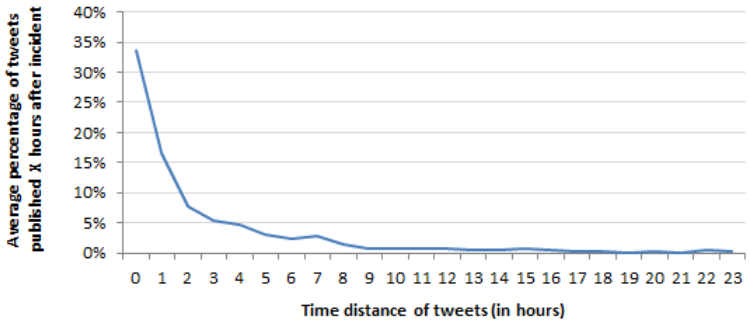

Figure 4.5: Tweet distribution after correcting tweet gap during night

4.3 Tweet content interpretation

This section will go one step further by actually analyzing the contents of tweets. This is done by means of an interpretative analysis. In order to present a clear and coherent analysis, one incident scenario will be selected for analysis. Subsequently, a number of content features will be chosen to give the analysis more structure. Finally, information needs will be derived.

4.3.1 Selecting an incident scenario

We consider the Moerdijk chemical fire incident most applicable for analysis, as it occupies 89% of the tweets in the data collection. In the following paragraph the course of events of this incident are described and the tweeting activity will be clarified.

Incident scenario A fire started on January 5, 2011 at 2:27 PM at the chemical storage and processing company Chemie-Pack in the port and industrial area of Moerdijk[5]. Large quantities of chemicals were exposed and smoke clouds passed over Dordrecht

[5]http://www.nrc.nl/nieuws/2011/01/05/grote-brand-chemisch-bedrijf-moerdijk/

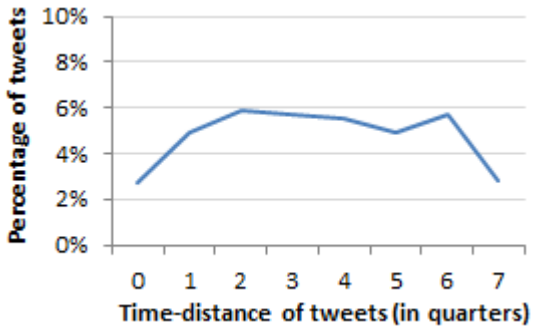

Figure 4.6: Tweet distribution at quarter x after incidents (500+ tweets)

and surrounding areas. Only six minutes after the fire started a complete fire brigade company was called (over 80 firemen), which indicate the seriousness of the incident and the potential exposure to hazardous substances. Three safety regions were involved: region Midden- en West-Brabant (scaled to GRIP4), region Zuid-Holland-Zuid (scaled to GRIP4), and region Rotterdam-Rijnmond (scaled to GRIP2). About 20 emergency responders were treated in the hospital regarding health problems, besides another 150 citizens who were present in the area at that time. During the fire citizens were advised to close windows and doors, to keep pets and cattle indoor and to eat no fruit and vegetables from own garden due to harmful soot. At 12:15 AM the fire was mastered.

Tweet distribution Figure 4.7 shows the tweet distribution of the Moerdijk incident. The pattern completely differentiates from the ones presented previously. At $x = 0$ many members of the public post that they see large black smoke clouds rising from the incident site. At $x = 1$ members of the public tweet that civil air sirens are activated in Dordrecht and surrounding areas. Soon it is understood that the smoke might be harmful as it comes from a chemical storage company with the name Chemie-Pack. The situation worsens at $x = 2$ when the Dutch national crisis website[6] goes offline, raising disbelief in the Twitter community. At $x = 3$ the incident is all over the national news becoming a trending topic. Between $x = 4$ and $x = 5$ we assume that most working people (try to) go home and have dinner, but we also observe that an unusual smell is perceived in multiple cities and breathing problems are reported. When at $x = 5$ authorities announce that the smoke is not harmful (according to measurements), the distrust of authorities is growing within the Twitter community. At $x = 6$ people see particles falling from the sky, questioning whether or not these are harmful. Only until $x = 10$, when the fire is mastered and midnight has started, the tweeting community becomes silent.

[6] http://www.crisis.nl

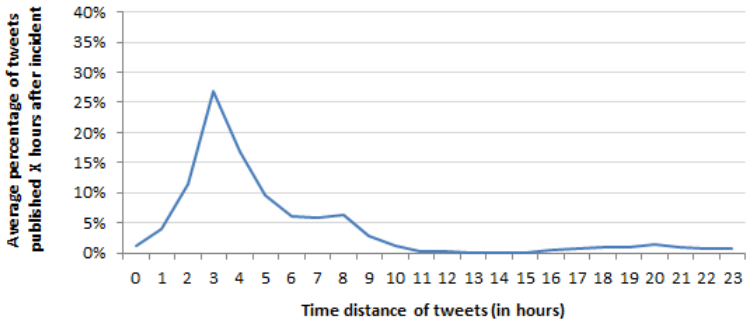

Figure 4.7: Tweet distribution at hour x (no time-shift applied)

4.3.2 Analyzing tweet content

In each of the following paragraphs a content feature will be presented and motivated, some which are characteristic for Twitter. These content features are then subject of an interpretative analysis. The objective of these analyses is to better understand what kind of content is shared via Twitter and eventually derive information needs during incidents from tweet contents.

Hashtags A hashtag is a convention among Twitter users to create and follow a thread of discussion by prefixing a word in a tweet with a `#` character. Our objective of analyzing hashtags in a set of tweets is to find patterns in the tweet content, i.e. find topics which are trending at the time of the incident.

Table 4.1 displays a top ten ranking of most frequently used hashtags at hour x[7]. For instance, at $x = 0$ one can observe that users mention `#dordrecht` in tweets, at $x = 2$ users mention `#toxicCloud` and at $x = 4$ the `#cnn` hashtag is popular. Without looking at the complete content of these tweets, it is difficult or simply impossible for an outsider to interpret them correctly. Consider the hashtag `#dordrecht`. One can only guess why this city name is mentioned during the Moerdijk fire. Hence, a term frequency analysis is carried out to find additional words that users mention in tweets containing the hashtag. Stop-words such as "the", "in" etc.[8] are filtered out. As a result one can see that words like `Air sirens`, `smoke` give background information about what is seen and heared by Twitter users, while words `close`, `windows` and `doors` comprehend a public safety instruction given by authorities. Both analyses suggest that an concise overview of an incident can be extracted from the tweet content.

URLs An URL is a means for a user to add more information to a tweet, e.g. by referring to pictures, news, blogs, etc. Our objective is to analyze these URLs and

[7] Author's note: a translation to English has been carried out to make the table more readable. The original data is in Dutch. For example, the hashtag `#fire` is a translation from the original Dutch hashtag `#brand`.

