Time-aware online reputation analysis

Peetz, M.-H.

Publication date
2015
Document Version
Final published version

2 Background

In this chapter we introduce the underlying concepts and background needed in later chapters of the thesis. The interdisciplinary of this thesis surfaces in the different sections. As we are working on social media, we are introducing social media and its data analysis in Section 2.1. The overall task of this thesis is the analysis of reputation; we define reputation and measures of reputation in Section 2.2. Section 2.3 and 2.4 survey the background material on information retrieval and entity filtering, respectively. Some related work is reviewed locally in the respective chapters. In Chapter 4 we review methodologies to study user behaviour in Section 4.1. Chapter 7 introduces and discusses cognitive models for memory retention in Section 7.1, while Chapter 8 introduces related work on active learning in Section 8.1.

2.1 Social Media

According to the Oxford dictionary,[1] *social media* is defined as

> Websites and applications that enable users to create and share content or to participate in social networking.

The main points in this definition are *websites and application*, denoting the virtuality of social media, *create and share content*, denoting the spreading of user-generated content, and *social networking*, denoting the social aspect and the connectivity of the users of social media. While in common language, social media mostly refers to social networking sites such as Twitter, G+, and Facebook, more examples fall under this definition. Social media applications that are mainly focussed on the social and collaborative aspect are online games and virtual worlds like World of Warcraft or SecondLife. Forums and newsgroups are the oldest form of social media, sometimes the content is more important (like Stackoverflow) while for some forums the social aspects are prominent (like support groups). LinkedIn and Xing as professional networking sites are mainly based on densifying the social graph, while G+ and Facebook combine the social graph with sharing content. Twitter has a special role, as the social graph is bi-directional: you can follow someone without being befriended. This results in Twitter being a content generating

[1] `http://www.oxforddictionaries.com/definition/english/social-media`

site with mass communication of information—it is not only a conversational social network but can also be considered an information highway for news [136]. Additionally, their terms of service offer easy data access for researchers and commercial parties alike. While tweets are short (140 characters), blogs are longer. They are social in so far as bloggers are connected via blogrolls or other automatic feeds. Wikipedia has the greatest focus on content generation, while this is collaborative, the networking between authors is not explicit. A successful online and social media presence depends on a skilled combination of different social media applications [85]. Additionally, the more social the applications become, the more the temporal aspects come into play. Old information is simply not as interesting as new.

Based on the two aspects, the social networking and the creation of content, different research areas emerge. While social networking analysis is an important area of research, this thesis is about the *content* generated in a network and its *changes over time*.

As we will see later in Section 2.2, one of the key assumptions of reputation monitoring is the similarity between the real-life and the virtual life. Early studies of Facebook emphasize that it is hard to compare findings from online social networks to offline social networks because not everyone lives out their social lives on Facebook to the same extent [143]. Nevertheless, they find a high similarity of tastes ("likes") between Facebook friends that occur in the same pictures and if they are in the same housing group. For Twitter, Huberman et al. [111] find that the friends network compared to the following network is more sparse. Unlike for the number of followers, the more friends users have, the more they post and share.

We now look at network similarities between online social networks and offline social networks. The most famous experiment on the connectivity of social networks is the *degrees of separation* experiment by Stanley Milgram [164, 244] where individuals with a large social and regional distance had to send postcards to people they knew to be connected to each other. Milgram found that the degree of separation is between five and a half and six. Backstrom et al. [15] and Huberman et al. [111] find that while the degrees of separation in an early Facebook is six, in all of Facebook in 2012 it is 3.74 with the degree of separation converging over time. For Twitter, the degree of separation was found to be 3.43 [17].

Looking at the content of conversations, Java et al. [119] find that users talk about their everyday life and activities just as well as they seek or share information. Building on that, Naaman et al. [170] identifies two groups of people, *meformers* who mainly talk about themselves and keep the networking ties (and friendships) together, as well as *informers* who share interesting information and are prone to be followed. One interesting subspace of a social network is the workplace and Twitter can and has been used to complement work-related conversations [284]. Twitter users also talk about brands [117], and their utterances are also a kind of electronic word of mouth: 19% of the users direct a post to a brand [117], but as a part of a discussion and spreading news [227]. Additional research identifies peaky topics and lingering conversations [223]: some topics peak at a certain time, while other topics are conversational and persistent. Peaky topics are topics that often propagate through the network.

