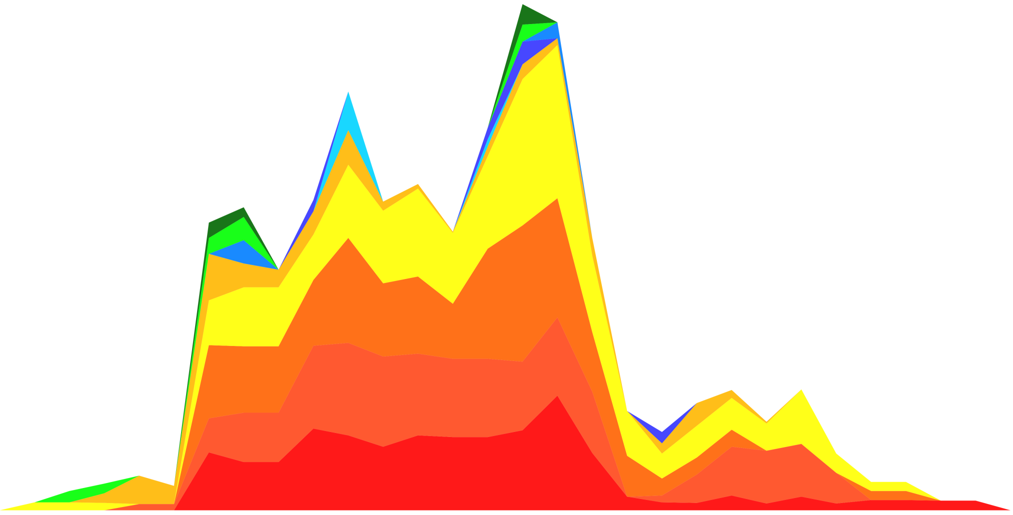
ANALYSIS AND FORECASTING OF TRENDING TOPICS IN ONLINE MEDIA

CHRISTOPHER TIM ALTHOFF



Computer Science Department
University of Kaiserslautern
Germany

April 2013

Master's Thesis
Department of Computer Science
University of Kaiserslautern
P.O. Box 3049
67653 Kaiserslautern
Germany

SUPERVISORS:
Prof. Dr. Prof. h.c. Andreas Dengel
Damian Borth, M.Sc.

DECLARATION

I declare that this document has been composed by myself, and describes my own work, unless otherwise acknowledged in the text. It has not been accepted in any previous application for a degree. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Kaiserslautern, April 29, 2013

Christopher Tim Althoff

ABSTRACT

Among the vast information available on the web, social media streams capture what people currently pay attention to and how they feel about certain topics. Awareness of such trending topics plays a crucial role in many application domains such as economics, health monitoring, journalism, finance, marketing, and social multimedia systems.

However, optimal use of trending topics in these application domains requires a better understanding of their various characteristics in different social media channels. To this end, we present the first comprehensive study across three major online and social media channels, Twitter, Google, and Wikipedia, covering thousands of trending topics over an observation period of an entire year. Our results indicate that depending on one's requirements one does not necessarily have to turn to Twitter for information about current events and that some media channels strongly emphasize content of specific categories. As our second key contribution we further present a novel approach for the challenging task of forecasting the life cycle of trending topics in the moment they emerge. Our fully automated approach is based on a nearest neighbor forecasting technique exploiting the observation that semantically similar topics exhibit similar behavior.

We further demonstrate on a large-scale dataset of Wikipedia page view statistics that forecasts by the proposed approach are about 9-48k views closer to the actual viewing statistics compared to baseline methods and achieve a mean average percentage error (MAPE) of 45-19% for time periods of up to 14 days.

ZUSAMMENFASSUNG

Soziale Medienkanäle im Web erfassen welchen Themen Menschen derzeit Aufmerksamkeit schenken und spiegeln das momentane Stimmungsbild wider. Das Verständnis des Verhaltens von sogenannten Trending Topics, Themen von besonderer aktueller Relevanz, spielt eine entscheidende Rolle in Anwendungsdomänen wie Wirtschaftswissenschaften, Gesundheitsüberwachung, Journalismus, Finanzwesen, Marketing, und sozialen Multimedia-Systemen.

Eine optimale Nutzung von Trending Topics in diesen Domänen setzt ein gutes Verständnis von deren Eigenschaften und Verhaltensmerkmalen in verschiedenen sozialen Medienkanälen voraus. Zu diesen Zweck präsentieren wir die erste umfassende Studie von tausenden von Trending Topics in drei bedeutenden Online-Medienkanälen, Twitter, Google und Wikipedia, über einen Beobachtungszeitraum von einem Jahr. Unsere Ergebnisse zeigen, dass man je nach Anforderungen nicht notwendigerweise Twitter als Informationsquelle für aktuelle Geschehnisse heranziehen muss und dass manche soziale Medienkanäle bestimmte Inhaltskategorien besonders betonen. Der zweite Hauptbeitrag dieser Arbeit ist ein neuartiges Prognoseverfahren, um den komplexen Lebenszyklus von Trending Topics zum Zeitpunkt deren Entstehung vorherzusagen. Der vollautomatische Ansatz basiert auf einem Nearest-Neighbor-Prognoseverfahren und nutzt aus, dass wie in dieser Arbeit beobachtet semantisch verwandte Themen ein ähnliches Verhalten aufweisen.

Auf einem umfangreichen Datensatz von Anzeigestatistiken von Wikipedia-Artikeln zeigen wir empirisch, dass der vorgeschlagene Ansatz 9-48k Betrachtungen näher an der Wirklichkeit liegt als vergleichbare Baselines und autoregressive Modelle, und dass dieser Ansatz Vorhersagen für bis zu 14 Tage mit einem Mean Average Percentage Error (MAPE) von 45-19% ermöglicht.

Computers are good at following instructions,
but not at reading your mind.
— Donald E. Knuth [53]

ACKNOWLEDGMENTS

This thesis originates from my work as a student research assistant at the German Research Center for Artificial Intelligence (DFKI), where I have been a member of the Multimedia Analysis and Data Mining research group from May 2009 until May 2011 and after my time abroad again since January 2013. I would like to thank Prof. Andreas Dengel for the many opportunities I've been given at DFKI and constant support during my studies and academic career.

I'm heartily thankful to my supervisor, Damian Borth, Master Mind, who from nightly discussions about trending topics on the streets of Nara, Japan to the last draft of my thesis provided me with opportunities, motivation, instruction and insight. I am grateful for the freedom in researching various aspects of trending topics and forecasting, a great time in Japan, and countless discussions on soccer, hair transplants, and other things.

I offer my regards to all of those who supported me in any respect during the completion of the project. Particularly, I feel very grateful to Jörn Hees for many helpful discussions as well as paper writing, reviewing, coding, and microgolf sessions, to Andreas Hoepner for insighful discussions on time series analysis, and to Wolfgang Schlauch and Klaus-Dieter Althoff for proof-reading the manuscript. I am further thankful for the support through the PiCloud Academic Research Program Grant that enabled me to conduct this research on their computing cluster free of charge.

Lastly, my deepest gratitude goes to my family, and especially my parents, for all the opportunities they have created for me and for giving me the best possible support, careful correction, and enabling freedom.

Christopher Tim Althoff

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Currently, as multimedia content consumption is rapidly increasing on the internet we can consider ourselves as living in the zettabyte era [82]. One major driving factor is increasing amounts of rich media content, with users downloading video in 2016 at an estimated rate of 1.2 million video minutes per second. Another factor is people increasingly using the web for communication. Online social networks are already the primary method for communication among U.S. college students [42] and Twitter users communicate through over 340 million short text messages each day which often include rich media content such as images or links to videos or news articles [88]. To a large degree, this vast amount of information is created, shared, and exchanged by the users themselves in such social networks and multimedia systems. These systems provide us with channels of information that capture how we are spending our time and what we are talking about. Over time, different topics arise in these media channels reflecting changing interests of groups of individuals. In this sense, social media channels mirror our society [9]. These channels contain rich information and immediate feedback about what people currently pay attention to and how they feel about certain topics.

This data capturing human behavior on the web is of particular interest in fields such as computational social science and represents one instance of the emerging *Big Data* trend – "high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [11]." It is believed that these large-scale datasets will drive research and allow for many new insights in the coming future. Particularly in the field of social science and social media, there is an increasing interest to analyze and harness these datasets for a better understanding of human behavior, society, as well as information creation, consumption, and diffusion. However, current commercial applications such as

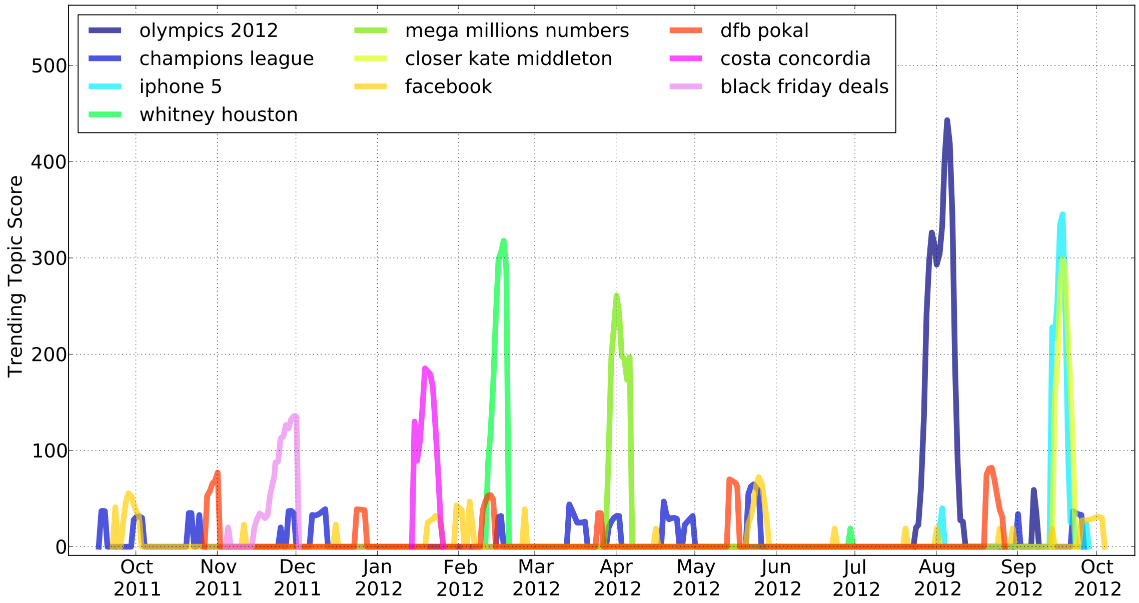


Figure 1: Top 10 trending topics over the course of one year. Note that different trending topics exhibit different behavioral patterns such as the death of Whitney Houston causing a single spike whereas Champions League recurs many times.