[8] A complete overview of Dutch words can be found on the website of the University of Leipzig: http://wortschatz.uni-leipzig.de/Papers/top1000nl.txt

$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
#moerdijk	#moerdijk	#moerdijk	#moerdijk	#moerdijk	#moerdijk
#fire	#fire	#fire	#fire	#fire	#fire
#grip2	#grip4	#rtl7	#rtl7	#toxicCloud	#nos
#chemiePack	#grip2	#grip4	#fireMoerdijk	#fail	#fireMoerdijk
#Dordrecht	#dordrecht	#fireMoerdijk	#toxicCloud	#fireMoerdijk	#toxicCloud
#news	#chemiePack	#dordrecht	#nos	#crisis	#dareToAsk
#vlasweg	#airAlarm	#chemiePack	#dareToAsk	#dareToAsk	#fail
#dateToAsk	#fail	#moerdijkFire	#rtl4	#big	#cnn
#photo	#alarm	#toxicCloud	#fail	#rtl4	#big
#bigFire	#omroepBrabant	#dareToAsk	#grip4	#cnn	#dordrecht

Table 4.1: Ranking of most frequently used hashtags at hour x.

Words
smoke
windows
doors
air sirens
close
smoke clouds
Brabant
Omroep
CNN
police

Table 4.2: Ranking of words from tweets mentioning #dordrecht

discover what kind of websites are frequently posted and what information is provided by these sources.

A straightforward strategy to find tweets with URLs is to filter tweets that contain the text “http:” or “https:”. In total 11,198 URLs are encountered in 23,841 tweets (including re-tweets), covering 863 Internet domains. We converted all shortened URLs to their original URL and extracted the Internet domains from the original URLs. The most popular domains are listed in Table 4.3. One can observe that mainly news media and image sharing services are listed in the top ten. The top five shows that particularly image content is popular to share during incidents on Twitter.

Additionally, the top three most mentioned URLs are shown in Table 4.4 and again both image sharing and news media websites are encountered.

Domain	Content type	Number of tweets	Number of URLs
twitpic.com	Photo sharing	2,093	549
nu.nl	News media	1,300	172
youtube.com	Video sharing	916	253
yfrog.com	Photo sharing	826	468
gigapica.geenstijl.nl	Photo sharing	425	8
rtl.nl	News media	409	146
nos.nl	News media	406	176
omroepbrabant.nl	News media	385	99
telegraaf.nl	News media	321	123
volkskrant.nl	News media	301	103

Table 4.3: Top 10 Internet domains mentioned in tweets

URL	Number of mentions
http://gigapica.geenstijl.nl/[...]	411
http://www.nu.nl/[...]	321
http://twitpic.com/[...]	165

Table 4.4: Top 3 URLs mentioned in tweets (URLs are partly displayed)

Geo-locations A geo-location of a tweet indicates the location of a user at the time tweeting. Our objective is to analyze a large set of geo-locations to find relations between the tweeting location and impact area of an incident over time.

Since tweets have either a geo-location annotation or not, the annotated tweets can be easily selected. In total 1,246 such tweets are encountered. Figure 4.8 displayes a plotting of these tweets on a map. One can observe that the highest density of tweets is north of Moerdijk. At $x = 0$ plots nearby Moerdijk are visible. At $x = 1$ and $x = 2$ the mayority of tweets originates from neighboring cities such as Dordrecht and Rotterdam; cities which are covered under the smoke cloud. Sakaki et al. [24] devised strategies to compute the impact area based on the coordinates of the geo-locations. These will not be applied in this research.

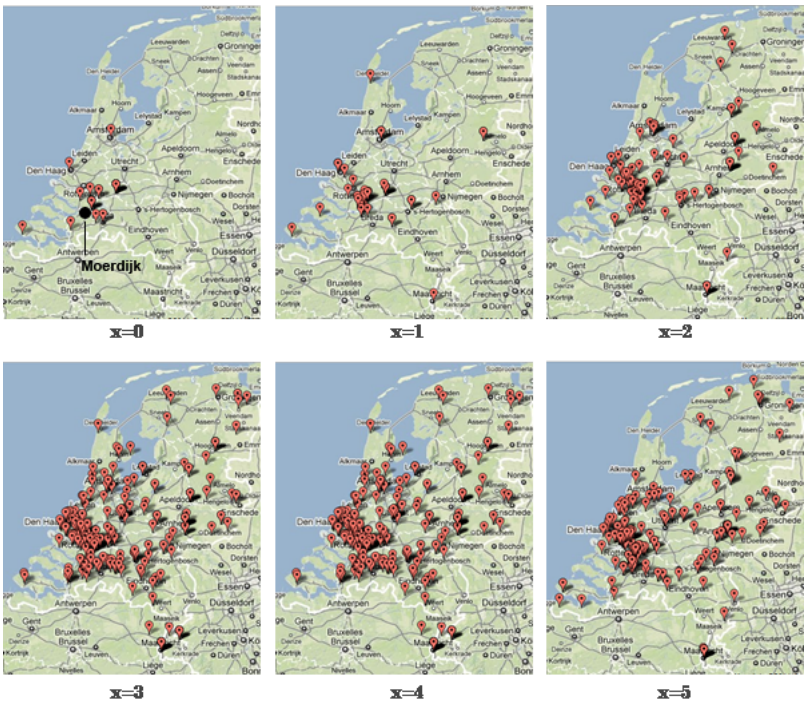

Figure 4.8: Tweets plotted on Google Maps at hour x after the incident.

Questions Questions in tweets indicate what information is requested by Twitter users during incidents. This request can be interpret as an attempt from a user to report a particular information need. Our objective is to analyze these questions and try to identify the information requests that are recurring in these questions.

A straightforward strategy to find tweets with questions is to filter tweets that contain a question mark. We choose not to include tweets that potentially resample emotional utterances (e.g. indicated by ??, ?! or !?), nor do we include tweets that have question marks within URLs. In total 11,531 such tweets (excluding re-tweets) are encountered in the Moerdijk fire incident. We have taken a sample of these questions and grouped similar ones into topics. Table 4.5 displays seven topics and typical questions found in tweets. The topic names are chosen in a way that they clarify the subject of the question.

Topics	Sample question
Status	What is currently happening / has happened?
Impact area	What locations are / will be impacted by the incident?
Risk	Does the incident form a risk to members of the public?
Related news	Where can I find the latest online news stories?
Casualties	What is the number of casualties?
Image reporting	What photos and videos show the incident site?
Instructions	What should I do?

Table 4.5: Classification of questions in tweets

4.3.3 Analyzing re-tweets

In the previous section an interpretative analysis was conducted on the contents of tweets. In this section re-tweets are subject of analysis. A re-tweet is simply a re-posting of a tweet from another user, making that content visible to a user's own followers. Generally, when a user encounters content that it wants to share with its own followers, it can post a re-tweet.

Extracting re-tweets Twitter Search API does not return the number of times a tweet has been re-tweeted. Therefore, this information is extracted from the dataset of incident tweets. When a tweet matches the regular expression `( RT |^RT )` it is considered to be a re-tweet. Subsequently the username of the re-tweeted user is extracted (indicated by `RT @username`) together with a signature of the tweet (we use the first 20 characters after the username). Both the username and signature are then matched against a tweet in the dataset of incident tweets.

What is re-tweeted? Cha et al. [3] argue that the number of re-tweets represents the content value of a tweet. The more people re-post a particular tweet the higher the content value is. We consider the number of re-tweets as a tweet's popularity or likeness indicator. This measure is useful when comparing tweets that report on a similar topic. For example, when two tweets report on the risk related to an incident and one tweet is re-tweeted sixty times, while another tweet is never re-tweeted, we consider the former tweet to be more relevant to the risk topic. This measure of relevance comes in handy when tweets are ranked based on relevance, as will be described in Chapter 5.