The process of how information propagation works is not entirely understood. Bernstein et al. [29] find that on average, Facebook users reach 35% of their friends with each post and 61% of their friends over the course of a month. Aral and Walker [12] and

Romero et al. [208] find, however, that even though one might have a lot of followers or friends, it does not necessarily mean that one influences them. Users input can also be more than one topic [208]. When it comes to who influences whom, Aral [11] finds that men are more influential than women, younger people are more susceptible, while married people are least susceptible. In general, understanding information propagation is still an open topic [33, 209].

There is a subset of content on social networks that is *viral*, i.e., it is spreading through the network like a virus. Online social networks allow for easy information propagation and viral spreading of information. Early, offline, viral letters were often pyramid schemes like the send-a-dime letter[2] in the 1930's, where dissemination of the letter was enforced by "bad luck" lingering over the chain breaker. A similar propagation happened with emails. Hoaxes, such as *Bonsai Kitten*[3] or promises for money, such as the *Bill Gates will give you $245*[4] mail, could easily be sent to the entire address book, instead of photocopying letters. With the rise of social networks, re-posting (or re-tweeting) on a timeline allowed internet memes such as the LOL cats to spread even easier. Not only funny memes were spreading, also serious games such as *Take This Lollipop*:[5] an interactive horror movie teaching Facebook users to keep their data private. The earlier examples were mainly restricted to online media. A new area of information propagation happened with the Arab spring, where Twitter helped spreading information to western societies as well as to demonstrators involved: providing uncensored information as to where e.g., demonstrations were happening [152]. Other examples on real-life was the *Project X* in Haren [251], were a private party turned into a public party in the Netherlands, or the hashtag *#Aufschrei* that opened a public discussion on the normality of sexual harassment in Germany. Viral information spreading is like a dream come true for marketeers: no costs for advertisement space on television, radio, or print media but still virtually complete coverage of entire population groups. But it can also be a nightmare for marketeers: unscheduled events that harm the reputation can spread just as quickly as positive events. Viral marketing campaigns include the funny campaign for BlendTec[6] where expensive or seemingly unblendable items, such a smartphones, were blent in a BlendTec blender. The *Dove Real Beauty* campaign allowed customers to see themselves as models, therefore changing the beauty image imposed by traditional cosmetics companies. The counter-campaign by Greenpeace pointing out the use of palm oil turned into a nightmare for the marketeers [120] with nearly 2 million viewers as of November 2014. The *KLM surprise* campaign combines viral with webcare, users posting about KLM were awarded with little surprises on their flight. However tempting those marketing campaigns are, they are also dangerous. Hyundai tried to market a car with an attempted suicide video. The video went viral, but with very negative sentiment. People were devastated and shared stories and notes of their own family members' suicide [42]. Also, not only news approved and released by companies themselves can get viral. Missing important tweets and news items about an entity of interest can potentially

[2] http://www.mortaljourney.com/2010/11/1930-trends/the-prosperity-or-send-a-dime-chain-letter-fad

[3] http://bonsaikitten.com/bkintro.php

[4] http://archive.wired.com/wired/archive/12.07/hoax.html

[5] http://en.wikipedia.org/wiki/Take_This_Lollipop

[6] http://www.willitblend.com/

be disastrous and expensive: when users on Twitter found out about H&M deliberately destroying perfectly wearable winter jackets this incident hyped and caused bad publicity [195]. Viral spreading of information can therefore have a positive or negative impact on the reputation of a company.

2.2 Reputation

Corporate reputation as a vital part of brand definition, was important from the onset of advertisement. David Ogilvy (1955, in [254]) considered a *brand* as:

> The intangible sum of a product's attributes: its name, packaging, and price, its history, its reputation, and the way it's advertised.

In this definition, reputation is the only intangible feature. Further research has been aimed at defining reputation, van Riel and Fombrun [254, p. 43] define reputation as

> Reputations are overall assessments of organizations by their stakeholders. They are aggregate perceptions by stakeholders of an organization's ability to fulfill their expectations [...]

Let us explain some of the key points in this quote. The authors use the term *organization* in the definition, but it may as well apply to sub-brands (*Coke Zero* being sub-brand of *Coca-Cola*). *Stakeholder* is everyone who has something to do with the company, be it employees, customers, deliveries, or law-makers. Finally, *aggregate* means that reputation is no single point: it is an *aggregation* over stakeholders, but also over their *expectations, attitudes, and feelings* [243]. This and other definitions [14, 67, 243] focus on the transitivity of reputation: without stakeholders and their perception, companies do not have a reputation.