Raytheon Riot (Rapid Information Overlay Technology)[1] or Recorded Future[2] tend to focus on aggregation and visualization of the underlying data and seldom provide in-depth analyses or actual forecasting capabilities. In contrast, this work presents a multi-channel analysis of trends in online and social media as well as a fully automatic forecasting approach.

1.1 TRENDING TOPICS

Trends in online media are commonly called *trending topics* [63]. In the scope of this work, a *topic* refers to a subject matter of discussion or conversation that is agreed on by a group of people (usually represented by a textual label). Such a topic becomes *trending* when it experiences a sudden spike in user interest or engagement. Therefore, the ability to measure user interest and engagement is necessary. In the context of this work, these measures are approximated by the

[1] Retrieved from `http://www.raytheon.com/` on April, 16.

[2] Retrieved from `https://www.recordedfuture.com/` on April, 16.

amount of attention that is paid to a topic in online and social media channels. As more and more people use the web for news, information, and communication, their online activity becomes an good proxy for the interests of the global population. Concrete statistics and measures for such online activity include the number of search queries, posts, re-posts, shares, or comments related to a given topic or the rank of that topic within a ranking of trends in different media channels (more details will be given in Section 4.1). Observing several online and social media channels over the course of one year led to the discovery of many different trending topics. Ten of the most important trends in the observation period are shown in Figure 1. Generally, trends live for a limited amount of time with most trends vanishing after a couple of days. Further note that the topics exhibit very different behavior leading to different patterns in the plot. For example, the Summer Olympics 2012 (dark blue) were a trend for a relatively long time and we can identify two separate peaks that correspond to the opening and closing ceremony of the event. There are also several trends that feature a single spike in user interest such as Whitney Houston's death (green). The interest in Black Friday deals (light pink) increases gradually before it drops sharply after Black Friday. Product releases such as the iPhone 5 (light blue) feature small peaks around the time where rumors are spread and release dates are being discussed and one large peak when the product is finally released. There are also trends that occur multiple times such as Champions League (blue), a sports topic discussed whenever important games are played (note that nobody talks about it during the summer break, see Section 4.2).

As illustrated by Figure 1, there is a wide variety of trending topics in online media and most of them were found to correspond to real world events such as sport events (Olympics 2012), product releases (iPhone 5), celebrity news (Steve Jobs' death), incidents (Sinking of the Costa Concordia), political movements (Occupy Wall Street), and entertainment (Academy Awards). Note that these trending topics can subsume multiple related stories such as the "Olympics 2012" representing events in different sports (running, hurdles, swimming, gymnastics, etc.) or the political "Occupy" movement referring to several events in different locations across the United States of America (New York City, Oakland, Washington D.C., etc.). An example for all the individual topics, stories, and events related to a trending topic is given in Figure 2 for the "Olympics 2012". Note that it unifies interest in the current medal count, the TV channels showing the Olympics, or different sports such as hurdles, swimming, canoe, or fencing.

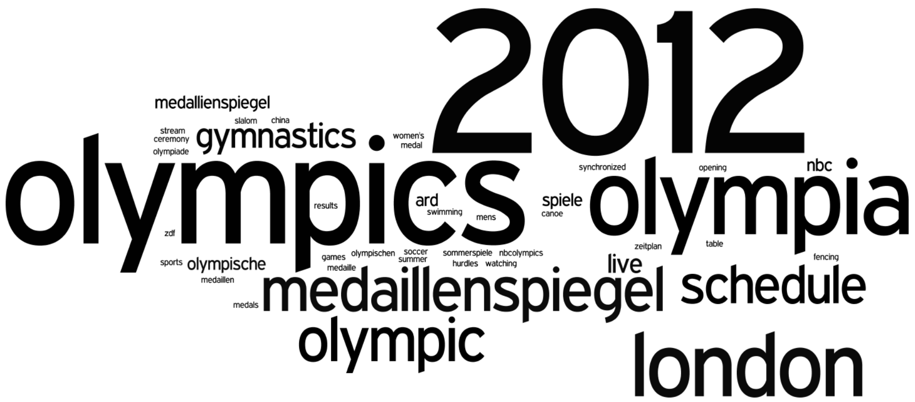


Figure 2: A word cloud representing individual topics of the cluster "Olympics 2012" (larger font size indicates higher term frequency). Note that the cluster contains diverse sports such as (swimming, canoe, hurdles, fencing) as well as terms in different languages (schedule, zeitplan).

Today, a large variety of social multimedia systems is available to users on the web including Twitter, YouTube, Facebook, Flickr, Pinterest, Tumblr, Google News/Trends, and Wikipedia serving different kinds of needs such as information demand, social communication, as well as sharing and consumption of multimedia content (for more details please refer to Section 2.1). Therefore, many users turn to multiple of these online platforms, depending on their present individual needs, creating a heterogeneous multi-channel environment within the social media landscape. While researchers have a general intuition about the nature of these channels (e.g. many current event detection systems are based on Twitter feeds [66, 92]), studies of the exact distribution of information across multiple channels have not been in the scope of the research community until now. For example, the trending topic "Olympics 2012" manifests itself in multiple media channels which clearly exhibit different behavior (as illustrated in Figure 3). For instance, this topic is only trending on Twitter (blue) or Wikipedia (green) for short periods of time while it continues to grow on Google (red). This raises questions such as whether the individual channels have different temporal characteristics, whether some channels are always faster than others and pick up trends earlier, or whether these characteristics depend on the needs served by the respective channels. These types of questions have direct impact on the design of trend-aware information systems in various application domains as illustrated below.

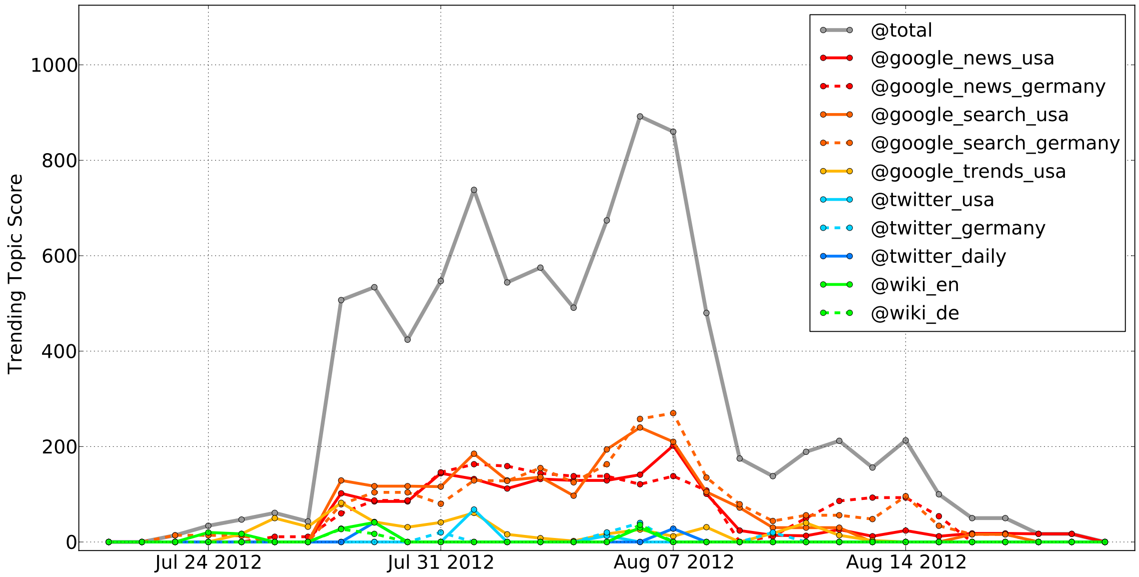


Figure 3: The trending topic "Olympics 2012" during summer 2012. The colored curves represent the contribution of the different online and social media channels. Note how these channels behave differently, e.g. that the trend rises and vanishes quickly on Twitter or Wikipedia but continues as a trend on Google for longer periods of time.