Table 4.6 displays the top five re-tweets of the Moerdijk incident[9]. One can observe that both amusing (tweet 1 and 4) and informative (tweet 2, 3 and 5) tweets are re-posted.

[9]The tweet contents are translated to English.

User	Original tweet content	Number of re-tweets
SuperrKelsey	Dear citizens of Moerdijk. What you hear next are NOT, I repeat NOT, the civil air sirens. It is Jennifer Ewbank on RTL.	269
NicoletteVanDam	Violent pictures #Moerdijk http://bit.ly/h3ipCH /via geenstijl.nl	250
DordrechtNL	Civil air sirens are activated in Dordrecht due to fire in #Moerdijk. Close windows and doors and deactivate automatic ventilation.	107
Koningin_NL	Sir #Wilders is concerned whether he can still bleach his hair, since all chemicals from #Moerdijk are burned.	103
GroteIncidenten	The civil air sirens were just activated in Dordrecht. If you live in this area, please close windows and doors and deactivate automatic ventilation.	93

Table 4.6: Top five re-tweets Moerdijk

Who is re-tweeted? Table 4.7 displays the top five re-tweets in the Moerdijk incident at $x = 3$ for the Risk topic. In order to find tweets that relate to the Risk topic, a query was designed based on words that users mention when referring to the risk associated with an incident, e.g. risk, danger, harmfull, health complaints. Here, one can observe that only informative tweets are present. Furthermore, we notice that four of the re-tweeted users are dedicated news reporting accounts, whereas "johnvandertol" is a person who works for a news reporting organization, more specifically a national broadcasting company. These examples reveal that the type of user publishing the tweet matters to those users that re-tweet it. In Chapter 5 we will further discuss why we believe users re-tweet other users.

User	Original tweet content	Number of re-tweets
AT5	The smoke clouds originated from the chemical fire in Moerdijk pose no threath to Amsterdam. http://at5.nl/s/emZ	59
112twente	The smoke of the Moerdijk fire moves North! http://is.gd/katNP No danger for Twente at this moment! Please RT!	22
Regio_Waterland	Fire Moerdijk not dangerous for Noord-Holland http://bit.ly/hEjgqf	20
johnvandertol	Police: Smoke from Moerdijk causes no danger for Alphen. Rain might change that #grip4 #moerdijk	12
omroepzeeland	Toxic cloud Moerdijk not dangerous for Zeeland http://bit.ly/eanM61 #omroepzeeland	7

Table 4.7: Top five re-tweets Moerdijk at $x = 3$ for topic Risk

4.3.4 Deriving information needs

An information need is an individual's or group's desire to locate and obtain information to satisfy a conscious or unconscious need [15].

In Section 4.3.2 an inital step was taken to derive information needs by analyzing questions in tweets. We argued that questions in tweets represent information needs. Similar questions were then grouped into topics of information needs.

Here, we note that a kwalitative study (i.e. questioning people involved in incidents about their information needs) could potentially provide a more complete overview of information needs, as one cannot be sure that all (latent) information needs are posted on Twitter during incidents and that the group of people joining the conversation is representative. Yet, by investigating a large sample of relevant tweets from different users of nearly 80 incidents, we do argue that this strategy is anyhow representative, and saves time compared to conducting a kwantitative study.

Typical information needs Information needs are thus derived from the tweet contents in the dataset of tweets. Similar questions are grouped and typical information needs are extracted. The result is similar to Table 4.5, yet an additional topic was identified: recovery time. Each of the topics is then mapped to the four incident types covered by the dataset of incident profiles, as displayed in 4.8.

	Incident types			
Topic of information need	Fire	Accident	Gas leak	Power failure
Status	x	x	x	x
Impact area	x	x	x	x
Risk	x		x	
Related news	x	x	x	x
Casualties	x	x		
Image reporting	x	x		
Instructions	x		x	
Recovery time				x

Table 4.8: Matrix of information needs during incidents

Particularly during power failures people want to know when their household power is expected to be recovered. Based on this knowledge one can decide what to do (e.g. stay home and wait, or visit a place where the heating actually works). The nature of an incident determines largely the set of information needs that users have, e.g. during a gas leak members of the public want to know the risk involved and get clear instructions: is it dangerous and should I evacuate?

Fulfilling information needs Can Twitter actually be exploited to fulfill information needs during incidents? The answer to this questions can be found in the contents of the tweets. We use the Moerdijk example again to manually match sample answers to the information needs. Table 4.9 displays such answers at $x = 3$. One can observe that three hours after the incident occurrence no casualties are found, and recovery time is not applicable to this type of incident. The sample answers show that Twitter can actually be exploited to fulfill information needs during incidents. In Chapter 5, a strategy will be proposed to predict which tweets can fulfill these typical information needs.

Information need	Sample answer (at $x = 3$)
Status	The flames are getting bigger. #moerdijk
Impact area	Smoke cloud moves to Amsterdam and IJmuiden
Risk	No danger for citizens in North Holland http://ow.ly/3yKZk
Related news	Fire Moerdijk strikes second company http://bit.ly/dLeva4 #news
Casualties	
Image reporting	First images Moerdijk: http://bit.ly/f64n7l #moerdijk
Instructions	Close doors and windows and turn off automatic ventilation #dordrecht
Recovery time	N/A

Table 4.9: Sample answers for each classification

4.4 Conclusion

The objective of this chapter was to identify information needs of Twitter users during incidents by analyzing the tweeting activity and tweet contents. Based on these information needs, Chapter 5 can present a strategy to fulfill them by ranking tweets based on relevance.

First we investigated how a data collection of incident profiles and incident tweets could be built. We showed that a custom archiving application was the best solution to store a large quantity of messages (paging messages and tweets). Statistics showed that mainly fire incidents are covered in the dataset of incident profiles. Next, we started analyzing the tweeting activity during incidents and showed that the majority of tweets is published in the first few hours, which is representative for low scaled/sized incidents. Large scale showed distinct tweet distributions, which can be explained by actually interpreting the tweet content. An interpretative analysis was carried out on the Moerdijk fire incident from January 5, 2011. The first analysis showed that the most frequently used hashtags and words could provide an concise overview of topics discussed during an incident. Analysis of URLs showed that image sharing services and news media are the most popular sources to link to. Furthermore, plots of tweets with geo-location on a map could reveal the progress of the incident observed by Twitter users: initially tweets originate from Moerdijk where smoke clouds are detected, but after a few hours tweets popup from surrounding towns where the smoke cloud is heading. Finally, the analysis and classification of questions in tweets reveal the information needs that Twitter users have during incidents. Eight typical information needs were identified and further analysis proved the feasibility of exploiting Twitter to fulfill these information needs. What remains, is devising a strategy that can actually fulfills these typical needs.

Chapter 5

Ranking incident tweets based on relevance

The final step in this research is to device a strategy that can locate incident tweets which fulfill information needs of Twitter users at a particular time. The information needs have been identified, therefore the objective of this chapter is to find relevant tweets. When this final objective is achieved, the main research goal will be met as well.