The reputation of a company is important for both the stakeholders and the company itself. For the stakeholders, the image of a company may help them to cast decisions about the company and its products faster than without previous experiences [197],[7] in other words providing a mental shortcut for decision making [200]. For the company, reputation can be an asset: it attracts stakeholders [85] and can therefore be a buffer from economic loss [123].

2.2.1 Measuring Reputation

The widely-published first ranking of companies was Fortune's *America's Most Admired Companies* (AMAC) survey in 1982. They publish an aggregation of reputation over different dimensions based on the opinion of industry professionals. They heavily rely on early survey methodologies [45]. Those methods for measuring reputation on a dimension include the Kelly repertory grid [131], natural grouping [259], Q-sort [235], card sorting, attitude scales [82], and questionnaire based surveys. For an elaborate discussion of the individual methodologies we refer to [84] and [254]. The dimensions over which reputation was measured in the AMAC survey include: the quality of management, products or services; the financial soundness; and innovativeness. The problem

[7] Citation via [254].

with respect to different dimensions is that they have to be statistically independent to be aggregated in a sound way.

Several different professional measures spawned from the approaches of the AMAC ranking. The *Brand asset valuator* by the Young & Rubicam agency is consumer-based, reporting on authority and strength, while the *Leveraging corporate equity* approach [89] used a broader range of professionals. The *Reputation Quotient* [86] took more stakeholder types into account to find the attitude towards less, but independent dimensions of reputation. Newell and Goldsmith [172] introduce the first standardized and reliable measure of a company's *credibility* from a consumer perspective, solely based on a questionnaire as survey methodology. Davies [64] sees a company as a personality. This measure uses personality traits as dimensions, and assigns a company a *corporate personality*. Following this approach, the reputation of a company is dependent on personality matching of the stakeholders personality and the companies personality. This assumes that some people with a certain personality find other personalities, of brands or people, more appealing, e.g., risk averse people prefer "safe" companies while other people might consider this type of company boring. Stacks [234] and Fombrun and Van Riel [85] find a connection between indicators (e.g., reputation, trust, and credibility) and financial indicators (e.g., sales, profits). They find that reputation, being intangible, still has tangible assets. A very successful measurement framework of the last years is *RepTrak* [83], which is based on the Reputation Quotient. This framework is also used by the analysts in the following chapters, so we elaborate below what it entails. The framework has seven dimensions, which in total have 23 attributes. Table 2.1 describes the dimensions. Similar to the RQ they consider the attitudes of different stakeholders: Consumers, Executives, Media, Investors, Employees, Government, Others. While the original data used for RepTrak is again based on the above mentioned survey methodology, the state of the art integrates media analysis. Media analysis is different from the previous methodology, because analysts do not directly ask stakeholders in questionnaire, but analyse media. Media can be newspapers and social media, but just as well TV and radio broadcasts. The analysis usually involves consuming the media and categorize it according to stakeholder and reputation polarity. Recently, with the rise of user-generated content online, social media analysis is gaining importance as a proxy to *people's* opinion, giving birth to the field of *online reputation analysis* [135].

2.2.2 Reputation Polarity

In online reputation analysis, the aggregated reputation of a company (or brand or entity) in general is based on the influence of single tweets on this reputation. This influence is called the *reputation polarity* of a tweet. More specifically, polarity for reputation implies that a tweet has negative or positive implications for the reputation of the firm. For analysis purposes, this reputation polarity is then split into dimensions and stakeholders (see the previous section). Social media analysts rely on commercial sentiment analysis tools that tend not to distinguish between reputation and sentiment analysis. Most tools provide an option to train the sentiment classifier using manual input and manually annotated texts. For example, the social media monitoring tool from Crimson Hexagon aims to give only aggregated numbers of the proportion of documents that are negative or positive for the company's reputation, instead of classifying individual documents or

Table 2.1: Dimensions along which the reputation of a brand is being analysed according to RepTrak along with (somewhat simplified) descriptions. Summarised from [254, Figure 10.13].

Dimension	Description
Performance	The financial performance, now and in future
Products/Services	Quality of products and customer service
Innovation	Product innovation and quick adaptation
Workplace	Employee satisfaction
Governance	Transparency, ethical awareness and business values
Citizenship	Enviromental and societal responsible
Leadership	Management is well organised, structured, and has a clear vision for the future

tweets [108]. For many types of analysis done by marketing analysts and social scientists, there is no need for accurate classifications on the individual document level, as long as the category proportions are accurate. Crimson Hexagon achieves an average root mean square error of less than 3 percentage points when at least 100 hand coded documents are available.

