Semantics + Filtering + Search = Twitcident
Exploring Information in Social Web Streams

Fabian Abel, Claudia Hauff, Geert-Jan Houben, Ke Tao
Web Information Systems, TU Delft
PO Box 5031, 2600 GA Delft, the Netherlands
{f.abel,c.hauff,g.j.p.m.houben, k.tao}@tudelft.nl

Richard Stronkman
Twitcident.com
Koningin Wilheminaplein 400, 1062 KS
Amsterdam, the Netherlands
richard@twitcident.com

ABSTRACT

Automatically filtering relevant information about a real-world incident from Social Web streams and making the information accessible and findable in the given context of the incident are non-trivial scientific challenges. In this paper, we engineer and evaluate solutions that analyze the semantics of Social Web data streams to solve these challenges. We introduce Twitcident, a framework and Web-based system for filtering, searching and analyzing information about real-world incidents or crises. Given an incident, our framework automatically starts tracking and filtering information that is relevant for the incident from Social Web streams and Twitter particularly. It enriches the semantics of streamed messages to profile incidents and to continuously improve and adapt the information filtering to the current temporal context. Faceted search and analytical tools allow people and emergency services to retrieve particular information fragments and overview and analyze the current situation as reported on the Social Web.

We put our Twitcident system into practice by connecting it to emergency broadcasting services in the Netherlands to allow for the retrieval of relevant information from Twitter streams for any incident that is reported by those services. We conduct large-scale experiments in which we evaluate (i) strategies for filtering relevant information for a given incident and (ii) search strategies for finding particular information pieces. Our results prove that the semantic enrichment offered by our framework leads to major and significant improvements of both the filtering and the search performance. A demonstration is available via: `http://wis.ewi.tudelft.nl/twitcident/`.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms, Design, Experimentation

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.

Keywords

Semantic Enrichment, Social Web Streams, Filtering

1. INTRODUCTION

During crisis situations such as large fires, storms or other types of incidents, people nowadays report and discuss about their observations, experiences and opinions in their Social Web streams. Therefore, valuable information that is of use for both emergency services and the general public is available online. Recent studies show that data from the Social Web and particularly Twitter helps to detect incidents and topics [15, 19, 23] or to analyze afterwards the information streams that people generated about a topic [7, 13, 18]. However, (i) automatically filtering relevant information about an incident from Social Web streams and (ii) making the information accessible and findable for people who are demanding information about an incident are two fundamental challenges that have not been answered sufficiently by literature yet.

In this paper, we tackle these two challenges and present Twitcident, a framework for filtering, searching and analyzing Twitter information streams during incidents. Here, filtering refers to an automatic process while search involves a user who is issuing a query. We showcase our framework and present the Twitcident system[1] which monitors emergency broadcasting services and automatically collects and filters Twitter messages whenever an incident occurs. Incidents are thus primarily events that typically require actions of emergency services. Twitter messages (tweets) as well as other types of Social Web status messages are typically very short—e.g. tweets are limited to 140 characters—which makes it difficult to identify relevant tweets. Initiatives such as the TREC task on filtering micro-blogging data[2] illustrate that there is currently a high demand in solving these filtering and search problems.

We approach these problems by enriching the semantics of short messages which includes named entity recognition, tweet classification as well as linkage to related external Web resources. Semantic enrichment also builds the basis for the search and analytics functionality that is provided by our Twitcident framework. Given the semantically enriched Social Web content about an incident, we allow users to explore information along different types of information needs (e.g. damage, casualties). Therefore, we inte-

[1] `http://twitcident.com`

[2] `http://sites.google.com/site/trecmicroblogtrack/`

grate faceted search strategies [1] that go beyond traditional keyword search as offered by Twitter[3] or topic-based browsing as proposed by Bernstein et al. [2]. Moreover, users can overview information by exploiting Twitcident realtime analytics that allow users to get an understanding of how different types of information are posted over time. The main contributions of this paper can be summarized as follows.

- We introduce a framework for incident-driven information filtering and search on Social Web streams. Our framework features automated incident profiling, aggregation, semantic enrichment and filtering functionality. Furthermore, it provides advanced search and analytics functionality that allows users to find and understand relevant information. (Section 3)
- We propose and evaluate strategies for solving two fundamental research challenges: (1) information filtering and (2) search on Social Web streams.
  1. We compare different stream filtering strategies on a large Twitter corpus and prove that the semantic filtering strategies of our Twitcident framework lead to major improvements compared to keyword-based filtering. (Section 4)
  2. We employ faceted search strategies that enable users to find relevant information in Social Web streams. Our evaluation confirms that the semantic faceted search strategies, which are applied on top of the filtered streams, enhance keyword-based search significantly. Contextualization (adapting to the temporal context of a search activity) and personalization (adapting to the interests of the user who performs the activity) gains further improvements. (Section 5)
- We apply our Twitcident system to incidents that happen during everyday life (mainly targeted towards the Netherlands) and discuss experiences and insights we gained from running Twitcident in practice. (Section 6)

2. RELATED WORK

In the last decade, Social Web platforms such as Twitter gained huge popularity and researchers started to investigate the motivation for using these systems [8], user behavior [12, 17], emerging network structures [10, 16], information propagation principles [10, 13, 18] and event detection based on Social Web streams [19, 23]. Yet, supporting users in finding information in Twitter streams has not been studied extensively yet. Chen et al. proposed strategies for recommending entire conversations on Twitter [4] as well as URLs that are posted in tweets [5]. Dong et al. [6] exploited Twitter data to improve the ranking of fresh URLs in search engines [6].

However, yet there exists little research on engineering search and retrieval of relevant information from Social Web streams. Marcus et al. [14] studied how to visualize Twitter streams. Bernstein et al. [2] proposed a topic-based browsing interface for Twitter in which a user can navigate through her personal Twitter stream by means of tag clouds. So far, there is a lack of research on how messages posted in Social Web streams can satisfy information needs of individual users. In fact, Teevan et al. [22] confirm studies that emphasize Twitter's role as news source [10, 20] and reveal that there are significant differences in the search behavior on Twitter compared to traditional Web search: Twitter users are specifically interested in information related to events and often use the rudimental search functionality of Twitter to monitor search results. With Twitcident, we introduce a framework that automates the process of monitoring relevant information published in Social Web streams and therefore reduces the efforts that users need to invest to satisfy their information needs. On top of the automatically filtered streams, Twitcident provides faceted search functionality as introduced in previous work [1].