1.2 APPLICATION DOMAINS

As opportunities become more and more abundant, data analysis becomes increasingly important for corporations, government, politics, and marketing. A substantial amount of data that is analyzed to provide new insights is proprietary and private data typically owned by corporations that collected it. Supported by movements such as the Open Data Initiative[3] there is also a trend of publicly releasing data to provide more transparency for example in the government sector. Today, citizens have access to public data and reports about health, transport, education, agriculture, or population statistics. However, writing such reports and releasing the corresponding data can take some time. Therefore, reports and releases are often delayed. As will be illustrated below, many application domains benefit or even require more real-time access to people's current interests and intentions. In this case, analyzing social media streams is a promising option as they increasingly mirror present developments of our society. As people more and more turn to the Internet for news and information, online activity can be seen

[3] Retrieved from `http://www.opendatainitiative.org/` on April 16, 2013.

as a snapshot of the collective consciousness reflecting current interests, concerns, and intentions of the global population [5, 40, 36]. For example, if somebody is planning on buying a new car they will often research different options online by searching the web for current models and specifications, or clicking on search results promising good deals. They might also ask their peers for advice by requesting help on online social networks or discussing their purchase in online messaging boards. Therefore, our current interests and opinions are well represented by our behavior in online and social media channels. Detecting certain patterns of behavior can be very lucrative, for instance for a car dealer that would like to sell a car to the likely buyer of the previous example. This led researchers to detect and exploit such patterns based on the observation that what people are searching for today can be indicative of their current and future situations and interests. For instance, it was found that when people become ill they tend to search for specific diseases and related keywords. This has been exploited to estimate the number of flu infections in the USA (see Figure 4). The red curve (ILI: influenza-like illness) illustrates the official number of reported flu infections while the blue curve is an estimate based on how many people queried Google for flu-related terms such as "flu remedy", "body aches", or "sore throat". Note how well the search behavior approximates the actual number of flu infections. Clearly, people's online behavior reflects their current situation and needs in this case. A similar approach was used to predict the current number of unemployment claims based on the number of people that searched the web for unemployment benefits or to find a new job. The official statistics by the US Department of Labor and the prediction based on search data are shown in Figure 5. Again, the online behavior significantly matches the real-world behavior. Note that these two examples predict the present rather than the future (often referred to as nowcasting for that reason). To predict today's flu infections or unemployment claims they require today's (!) search activity. However, some of these signal feature recurring patterns or seasonality such as the yearly seasonality in the flu example. If past behavior is similar to current behavior patterns this suggests that predicting the future, i.e. actual *forecasting*, might be possible as well. Such forecasts could be used to estimate necessary flu vaccine production levels. Of course, the same idea applies to other domains as well. For example, one could estimate the popularity of a new movie based on how many people search for trailers or movie theaters, or how many products a company can expect to sell based on how many people talk about them or search for reviews online. Obviously, knowing about such

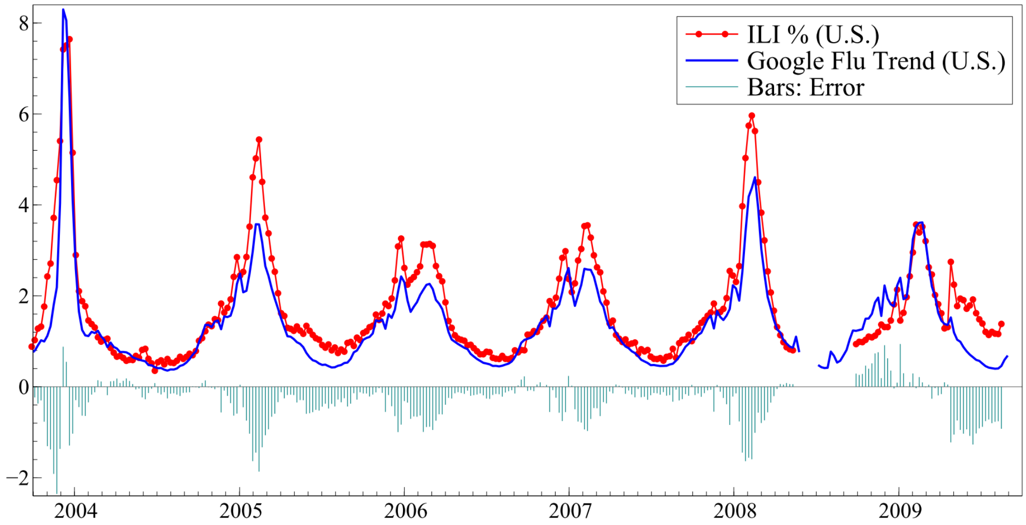


Figure 4: Comparison of ILI% as reported by the US Centers for Disease Control and Prevention and the Google Flu Trend indicator based on search behavior. The bars show the difference between the CDC data and Google Flu Trends. Figure taken from [27].

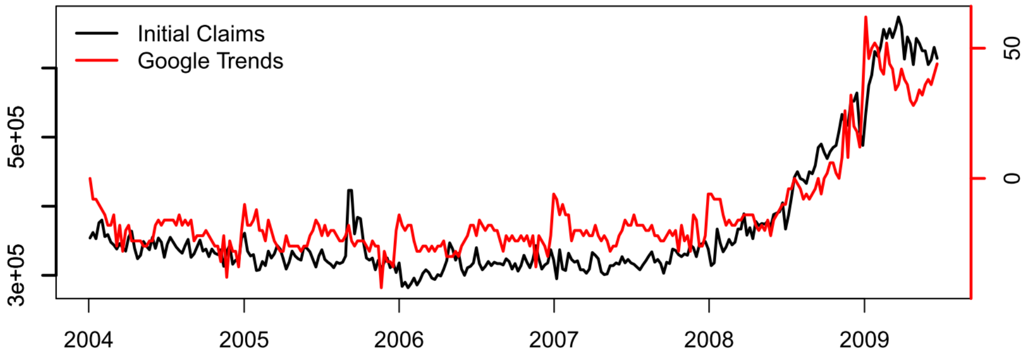


Figure 5: Comparison of initial unemployment claims as reported by the US Department of Labor and a Google Trends indicator based on terms of the "Welfare & Unemployment" category. Figure taken from [18].

trends has strong implications for product placement and monitoring strategies as well as advertising.

Awareness of trends in online media is also of increasing interest in finance, marketing, and journalism to support human decision making. Financial investors are interested in trends as a real-time signal for trading and portfolio management or as indicators of shifting collective interests and opinions. Knowing that topics such as climate change will receive a lot of attention in the future can put investors with early access to such forecasts in a fortunate position. Further, companies are monitoring online and social media channels to learn about their consumers as well as detect potential scandals. The earlier such problems are detected the more time these companies have to judge the potential impact and to react early. Reacting quickly has become more important in marketing as well. The most prominent example of real-time marketing might be the power outage during the 2013 Super Bowl. Several companies ran ads only minutes after the power went out. Oreo's Twitter ad "No power? No problem. You can still dunk in the dark." was shared (i.e. retweeted) 10,000 times within one hour creating even more user engagement than its much more expensive TV ad [62, 49].

Many journalists aim to cover stories and write articles that interest a large group of people. Therefore, they are interested in what is currently going on in the world, what people would like to be informed about, and who started certain trends to help motivate and guide what stories they are covering. Particularly, the subfield of data-driven journalism refers to the journalistic process that is largely based on analyzing and filtering large data sets for the purpose of creating a new story. A major player in this field is the Open Knowledge Foundation[4] that promotes open knowledge, including open content and open data. The importance of real-time news coverage was further emphasized by the Pulitzer Prize 2013 which was awarded to the Denver Post staff for its "comprehensive coverage [...] using journalistic tools, from Twitter and Facebook to video and written reports, both to capture a breaking story and provide context [68]." The Denver Post Staff had made extensive use of Facebook updates and Tweets to cover the mass shooting at a movie theater in Aurora, Colorado, particularly during the first 24 hours of coverage [41]. As writing detailed and insightful articles requires time, journalists are further interested in what topics people will be interested in a couple of days in the future and would therefore greatly benefit from forecasts that predict the future relevance of potential topics.

Understanding the temporal characteristics of trending topics plays a key role in building social multimedia systems. As illustrated before people turn to the web to read news articles, participate in discussions in online social networks, or watch videos on video sharing platforms. Often, these people have specific tasks or interests in mind when they turn to such social multimedia systems. They may want to learn about new topics on the news such as recent award ceremonies or upcoming sport events, or participate in discussions on political topics they are passionate about, or they simply want to be entertained by yet

[4] Retrieved from `http://okfn.org/` on April 15, 2013.

another group producing a "Harlem Shake"-type video (a viral video trend which has sparked over 100k imitations and garnered nearly a billion views [44]). In all these cases, the users' needs are related to and driven by current and future events. Therefore, their multimedia consumption experience could be enhanced by explicitly recommending and serving content such as videos, photos, and news stories of current interest. To this end, it is necessary to be aware of current trends for example by detecting what people increasingly pay attention to in social media and social networks. Given a number of trends from such a trend detection system one needs to find related content that could be shown to the user. Therefore, trend detection systems have been developed and integrated with multimedia recommendation engines [26, 74]. However, finding related multimedia content can be a challenging problem if that content is poorly annotated and lacks textual description. This would require concept detection systems trained on current trends that are able to automatically recognize further instances of trend-related content [15]. To then efficiently serve such multimedia content to the end user in a timely manner, video platforms such as YouTube create content delivery networks in which they distribute and replicate specific videos to their servers all around the world. This raises questions such as when and where to store and replicate videos of certain topics. Obviously, very popular videos should be replicated more often as by definition they will be requested more frequently. Therefore, it is very important to estimate the current and future popularity of a video to optimize content delivery networks [90]. This popularity has been estimated for example by counting how often a particular video has been mentioned or shared on online social networks or microblogs. Essentially, this tries to estimate the eventual video popularity through some measure of early popularity. When a new video is created such information might not be available, yet. However, the popularity of a video in part depends on the popularity of the video content or associated topic. Therefore, if the topic popularity was known one would be able to make a reasonable guess about how many people might be interested in watching that particular video. Unfortunately, forecasting the popularity on a topic-level is very challenging because such a topic often subsumes the popularity of multiple different events. For example, while it might be possible to model the popularity of a particular video of the 2012 Olympics 100m dash, associated topics such as the Olympics 2012 subsume other events such as the opening ceremony, gymnastics, or weight-lifting competitions that all have to be taken into account. Here, forecasting topic-level popularity