Besides analyzing what is being said online, another aspect of online reputation management is webcare, i.e., responding to consumer comments online to handle complaints, answer questions, and proactively post information. Consumers evaluate a brand more positively when it responds to complaints online [252].

The growing need for social media analytics leads to the development of technology that is meant to help analysts deal with large volumes of data. The first problem to tackle here is identifying tweets that are relevant for a given entity. Looking at previous work on sentiment analysis, polarity detection of reputation seems to have evolved naturally from sentiment analysis. Much work has been done in sentiment analysis; extensive treatments of the topic can be found in [180] and [147]. Subtasks of sentiment analysis relevant to this chapter are *sentiment extraction* and *opinion retrieval*; following Pang and Lee [180], we use the terms sentiment and opinion interchangeably.

Sentiment extraction is the identification of attitudes and their polarities in text of arbitrary length. The most relevant work in sentiment extraction analyses how polarity changes with context [205, 273]. *Opinion retrieval* is the identification of people's attitudes and their polarities towards a topic. On data from the Blog Track at TREC [176] (see Chapter 3), Jijkoun et al. [122] see the need for learning topic specific sentiment lexicons.

While features may range from purely term-based features (1-grams) to part of speech tags and syntactic information, a sentiment classifier needs to be able to handle negation [180]. Additionally, Pang et al. [181] find that some discourse structure analysis is important to understand the sentiment. Thelwall et al. [240] provide a sentiment classifier for social media that combines negation, emotion, and emoticons. With employee satisfaction being one of the dimensions for reputation, the work by Moniz and de Jong [167] uses computational methods to extract the sentiment polarity of employees and looks at

their influence on firm earnings.

The shortcomings of pure sentiment polarity as a substitute for reputation polarity are apparent and manual annotation approaches are currently prevailing [57]. Recently, there are efforts towards fully automating the annotation [6, 7], or to semi-automate the annotation using active learning [280] or other annotator input [65]. RepLab at CLEF [6, 7] provides annotated datasets (described in Chapter 3) for reputation polarity for different entities (companies, brands, universities, etc). They also feature an annual meeting where researchers can exchange problems and ideas towards automating the estimation of reputation polarity. Several approaches to reputation polarity detection were followed at RepLab 2012 [6]. In the following we sketch the general directions taken; for a more in depth treatment we refer to [6] or the papers themselves. Many papers follow the intuition that reputation polarity can be approximated with sentiment. Balahur and Tanev [18] train a sentiment classifier with additional training data, while other groups add more features. Carrillo-de Albornoz et al. [44] focus on emotion detection and Yang et al. [278] add a *happiness* feature. Kaptein [127] uses SentiStrength together with some user features. Similarly, Peetz et al. [185] add textual and user features. For training, not all groups rely on the original training dataset, but bootstrap more data from the background data: Chenlo et al. [50] learn hashtags denoting positive or negative sentiment, while Peetz et al. [185] assume that reputation is captured by the sentiment in reactions to a tweet. Other approaches treat the problems as a text classification problem [93] or select correlating words using feature selection [121]. With Karlgren et al. [128] and Villena-Román et al. [260], two very knowledge-intensive commercial systems led the ranks of best performing systems. Karlgren et al. [128] positioned each tweet in a semantic space using random indexing. Villena-Román et al. [260] based their results on a strong sentiment analysis tool, using linguistic features to control the scope of semantic units and negation.

In the following year, at RepLab 2013 [7], sentiment and additional textual features remain successful. One of the new systems used KL-divergence to build up a discriminative terminology for the classification [48]. With the best performing system for the polarity task, Hangya and Farkas [99] used engineered textual features such as the number of negation words, character repetitions, and n-grams. Similarly, Filgueiras and Amir [81] used sentiment terms and quality indicators similar to [266], while Saias [212] and Mosquera et al. [168] mainly based their approaches on sentiment classifiers and lexicons. Cossu et al. [58] used a combination of TF-IDF with support vector machines. Based on their earlier approach at RepLab 2012 [44], Spina et al. [230] used emotion words and domain-specific semantic graphs. The tool used for the annotations was provided by de Albornoz et al. [65]. The good results of sentiment analysis tools at RepLab 2012 and RepLab 2013 show that sentiment analysis is a sensible starting point to capture reputation polarity. In Chapter 5, therefore, we build on work from sentiment analysis. We classify reputation polarity by incorporating strong, word list-based, sentiment classifiers for social media [240] with social media features such as authority [1], and recursive use of discourse structure in Twitter.