In this chapter the following research question will be answered: how can tweets, that fulfill information needs during incidents, be located? To find an answer to this question, the following subquestions have been derived: what concepts from Information Retrieval can be applied to rank documents? What strategies can be applied to rank tweets? How can the effectiveness of these strategies be evaluated and compared?

5.1 Information Retrieval

In this section basic notions and techniques from Information Retrieval (IR) will be explained, as they serve as introduction to the ranking problem that will be elaborated on in a subsequent section.

5.1.1 Introduction

Ranking is a subject extensively covered in IR studies. Manning et al. [15] define IR as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Kowalski [11] states that the general objective of an IR system is to minimize the overhead of a user locating needed information. Needed information is then defined as all information that is in the system that relates to a user's need. However, in the case of large document collections the resulting number of matching documents can far exceed the number a human user could possibly sift through. The needed information refers to retrieving sufficient information to complete a task, allowing for missed data.

A common example of an IR system is a web search engine, such as Google[1] which retrieves URLs, or an online university libary catalogus[2] which retrieves pointers to online books, journals, etc.. What these systems have in common is that they reduce what has been called "information overload". In the examples users are not likely interested in all information that relates to their information need, but merely the most relevant documents. In reality the definition of relevance is not a binary classification but a continuous function, where the term relevant document is used to represent a document containing the needed information.

5.1.2 Ranking documents

When facing large document collections it is essential for an IR system to rank-order the documents matching a query. Therefore, a system should compute a score for a document matching the query. Here, a query expresses an information need in machine readable language. This section will first elaborate on a particular ranking strategy and subsequently present measures to evaluate ranking strategies.

Ranking with TF-IDF A document that mentions a query term more often is more relevant to a query and should therefore receive a higher score. The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j. Thus we have the term frequency (tf), defined as follows [15]:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j})}$$

where $n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $\left|d_j\right|$.

The inverse document frequency (idf) is a measure of the general importance of the term. This factor is motivated from the fact that if a term appears in all documents in a set, then it loses its distinguishing power. The idf can be obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient:

$$idf_i = log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|}$$

With:

- $|D|$ the total number of documents in the collection.

[1] http://www.google.com
[2] http://library.tudelft.nl

- $|\{j : t_i \in d_j\}|$ the number of documents where the term t_i appears (that is $n_{i,j} \neq 0$). If the term is not in the collection, this will lead to a division-by-zero. It is therefore common to add 1.

Tf-idf is then defined as:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_j$$

Next, the overlap score measure of a document is the sum, over all query terms of query q, of the number of times each of the query terms occurs in d. Thus, not the number of occurrences of each query term t in d, but instead the tf-idf weight of each term in d.

$$score_{q,d} = \Sigma_{t \in q}(tf - idf)_{t,d}$$

Evaluating retrieval results With respect to an information need, a document in a test collection is given a binary classification as either relevant or nonrelevant (a wider classification scheme could be applied too). This decision is referred to as the ground truth judgment of relevance. Assessing this relevance is often a time-consuming and expensive process that involves human judges. By increasing the number of judges, the relevance assessment for a document with respect to an information need will likely converge.

The two most frequent and basic measures for IR effectiveness are precision and recall [15]. These are first defined for the simple case where an IR system returns a set of documents for a query (not in rank-order).
Precision is the fraction of retrieved documents that are relevant:

$$Precision = \frac{\#(relevant\_documents\_retrieved)}{\#(retrieved\_documents)}$$

Recall is the fraction of relevant documents that are retrieved:

$$Recall = \frac{\#(relevant\_documents\_retrieved)}{\#(relevant\_documents)}$$

A perfect precision score of 1.0 means that every result retrieved by a search was relevant, whereas a perfect recall score of 1.0 means that all relevant documents were retrieved by the search.

In a ranked retrieval context, particularly in search engines, what often matters is how many good results there are on the first page or the first three pages. This leads to measuring precision at fixed numbers of retrieved results, such as 10 or 30 documents. This is referred to as *Precision at k*, for example "Precision at 10". The disadvantages of this measure is that it does not average well, since the total number of relevant documents for a query has a strong influence on precision at k. Alternatively, *average precision* could be applied to compute the average of precisions at the point of each of the relevant documents in the ranked sequence. Finally, the *mean average precision* is the mean of the average precision scores for each query.

5.2 Incident tweet ranker

In the previous section an introduction was given to IR systems. The objective of this section is to design a ranking strategy for incident tweets such that the most relevant tweets fulfill the typical information needs of Twitter users during incidents. First, relevance assessments will be obtained and queries will be derived for the typical information needs. Next, the tf-idf technique will be put into practice to create an initial IR system that can return tweets in ranked order. Finally, a custom strategy based on tweet features will be designed, evaluated and compared to tf-idf based ranking.

5.2.1 Assessing relevance

In Section 4.3.3 re-tweets were analyzed. There we argued that whenever one tweet is re-tweeted more frequently than another tweet and both tweets report on the same topic (match the same query), the first tweet is considered more relevant in respect to the topic. Hence, we consider re-tweeting as a rating mechanism and the number of re-tweets as the ground truth judgement of relevance. We do notice that distinctive reasons exist for re-tweeting, e.g. tweets can be informative or amuzing. This should be taken into account when designing a custom ranking strategy, which should not return the amuzing tweets. The aim of a ranking strategy is then to predict which tweet will be re-tweeted, or in other words, to predict which tweet will likely be judged relevant.

5.2.2 Designing queries

In Table 4.8 information need topics were derived from the questions analyzed for four incident types covered in the dataset of incident profiles. In this section queries will be designed for each of these typical information needs. These queries will not match tweets with questions (indicated by "?") nor re-tweets (indicated by "RT"). Table 5.1 displays the queries for each of the information needs. Word conjugations are omitted and queries are stated in English to improve the readability.

Information need	Query
Image reporting	$photo \vee video \vee picture \vee image$
Risk	$danger \vee harm \vee health \vee complaint \vee risk$
Casualties	$dead \vee casualty \vee deceased \vee injured \vee victim \vee sick$
Instructions	$advise \vee recommend$
Related news	$news \vee broadcast$
Impact area	$see \vee hear \vee smell \vee taste \vee feel$
Recovery time	$what\,time \vee how\,long \vee recover$
Status	$at\,this\,moment \vee right\,now$

Table 5.1: Query for each typical information need.

In order to find tweets that report on the *impact area* we propose to locate tweets that are field observations. By field observation is meant that users are actually reporting what they see/hear/smell/taste/feel during incidents. We assume that whenever

they do such observation, they will mention a location as well. Furthermore, we acknowledge another approach would suit as well, e.g. computing impact area based on geo-locations of tweets, as proposed by Sakaki et al. [24], or by simply plotting geo-locations on a map such that users themselves can assess the impact area, as presented in Figure 4.8.

For information need *status* we have chosen to find tweets that refer to something that is actually happening right now. Twitter users tend to use the words "at this moment" and "right now" to mention these situations.

We do not claim that these queries return all tweets that could fulfill information needs, however we do claim that relevant tweets are found by these queries and that these tweets fulfill information needs.