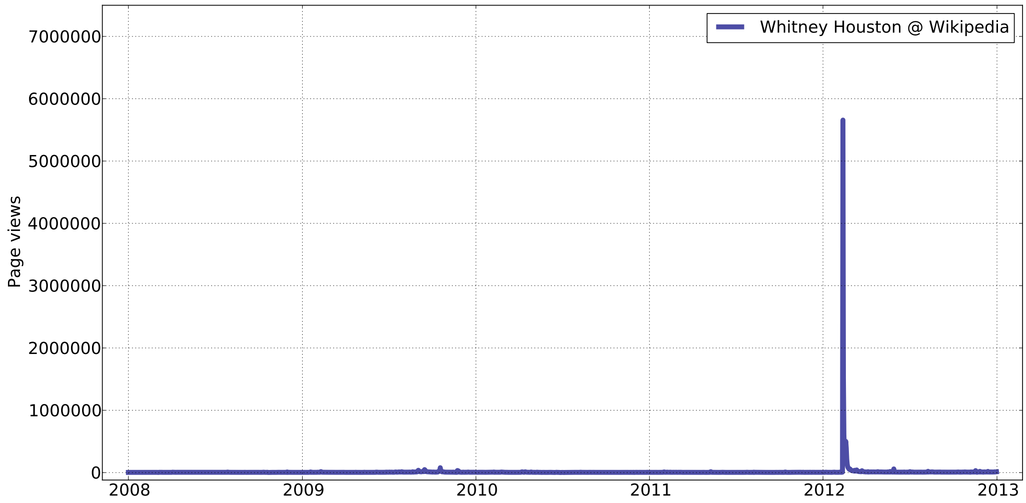


Figure 6: Structural breaks often occur when topics become trending. Here, the page views for the Whitney Houston article on Wikipedia show a 2000-fold increase following the actress's death on February 11, 2012.

is equivalent to forecasting the life cycle of a trending topic, i.e. forecasting the amount of corresponding user engagement in the very moment it emerges, and represents one main object of study of this thesis. While obviously of great use, forecasting trending topics is a very challenging problem since the corresponding time series tend to exhibit highly irregular behavior. These irregularities are also known as structural breaks that can lead to large forecasting errors and unreliable models [21] (for more detail please refer to 2.2.6). A good example for such a structural break is Whitney Houston's death in February 2011 that caused 2000 times (!) more people to access her Wikipedia article than usual (see Figure 6). The complexity of this task will be further illustrated in more detail in Chapter 5.

1.3 GOALS AND OUTLINE

As motivated, many application domains utilize trending topics and therefore require a good understanding of their behavior across different online and social media channels. Therefore, this work presents a multi-channel analysis of trending topics including temporal and topical characteristics of online and social media channels over a one year observation period. To the best of the author's knowledge, there does not exist a similar study prior to this work.

Furthermore, in several application domains there is an increasing demand to predict future behavior of people, i.e. upcoming trends and developments, rather than analyzing past behavior. These forecasts then enable us to anticipate changing information needs of users ahead of time and react accordingly, for example by changing product placement strategies, writing certain news articles, or optimizing resource allocation in content delivery networks. This work presents a novel forecasting approach exploiting the observation that semantically similar topics exhibit similar behavior in a nearest neighbor framework. Semantically similar topics are discovered by mining topics of similar type and category on DBpedia [6], a structured representation of the information on Wikipedia. The fully automatic approach is evaluated on a large-scale dataset of billions of views of several million articles on Wikipedia.

Summarizing, the contribution of this thesis is three-fold:

1. The first comprehensive study of trending topics in three major social and online media channels with an observation period of one year.

2. A fully automatic forecasting technique for these trending topics based on a nearest neighbor approach exploiting semantic relationships between topics.

3. An empirical evaluation of our forecasting model over real-world user behavior on a large-scale Wikipedia dataset.

A summary of the main insights from the analysis and evaluation is further presented in the conclusion (Chapter 7).

The reminder of this thesis is organized as follows. The second chapter provides background information on online and social media as well as introduces standard modeling and forecasting approaches for time series data. In Chapter 3, we describe previous work related to trending topics in multimedia and social media-based forecasting. The multi-channel analysis of trending topics is presented in Chapter 4. Our novel forecasting technique is described in Chapter 5 before evaluating the approach in Chapter 6. Lastly, Chapter 7 concludes this thesis with a summary and an outline for future work.

2 BACKGROUND

This chapter summarizes the impact of online and social media on our everyday lives, reasons about how addressing basic human needs could explain this impact, introduces three major platforms (Wikipedia, Twitter, and Google), and describes standard forecasting approaches for time series data.

2.1 ONLINE AND SOCIAL MEDIA

Online media refers to the means of mass communication of digital media (which includes news, photos, videos, and music) distributed over the Internet. Examples for the individual categories include The Huffington Post[1] (news), Flickr[2] (photos), YouTube[3] (videos), and Spotify[4] (music). Social media further includes the interactions among people with a common definition being "the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks [2]." Popular examples for social media include the social networks Facebook[5], LinkedIn[6], and Twitter[7].

The following statistics and examples illustrate that social media has had a tremendous impact on society and has arguably changed our lives and how we interact with one another and the world around us. Communication between people increasingly takes place online with, for example, the social network Facebook being the primary method for communication by college students in the United States now [42]. If Facebook were a country it would be the third

[1] Retrieved from http://www.huffingtonpost.com// on April 18, 2013.
[2] Retrieved from http://www.flickr.com/ on April 18, 2013.
[3] Retrieved from http://www.youtube.com/ on April 18, 2013.
[4] Retrieved from http://www.spotify.com/ on April 18, 2013.
[5] Retrieved from http://www.facebook.com/ on April 18, 2013.
[6] Retrieved from http://www.linkedin.com/ on April 18, 2013.
[7] Retrieved from http://www.twitter.com/ on April 18, 2013.

largest in the world by population (with about 1 billion users) [69]. Market research shows that social media is already used by over 90% of US companies (with 100+ employees) for marketing purposes with Facebook, LinkedIn, Twitter, and YouTube being the most important channels [28]. Social media marketing is becoming more important as people spend more time on social networks than any other category of sites – 20% of their time spent on PCs and 30% of their mobile time [45]. There is also a trend to use social media to report on one's personal life and to share updates with friends. For instance, nearly a quarter of people aged 18-34 use social media to comment on what they like/dislike about a storyline while watching TV [45].

Networks like Facebook and Twitter give citizens a sounding board to be heard by the rest of the world. While online chatter is often just noise, groups forming for a common cause can be heard collectively and make world news. Activists, for instance during the Arab Spring 2011 or the Egyptian elections 2012, used technology and social media to share ideas and tactics [60]. While there is some debate over the actual impact of social media in the facilitation of change, it is certain that it helped make these issues known to the world in a way that would not have possible before. Furthermore, social media is sometimes breaking news ahead of traditional media. For instance, Twitter users in New York City were seeing tweets about the 2011 Washington D.C. earthquake half a minute before they could even feel it [60]. Therefore, Soren Gordhamer [38] concludes that social media impacted where we get our news, how we start and do businesses, how we meet and stay in touch with people, what we reveal about ourselves, and what we can influence.

2.1.1 *Reasons for Participating in Social Media*

Online social media provides numerous opportunities to study the behavior of people and society at an unprecedented scale. Particularly, social networks are interesting from a societal point of view as they allow for studies how people connect, communicate, and interact with each other, possibly across geographic, national, political, linguistic, and cultural boundaries [55]. Online social media and networks just started around a decade ago with the advent of platforms such as Friendster (2002), LinkedIn (2003), MySpace (2003), and Facebook (2004). It is an interesting question how and why a recent phenomenon such as social

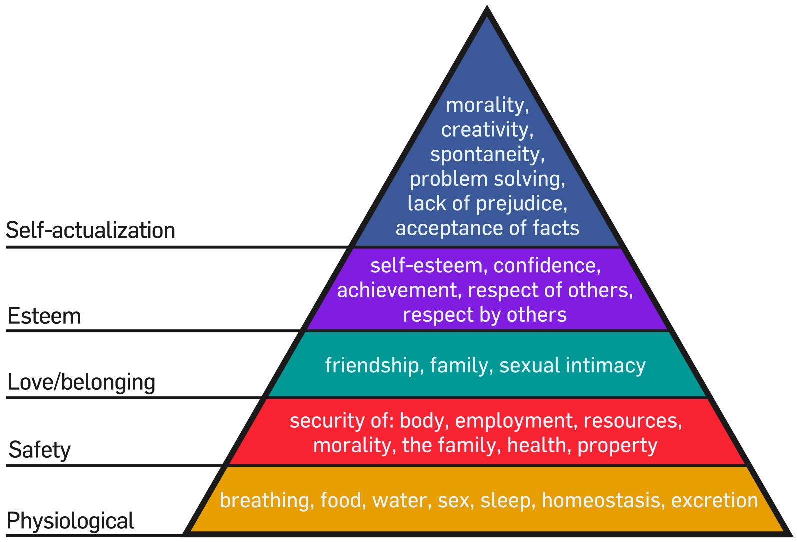

morality,
creativity,
spontaneity,
problem solving,
lack of prejudice,
acceptance of facts
Self-actualization
self-esteem, confidence,
achievement, respect of others,
respect by others
Esteem
friendship, family, sexual intimacy
Love/belonging
security of: body, employment, resources,
morality, the family, health, property
Safety
breathing, food, water, sex, sleep, homeostasis, excretion
Physiological