The performance of systems estimating reputation polarity is highly dependent on the error aggregation of classifiers, as well as filtering and retrieval approaches in the pipeline [230]. Data used for monitoring and estimating the reputation for a company or brand needs to be found and retrieved. Simple keyword matching is not enough [230].

Entities can be ambiguous (like the band A[8]) or omni-present (like Kleenex[9]). A typical pipeline for finding documents that are used for monitoring first retrieves a ranked list of documents using traditional retrieval algorithms adjusted to the media type and temporal changes, and then filters this list using entity filtering approaches [6, 7, 57]. The first step is more recall-oriented, while the second step is more precision-oriented. In the following we describe traditional information retrieval algorithms (Section 2.3) and proceed with its adjustments for temporal information needs (Section 2.3.1). Section 2.4 elaborates on state-of-the-art approaches to entity filtering.

Chapter 4 focuses on properly understanding the indicators used to annotate the reputation polarity of a tweet. These insights, this can be incorporated into (semi)-automatic algorithms. We later show algorithmically in Chapter 5, that the features are company (or entity) dependent and that training classifiers per company performs better than using more data but not training per company. In Chapter 7 we show some approaches to filter the right tweets that can be used for reputation aggregation.

2.3 Information Retrieval

We have just seen how the retrieval of the correct information is important for the estimation of reputation. But what is information retrieval?

Manning et al. [155] define information retrieval as:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

In essence, while the term *information retrieval* stems from the 1950's [90], the idea to find information in large, non-digital, collections was already present in the 3rd century BC [76], when the Greek poet Callimachus created a library catalogue. Later approaches include microfilm and punchcards. Together with the increase of scientific literature and the advent of computers, the field of IR emerged: particularly in library science. The field clearly advanced, introducing features like indexing and ranked retrieval, and new algorithms that were tested on data provided by the Text REtrieval Conference (TREC) [100]. The field truly changed in the 1990's, with the emergence of the World Wide Web and the first web search engines. Two major changes happened: for one, the search that was previously left to experts was now open to everyone with web access, and secondly the nature of data changed. There was more data (and steadily increasing [55]) and the data was interlinked [217]. PageRank [178] and HITS [133] were the first algorithms making use of this information, the earlier being the backbone of the early Google system. For a more detailed treatment of the history of information retrieval, we refer to Sanderson and Croft [217], who give an excellent overview.

Retrieval models

With early, boolean retrieval systems, documents were represented as a list of terms—only if exactly the query term was present, the document would be retrieved [155]. Salton

[8] http://en.wikipedia.org/wiki/A_(band)
[9] http://en.wikipedia.org/wiki/Kleenex

Table 2.2: Notation used in this thesis.

Notation	Explanation
q	query
D, D_j	document
$w \in D$	term in document D
$w \in q$	term in query q
C	background collection
$\text{time}(D)$	normalized publishing time of document D
$\text{time}(q)$	normalized publishing time of query q
λ	interpolation parameter
β	decay parameter
μ	average document length in collection C

et al. [215] introduced the *vector space model* to represent a document D and query q as a vector, where each term represents a dimension. Documents are then *ranked* according to their spatial distance to the query. The actual value of each dimension or word would be the term frequency (*tf*) of the term in the document. Jones [124] introduced the inverse document frequency (*idf*) for a term, the inverse number of documents a term occurs in: a term frequently occurring in all documents (such as *a*) has a lower *idf* than terms occurring in only one document. The combination *tf* $\cdot$ *idf* is a commonly used statistic for vector space models.

The *query likelihood model* [155, 198] ranks documents D by the likelihood $P(D \mid q)$ for a query q; using Bayes' rule and the assumption that $P(q)$ is uniform, so one obtains $P(D \mid q) \propto P(q \mid D)P(D)$. The prior $P(D)$ is usually set to be uniform and documents are ranked by the probability that their model generates the query. More formally, $P(q \mid D) = \prod_{w \in q} P(w \mid D)$, where w is a term in a query. The most intuitive approach to estimate $P(w \mid D)$ is to use $\hat{P}(w \mid D)$, the maximum likelihood estimate of D. However, as $P(q \mid D)$ is a product of all terms it will be 0 whenever one single query term does not occur in D. One therefore employs smoothing techniques.