5.2.3 Ground truth rankings

In the previous sections a ground truth of relevance judgements was established and queries for typical information needs were defined. Based on these findings ground truth rankings can be established. However, we notice that time is an important aspect in relation to incidents, as incidents are highly dynamic events. A continuous stream of tweets produced during an incident might bring new information that better fulfills information needs. Therefore, considering this time-dependency, we choose timeslots of one hour by which the dataset of incident tweets can be split. Then, for each timeslot and incident a ground truth ranking will be established. Since not all timeslots contain tweets that are both re-tweeted and match the information need query, we decide to establish a ranking only for those timeslots which have both. More specifically, a ranking will only be built if at least five relevant tweets are found.

Table 5.2 shows that 54 ground truth rankings are established, for each timeslot one. In case of the *risk* information need 19 relevant timeslots are included from 5 incidents. Within these 19 timeslots 54,536 tweets are published, of which 962 are re-tweeted. Note that the number of tweets does not include re-tweets and questions, as mentioned previously.

Information need	Number of timeslots	Number of incidents	Number of tweets	Number of relevant
Image reporting	15	3	53,678	400
Risk	19	5	54,536	962
Casualties	1	1	14,126	11
Instructions	5	1	36,699	112
Related news	7	1	46,558	79
Impact area	5	2	23,622	63
Status	2	1	18,371	12
Recovery time	0	0	0	0

Table 5.2: Statistics of ground truth rankings

5.2.4 Ranking strategy 1: tf-idf

The goal of this chapter is to devise a strategy that can rank tweets in such a way that the contents of the tweets fulfill typical information needs during incidents. In the previous section ground truth rankings were established. These rankings will be used to evaluate the effectiveness of ranking strategies. In this section an initial strategy will proposed, based on tf-idf . We have chosen tf-idf as it is a weighting scheme often used by search engines as a central tool in scoring and ranking a document's relevance given a query. The effectiveness of this strategy will be evaluated by means of precision and recall measures.

Computing scores For each tweet in the ground truth ranking a tf-value can be computed. Computing idf-values requires more consideration, since the general importance of a term may differ dramatically for each incident type. At gas leaks the term "smell" may be popular while at fires the term "see" might be more popular. By computing the idf-value based on the tweets of the particular incident type, the general importance of a term is better computed.

Given a tweet and query the tf-idf value of each query term is computed. Since each query term in a query is seperated by a logical *OR* operator the actual score is simply the highest tf-idf value of all query terms. Tweets are then ranked according to their score.

Evaluating effectiveness The strategy can be evaluated based on precision and recall measures, where we compare each of the ground truth rankings with the tf-idf based rankings. This is done at fixed points: precision at 5, 10 and 20. Table 5.3 shows the average precision and recall values for each of the information needs.

	$k = 5$		$k = 10$		$k = 20$	
Information need	*AP*	*AR*	*AP*	*AR*	*AP*	*AR*
Image reporting	0.00	0.00	0.00	0.00	0.00	0.00
Risk	0.02	0.01	0.03	0.02	0.09	0.09
Casualties	0.00	0.00	0.00	0.00	0.00	0.00
Instructions	0.00	0.00	0.00	0.00	0.00	0.00
Related news	0.00	0.00	0.00	0.00	0.00	0.00
Impact area	0.00	0.00	0.00	0.00	0.00	0.00
Status	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.3: Average precision and recall for each information need at k=5,10,20

It does not take much time to realize that tf-idf does not apply, at all. Table 5.4 displays the top five tweets predicted by the tf-idf based strategy for one of the nineteen rankings of the risk information need.

The contents of these tweets are indeed not very informative, which explains that none of these tweets are re-tweeted. What these tweets have in common is the limited number of terms. Hence, term frequency is influenced by the document size (in terms). Particularly, in short documents term frequency is sensitive to small variations

User	Tweet content	#RT
RianneElisabeth	What a danger! #moerdijk	0
McRenato1	Wow, dangerous toxic cloud #moerdijk	0
bobasbreuk	No dangerous substances measured #moerdijk	0
naaomi2210	Fire is not healthy #moerdijk	0
Anket020	Fire of Moerdijk seems unhealthy	0

Table 5.4: Sample top five tweets predicted by tf-idf based strategy

in document sizes. Hence, we argue that tf-idf does not apply since fewer terms do not make a tweet more relevant. On the contrary, full length tweets are often more relevant since they contain the maximum amount of information, in the even so small amount of space in tweets. Furthermore, in order to fit a long message in a tweet a user has to carefully choose terms, which implicitly increases the quality of the message. Finally, tf-idf does not take into account social factors, e.g. who is tweeting? In Section 4.3.3 it was already shown that typical users, such as news media, are re-tweeted more often. This calls for a new strategy.

5.2.5 Ranking strategy 2: tweet features

In the previous section an initial ranking strategy was proposed based on tf-idf. In this section a relevance metric is proposed that considers tweet features. The aim is to identify the key tweet features that influence a tweet's relevance. The metric is used to build a custom scoring function, which can be evaluated based on effectiveness measures, again by computing precision and recall.

Relevance metric Why is a tweet perceived as relevant? The findings in Section 4.3.3 suggest that the informativeness of a tweet's content as well as the kind of user influences relevance judgements of tweets. Among the most re-tweeted users particularly news media and journalists are found. One can argue that these users are more trusted to be reporters of reliable and accurate content, since it is their professional job (i.e. they verify news sources before they publish a news item). Furthermore, from a social perspective one will naturally trust a police man, mayor or other public safety representative or spokesman more easily than some thirdteen year old kid when discussing an incident on Twitter. It is the responsibility of these people (both news media en public safety representatives) to inform members of the public timely and accurately during incidents and they will likely not take the risk to act the opposite.

The identity of these authoritive Twitter users can be determined by examining the biography on Twitter. When words like "news media", "broadcasting", "public safety officer" (a complete list of authority terms is found in Appendix E) are encountered, one can judge that these users represent authoritive persons or organizations in real-life. However, there is no guarantee that these Twitter accounts actually represent the persons or organizations they claim to be. Yet, by assessing the social reputation of a Twitter user this issue may be solved. A Twitter user with 20.000 followers claiming to be the official Twitter account of a national newspaper is more likely to be the legitimate

owner than a user who has only 10 followers. Given the characteristics of Twitter a user cannot so easily enforce other users to follow them, without first making a sincere and real impression.

Thirunarayan et al. [27] argue that a user can be given a global trust value (more appropriately called reputation) and a trust value that depends on a trust scope (e.g. authority on a topic). The trust scope captures the domain/context/task/function for which a trust relationship is applicable. For instance, user $a1$ may trust user $a2$ for user $a2$'s ability to produce quality content in a trust scope because user $a2$ is knowledgable in that trust scope.

Based on these notions and previous rationale, we propose a relevance metric that considers the following dimensions: informativeness and trustworthiness, whereas trustworthiness is given by reputation and authority. Table 5.5 shows these three dimensions and states how they can be measured. We argue that a tweet is likely to be informative when it matches an information need query, as it might contain content that fulfills the information needs. Furthermore, we argue that users with a large audience will likely publish more valuable and/or reliable content than users with a small audience, since a user's popularity is somehow earned by previous actions on Twitter. Finally, we argue that a user's background, given by its biography, will influence the perceived authority.