Figure 7: Maslow's hierarchy of needs with the most basic needs at the bottom of the pyramid and the more complex needs located at the top. Maslow's theory is often used to explain the incentives driving human behavior. Figure from [31].

media could have such an impact on our lives in such a short amount of time (as illustrated by the previous section). Psychologists often explain human behavior through incentives that motivate individuals to perform certain actions. In the case of the rapid adoption of social media technologies, researchers believe the main reason to be that these platforms help satisfy basic human needs [50]. A classic theory of human needs in psychology is Maslow's hierarchy of needs, proposed by Abraham Maslow in his 1943 paper "A Theory of Human Motivation" [59]. Maslow's hierarchy of needs is often portrayed in the shape of a pyramid (see Figure 7) with the most basic needs at the bottom while the more complex needs are located at the top – starting off with physiological needs, then safety, love/belonging, esteem, and finally, self-actualization needs. The hierarchy suggests that people are motivated to fulfill their basic needs first before moving on to the more complex ones. Maslow's hierarchy is often used to explain human behavior, including for example what human needs are served by social media products and why some of them are so successful [50]. Generally, such needs are related to social interaction and map to the higher levels in the pyramid: love/belonging, esteem, and self-actualization.

First of all, online social networks can be used to reconnect with old friends and to get to know new people as well. All major networks employ algorithms that are supposed to help the users to connect with friends and friends of friends. These connections and friendships map the love/belonging level in the hierarchy of needs. Further, as a large number of people spends increasingly more time on online social networks they can also be used to learn more about other people they already know and to grow relationships. A driving factor in the success of Facebook was the introduction of the relationship status in user's profile. This clearly relates to most people's need to find a partner in life. These networks also can give users a sense of belonging as users can become a part of various groups that share certain information only within that group or organize events.

As job search is progressively moving online and hiring committees gain access to personal profiles, every user's profile page can be seen as a part of that individuals online presence which is therefore becoming more and more important. Forbes already argued in 2011 that in ten years one's online presence will have replaced one's resume [78]. Therefore, online profiles are becoming more important and are used to portray a particular positive image of a person. Moving up in hierarchy to the esteem level, people use their online presence to build their self-esteem and self-confidence as they largely control what information about them is linked to their identity. Furthermore, users often try to draw a picture of their current situation that will gain their peers or followers interest and respect, for example by posting pictures of their family, holidays, cars, and other status symbols. People can also gain a sense of achievement if they accumulate a large number of followers that are interested in what they have to say or appreciate the content that is shared with them [51].

On the level of self-actualization, social media can support its users to develop and spread their creativity by creating and distributing new content. Such content includes short text messages and status updates, pictures shared with friends, blog posts or news articles, and self-produced videos. Furthermore, being connected to other people all around the clock through the Internet can allow for spontaneous actions that would never have been possible without rigorous planning and a few phone calls. For example, several services allow users to share their location with friends to enable spontaneous meet-ups whenever two friends happen to be in the same area. Lastly, social media can enable its users to empathize with others through spreading the news about humanitarian causes, environmental problems, economical issues, or political debates and help connect people on deeper levels for a common cause.

Note that, particularly in the area of social networks, Maslow's hierarchy has been criticized for imposing a hierarchical order on needs. Ruthledge [75] argues that "none of Maslow's needs can be met without social connection" and that therefore all human needs need to be "anchored in our ability to make social connections" as "belongingness is the driving force of human behavior, not a third tier activity" [75] referring to love/belonging as only the third stage in the hierarchy.

However, the increased usage of social media has its drawbacks as well. Some users have become addicted to social media services and even show withdrawal symptoms when forced to quit their habit [81]. In the light of the previous analysis of needs, Huffington Post writer Sam Fiorella arrives at the conclusion that social media has "gamified the [social] experience to appeal to our human needs so well that Maslow himself would weep with pride [34]." He points to increasing pressure and distraction to spend more and more time with social media, dangerous emotional investments in a culture of "over-sharing", and the loss of privacy as the price people are willing to pay for access to social networking.

Three very successful online media channels that because of their widespread use capture the behavior of a large fraction of the general population are Wikipedia, Twitter, and Google. Their success seems to indicate that they all cater well to the needs of these people. For example, Wikipedia satisfies our thirst for knowledge, Twitter our communication demand, and Google captures the search patterns of billions of people. In the following, these three particular media channels will be described in more detail before they are utilized to discover trending topics and analyzed for their various temporal and topical characteristics in Chapter 4.

2.1.2 *Wikipedia*

Wikipedia is a collaboratively edited, multilingual, free Internet encyclopedia featuring over 25 million articles in 285 languages (over 4.1 million articles in the

English Wikipedia alone)[8]. This makes Wikipedia arguably the most prominent example of user-generated content on the web. Everyone with access to the site can edit its articles that have been created by about 100,000 active contributors. Until today, about 1.8 billion page edits[9] have been made. Wikipedia is ranked sixth globally among all websites on Alexa[10] indicating the number of people that regularly use the encyclopedia. Wikipedia is supported by the non-profit Wikimedia Foundation which also operates other collaborative wiki projects including Wiktionary[11] (multi-language dictionary and thesaurus), Wikiquote[12] (encyclopedia of quotations), or Wikibooks[13] (collection of free content textbooks and annotated texts).

Today, Wikipedia plays a key role for online users with information needs. To give just one example from politics, The Washington Post [89] reported on the importance of Wikipedia in the 2008 U.S. election campaign: "Type a candidate's name into Google, and among the first results is a Wikipedia page, making those entries arguably as important as any ad in defining a candidate. Already, the presidential entries are being edited, dissected and debated countless times each day." This further illustrates that people's interest in specific Wikipedia articles is correlated with real-world events. Another prominent example is Michael Jackson's death on June 25, 2009. Due to the increased traffic and the bombardment of requests to update the information, the corresponding Wikipedia page even had to be brought down. As conflicting news reports were released, users and fans began an editing war on his page trying to get the latest and correct information up [73].

As people tend to look up information regarding people, events, and general topics of current interest on Wikipedia, this channel captures the information demand of people. This satisfies a common need of people using the web and is the reason Wikipedia was included in the context of this work. Trending topics can form on Wikipedia if many people access certain articles that are usually not viewed that frequently (thereby satisfying the requirement of spiking user interest). Wikipedia also has the great advantage that page view statistics for

[8] Retrieved from `http://en.wikipedia.org/wiki/Wikipedia` on April 1, 2013.

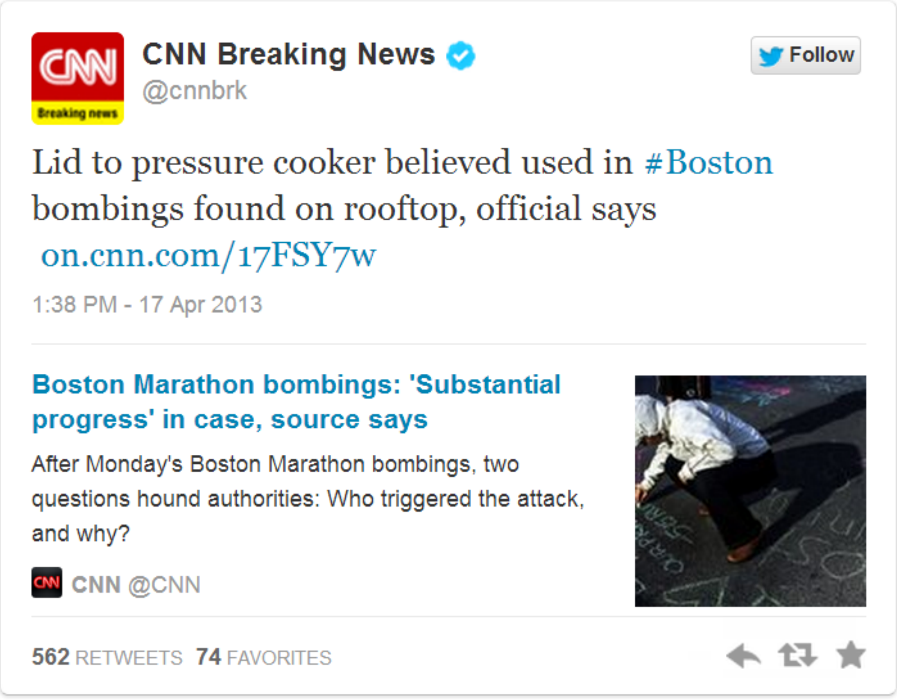
[9] Retrieved from `https://toolserver.org/~emijrp/wikimediacounter/` on April 1, 2013.

[10] Retrieved from `http://www.alexa.com/topsites` on April 1, 2013.

[11] Retrieved from `http://www.wiktionary.org/` on April 17, 2013.

[12] Retrieved from `http://www.wikiquote.org/` on April 17, 2013.

[13] Retrieved from `http://www.wikibooks.org/` on April 17, 2013.