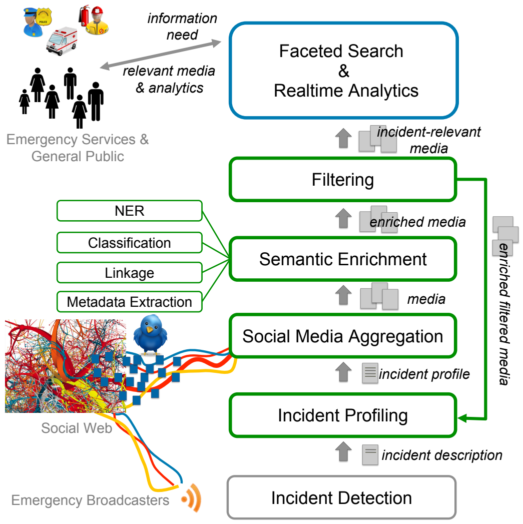


Figure 1: Twitcident architecture: (i) *incident profiling* and *filtering* of social media that is relevant to an incident (green boxes) and (ii) provide *faceted search* and *realtime analytics* functionality to explore and overview the media (blue box). Both types of components benefit from the *semantic enrichment*.

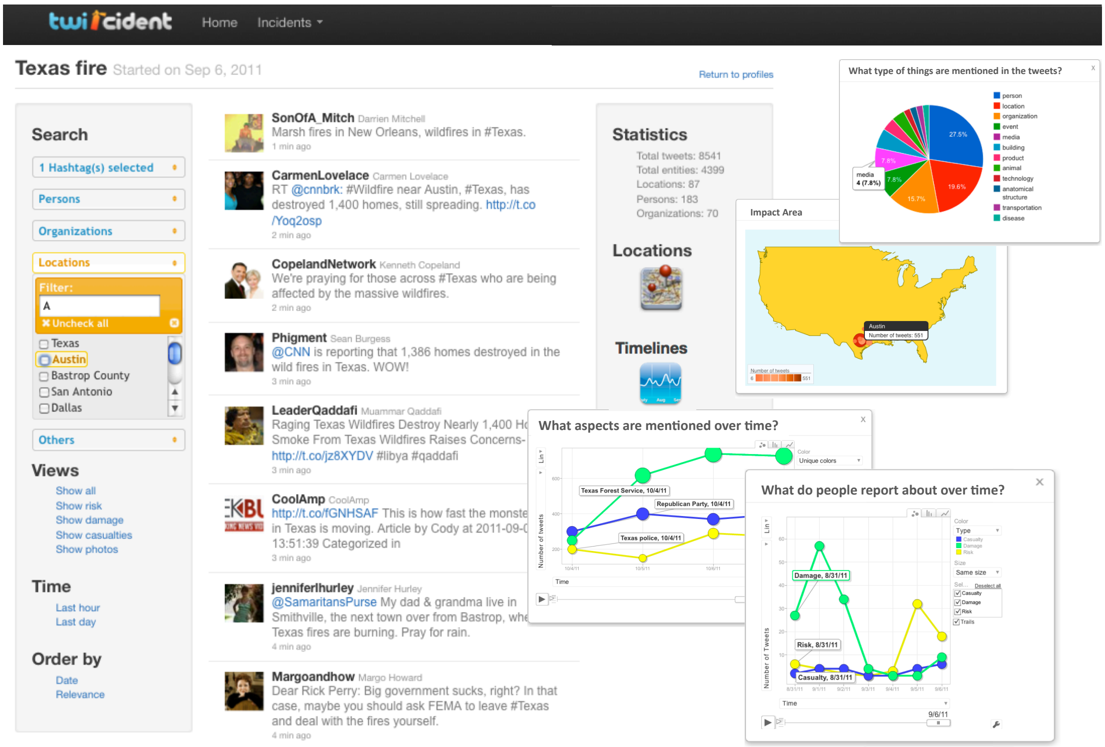
3. TWITCIDENT

In this section, we will overview the architecture of the Twitcident framework and detail its key components that allow for filtering, searching and analyzing information available in Social Web streams. The Web-based front-end of the Twitcident system is depicted in Figure 2 and allows users to explore and analyze information from Social Web streams during incidents such as natural disasters, fires or other types of emergency events.

3.1 Architecture

The Twitcident framework architecture is summarized in Figure 1. The core framework functionality is triggered by an incident detection module that senses for incidents being broadcasted by emergency services. Whenever an incident is detected, Twitcident starts a new thread for profiling the incident and aggregating social media and Twitter messages from the Web. The collected messages are further processed by the semantic enrichment module which features named entity recognition (NER), classification of messages, linkage of messages to external Web resources and further metadata extraction. The semantic enrichment is one of the key enabling components of the Twitcident framework as it (i) supports semantic filtering of Twitter messages to identify those tweets that are relevant for a given incident, (ii) allows for faceted search on the filtered media and (iii) gives means for summarizing information about incidents and providing realtime analytics.

[3] http://twitter.com/search



Texas fire Started on Sep 6, 2011

SonOfA_Mitch Darrien Mitchell
Marsh fires in New Orleans, wildfires in #Texas.
1 min ago

CarmenLovelace Carmen Lovelace
RT @cnnbrk: #Wildfire near Austin, #Texas, has destroyed 1,400 homes, still spreading. http://t.co/Yoq2osp
2 min ago

CopelandNetwork Kenneth Copeland
We're praying for those across #Texas who are being affected by the massive wildfires.
2 min ago

Phigment Sean Burgess
@CNN is reporting that 1,386 homes destroyed in the wild fires in Texas. WOW!
3 min ago

LeaderQaddafi Muammar Qaddafi
Raging Texas Wildfires Destroy Nearly 1,400 Ho
Smoke From Texas Wildfires Raises Concerns-
http://t.co/jz8XYDV #libya #qaddafi
3 min ago

CoolAmp CoolAmp
http://t.co/fGNHSAF This is how fast the monste
in Texas is moving. Article by Cody at 2011-09-0
13:51:39 Categorized in
3 min ago

jenniferlhurley Jennifer Hurley
@SamaritansPurse My dad & grandma live in
Smithville, the next town over from Bastrop, whe
Texas fires are burning. Pray for rain.
4 min ago

Margoandhow Margo Howard
Dear Rick Perry: Big government sucks, right? In that case, maybe you should ask FEMA to leave #Texas and deal with the fires yourself.
4 min ago

Statistics
Total tweets: 8541
Total entities: 4399
Locations: 87
Persons: 183
Organizations: 70

(a) Search (b) Filtered Social Web stream (c) Analytics

Figure 2: Screenshot of the Twitcident system: (a) search and filtering functionality to explore and retrieve particular Twitter messages, (b) messages that are related to the given incident (here: fires in Texas) and match the given query of the user and (c) realtime analytics of the matching messages.

In the Twitcident system, both faceted search and realtime analytics are made available to client users via a graphical user interface that is displayed in Figure 2. The search functionality allows end-users to further filter messages about an incident while analytics deliver diagrams and gadgets that enable users to analyze and overview how people report about the incident on the Social Web. We now discuss each of the components of our architecture in detail.