Dimension	Measurable tweet feature
Informativeness	Term count of query terms in tweet content
Reputation	Number of followers of user
Authority	Term count of authority terms in user biography

Table 5.5: Tweet feature per relevance dimension

Designing a scoring formula The next step is to design a scoring formula based on the tweet features. Basically, the score for tweet d on query q is determined by the multiplication of its informativeness and user's trustworthiness. We choose the following scoring formula given tweet d and query q:

$$score_{q,d} = informativeness_{q,d} \times trustwortiness_d$$

Where:

$$trustworthiness_d = c1 \times reputation_d + c2 \times authority_d$$

And:

$$informativeness_{q,d} = nor(\Sigma_{i \in q} tc_{i,d})$$

$$reputation_d = nor(fc_d)$$

$$authority_{q,d} = nor(\Sigma_{j \in A} tc_{j,d})$$

With:

- c_1 and c_2 the constants for weighting tweet features reputation and authority,
- $tc_{i,d}$ the term count of term i in document d,
- fc_d the follower count for the user that published document d,
- $nor(x)$ the normalization function applied to x such that x is between 0 and 1,
- A is the set of authority terms.

The formula shows that a user can be trustworthy either because of its reputation or because of its authority. A tweet, on the other hand, will be relevant when it is both informative and published by a trustworthy user. Indeed, an uninformative tweet should by default obtain a low relevance score, as well as tweets published by untrustworthy users. The formula shows these distinct cases by respectively a weighted summation and multiplication of these normalized tweet feature scores.

Assigning weights to tweet features In the previous section two unknown weighting factors were identified, c_1 and c_2 for respectively reputation and authority. In this section these constants will be computed. Basically, an optimal value can be calculated given all 54 ground truth rankings. However, to evaluate the effectiveness of the ranking strategy a distinct testset should be used. Therefore, the set of ground truth rankings will be split into a trainingset and testset of rankings. We choose 36 rankings covering all information needs, whereas the 18 rankings in the testset cover most information needs as well (casualties only has one ranking).

An optimal value for both constants is then calculated by means of an linear support vector machine (SVM). SVMs have become one of the most prominent machine learning techniques for high-dimensional sparse data commonly encountered in applications like text classification. One such tool, SVM^{rank} [8], can process a large set of rankings and ranking features to compute weights for each of the features. These computed weights complete the ranking formula.

Evaluating effectiveness The final step is to evaluate the ranking strategy based on precision and recall measures. The test dataset contains 18 ground truth rankings which are compared with the rankings produced by the custom ranking strategy, at fixed points: precision at 5, 10 and 20. Table 5.6 shows the average precision and recall values for each of the information needs.

These results show clearly that our custom ranking strategy is more effective than tf-idf based ranking. Now, to better understand what these results mean, Table 5.7 presents the top five tweets for the risk information need predicted by the custom ranking strategy for one of the nineteen ground truth rankings.

5.2.6 Comparing ranking strategies

Comparing Table 5.7 to the top five tweets produced by tf-idf based ranking, the custom strategy produces clearly better results: all tweets are informative. The first three

	$k = 5$		$k = 10$		$k = 20$	
Information need	*AP*	*AR*	*AP*	*AR*	*AP*	*AR*
Image reporting	0.35	0.07	0.37	0.14	0.30	0.23
Risk	0.27	0.03	0.32	0.06	0.32	0.13
Casualties	0.40	0.18	0.30	0.27	0.25	0.45
Instructions	0.08	0.02	0.18	0.08	0.22	0.20
Related news	0.34	0.15	0.20	0.18	0.18	0.32
Impact area	0.40	0.16	0.34	0.27	0.35	0.56
Status	0.60	0.50	0.40	0.67	0.22	0.75

Table 5.6: Average precision and recall for each information need at k=5,10,20

User	Tweet content	#RT
nosheadlines	According to the mayor of Moerdijk, no dangerous substances are measured.	26
nrc	Mayor: no danger public health, no casualties. Fire fighters: fire mastered within half hour http://bit.ly/gKLvRG #moerdijk	10
nu_nl	No dangerous substances measured: MOERDIJK - By the fire in Moerdijk no dangerous substances are ... http://bit.ly/f05VlB	1
RoelJewel	TV Rijnmond is more clear about the smoke and warns that it could be dangerous after all, due to the rain. #Moerdijk	0
Bas_Taart	There are NO harmfull chemical substances released, says a spokesman. This one, I guess: http://bit.ly/ckY8ut #moerdijk	20

Table 5.7: Sample top five tweets predicted by custom ranking strategy

tweets are from official news media, scoring high on authority. The last two tweets are published by users with a high reputation on Twitter.

These results are promising, since the contents of these tweets fulfill a typical information need. Hence, a user of this IR system would instantly understand the risk associated to this incident. Finally, the relevance metric applied to build the ranking strategy gives confidence that the located tweets are both informative and published by a trustworthy user.

5.3 Conclusion

The final step in this research was to device a strategy that can locate incident tweets which fulfill information needs of Twitter users at a particular time.

First we introduced concepts from Information Retrieval studies. We showed how

documents can be ranked based on tf-idf and how effectiveness of ranking strategies can be measured based on precision and recall. Subsequently, we derived a ground truth of relevance judgements based on re-tweet counts. We argued that whenever one tweet is re-tweeted more frequently than another tweet and both tweets report on the same topic (match the same query), the first tweet is considered more relevant in respect to the topic. We designed queries for all typical information needs and created 54 ground truth rankings, one for each query and timeslot of an incident. The aim was to design two ranking strategies and compare them based on effectiveness measures. First, we applied the tf-idf strategy, where we computed the idf per incident type. Unfortunately, the effectiveness of tf-idf was poor and we learned that term frequency has a negative impact on computing relevance scores. The second strategy was based on a relevance metric based on informativeness of tweets and trustworthiness of users. We showed that a user's trustwortiness is either gained by reputation or authority within a domain. The key concept was to identify that particular users (news media and public safety persons/organizations) are generally trusted more when reporting on incidents compared to others (e.g. thirdteen year old kid), due to their professional responsibility. Based on these relevance dimensions tweet features were devised that quantify these dimensions. To assess the importance of each tweet feature, SVM^{rank} was used based on a training dataset of 36 rankings. Subsequently, the obtained scoring formula was applied on the test dataset and it was shown that the custom ranking strategy ranks informative tweets from trustworthy users higher, returning tweets that actually fulfill information needs at particular times.

Part III

Conclusions and future work

Chapter 6

Conclusions

In this chapter our research will be concluded. First a summary will be given in which the research goal and questions will be evaluated. Subsequently, some limitations of this research will be highlighted and discussed, as well as ethical issues which were not yet covered by this research. Finally, unsolved problems and uncovered topics are mentioned and proposed as future work.

6.1 Summary

In the introduction chapter the research goal and research questions were stated. In each of the subsequent chapters we tried to answer a research question. In this section it will be reviewed to what extent these questions are answered and whether or not the research goal is met. In the introduction chapter we proposed three process steps as depicted in Figure 6.1 to complete the research goal and we proposed a structured approach to build and evaluate each of the three components: incident profile builder, incident tweet filter and tweet ranker. The research questions relate to the different components (or input) of this model.