CNN Breaking News
@cnnbrk

Lid to pressure cooker believed used in #Boston bombings found on rooftop, official says on.cnn.com/17FSY7w

1:38 PM - 17 Apr 2013

Boston Marathon bombings: 'Substantial progress' in case, source says

After Monday's Boston Marathon bombings, two questions hound authorities: Who triggered the attack, and why?

CNN @CNN

562 RETWEETS 74 FAVORITES

Figure 8: An example tweet by CNN Breaking News reporting on the Boston Marathon bombings in April 2013. Retrieved from `https://twitter.com/NewsFromWebSite` on April 17, 2013.

all articles over the last five years are public. This distinguishes Wikipedia from many commercial platforms such as Twitter, Google, or Facebook.

2.1.3 *Twitter*

Twitter is an online social networking and microblogging service created in March 2006. It enables its users to send and read short text messages of up to 140 characters called "tweets". Twitter's service has rapidly gained worldwide popularity with over 500 million registered users as of 2012 generating over 340 million tweets daily and handling over 1.6 billion search queries per day [57, 84, 88]. During the night of Super Bowl XLVII a record 24.1 million tweets about the event were posted [65].

The tweets posted on Twitter are public to anyone but only registered users are able to post new ones. An example tweet by CNN Breaking News (Account @cnnbrk) reporting on the Boston Marathon Bombings from April 15, 2013 is shown in Figure 8. The timestamp below the message shows that the tweet was posted on April 17, 2013. Further, the tweet employs a hashtag to indicate the Tweet's topic #Boston. Using such hashtags, words or phrases prefixed with a "#" sign, is a very common way to assign a topic label to a tweet. Hashtags are the main way of aggregating or searching for tweets of a specific topic. The example tweet also uses a shortened URL to link to richer media content such as a news article or video. Tweets can be shared with one's followers through Twitter's retweeting mechanism as well as added to one's favorites. The tweet in the example was already retweeted over 500 times. Twitter also allows to add a location to the tweet as well as a picture (not used in example).

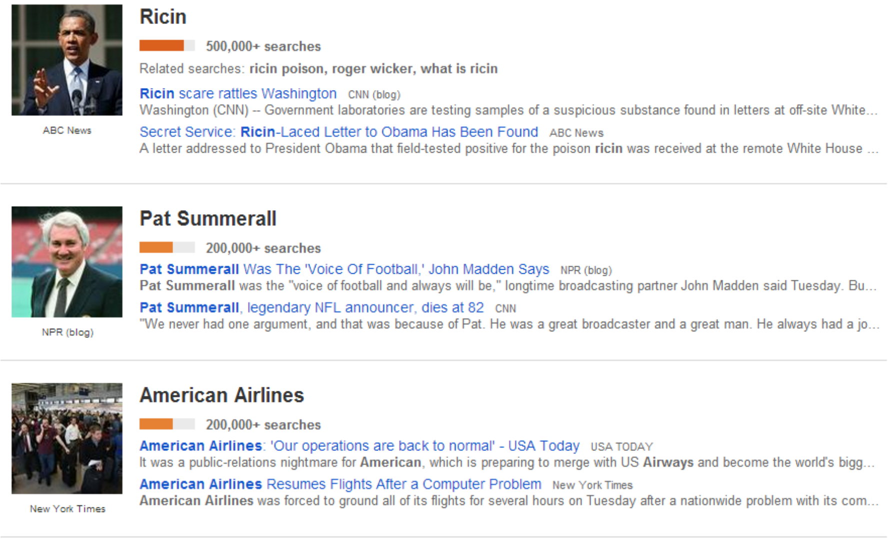
Twitter is often viewed as a social network in which follower relationships are represented by directed edges connecting different Twitter accounts. Some individuals are followed by as many as 37 million people as in the case of Justin Bieber [87]. The highest ranked account representing a politician is Barack Obama with about 29 million followers as of April 2013.

Trends on Twitter occur whenever certain hashtags start to be used frequently, i.e. receive a spike in user attention. This happens if many people tag their own tweets with such hashtags or if tweets containing the hashtag are retweeted many times. Twitter ranks those trends and allows the user to view recent Tweets on these topics. Twitter was included in this thesis to capture the communication demand of people who by tweeting explicitly and intentionally express to their followers that they care about a certain issue or read a specific article.

2.1.4 *Google*

Google Inc. is a corporation specializing in Internet-related services and products including search, cloud computing, software and online advertising technologies. It is most well known for its web search engine, the most visited website in the world according to Alexa[14] and the most dominant search engine in the United States market with a market share of 65.6% [56]. The corporation has been estimated to process over one billion search requests [54] and about twenty petabytes of user-generated data each day [80]. To provide fresh search results, Google downloads the web continuously by processing at least hundreds of

[14] Retrieved from `http://www.alexa.com/topsites` on April 5, 2013.



Ricin

500,000+ searches

Related searches: **ricin poison, roger wicker, what is ricin**

Ricin scare rattles Washington CNN (blog)

Washington (CNN) -- Government laboratories are testing samples of a suspicious substance found in letters at off-site White...

Secret Service: **Ricin**-Laced Letter to Obama Has Been Found ABC News

A letter addressed to President Obama that field-tested positive for the poison **ricin** was received at the remote White House ...

ABC News

Pat Summerall

200,000+ searches

Pat Summerall Was The 'Voice Of Football,' John Madden Says NPR (blog)

Pat Summerall was the "voice of football and always will be," longtime broadcasting partner John Madden said Tuesday. Bu...

Pat Summerall, legendary NFL announcer, dies at 82 CNN

"We never had one argument, and that was because of Pat. He was a great broadcaster and a great man. He always had a jo...

NPR (blog)

American Airlines

200,000+ searches

American Airlines: 'Our operations are back to normal' - USA Today USA TODAY

It was a public-relations nightmare for **American**, which is preparing to merge with US **Airways** and become the world's bigg...

American Airlines Resumes Flights After a Computer Problem New York Times

American Airlines was forced to ground all of its flights for several hours on Tuesday after a nationwide problem with its com...

New York Times

Figure 9: Google Trends shows currently popular search queries entered into their search engine along with related search terms and news stories. Retrieved from `http://www.google.com/trends/hottrends` on April 17, 2013.

thousands of pages per second collecting updated page information and re-processing the entire web-link graph several times per day [39]. The last published estimate (2008) of the size of the web was 1 trillion unique URLs [4]. Google further publishes[15] how often a particular term is entered into their search engine relative to the total search volume across various regions and languages. This product, called Google Trends, can be used to analyze past and current trends in the world that capture rising information demands by large groups of individuals. Note that this product, unlike the present work, does not support forecasting of this kind of user behavior for most terms. An example for trends from Google's "hot trends" category is given in Figure 9. Newly popular search queries are displayed along with related terms and news articles. Note that such trends are being searched for hundreds of thousands of times on a single day. Search patterns on such a large scale are likely to reflect the interests of the general population. Therefore, we include trends from Google in the trending topic analysis presented in this thesis.

[15] Retrieved from `http://www.google.com/trends/` on April 18, 2013.

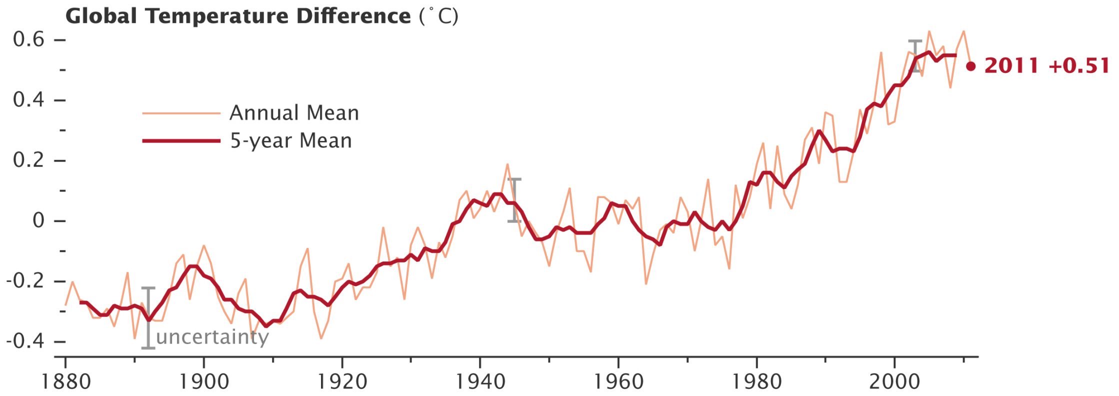


Figure 10: Change in global surface temperature compared to the average global temperature from the mid-20th century released by NASA's Goddard Institute for Space Studies (GISS). Figure taken from [58].

2.2 TIME SERIES FORECASTING

In this section, we will introduce the concept of a time series along with standard modeling approaches and challenges for time series forecasting.

2.2.1 *Time Series*

A time series is a sequence of data points, measured typically at successive points in time, spaced at uniform time intervals. Time series are used in various fields such as statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, and earthquake prediction. Much data can be thought of as time series, such as the daily closing value of a stock market index, the monthly mean temperature, or the number of daily visitors to a website. Example time series are given in Figure 10 and Figure 11 that depict the change in global surface temperature and the Google stock price, respectively.

A common notation specifying a time series X that is indexed by the natural numbers is:

$$X = X_1, X_2, \dots, X_i, \dots$$

where $i \in \mathbb{N}$ and typically $X_i \in \mathbb{R}$ or $X_i \in \mathbb{N}$.