3.2 Incident Detection

For detecting incidents, the Twitcident system relies on emergency broadcasting services. In the Netherlands, incidents which require the police, fire department or other public emergency services to take an action and which are moreover of interest to the general public, are immediately published via the P2000 communication network and describe what type of incident has happened, where and when it happened and also what scale the incident is classified as. Figure 3(a) shows an example P2000 message informing about a large fire incident that happened in the city of Moerdijk, the Netherlands[4]. The figure visualizes the automatic workflow that is triggered whenever a new incident is reported. For a given incident it may happen that several different P2000 messages are broadcasted which requires Twitcident to first perform duplicate detection before starting a new incident monitoring thread. Therefore, the incident detection component compares the location, starting time and type of the newly reported incident with the incidents that are already monitored by Twitcident. If a new incident is detected then the Twitcident framework translates the broadcasted message into an initial incident profile that is applied as query to collect relevant messages from the Social Web and Twitter in particular. All incidents that are monitored by the Twitcident system are listed on the dashboard that is depicted in Figure 3(b).

3.3 Incident Profiling and Filtering

While monitoring an incident, Twitcident continuously adapts the incident profiling to improve the filtering of messages. This process is realized via the following components (see Figure 1): (i) incident profiling, (ii) social media aggregation, (iii) semantic enrichment and (iv) filtering.

3.3.1 *Incident Profiling*

Based on the initial incident description and the collected, enriched Social Web messages, the incident profiling module generates an incident profile that is used to refine the media aggregation and the filtering. An incident profile is a set of weighted facet-value pairs that describe the characteristics of the incident:

Definition 1. An *incident profile* of an incident $i \in I$ is a set of tuples $((f,v), w(i,(f,v)))$ where (f,v) is a facet-value pair that describes a certain characteristic f of the incident and $w(i,(f,v))$ specifies the importance of the facet-value pair for the incident that is computed by a weighting function w:

$$P(i) = \{((f,v), w(i,(f,v)))|(f,v) \in FVPs, i \in I, \\ w(i,(f,v))) \in [0..1]\} \quad (1)$$

Here, $FVPs$ and I denote the set of facet-value pairs and incidents respectively. A facet-value pair characterizes a cer-

[4] http://nl.wikipedia.org/wiki/Brand_Moerdijk_5_januari_2011

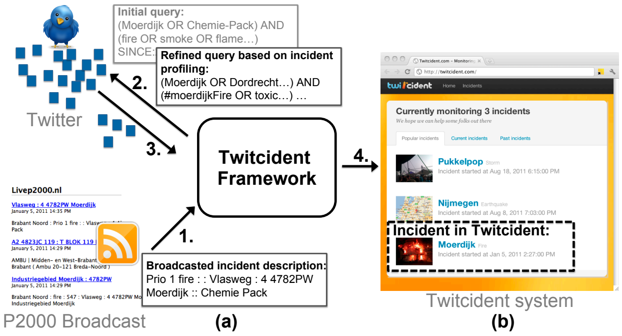

Initial query:
(Moerdijk OR Chemie-Pack) AND
(fire OR smoke OR flame…)
SINCE:
Refined query based on incident profiling:
(Moerdijk OR Dordrecht…) AND
(#moerdijkFire OR toxic…) …
2.
Twitter
3.
Twitcident Framework
4.
Livep2000.nl
Vlasweg : 4 4782PW Moerdijk
January 5, 2011 14:35 PM
Brabant Noord : Prio 1 fire : : Vlasw
Pack
A2 4823JC 119 : T BLOK 119
January 5, 2011 14:29 PM
AMBU | Midden- en West-Brabant
Brabant (Ambu 20-121 Breda-Noord)
Industriegebied Moerdijk : 4782PW
January 5, 2011 14:29 PM
Brabant Noord : fire : 547 : Vlasweg : 4 4782PW Mo
Industriegebied Moerdijk
1.
Broadcasted incident description:
Prio 1 fire : : Vlasweg : 4 4782PW
Moerdijk :: Chemie Pack
P2000 Broadcast
(a)
Twitcident.com – Monitoring
http://twitcident.com/
twitcident
Home
Incidents
Currently monitoring 3 incidents
We hope we can help some folks out there
Popular incidents
Current incidents
Past incidents
Pukkelpop Storm
Incident started at Aug 18, 2011 6:15:00 PM
Nijmegen Earthquake
Incident started at Aug 8, 2011 7:03:00 PM
Incident in Twitcident:
Moerdijk Fire
Incident started at Jan 5, 2011 2:27:00 PM
Twitcident system
(b)

Figure 3: Incident detection: (1) as soon as an incident is broadcasted via the P2000 network, the Twitcident framework (2) transforms the encoded P2000 message into an initial incident query to (3) collect Twitter messages that are possibly relevant for the incident so that (4) information about the incident can be accessed via the Twitcident system. Over time, the incident profiling effects refinements of the queries that are used to collect tweets. The screenshot shows the dashboard of popular incidents that are (and have been) monitored by Twitcident.

tain attribute (facet) of an incident with a certain value. Twitcident allows for various types of facets including locations, persons, incident classes or keywords. Therefore, the aforementioned fire that happened in Moerdijk may have the following incident profile: $P(i_{moerdijk}) = \{((location, Moerdijk), 1.0), ((location, Dordrecht), 0.73), ((type, Fire), 1.0), \ldots\}$. The weight that is associated with each facet-value pair ranges between 0 and 1: the higher the weight, the more important the facet-value pair for the incident. We apply the relative occurrence frequency as basic weighting strategy, i.e. the fraction of messages about the incident that mention the given facet-value pair. Incident profiles are continuously updated to adapt to topic changes that arise within an incident. To prevent topic drift, we combine the current profile with the initial incident profile following a classical mixture approach: $P(i) = \lambda P_{initial}(i) + (1 - \lambda)P_{current}(i)$ where we experimented with $\lambda \in [0..1]$ ranging between 0.25 and 0.5.

3.3.2 Social Media Aggregation

Based on the incident profiling, the Twitcident system exploits the social media aggregation component to collect Twitter messages as well as related pictures and videos that are posted on platforms such as Twitpic or Twitvid[5] respectively. Twitcident utilizes both the REST API and the Streaming API of Twitter[6] to collect messages. The REST API allows for querying Twitter messages that have been published within the last seven days and therefore enables Twitcident to collect those incident-related tweets that have been posted before Twitcident detected the incident. The Streaming API does not allow for querying previously published tweets but allows Twitcident to continuously listen for current tweets that mention keywords related to an incident.

3.3.3 Semantic Enrichment

The aggregated Social Web content (Twitter messages) is processed by the semantic enrichment component of Twitcident which features the following functionality.

NER. The named entity recognition (NER) module assembles four different services for detecting entities such as persons, locations or organizations that are mentioned in tweets: DBpedia spotlight, Alchemy, OpenCalais and Zemanta[7]. As those entity recognition services only function for English texts, Twitcident translates non-English tweets to English[8]. The extracted entities are mapped to concepts in DBpedia [3], the RDF representation of Wikipedia, and the type of an entity is utilized to specify the facet of the corresponding facet-value pair. For example, given a Twitter message such as *"#txfire is approaching Austin, 50 houses destroyed already http://bit.ly/3r6fgt"*, the NER module allows for detecting the facet-value pair *"(location, dbpedia:Austin_Texas)"*[9].