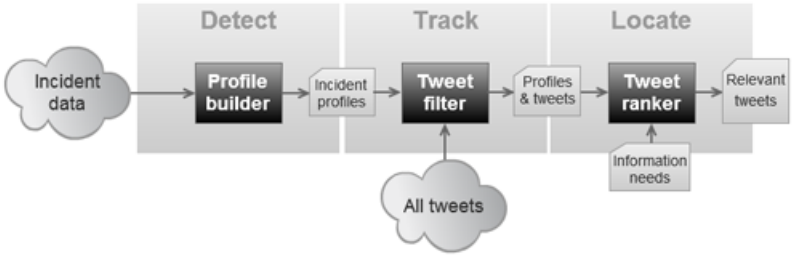

Figure 6.1: System perpective showing three main components

Reviewing research questions A summary per research question is given next.

- *How can incidents be detected based on unstructured data sources?*

We investigated what online data source contained reliable and timely information about incidents. We discovered that the Dutch public safety paging network was the best option as it meets both criteria. We then analyzed the paging message content and identified what incident information is contained. We designed incident profiles and proposed a strategy to extract information from paging messages and map it to incident profiles. Here, we chose to use regular expressions as they provide a concise and flexible means for matching strings of text. Finally, we measured the effectiveness of this strategy and showed that 88% of 6,521 test paging messages were correctly matched. For the remainder 12% additional case-based expressions could be created to increase effectiveness. Hence, we have shown that we can detect incidents and create incident profiles based on the unstructured data published by the public safety paging network.

- *How can tweets, that report on an incident, be tracked?*

We investigated what characterizes Twitter and how tweets can be retrieved from the micro-blogging service. We provided a definition of incident tweets and presented tweet examples to clarify the difference between relevant and irrelevant tweets. We proposed a strategy to build Twitter search queries based on incident profiles, by exploiting the incident properties. For instance, the classification property was used to guess the words Twitter users will likely use when referring to such type of incident. Finally, we measured the effectiveness of this strategy and showed that incidents with a small number of tweets (<1,000) score well (94%), but incidents with a large number of tweets (50,000+) score poor (44%) as time increases. An optimization step, based on query expansion, was proposed to improve these results. Hence, we have shown that tweets can be tracked given an incident profile.

- *What information needs do users have on Twitter during incidents ?*

We built a data collection of incident profiles and incident tweets, containing mainly fire incidents. We started analyzing the tweeting activity and conducted an interpretative analysis on the Moerdijk fire incident. We showed that the analysis and classification of questions in tweets reveal the information needs that Twitter users have during incidents. Eight typical information needs were identified and further analysis proved the feasibility of exploiting Twitter to fulfill these information needs.

- *How can tweets, that fulfill information needs during incidents, be located?*

First we introduced concepts from Information Retrieval studies. We showed how documents can be ranked based on tf-idf and how effectiveness of ranking strategies can be measured based on precision and recall. Subsequently, we derived a ground truth of relevance judgements based on re-tweet counts, we designed queries for all typical information needs and created 54 ground truth rankings, one for each query and timeslot of an incident. We then applied the tf-idf strategy and showed that the effectiveness of tf-idf ranking was poor and we learned that term frequency had a negative impact on computing relevance scores. We then applied a custom strategy based on a thoroughly devised relevance metric, wihch includes informativeness of tweets and trustworthiness of users. We showed that a user's trustwortiness is either gained by reputation

or authority within a domain. The key concept was to identify that particular users (news media and public safety persons/organizations) are generally trusted more when reporting on incidents compared to others (e.g. thirdteen year old kid), due to their professional responsibility. We quantified these dimensions by means of measurable tweet features and assessed the importance of each tweet feature using SVM^{rank}. Finally, we showed that this proposed ranking strategy returns tweets that actually fulfill information needs at particular times.

Reviewing research goal By having answered these research questions the main goal of this research can be reviewed. The research goal was to:

> Develop a strategy to locate tweets that fulfill information needs during incidents, by detecting incidents from unstructured data sources and tracking tweets that report on an incident.

The intended strategy has been developed and evaluated, and we can therefore state that the research goal has been achieved.

6.2 Discussion

In this research we have made several decisions that directed our research. At this point we can reflect on these decisions and discuss the impact and alternatives.

Selecting an incident data source When we decided to use the public safety paging network as input data source for the incident profile builder, we were not fully aware that some incident types are not covered by this media, such as crimes (e.g shootings, hostages). This limitation should be known when the intended system is used for real-life incident detection, as not all incident types can be detected.

Ethics of technology The engineer that will develop this application should consider moral principles that apply to incident information. The author of this paper strongly recommends that an application should not assist Twitter users to fulfill particular information needs. Consider photos of accidents. During this research, we were able to locate tweets that point to cruel photos of car accidents in which young people died. The author believes that respect and privacy should be paid to the relatives and families of these victims and any application should discourage anyone to locate such information.

6.3 Future work

Our research is directed by the decisions we made. Below we present various other paths that could be taken by other researchers.

Exploiting Twitter to detect incidents In Chapter 2 we decided to use the public safety paging network as incident data source for detecting incidents. Here, it might be interesting to see if Twitter can replace this incident data source, whereas one should find a strategy to deal with potential unreliable and unrelated content. Sakaki et al. [24] showed for example that earthquakes could be detected with a high accuracy, and their approach might be feasible to other incident types as well.

False rumor propagation on Twitter In Chapter 5 we created a relevance metric that was built upon the idea of trustworthiness of a user. Another approach might be to investigate the reliability of content. Mendoza et al. [17] showed that by analyzing tweets and replies, a level of doubt could be measured, whereas a high level indicates a higher uncertainty of reliability. This approach could be another implementation of the relevance metric that we proposed.

Measuring trustworthiness on Twitter It might be interesting to experiment with different strategies to measure trustworthiness. In the relevance metric we simply exploited "number of followers" as a way to compute reputation. More comprehence strategies could be applied, such as PageRank[20] or TwitterRank[30].

Bibliography

[1] Abel, F., Q. Gao, G.-J. Houben, and K. Tao. 2011 (May). Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *Extended Semantic Web Conference (ESWC), Heraklion, Greece*. Springer.

[2] Berghuijs, J. D. 2009. Eindrapportage expertgroep C2000. Technical report, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

[3] Cha, M., H. Haddadi, F. Benevenuto, and K. P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[4] Diakopoulos, N. A. and D. A. Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, New York, NY, USA, pp. 1195–1198. ACM.

[5] Fielding, R. T. 2000. *Architectural styles and the design of network-based software architectures*. Ph. D. thesis. AAI9980887.

[6] Gillmor, D. 2004. We the media: The rise of citizen journalists. *National Civic Review* 93, no. 3: 58–63.

[7] Java, A., X. Song, T. Finin, and B. Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, New York, NY, USA, pp. 56–65. ACM.

[8] Joachims, T. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, New York, NY, USA, pp. 217–226. ACM.

[9] Julie Letierce, Alexandre Passant, J. B. S. D. 2010 (4). Understanding how Twitter is used to spread scientific messages. In *Web Science Conference (WebSci10)*, Raleigh, NC, USA.

[10] Kaplan, A. M. and M. Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53, no. 1: 59 – 68.