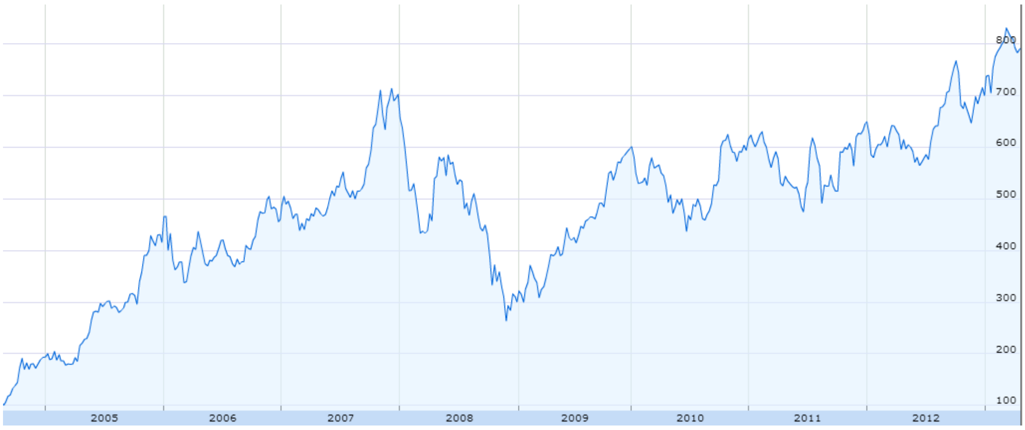


Figure 11: Google (GOOG) stock price over the last nine years. Figure retrieved from `https://www.google.com/finance?q=GOOG` on April 16, 2013.

2.2.2 *Forecasting*

Forecasting is the process of making statements about events whose actual outcomes (typically) have not yet been observed. In the context of time series, this means predicting future values of the time series, i.e. predicting X_t for $t > t_0$ at time t_0. Usually, previously observed values are used to model the behavior of the time series mathematically. Then, this model is used to generate predictions for future points in time. For the two examples given in the previous section this would mean to forecast future values of Google's stock price or future differences in global surface temperature. Both weather and finance are standard examples where forecasting methods are employed. However, while weather forecasts often predict the weather days in advance stock forecasting may operate on a microsecond time scale to make profits on today's stock markets. In the following, basic and more advanced models will be introduced that are commonly used for time series forecasting.

2.2.3 *Basic Forecasting Models*

Very simple models have been applied to forecast time series. A common baseline (e.g. [71]) and arguably the most simple model is the "naive" forecast that simply repeats the last observed value X_{t-1}:

$$X_t = X_{t-1}$$

This can be a reasonable forecast when the subject of interest rarely changes. For example, given the sunny weather in California it is reasonable to assume that today's weather equals yesterday's weather in this particular state. Since it rarely rains this naive forecasts will already lead to a reasonable accuracy.

Another common model is to fit a linear trend to the values observed over the last few periods. Formally,

$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0), \quad t > t_0,$$

where we estimate the slope of our linear trend from the last observed value X_{t_0} and the value d periods before that X_{t_0-d} (common values are $d = 7$ or $d = 24$ representing seven days in a week or twenty-four hours in a day).

We will compare against both of these baselines in the evaluation of our proposed approach (Chapter 6). Example forecasts for the two basic forecasting models are shown in Figure 12. Note that both basic models are unable to capture the complex behavior of the time series such as its seasonality. A standard time series model capable of this is the autoregressive models (with several extensions) that will be introduced below.

2.2.4 *Autoregressive Models*

Regression analysis is a statistical technique for estimating the relationships among variables. Over the years, many different regression models have been proposed such as linear regression, logistic regression [64], or LASSO regression [86]. Applying regression analysis to time series forecasting estimates the relationship between the current value of the time series and its own previous values. Therefore, such methods are called *auto*regressive models and are frequently used

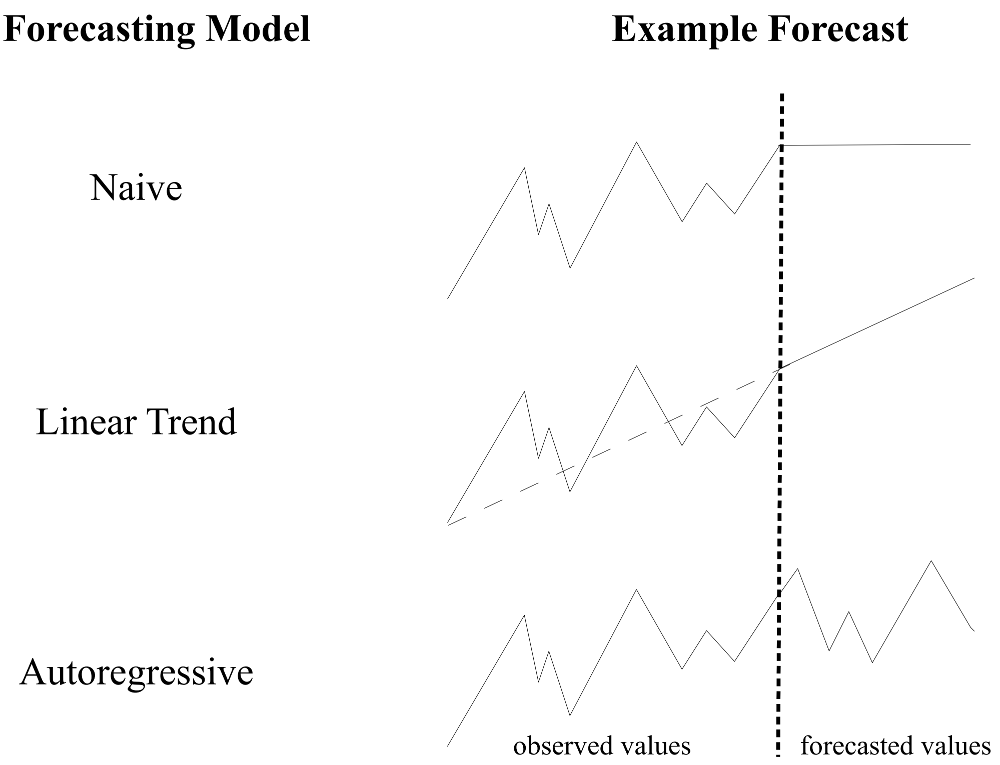


Figure 12: Example forecasts for the naive, linear trend, and autoregressive models. Note that the basic models cannot capture the complex behavior of the time series.

to forecast time series in economics, econometrics, and finance. The autoregressive (AR) model was popularized by Box and Jenkins [16] and specifies that the output variable depends linearly on its own previous values. Formally, the AR(p) model is defined as

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, c is a constant, and ε_t is white noise (the source of randomness). The latter is assumed to have the following characteristics

$$\begin{aligned} E[\varepsilon_t] &= 0\,, \\ E[\varepsilon_t^2] &= \sigma^2\,, \\ E[\varepsilon_t \varepsilon_s] &= 0 \quad \text{for all } t \neq s\,, \end{aligned}$$

that is the error term has zero mean, a fixed variance, and is uncorrelated across time.

The AR model is generalized in the autoregressive-moving-average (ARMA) model that further includes a moving average term over the past error terms ε_t. Formally, the ARMA(p,q) model was first described by Peter Whittle [93] in 1951 before it was popularized by Box and Jenkins who presented an iterative (Box-Jenkins) method for choosing and estimating its parameters [16].

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} ,$$

where $\theta_1, \dots, \theta_q$ are additional parameters of the moving-average part of the model. Conceptually, a moving-average model is a linear regression of the current value of the series against current and previous (unobserved) white noise error terms. Fitting these MA estimates is more complicated than with autoregressive models because the lagged error terms are not observed. Therefore, iterative non-linear fitting procedures need to be used instead of for example linear least squares.

Many more extensions of this model have been proposed such as autoregressive-integrated-moving-average (ARIMA) model dealing with non-stationary time series (loosely speaking time series for which mean and variance change over time), seasonal ARIMA (SARIMA) model including seasonal effects, ARIMA with exogenous factors (ARIMAX) which model external influences (e.g. "dummy" variables indicating weekends or national holidays [43, 29]), and vector ARIMA (VARIMA) models that generalizes ARIMA models to the multivariate case.

In all these cases, the models parameters have to be estimated to fit the underlying time series. For this, least squares regression or maximum likelihood estimates are often used. For more detail regarding the estimation process please refer to Hyndman and Khandakar [47]. They further describe a series of statistical tests for model selection that can guide a particular choice of the above models. This meta-model is called AutoARIMA and conducts a search over all possible models based on the Akaike information criterion (AIC) and Bayes information criterion (BIC) [47]. It has been shown to provide comparable performance to the best known methods in several forecasting competitions while being fully automatic [48].

Once the model parameters are estimated, the model can be used to forecast an arbitrary number of periods into the future. At some point, variables such as X_t or ε_t are required that are unknown (since future values are unobserved) in which case they are substituted for their expected values. Instead of the actual value for X_t the previously generated forecast for X_t is used and the error term ε_t is set to its expected value of zero (as specified above). Therefore, note that out-of-sample forecasts by a MA(q) model greater than the order of the model (q) become constant.

While these autoregressive models are often used for forecasting in econometrics it is important to note that the introduced ARMA models assume stationary time series, i.e. the underlying distribution of the data must not change over time. This critical assumption is regularly violated by the sudden spikes in user attention when a topic becomes trending (recall Figure 6 on page 10). At these points (sometimes referred to as structural breaks; more detail in Section 2.2.6) these models fail to model and forecast behavioral dynamics on the web. Further, parameter estimation fails if little historic information is available, e.g. for new topics such as "Costa Concordia" that did not attract any attention until they became trending. In the context of this thesis, autoregressive models were found to exhibit (non-trivial) predictive power mostly if similar, and therefore predictive, time series are included as exogenous factors. This fact leads us to consider nearest neighbor forecasting models that more directly exploit information in similar time series and prove to be more stable (since no model parameters are estimated from data). This class of models will be introduced in the following section.