To obtain $P(w \mid D)$, ones can use Jelinek-Mercer smoothing, defined as a linear interpolation between $\hat{P}(w \mid D)$, the maximum likelihood estimate of D, and $P(w \mid C)$, the estimated probability of seeing w in the collection C [74, 155]:

$$P(w \mid D) = (1 - \lambda)\hat{P}(w \mid D) + \lambda P(w \mid C). \tag{2.1}$$

For $\lambda = 0$ no smoothing is performed, while for $\lambda = 1$ the retrieval of the document is document-independent. Dirichlet smoothing generally performs better [282] as the interpolation with the background corpus is document-dependent. Here,

$$P(w \mid D) = \frac{\hat{P}(w \mid D) + \mu\lambda P(w \mid C)}{|D| + \mu}, \tag{2.2}$$

where μ often is set to be the average document length of the collection [151].

Table 2.2 introduces some of the basic notation used in this thesis. We use the multinomial unigram language model to estimate $\hat{P}(w \mid D)$ and, unless otherwise stated, use

both smoothing methods. We therefore assume term independence in documents.

Query Modeling One thing most search systems have in common is the query, which is assumed to be representing the user's underlying information need. As a query often consists of only a few keywords, this may or may not be adequate. Query modeling aims to transform simple queries to more detailed representations of the underlying information need. Among others, those representations can have weights for terms or may be expanded with new terms. There are two main types of query modeling, global and local. Global query modeling uses collection statistics to expand and remodel the query. An example of global query modeling can be found in [202], using thesaurus and dictionary-based expansion, and Meij and de Rijke [160] perform semantic query modeling by linking queries to Wikipedia. Local query modeling is based on the top retrieved documents for a given query. Typical local query expansion techniques used are the relevance models by Lavrenko and Croft [139].

Relevance models [139] re-estimate the document probabilities based on an initial feedback set. First, the top-N documents ($\mathcal{D}$) for a query q are retrieved using a simple retrieval method (e.g., Eq. 2.2). A model M_D over a document D is the smoothed maximum likelihood distribution over the term unigrams in the document D. The set of all models M_D where $D \in \mathcal{D}$ is $\mathcal{M}_D$. For all documents D, the final score is computed as

$$\text{score}(D) = \prod_{w \in D} \frac{P(w \mid R)}{P(w \mid N)}, \tag{2.3}$$

where R is a model of relevance and N of non-relevance. The term $P(w \mid N)$ can be based on collection frequencies. As to $P(w \mid R)$, Lavrenko and Croft [139] assume that the query was generated from the same model as the document. The model of relevance R is then based on the query and

$$P(w \mid R) = \lambda \frac{P(w, q)}{P(q)} + (1 - \lambda) P(w \mid q), \tag{2.4}$$

where $P(q)$ is assumed to be uniform, $P(w \mid q)$ is defined as the maximum likelihood estimate of w in the query, and $\lambda \in [0, 1]$. Interpolation with the query model was shown to be effective [116]. We use the first relevance model (RM-1), i.i.d. sampling of the query terms with a document prior [139], to estimate $P(w, q)$:

$$P(w, q) = \sum_{M_j \in \mathcal{M}} P(M_j) P(w \mid M_j) \prod_{w' \in q} P(w' \mid M_j). \tag{2.5}$$

The relevance model is then truncated to the top-N_{RM} terms. The resulting relevance model is often called RM-3 [116]. Richer forms of (local) query modeling can be found in work by Balog et al. [21, 23].

Ounis et al. [176] organised the first platform for researchers to exchange ideas and approaches for information retrieval on social media at TREC 2006, 2007, and 2008 [153]. We detail the dataset in Section 3.2. The main task was the opinion retrieval task. The opinion retrieval task asked participants to retrieve *What do people think about X*, with X being the target. As the task was mainly solved in two stages,

the retrieval task and the opinion ranking task, the retrieval task can be reviewed separately. While it was hard to perform better than the baselines, query expansion and modeling [118, 267, 268, 283] proved to be a recurring approach every year (see below). For blog (post) retrieval, one often uses large external corpora for query modeling [66]. Several TREC Blog track participants have experimented with expansion against a news corpus, Wikipedia, the web, or a mixture of these [118, 267, 268, 283]. Weerkamp [264] provides an excellent overview over different approaches to information retrieval in blog search and offers new techniques. For blog retrieval, the motivation underlying most of these approaches is to improve the estimation of the query representation, often trying to make up for the unedited nature of the corpus from which posts or blogs need to be retrieved. Our approach, presented in Chapter 6, tries to address this problem by focusing on bursts in the collection. There we use Dirichlet smoothing, Jelinek-Mercer smoothing, as well as relevance models as baselines for our new approaches.