Classification. Twitcident classifies the content of Twitter messages into reports about casualties, damages or risks and also categorizes the type of experience that is reported in a tweet, e.g. whether the publisher of a tweet is seeing, feeling, hearing or smelling something. The classification is done by means of hand-crafted rules (e.g. *if a tweet mentions* $(X_1\ AND\ X_2 \ldots)\ OR \ldots$ *then classify as* Y) that operate on both the facet-value pairs and the plain words that are mentioned in a tweet.

Linkage. By following links that are posted within messages, Twitcident further contextualizes the semantics of a message. Therefore, the semantic enrichment module extracts the main content of the Web resource that is referenced from a tweet using Boilerpipe[10] and processes it via the NER module to further enrich the Twitter message with facet-value pairs that describe its content. For the aforementioned tweet which lists *"http://bit.ly/3r6fgt"*, one may extract additional facet-value pairs such as *"(location, dbpedia:Bastrop_Texas)"* or *"(organization, dbpedia:Texas_Forrest_Service)"*.

Metadata extraction. Twitcident also collects and infers additional metadata about Twitter messages such as pictures referenced from the tweet or background information about the publisher of a tweet; for example, the profile picture, number of followers, number of tweets published during the incident or the location of the user when publishing her tweets. Such provenance data is important for end-users to assess the trustworthiness of a tweet and is moreover exploited by the Twitcident system when tweets that match the current query are sorted according to their relevance (see the search in Figure 2(a)).

Enriched Twitter messages can therefore also be represented by means of a set of weighted facet-value pairs. In line with Definition 1, the profile $P(t)$ of a Twitter message $t \in T$ can therefore specified as: $P(t) = \{((f, v), w(t, (f, v)))| (f, v) \in FVPs, t \in T, w(t, (f, v))) \in [0..1]\}$.

3.3.4 Filtering

The goal of the filtering step is to automatically identify those tweets that are relevant to an incident. Therefore, the Twitcident filtering component first detects the

[5] http://twitpic.com and http://twitvid.com
[6] http://dev.twitter.com/docs

[7] http://dbpedia.org/spotlight, http://alchemyapi.com, http://opencalais.com, http://zemanta.com
[8] Language detection: http://code.google.com/p/language-detection/ Translation: http://code.google.com/apis/language/translate/overview.html
[9] The namespace abbreviation "dbpedia" points to: http://dbpedia.org/resource/
[10] http://code.google.com/p/boilerpipe/

language of a Twitter message and filters out all tweets that do not match the target language(s). In the deployed Twitcident system, we only consider Dutch or English tweets as relevant and discard Twitter messages for which we detect another language. Based on this pre-processing, the Twitcident framework features two core filtering strategies: (i) semantic filtering and (ii) semantic filtering with news contextualization.

3.3.4.1 Semantic Filtering.

Given the current incident profile $P(i)$ and the set of semantically enriched Twitter messages $P(t)$, the core challenge is to decide whether a tweet t is relevant for an incident i. The semantic filtering strategy therefore exploits the set of alternative labels of a DBpedia URI v that is mentioned in the facet-value pairs (f, v) of $P(i)$. If an alternative label is mentioned in the content of a Twitter message t then the corresponding facet-value pair (f, v) is added to the tweet profile. Given the further enriched tweet profile—denoted as $\bar{P}(t)$—and $P(i)@k$, the top k weighted facet-value pairs of the incident profile $P(i)$, the semantic filtering strategy computes the similarity between $P(i)@k$ and $\bar{P}(t)$ and considers a tweet t relevant to an incident i if $filter_{sem}(P(i), P(t)) = 1$:

$$filter_{sem}(P(i), P(t)) = \begin{cases} \mathit{1} & \text{if } sim(P(i)@k, \bar{P}(t)) > \delta \\ \mathit{0} & \text{otherwise} \end{cases} \quad (2)$$

In our experiments in Section 4, we use $P(i)@20$, apply the Jaccard similarity coefficient to compute $sim(P(i), \bar{P}(t))$ and set $\delta = 0$ as threshold. A Twitter message t is thus relevant if at least one facet-value pair of $P(i)@k$ also occurs in $\bar{P}(t)$.

3.3.4.2 Semantic Filtering with News Context.

As Twitter users might be influenced by public news media, Twitcident also monitors popular news agencies. The semantic filtering with news contextualization therefore extends the semantic filtering by enriching the incident profile $P(i)$ with information from mainstream news media before generating $\bar{P}(t)$. In particular, $P(i)$ is complemented with facet-value pairs that are extracted from related news articles. A news article is considered to be related to an incident if it matches the initial incident profile $P(i)$. The expanded incident profile $\bar{P}(i)$ is then used to perform the semantic filtering as described above. A tweet t is considered to be relevant to an incident i if $sim(\bar{P}(i)@k, \bar{P}(t)) > \delta$.

3.4 Faceted Search and Analytics

Incident detection, incident profiling, media aggregation, semantic enrichment and filtering are automatic processes that deliver information about an incident as reported by people on the Social Web. However, in order to find information in the filtered Social Web streams, appropriate functionality for search and analysis has to be engineered as well. The Twitcident framework approaches the challenge of retrieving relevant information from Social Web streams by means of faceted search as proposed in [1]. In this section, we re-visit the different faceted search strategies provided by the Twitcident framework and detail Twitcident analytics.

3.4.1 Faceted Search Strategies

The faceted search functionality allows users to further filter incident-related messages by selecting facet-value pairs that should be featured by the retrieved messages. A faceted query q thus may consist of several facet-value pairs. Only those tweets that match all the facet-value constraints will be returned to the user. The ranking of the tweets that match a query is a research problem of its own and is, in the context of micro-blogging systems, usually solved by ranking according to recency [22]. Twitcident ranks the matching tweets according to their (i) creation time or (ii) relevance. The relevance is computed by exploiting various features including provenance information such as the authority score of the user who published a tweet [21].

A key challenge in engineering a faceted search interface is to support the facet-value selection as good as possible. Hence, the facet-value pairs that are presented in the faceted search interface (see Figure 1(a)) have to be ranked so that users can quickly narrow down the search result lists until they find the tweets that fulfill their information needs. The Twitcident framework provides different strategies that allow for ranking facet-value pairs and therefore generating query recommendations.

3.4.1.1 Frequency-based Faceted Search.

A straightforward approach is to rank the facet-value pairs $(f, v) \in FVPs$ based on their occurrence frequency in the current hit list H of Twitter messages that match the current query $q = \{(f, v)|(f, v) \in FVPs \textit{ selected as filter}\}$, i.e. messages that contain all facet-value pairs in q:

$$rank_{frequency}((f, v), H) = |H_{(f,v)}| \quad (3)$$

$|H_{(f,v)}|$ is the number of (remaining) messages that contain the facet-value pair (f, v) which can be applied to further filter the given hit list H. By ranking those facet values high that appear in most of the messages, $rank_{frequency}$ minimizes the risk of ranking relevant facet values too low. However, it might increase the effort that a user has to invest to narrow down search results: by selecting facet values which occur in most of the remaining tweets the size of the hit list is reduced slowly.