[11] Kowalski, G. 1997. *Information Retrieval Systems: Theory and Implementation* (1st ed.). Norwell, MA, USA: Kluwer Academic Publishers.

[12] Kwak, H., C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, New York, NY, USA, pp. 591–600. ACM.

[13] Lee, K., J. Caverlee, and S. Webb. 2010a. The social honeypot project: protecting online communities from spammers. In *WWW '10: Proceedings of the 19th international conference on World wide web*, New York, NY, USA, pp. 1139–1140. ACM.

[14] Lee, K., J. Caverlee, and S. Webb. 2010b. Uncovering social spammers: social honeypots + machine learning. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 435–442. ACM.

[15] Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[16] Marnix Heskamp, Roel Schiphors, K. S. 2009. *Cognitive Radio Communications and Networks: Principles and Practice*. Academic Press.

[17] Mendoza, M., B. Poblete, and C. Castillo. 2010 (July). Twitter Under Crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA '10)*. ACM Press.

[18] Naaman, M., J. Boase, and C.-H. Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, New York, NY, USA, pp. 189–192. ACM.

[19] O'Reilly, T. 2007 (March). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. MPRA Paper 4578, University Library of Munich, Germany.

[20] Page, L., S. Brin, R. Motwani, and T. Winograd. 1999 (November). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[21] Palen, L., S. R. Hiltz, and S. B. Liu. 2007. Online forums supporting grassroots participation in emergency preparedness and response. *Commun. ACM* 50, no. 3: 54–58.

[22] Palen, L. and S. Vieweg. 2008. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, New York, NY, USA, pp. 117–126. ACM.

[23] Ramage, D., S. Dumais, and D. Liebling. 2010. Characterizing Microblogs with Topic Models. In *ICWSM*.

[24] Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, New York, NY, USA, pp. 851–860. ACM.

[25] Shklovski, I., L. Palen, and J. Sutton. 2008. Finding community through information and communication technology in disaster response. In *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, New York, NY, USA, pp. 127–136. ACM.

[26] Tao, K., F. Abel, Q. Gao, and G.-J. Houben. 2011 (May). TUMS: Twitter-based User Modeling Service. In *International Workshop on User Profile Data on the Social Semantic Web (UWeb), co-located with Extended Semantic Web Conference (ESWC), Heraklion, Greece*.

[27] Thirunarayan, K., P. Anantharam, C. Henson, and A. Sheth. 2010 (may). Some trust issues in social networks and sensor networks. In *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*, pp. 573 –580.

[28] Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*.

[29] Vieweg, S., A. L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, New York, NY, USA, pp. 1079–1088. ACM.

[30] Weng, J., E.-P. Lim, J. Jiang, and Q. He. 2010. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, New York, NY, USA, pp. 261–270. ACM.

Appendices

Appendix A

Parsing results

Table A.1 shows the percentage of tweets that was parsed for each of the 25 safety regions, given 6,521 paging messages of which 240 messages mentioned a scaling.

Safety region	Paging messages		%Parsed	
	Total	Scaled	Total	Scaled
Amsterdam-Amstelland	221	3	81%	100%
Brabant Noord	239	24	96%	92%
Brabant Zuid en Oost	266	1	100%	100%
Drenthe	153	7	70%	71%
Flevoland	107	2	78%	100%
Friesland	160	6	74%	100%
Gelderland Midden	250	8	90%	89%
Gelderland Zuid	229	11	79%	91%
Gooi en Vechtstreek	128	5	78%	100%
Groningen	110	2	77%	100%
Haaglanden	577	9	83%	98%
Hollands Midden	276	5	96%	100%
IJsselland	202	16	93%	100%
Kennemerland	273	1	72%	100%
Limburg Noord	179	13	75%	22%
Limburg Zuid	315	0	99%	0%
Midden- en West-Brabant	382	20	80%	100%
Noord en Oost Gelderland	468	25	94%	92%
Noord-Holland Noord	243	5	99%	80%
Rotterdam Rijnmond	643	12	98%	100%
Twente	174	5	98%	100%
Utrecht	421	13	95%	100%
Zaanstreek-Waterland	119	5	71%	100%
Zeeland	145	26	93%	77%
Zuid-Holland Zuid	241	16	70%	81%
Total/average	6,521	240	82%	82%

Table A.1: Results of parsing public safety paging messages

Appendix B

Query terms per incident type

Table B.1 shows the list of terms that Twitter users mention when referring to a particular incident type. All terms are translated to English for readability reasons.

Incident type	Query terms
Fire	$fire \vee smoke \vee flame \vee cloud$
Accident	$accident \vee collision \vee crash$
Power failure	$power\,cut \vee power\,failure$
Gas leak	$gas\,leak \vee gas\,smell \vee pipeline$

Table B.1: Query terms per incident type

Appendix C

Expanded search query

Table C.1 shows the terms that were encountered in the tweets of the Moerdijk incident at different timeslots. To optimize the incident tweet filter, these terms are added to the search query to find additional tweets that report on the incident. Terms are translated to English.

Timeslot x (in hours)	Query terms
0	company, industrial area, chemical, grip2, vlasweg
1	strijen, dordrecht, air alarm, crisis control, emergency channel, grip4, smoke clouds, crashtender
2	amsterdam, alarm phase, rijnmond, rtl, defense, hoekse, carcinogenic, toxic, toxic cloud, ijmuiden, water tanks, omroepbrabant
3	fire works, rijnmond, explosions, corrosive, ned1, blowing, rotterdam
4	mega fire, extinguish, harmful, crisisnl, gas cloud

Table C.1: Additional query terms to find tweets of the Moerdijk incident

Appendix D

Most commonly used terms

Table D.1 shows a portion of the list of the 10.000 most commonly used terms in tweets (excluding incident related terms), as derived from a sample of Dutch tweets. The terms are not translated to English, nor are they sorted in any way.

ik	zo	doen	deze	zin	zeg
de	weer	heeft	daar	iemand	hem
is	zijn	gaat	nee	kom	moeten
een	bij	even	mee	jou	ons
in	als	meer	alleen	jullie	kijk
en	al	mij	maken	tijd	onze
je	we	gaan	dag	vind	andere
het	moet	zien	you	zie	2010
op	ze	door	heel	onze	welk
van	kan	dit	komt	zal	leren
niet	over	net	man	kunnen	krijg
met	mijn	hebben	wie	laat	alle
dat	om	hoe	na	alles	gehad
voor	the	gewoon	wordt	werk	binnen
maar	echt	uur	nou	bent	zeggen
ook	ja	zit	2011	tegen	halen
nog	uit	eens	iets	hebt	onder
die	was	veel	via	steeds	tweet
dan	ga	leuk	zou	haar	leuke
me	geen	hier	twitter	want	laten
heb	jij	waar	mag	doet	maakt

Table D.1: Most commonly used terms in tweets

Appendix E

Authority query terms

Table E.1 shows the query terms to identify both public safety and news media persons or organizations. All query terms are translated to English.

Authority	Query terms
Public safety	help, safety, environment, region, crisis, police, fire fighter, ambulance, municipality, minister, spokesman, mayor, government, 112
News media	news media, news paper, broadcasting, channel, tv, television, radio, news editor, news reporter, journalist, news presentor

Table E.1: Query terms per incident type