2.3.1 Temporal Information Retrieval

Time is an important dimension for IR [3], and plays an important part in the definition of relevance [104]. Temporal information retrieval (TIR) takes the temporal dimension into account for typical information retrieval tasks [3]. Alonso et al. [3] state the main challenges of TIR, ranging from extracting mentions of time within documents and linking them (like [258]), to spatio-temporal information exploration (e.g., [156]), and temporal querying (such as [174]). With respect to this thesis four open research questions deserve more attention:

1. What is the lifespan of the main event?

2. How can a combined score for the textual part and the temporal part of a query be calculated in a reasonable way?

3. Should two documents be considered similar if they cover the same temporal interval?

4. Can we improve bibliographic search instead of just sorting by publication date?

The main topic in this thesis is retrieval on social media with a temporal dimension. In the following we review how previous work approaches this problem with respect to an answer to the questions 1 to 4.

Early research efforts in TIR were based on news data. Under the assumption that more recent news documents are more likely to be read and deemed relevant, early work by Li and Croft [144] creates an exponential recency prior. Rather than having a uniform document prior $P(D)$, they use an exponential distribution (or decay function). Intuitively, documents closer to the query time $\mathrm{time}(q)$ have a higher chance of being read and are therefore more likely to be relevant. Hence, the prior $P(D)$ can be approximated by $P(D \mid \mathrm{time}(q))$, a query time $\mathrm{time}(q)$ dependent factor:

$$P(D \mid \mathrm{time}(q)) = \beta e^{-\beta(\mathrm{time}(q)-\mathrm{time}(D))}, \tag{2.6}$$

where $\mathrm{time}(D)$ is the document time. The exponential decay parameter β indicates how quickly news grows old and less relevant. The higher it is, the steeper the curve, causing more recent documents to be rewarded.

Corso et al. [56] rank news articles using their publication date and their interlinkage. Li and Croft [144] relied on a manually selected set of queries considered temporal. What if we can identify which queries are temporal? Jones and Diaz [125] classify queries according to the temporal distribution of result documents into temporal and non-temporal queries. Efron and Golovchinsky [74] expand on Li and Croft [144]'s recency prior by directly incorporating an exponential decay function into the query likelihood. In this thesis, we examine the performance of a range of cognitively motivated document priors.

Meanwhile, TIR gained more interest with the rise of social media data. Weerkamp and de Rijke [265] use timeliness of a blog post as an indicator for determining credibility of blog posts. For blogger finding, Keikha et al. [130] propose a distance measure based on temporal distributions and Seki et al. [221] try to capture the recurring interest of a blog for a certain topic using the notion of time and relevance. For blog feed retrieval, Keikha et al. [129] use temporal relevance models based on days and the publications in the blog feeds.

Later, microblogging emerged and introduced new challenges. For one, the documents are "micro", i.e., very short. In the case of Twitter 140 characters long. Further, the information need is recent and traditional ranking does not directly transfer. Much of the research in microblogging is focussed on Twitter data, due to its accessibility, legally and electronically. Several new problems surfaced. Under the assumption that the most recent tweets are the most relevant, Massoudi et al. [157] use an exponential decay function for query expansion on microblogs. According to Efron [72] the unit of retrieval is often not clear, opinion and subjectivity are present, and time and place are vital and should be taken into account. Teevan et al. [239] compares search behaviour on Twitter with web search and finds that users search for temporally relevant information and information related to people. Early approaches for microblog retrieval embraced the notion of time, be it as a prior [157] together with document quality, or to help for query modeling [74]. Ultimately, this lead to the introduction of a Microblog track at TREC 2011. In 2011 (and the following year) the task was to return relevant and interesting tweets for a given query Ounis et al. [177], but close to the timestamp of the query—essentially realtime search. The participants exploited typical characteristics of Twitter, e.g., the hashtags, the existence of hyperlinks, and the importance of time. Typically they were integrated into query expansion, filtering or learning to rank approaches [177, 228]. The 2011 dataset resulting from this track is being described in Section 3.3.

Efron [73] argues that query-specific temporal dynamic functions might be the key for microblog retrieval, and Efron et al. [75] use document expansion and incorporates a dynamic exponential prior. Khodaei and Alonso [132] propose to use temporal information, like recency of likes and friendships for social search. Whiting et al. [270] also makes use of the pseudo-relevant documents, creating a similarity graph between terms. The similarity is based on traditional term frequency as well as the terms temporal similarity. Metzler et al. [163] retrieve events instead of single messages, using the messages as a description of the happenings around the events. Similarly, Choi and Croft [53] select time windows for query expansion on microblogs based on social signals, such as

retweets. Independently, Lin and Efron [146] look at the temporal distributions of the pseudo-relevant and the relevant documents for microblog retrieval. They use an oracle based on the temporal distribution of the relevant documents to motivate why this should be done. At TREC 2013 a new, much larger corpus and dataset was introduced with again the task of realtime search [145].