3.4.1.2 Time-sensitive Faceted Search.

Topics that are reported and discussed on the Social Web about an incident may change over time [10, 13]. Hence, also the information demands of users who are seeking for relevant details about an incident are likely to shift. The time-sensitive faceted search strategy adapts to this behavior and promotes those trending facet-value pairs that are often mentioned in recent Social Web messages:

$$rank_{time}((f, v), H) = max(\{age(m)|m \in H\}) - \frac{\sum_{m \in H_{(f,v)}} age(m)}{|\{m \in H_{(f,v)}\}|} \quad (4)$$

Here, $age(m)$ is the age of a message $m \in H$ (and $m \in H_{(f,v)}$) with respect to the current time when the query is issued. $rank_{time}((f, v), H)$ thus calculates the temporal distance between the oldest message in the hit list and the average age of messages that contain the given facet-value pair (f, v). The younger the average age of messages that mention (f, v), the higher the ranking score.

3.4.1.3 Personalized Faceted Search.

Individual users may have different information needs that are reflected by their personal interests. To adapt the faceted search to the individual demands of a user, the Twitcident framework infers a user's interests from her Twitter activities, i.e. from the tweets a user published herself. The interest profile $P(u)$ of a user $u \in U$ can therefore be represented in the same way as incident or tweet profiles (cf. Definition 1), hence as a set of weighted facet-value pairs.

$$P(u) = \{((f,v), w(u,(f,v)))| \\ (f,v) \in \bigcup_{t \in T_u} P(t), u \in U, w(u,(f,v))) \in [0..1]\} \quad (5)$$

Twitcident analyzes the entire Twitter timeline of a user to construct a profile. It thus considers all the profiles $P(t)$ of tweets that the user published and weighs the facet-value pairs according to their occurrence frequency in the tweets. Given a facet-value pair (f, v), the personalized facet ranking strategy utilizes the weight $w(u, (f, v))$ in $P(u)$ to determine the ranking score:

$$rank_{pers}((f,v), P(u)) = \begin{cases} w(u,(f,v)) & \text{if } (f,v) \in P(u) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The Twitcident framework moreover allows to combine different faceted search strategies using their normalized ranking score so that $rank((f, v), H) \in [0..1]$. In our experiments in Section 5, we combine the personalized and time-sensitive ranking strategy with the frequency-based strategy and set $\lambda = 0.5$, for example: $rank_{combine}((f, v), H) = \lambda rank_{frequency}((f, v), H)+(1-\lambda)rank_{personalized}((f, v), H)$.

3.4.2 Realtime Analytics

Based on the semantic enrichment, the Twitcident framework provides functionality to analyze the current Social Web stream about an incident. Figure 2 shows some of the graphical gadgets that are delivered to the users such as the evolution of topics over time or the geographical impact area of an incident. Twitcident exploits the incident and tweet profiles to generate these diagrams. For example, the impact area of an incident is deduced from the geographical location of Twitter messages that report about experiences of users, e.g. in which people state that they see, hear or smell something. The analytical tools adapt furthermore to the current context of a user: if a user further filters the Social Web stream by means of faceted search then the diagrams summarize and visualize only that fraction of the information that matches the filter.

Having introduced the core functionalities of the Twitcident framework, we will, in the next sections, evaluate the two fundamental research challenges that we approach with the Twitcident framework: the automated filtering of relevant information from Social Web streams (see Section 4) and search within Social Web streams (see Section 5).

4. EVALUATION OF TWEET FILTERING

On Twitter, people publish around 200 million messages per day[11]. Automatically retrieving and filtering information about particular incidents from Twitter streams is thus a non-trivial problem. In this section, we evaluate and compare the different strategies that Twitcident provides in order to solve this challenge and investigate the following research questions:

1. Which filtering strategy performs best in retrieving messages that are relevant for a given incident? How do semantic filtering strategies perform in comparison to keyword-based approaches?
2. How are the filtering strategies affected by the characteristics of the (initial) incident description?

Corpus	#Elements
Crawled Twitter TREC 2011 Corpus	14,958,450
English Twitter Corpus	4,766,901
RSS News Feeds	62
News Articles	13,959
Entities extracted from English tweets	6,193,060
Entities extracted from News Articles	357,559

Table 1: Statistics of the Twitter corpus, the external news sources and the extracted named entities.

4.1 Experimental Setup

We evaluate the filtering strategies in context of the TREC microblog benchmarking task that was published this year, for the first time, at TREC[12]. The task is defined as retrieving the *interesting* and relevant Twitter messages for a given topic and a given time frame. As data, a corpus of sixteen million Twitter message IDs was released (which were posted on Twitter over a period of 2 weeks, from January 24, 2011 to February 8, 2011) together with 50 topics such as *Mexico drug war* or *Protests in Jordan*. In our experiments, we interpret these topics as incidents and consider the label of the topic (e.g. *Protests in Jordan*) as the initial incident description which the Twitcident framework exploits to perform incident profiling and tweet filtering (see Figure 1). The TREC topics are of different scale and most topics relate to a geographic location. Therefore, they have similar properties like the incidents that are monitored by Twitcident in practice.

For the top tweets returned by each filtering strategy for each topic, TREC provided relevance judgements indicating whether a tweet is considered to be relevant for a topic. On average, 58.35 tweets per topic were considered as relevant. Given these relevance judgments, we measure the performance via the mean average precision (MAP), precision within the top k returned items (P@k) and recall.

4.1.1 Dataset Characteristics

Table 1 gives an overview of the crawled dataset. Since over time, less tweets are available for public access, we were only able to crawl approximately fifteen million tweets (crawled in June/July), of which nearly five million tweets were detected to be written in English. Employing NER on the English tweets resulted in a total over six million named entities among which we find approximately 0.14 million distinct entities. The external news corpus was derived by extracting articles from 62 RSS feeds of prominent news media such as BBC, CNN or New York Times in the same time period as the Twitter posts.

4.1.2 Baseline: Keyword Filtering

We compare the semantic filtering strategies provided by the Twitcident framework with a keyword-based filtering baseline that interprets the label of a topic as a keyword query. The baseline evaluates a query and generates a ranking of tweets using language modeling with relevance model RM2 [11]. Apart from filtering out non-English tweets, the baseline also filters out re-tweets, tweets with less than 100 characters and tweets with words that contain a single letter three or more times in sequence (e.g., “oooooooooh”). It thereby aims to remove chatter from a stream of tweets.

[11] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

[12] Text REtrieval Conference: http://trec.nist.gov/

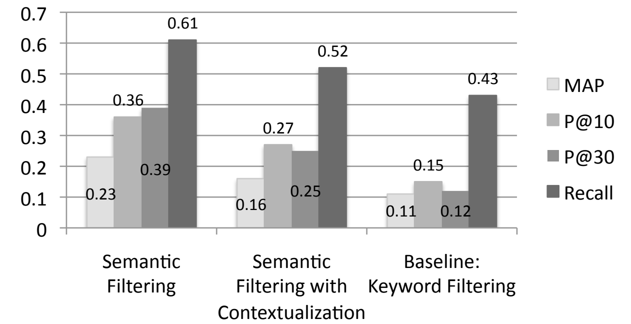


Figure 4: Result overview on the filtering strategies. Reported are the mean average precision (MAP), precision at k (P@10, P@30) and recall.

4.2 Experimental Results

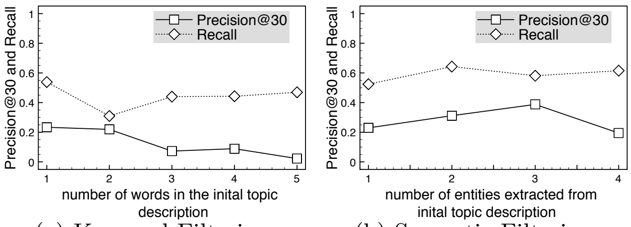
Figure 4 summarizes the results of our filtering evaluation and demonstrates that the semantic strategies of the Twitcident framework clearly outperform the keyword-based filtering regarding all metrics[13]. For example, the semantic filtering performs—with respect to MAP, P@10 and P@30—more than twice as good as the baseline and regarding recall it improves the filtering performance by 41.8%. News-based contextualization also leads to major improvements in comparison to the keyword-based baseline. However, it performs worse than the semantic filtering which does incident profiling solely on tweets. This indicates that facet-value pairs that are extracted from news articles, which report about the same incident/topic, seem to include noise in the incident profiling and filtering process.

Figure 5 illustrates the impact of the initial topic description on the filtering. The x-axis specifies the number of (a) words and (b) facet-value pairs that are extracted from the initial description while the y-axis marks precision@30 and recall. For keyword filtering, we observe that the precision almost gradually drops the more keywords are listed in the initial topic description so that for topics that feature six keywords, the average precision is just 0.03. In contrast, the semantic filtering, which does not consider all keywords from the topic description but considers only named entities for the topic profiling, is more robust and also achieves in the worst case a considerably higher average precision of 0.2. For both strategies, the recall increases slightly the more concepts are extracted from the initial topic description. Again, the semantic filtering performs better than the keyword-based filtering and features a more stable behavior when characteristics of the topic description vary.

4.3 Synopsis

In conclusion, we can therefore answer the research questions raised at the beginning of this section as follows.

1. Semantic filtering allows for the best filtering performance. It clearly outperforms the keyword-based strategy and more than doubles the mean average precision.
2. The complexity of a topic, measured by the number of concepts that can be extracted from the initial topic description, impacts the precision of the keyword-based strategy negatively: the higher the complexity the lower the precision. The semantic filtering strategy is more robust and also achieves high precisions for complex topics.



(a) Keyword Filtering (b) Semantic Filtering

Figure 5: Robustness of (a) keyword-based filtering and (b) semantic filtering: correlation between the number of (a) words and (b) semantic concepts that can be extracted from the initial topic description and the filtering performance (P@30 and Recall).

5. EVALUATION OF FACETED SEARCH

Based on the automatic filtering of Social Web streams for detecting messages that are relevant for a given incident, the Twitcident framework provides faceted search functionality that allows users to further filter the messages and retrieve information they are interested in. In line with the evaluations done in [1], we now evaluate also the quality of the faceted search strategies on top of the automatic filtering process and study the following research questions:

1. How well does faceted search supported by the Twitcident framework perform in comparison to keyword search?
2. What faceted search strategy supports users best in finding relevant Twitter messages?
3. What factors influence the performance of the faceted search strategies?

5.1 Experimental Setup

In order to answer the above research questions and evaluate the faceted search strategies (see Section 3.4.1), we applied an evaluation methodology introduced by Koren et al. [9] that simulates the clicking behavior of users in the context of faceted search interfaces. In a faceted search interface, a user can select a facet-value pair to refine the query and drill down the search result list until she finds a relevant document. We model the user's facet-value pair selection behavior by means of a *first-match user* that selects the first matching facet-value pair and continues to refine the query until no more appropriate facet-value pairs can be selected.

To evaluate the performance, we used again the TREC microblog dataset described in Section 4.1.1 and generated search settings by randomly selecting, for each of the 50 topics, 50 re-tweets which mention at least one hashtag—thus resulting in 2500 settings. Each search setting consists of (i) a target tweet (= the tweet that was re-tweeted), (ii) a user that is searching for the tweet (= the user who re-tweeted the tweet) and (iii) the timestamp of the search activity (= the time when the user re-tweeted the message). The set of candidate items is given by all those tweets which have been published before the search activity and are considered to be relevant to the corresponding topic based on the semantic filtering strategy of the Twitcident framework. We thus test—except for the incident detection—the entire pipeline of the Twitcident framework as depicted in Figure 1. The filtering delivered, on average, more than 5000 candidates per search setting while there is only exactly one Twitter message that is considered to be relevant, namely the Twitter message that was actually re-tweeted by the user.

[13] The keyword-based baseline deployed by the TREC microblog organizers features with 0.14 and 0.11 regarding MAP and P@30 respectively a similar performance as our keyword-based strategy.

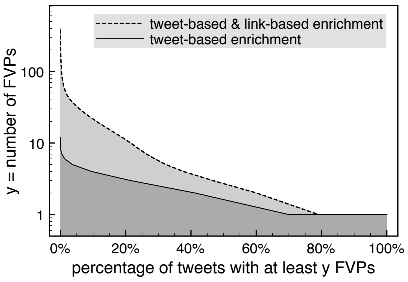


Figure 6: Impact of link-based semantic enrichment: the histogram shows the fraction of Twitter messages that feature at least y facet-value pairs (FVPs) for (i) semantic enrichment solely on tweets (tweet-based) and (ii) the link-based strategy that follows links which are posted in Twitter messages.

For measuring the performance of the search strategies, we use the mean reciprocal rank (MRR) of the target item in the search result ranking[14] when the user selects it. Furthermore, we utilize MRR of the first relevant facet-value pair and success at rank k (S@k) which is the probability that a relevant facet-value pair, that the user selects to narrow down the search result list, appears within the top k of the facet-value pair ranking. Both metrics are direct indicators for the effort a user needs to spend using the search interface: the higher MRR and S@k, the faster the user will find a relevant facet-value pair when scanning the facet-value pair ranking.

5.1.1 Dataset Characteristics

In the faceted search evaluation, we moreover experiment with the link-based semantic enrichment that is provided by the Twitcident framework (see Section 3.3.3). As depicted in Figure 6, we observe that the extraction of facet-value pairs from Web resources that are linked from a Twitter message allows to further extend the profile of the corresponding tweet. It therefore reduces the level of sparsity. For example, for the semantic enrichment, which is solely based on tweets, 41.2% of the messages feature at least two facet-value pairs while the additional link-based enrichment allows for representing 60.1% of the tweets with at least two facet-value pairs.