In general, the approach to detecting temporally active time periods (salient events) in the temporal distribution of pseudo-relevant documents proved successful in the news and blog setting [10, 26, 61, 62, 96, 129]. Berberich et al. [27] detect temporal information needs, manifested as temporal expressions, in the query and incorporate them into a language model approach. Amodeo et al. [10] select top-ranked documents in the highest peaks as pseudo-relevant, while documents outside peaks are considered to be non-relevant. They use Rocchio's algorithm for relevance feedback based on the top-10 documents. Dakka et al. [62] incorporate temporal distributions in different language modeling frameworks; while they do not actually detect events, they perform several standard normalizations to the temporal distributions. Building on our work [188], Berberich and Bedathur [26] and Gupta and Berberich [96] motivate the need for diversification of search results using twenty years of New York Times data. They combine models from [62] with diversification models. This is different from our work in so far as our datasets did not feature multiple salient events therefore rendering the need for diversification useless.

In Chapter 6 we use salient events for query modeling in news and blog data as well as for reranking for candidates in active learning for entity filtering on microblog data. In Chapter 7 we present different, cognitive, priors and use them similar to the exponential decay prior presented in Equation 2.6.

2.4 Entity Filtering

A different, but also important task for Online Reputation Analysis (ORA), is the filtering of social media for the relevance to an entity—entity filtering (EF). Filtering is different to retrieving documents: retrieval is the initial step of finding documents potentially relevant to an entity in a *large* document collection. The documents are often ranked according to relevance to the entity with a high recall being most important. The filtering of documents then casts a binary decision for relevant or not relevant, introducing higher precision. It is a vital preprocessing step for other tasks in ORA: If the performance of the EF module decreases, the performance of all subsequent modules is harmed [230]. Early filtering tasks at the TREC-4–TREC-9 conferences asked systems to filter a stream of documents according to a topic [206] to create profile. The topics could change over time.

WePS3 [5] is a community challenge for entity disambiguation for reputation management. The task was to disambiguate company names in tweets. From this emerged a large body of research on entity disambiguation [194, 245, 279]. The EF task has been part of evaluation campaigns at WePS-3 [5] and RepLab 2012 [6] and 2013 [7]. A similar task, not motivated by ORA, was introduced at TREC-KBA [87, 88]. Here, the focus is to present an entity as a semi-structured document that evolves over time and find important and relevant documents to improve the *profile* of an entity. In the ORA motivated task [5–7] *every* relevant document is interesting and needs to be looked at.

The RepLab 2013 dataset is provided with entity-specific training data, that can be used to build entity-oriented supervised systems and simulate the daily work of reputation analysts. The system that obtained the best results at RepLab 2013, POPSTAR [214], is based on supervised learning, where tweets are represented with a variety of features to capture the relatedness of each tweet to the entity. Features are extracted both from the collection (Twitter metadata, textual features, keyword similarity, etc.) and from external resources such as the entity's homepage, Freebase and Wikipedia. The best run from the SZTE_NLP group [99] applies Twitter-specific text normalization methods as a preprocessing step and combines textual features with Latent Dirichlet Allocation to extract features representing topic distributions over tweets. These features are used to train a maximum entropy classifier for the EF task. The best run submitted by LIA [58] is based on a k-nearest-neighbor classifier over term features. Similar to the official RepLab 2013 baseline, which labels each tweet in the test set with the same label as the closest tweet in the training set, LIA matches each tweet in the test set with the n most similar tweets. The best run submitted by UAMCLyR [216] uses a linear kernel SVM model with tweets represented as bags-of-word; this run is very similar to our passive initial model presented in Chapter 7, apart from differences in the preprocessing of tweets and the learner parameters. Finally, UNED_ ORM [230] report on the results of a Naïve Bayes classifier of bag-of-words represented tweets that performs similarly to the RepLab 2013 baseline.

Apart from the difference in task, the work in this thesis differs from earlier approaches in that (1) we do not use external data, only tweets are considered to learn a model, and (2) new labeled instances are directly added to the training set used to update the model. We present this approach in Chapter 7.