5.1.2 Baseline Strategies

We compare the faceted search strategies of the Twitcident framework (see Section 3.4.1) with two baseline strategies that exploit hashtags:

Hashtag-based Keyword Search. For this baseline strategy, the user randomly selects one of the hashtags that is mentioned in the Twitter message the user is searching for[15]. Given the messages that match this keyword query, the user starts scanning the result list.

Hashtag-based Faceted Search This strategy interprets hashtags as facet values and therefore ranks the hashtag-based facet-value pairs in the same way as the frequency-based faceted search strategy (see Section 3.4.1), i.e. according to their occurrence frequency in the current search result list. The selection of hashtag-based facet-value pairs is simulated according to the aforementioned procedure.

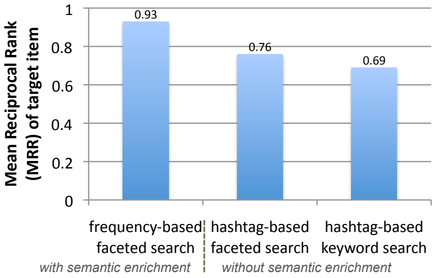


Figure 7: Result overview of search strategies: comparison of hashtag-based and semantic search.

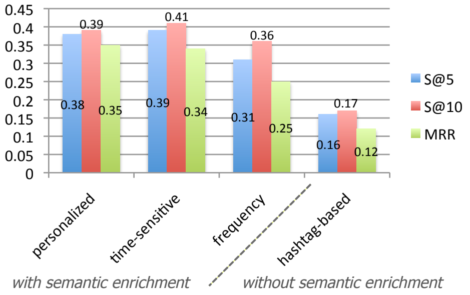


Figure 8: Result overview of the faceted search strategies. Reported are the mean reciprocal rank (MRR) of the first relevant facet-value pair (FVP) and success at k (S@5, S@10), i.e. the probability that a relevant FVP appears within the top k.

5.2 Experimental Results

Figure 7 compares the frequency-based faceted search strategy featured by the Twitcident framework with the hashtag-based search strategies. The comparison of the MRR scores reveals that the semantic faceted search strategy improves the search performance significantly by 34.8% and 22.4% over the hashtag-based keyword search and the hashtag-based faceted search strategy[16]. Interpreting hashtags as facet values leads to an improvement over the single keyword query as well. However, the semantic enrichment provided by the Twitcident framework proves to generate more valuable representations of the Twitter messages and therefore allows for faceted search functionality that clearly outperforms the two hashtag-based strategies.

The performance of the different faceted search strategies is listed in Figure 8. The performance of those strategies that benefit from the semantic enrichment significantly exceeds the performance of the hashtag-based strategy in predicting appropriate facet-value pairs. A detailed review of the results shows that a key success factor of the semantic faceted search strategies is given by their ability of disambiguating facet-value pairs. While the hashtag-based strategy would, for example, treat *#Tahrir* and *#TahrirSquare* as different facet values, the semantic faceted search strategies would—in context of the *"Egyptian evacuation"* incident which is one of the TREC topics—map both values to the same concept (namely *dbpedia:Tahrir_Square*) and therefore facilitate the faceted search for the user.

Figure 8 furthermore shows that both personalization and temporal contextualization lead to significant improvements over the frequency-based strategy. In fact, regarding MRR the performance of the personalized and time-sensitive strate-

[14] Tweets are ranked according to their creation time so that the latest tweets appear at the top of the ranking.

[15] To not discriminate the hashtag-based search strategies, we selected the search settings so that each target tweet contains at least one hashtag.

[16] Statistical significance was tested with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$.

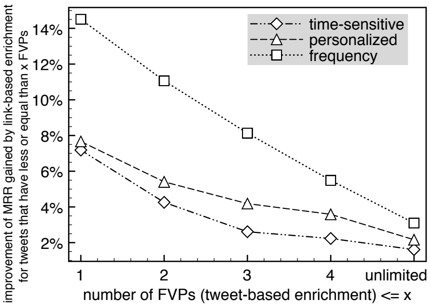
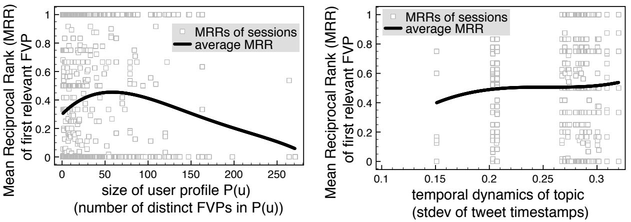


Figure 9: Impact of link-based semantic enrichment on faceted search performance. The y-axis shows the improvement with respect to the mean reciprocal rank (MRR) of the first relevant FVP that is gained when using link-based enrichment in addition to solely tweet-based enrichment averaged for those search settings where the target tweet features x or less than x FVPs.



(a) Personalized: impact of profile size

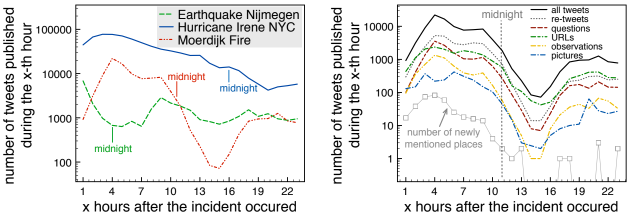
(b) Time-sensitive: impact of topic dynamics

Figure 10: Impact of (a) profile size on the search performance of the personalized faceted search strategy and correlation between (b) search performance and temporal dynamics of the topic within which a user is searching. Temporal dynamics is measured by means of the standard deviation of the timestamps of Twitter messages that are published within one topic, i.e. a high standard deviation indicates strong temporal dynamics.

gies is 39.7% and 36.8% better than the one of the faceted search strategy that ranks the facet-value pairs according to their occurrence frequency in the current search result set.

By enriching the tweet profiles with facet-value pairs extracted from external Web resources that are referenced from the Twitter messages (link-based semantic enrichment), one can further improve the performance of the semantic faceted search strategies (see Figure 9). The level of improvement depends on the characteristics of the tweet profiles. Those search settings where the target tweet contains exactly one facet-value pair benefit most from the link-based enrichment. For these settings, the performance increases by 14.5% for the frequency-based strategy and around 7% for the personalized and time-sensitive strategies.

Figure 10 allows us to study how the performance of the personalized and time-sensitive search strategies depends on the characteristics of the user and incident profiles. Therefore, Figure 10(a) plots the MRR scores of the personalized strategy in relation to the size of the profile of the user who performed the corresponding search activity. It is interesting to observe how the average performance varies with changing profile sizes: the average MRR for profiles with less then 10 distinct FVPs is 0.328. The personalized strategy achieves its maximum average MRR performance for profiles that feature between 50 and 70 FVPs while for the few user profiles which feature more than 150 FVPs the performance drops—possibly because those profiles feature too much diversity.



(a) Different incident types

(b) Moerdijk fire

Figure 11: Posting behavior about incidents within the first 24 hours of an incident: (a) comparison of different types of incidents and (b) type of information posted during a fire incident in Moerdijk.

The time-sensitive faceted search strategy, which promotes those facet-value pairs that are currently trending, performs best for those search settings that are performed within a topic that is characterized by strong temporal dynamics (see Figure 10(b)). Here, the dynamics of a topic are described by means of the standard deviation of the creation times of tweets which are considered to be relevant for the topic. Figure 10(b) depicts that the performance slightly increases the more a topic underlies temporal changes. Hence, the more distributed the messages are posted over time the more important it is to adapt to the temporal context.

5.3 Synopsis

Given the experimental results, we can answer the research questions raised at the beginning of this section:

1. Faceted search strategies allow for significantly higher search performance than the hashtag-based keyword search strategies. They enable users to more precisely filter tweets and therefore retrieve relevant information.
2. Personalized and time-sensitive faceted search strategies that adapt to the profile of a user and to the temporal context respectively allow for the best search performance and lead to significant improvements over the standard semantic faceted search strategy. Further exploitation of links posted in tweets allows us to further enrich the semantic representation of tweets and moreover induces additional improvements of the search performance.
3. The performance of the personalized faceted search strategy is influenced by the size of a user's profile and achieves the highest performance for medium-sized profiles. The quality of the time-sensitive strategy depends on the temporal dynamics within an incident: the more temporal changes the more important it is to adapt to the temporal context.

6. DISCUSSION

With Twitcident we introduce a system that allows users to explore, search and analyze information about incidents available on the Social Web and Twitter in particular. During the last ten months we tested the Twitcident system in practice to monitor various incidents, specifically to support emergency services such as the Dutch police and fire department. Given these experiences, we identify that different types of incidents imply different types of posting behavior on the Social Web. For example, Figure 11(a) compares the number of Twitter messages posted about three different types of incidents within the first 24 hours: a large-scale fire at a chemical factory in Moerdijk (Jan 5th 2011), an earthquake with its epicenter close to Nijmegen (Sep 8th 2011) and the so-called hurricane *Irene* which caused floodings in New York (Aug 28th 2011). One can see that all incidents reach their maximum peak within the first 4 hours after the incident occurred. For the fire and hurricane the amount of tweets gradually grows until it reaches its peak while for the unexpected earthquake most tweets are already published within the first hour after the incident. In fact, the hurricane Irene did not hit New York City unexpectedly, but was forecasted already weeks ahead which caused Twitter traffic already before the hurricane appeared.

Twitcident thus has to process huge amounts of messages within the first hours of an incident. To handle ten thousands of messages per hour, Twitcident parallelizes the semantic enrichment of Twitter messages which is the most time-intensive procedure. In particular, following URLs and processing the corresponding Web sites may take seconds. Therefore, Twitcident applies heuristics to decide whether the link of a tweet should be processed in realtime or marked for later processing (e.g. during the night when the amount of messages to be processed decreases; see Figure 11(a)). For example, URLs posted in tweets for which the tweet-based enrichment—which takes on average between 100 and 300 milliseconds—detects already two or more facet-value pairs are not processed immediately because for these tweets the link-based enrichment improves the search performance only slightly (see Figure 9).

Figure 11(b) illustrates for the fire at the chemical factory in Moerdijk the kind of information that is posted on Twitter within the first 24 hours after the fire started. It is interesting to see that the number of questions that are being asked is exceptionally high when the overall number of tweets reaches its maximum. At that point, questions such as *"What about the toxic cloud?"* or *"Is there a chance that the smoke is going to Leiden?"* are prominent and exceed the amount of URLs and pictures which may reveal answers to these questions. Emergency services are often interested in *new* information and question, for example, whether the impact area of an incident is exceeding (cf. "number of newly mentioned places" in Figure 11(b)).

Twitcident allows people to find answers to such questions and allows emergency services to analyze the information that people publish on the Social Web.

7. CONCLUSIONS

In this paper we introduced Twitcident, a framework for filtering, searching and analyzing information about incidents that people publish in their Social Web streams. Triggered by an incident detection module that monitors emergency broadcasting services, our framework automatically collects and filters relevant information from Twitter. It enriches the semantics of Twitter messages to adapt and improve the incident profiling and filtering over time. Semantic enrichment is also the foundation for faceted search and realtime analytics provided by the Twitcident framework. In our evaluations we proved that semantic enrichment boosts the performance of both the filtering of Twitter messages for a given incident and the search for relevant information about an incident within the filtered messages significantly.

Acknowledgements

This work is co-funded by the EU FP7 project ImREAL (`http://imreal-project.eu`).

8. REFERENCES

[1] F. Abel, I. Celic, G.-J. Houben, and P. Siehndel. Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In *ISWC*, pages 1–17, 2011. Springer.

[2] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *UIST*, pages 303–312, 2010. ACM.

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics*, 2009.

[4] J. Chen, R. Nairn, and E. H. Chi. Speak Little and Well: Recommending Conversations in Online Social Streams. In *CHI*, 2011. ACM.

[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proc. of CHI*, pages 1185–1194, 2010. ACM.

[6] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *Proc. of WWW*, pages 331–340, 2010. ACM.

[7] D. Gaffney. iranelection: quantifying online activism. In *WebScience*, 2010. ACM.

[8] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proc. of WebKDD/SNA-KDD*, pages 56–65, 2007. ACM.

[9] J. Koren, Y. Zhang, and X. Liu. Personalized interactive faceted search. In *WWW*, pages 477–486, 2008. ACM.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of WWW*, pages 591–600, 2010. ACM.

[11] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of SIGIR*, pages 120–127, 2001.

[12] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in Twitter. In *Proc. of WWW*, pages 1137–1138, 2010. ACM.

[13] K. Lerman and R. Ghosh. Information contagion: an empirical study of spread of news on Digg and Twitter social networks. In *Proc. of ICWSM*, 2010. AAAI Press.

[14] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proc. of CHI*, 2011. ACM.

[15] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proc. of SIGMOD*, pages 1155–1158, 2010. ACM.

[16] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We Know Who You Followed Last Summer: Inferring Social Link Creation Times In Twitter. In *Proc. of WWW*, 2011. ACM.

[17] M. Pennacchiotti and A.-M. Popescu. A Machine Learning Approach to Twitter User Classification. In *Proc. of ICWSM*, 2011. AAAI Press.

[18] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc. of WWW*, 2011. ACM.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of WWW*, pages 851–860, 2010. ACM.

[20] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proc. of GIS*, pages 42–51, 2009. ACM.

[21] R. Stronkman. Exploiting Twitter to fulfill information needs during incidents. Master thesis, TU Delft, 2011. `http://wis.ewi.tudelft.nl/twitcident/thesis.pdf`.

[22] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: a comparison of microblog search and web search. In *Proc. of WSDM*, pages 35–44, 2011. ACM.

[23] J. Weng and B.-S. Lee. Event Detection in Twitter. In *Proc. of ICWSM*, 2011. AAAI Press.