2.2.5 *Nearest Neighbor Forecasting Models*

Nearest neighbor methods have been applied to time series forecasting as a non-parametric counterpart to for example autoregressive models introduced above [32, 33]. They only use information local to the points to be predicted and do not try to fit a function to the whole time series at once (like AR models) [33]. Formally, a distance measure, generically represented by $d(\cdot,\cdot)$, is used to find best matches (nearest neighbors) to a given time series X (more concrete distance measures will be introduced in Section 5.4). Let $X^1, \dots, X^k$ be the k nearest neighbors to X, i.e. the smallest elements with respect to the distance metric $d(\cdot,\cdot)$. The nearest neighbor forecast for X at time t, X_t, then simply becomes

$$X_t = \mathtt{operator}\left(X_t^1, \dots, X_t^k\right),$$

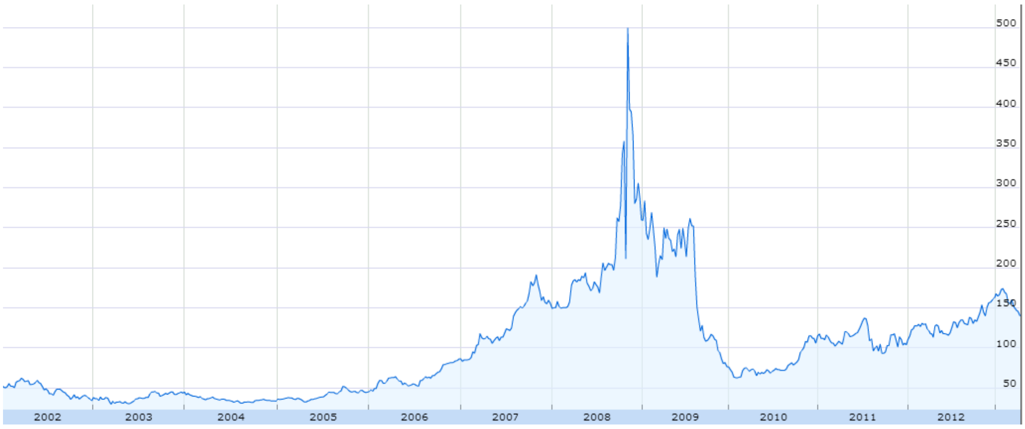


Figure 13: Volkswagen (VOW) stock price featuring a structural break in late 2008 when the stock temporarily tripled its value. Figure retrieved from `https://www.google.com/finance?q=ETR%3AVOW` on April 16, 2013.

where `operator` combines the values of the nearest neighbors at time t. A (weighted) average or the median are common choices of such operators. Note that we use the values of other time series at time t to predict the time series of interest at the same time (X_t) which is the case of nowcasting (as introduced in Chapter 1). In this case, it is impossible to make forecasts for future points in time as the values $X_t^1, \ldots, X_t^k$ will not have been observed, yet. Therefore, for actual forecasts historic values (prior to time t) of the nearest neighbor time series $X^1, \ldots, X^k$ need to be used (example situations will be illustrated in Section 5). In this case, the forecast becomes

$$X_t = \texttt{operator}\left(X_{t_1}^1, \ldots, X_{t_k}^k\right),$$

where $t_1, \ldots, t_k \leqslant t$ makes sure that historic values are used. Starting from time $t_0 = \max\{t_1, \ldots, t_k\}$ such a forecast will be possible.

2.2.6 *Forecasting in the Presence of Structural Breaks*

Unexpected shifts in time series are known as structural breaks (also described as parameter non-constancy) in the field of econometrics and can lead to large forecasting errors and unreliable forecasting models. Clements and Hendry [21] find that in practice, structural breaks are the main cause of large, unexpected forecast errors. One classic example for structural breaks is the Volkswagen stock price which temporarily tripled its value in October 2008 [37]. The corresponding time series is depicted in Figure 13 which shows the break in behavior in fall of 2008.

Recall that the autoregressive models introduced in Section 2.2.4 assume stationarity of the data which means that the mean and variance must not change over time. Obviously, this assumption is violated whenever a structural break occurs. Therefore, autoregressive models can be unstable and lead to large forecasting errors in such cases.

To address forecasting in the presence of structural breaks, pooling or combining forecasts from different models has often been found to outperform individual forecasts and to increase forecast robustness. For a more comprehensive overview on forecast combination in econometrics please refer to chapter 2.3 of [21]. Further, nearest neighbor techniques (see Section 2.2.5) for forecasting have been found to improve out-of-sample forecasts in exchange rate series that exhibit non-linear behavior [32]. Explicitly detecting break points in time series as well as incorporating this information in model specification have been studied extensively in econometrics literature, for example by using piecewise autoregressive (AR) processes [24].

3 RELATED WORK

The overview over related work is divided into four parts. We first refer to efforts on large-scale topic and event detection in online media streams. We further report work that analyzes or combines information from multiple media channels. Then, we focus on previous efforts on forecasting behavioral dynamics before finally presenting applications of trending topics in social multimedia systems.

3.1 TOPIC AND EVENT DETECTION

Detecting and tracking topics in news media has been studied since 1997 with the goal of understanding news in multiple languages and across multiple media channels [3]. In the past, this work largely focused on traditional media sources such as television an radio. Therefore, initial challenges included the segmentation of audio-visual news streams into individual stories, the identification of new topics in these streams, tracking all stories that discuss a particular topic, and identifying the first story that mentions a new topic [20].

Recently, topic and event detection has regained momentum because of the advent of social media which provides a wealth of information created and shared by billions of people online. Particularly Twitter has recently gained much attention in the area of online event detection. While event detection has long been a research topic, the characteristics of Twitter, such as the massive amount of tweets every minute, make it a non-trivial task [91]. Weng and Lee address this challenge by performing a wavelet analysis on the frequency-based word signals to detect new events and further cluster terms via a graph partitioning technique [91]. Twitter has also been used to detect earthquakes including their spatial localization in [76].

In contrast to these efforts, our study uses a diverse set of (textual) news streams that often already aggregate trends and news stories to a certain degree (such as Google News/Trends). Therefore, instead of "finding a needle in a haystack" by means of large-scale online event detection we identify common topics across channels and over time and present a multi-channel analysis of trending topics over an observation period of one year.

3.2 MULTI-CHANNEL ANALYSES

Ratkiewicz et al. [72] claim that while the dynamics of short-lived events such as the news cycle are relatively well understood, the popularity of a Wikipedia page or topic may be influenced by many news events over an indefinite time span. Therefore, the behavior of online popularity cannot in general be characterized by that of individual news-driven events. They illustrate this by considering the difference between the news story "Barack Obama inaugurated as U.S. President," and the Wikipedia article on "Barack Obama". The latter's popularity subsumes that of the former, and of potentially many other news stories. They further found that "bursts of traffic in Wikipedia often correspond to topics that have attracted sudden bursts of attention in the Web at large as measured by Google search volume [72]."

Wikipedia article views have also been correlated with behavior on Twitter to analyze Wikipedia's effectiveness of creating new content for breaking news (e.g. Japan earthquake) or updating pages at the moment of news (e.g. Oscar winners) [92] and to filter spurious events on Twitter [66]. An indirect result of the latter study is that Wikipedia lags behind Twitter by about two hours as measured by the textual similarity between tweets and "bursting" Wikipedia articles.

Adar et al. [1] correlate several behavioral datasets such as general queries from an observation period of one month from MSN and AOL search logs with the general idea of understanding past behavior to predict future behavior. They cluster search queries if they lead to very similar search results (using a standard web search engine) and propose a novel visual tool to analyze the temporal correlation within these clusters as well as differences in popularity and time delays. Note that Adar et al., in contrast to this thesis, does not explicitly focus on trending topics (as opposed to for example navigational queries for websites such as "facebook" or "bank of america") [1]. Further, our study also analyzes the correlation between topic categories and media channels and proposes a fully automated approach to predict future behavior.

The work of Yang and Leskovec [94] explores patterns of temporal variation in online media by clustering them to analyze the rise and fall of human attention with regard to certain phrases or memes with a similarity metric that is invariant to scaling and shifting. They report six distinct shapes of time series and analyzed which channels tend to mention the phrases first: Professional blogs, newspapers, news agencies, TV stations, or blog aggregators. Yang and Leskovec found that weblogs trailed main stream media by one or two hours for most of the considered phrases. They further predicted the cluster assignment (representing the pattern of temporal variation) based on which websites mentioned the phrase.

In contrast to the above, we explicitly focus our study on trending topics in online and social media channels over a long observation period. Similar to [25], we utilize trending topics published by the media channels themselves and focus on multi-channel analysis and forecasting. We further analyze the correlation between topic categories and media channels and propose a fully automatic approach of forecasting trending topics in terms of time series (as opposed to visualization tools [1] or predicting cluster assignment [94]).

3.3 FORECASTING BEHAVIORAL DYNAMICS

As people increasingly turn to the Internet for news and information, online activity can be seen as a snapshot of the collective consciousness reflecting current interests, concerns, and intentions of the global population [5, 40, 36]. This leads to the observation that what people are searching for today is indicative of their current and future situations and interests. Therefore, trends in online media have been utilized to forecast "external" variables of the real-world such as sales statistics or other economic indicators. Particularly search query volume has been correlated with such variables to "predict the present" (sometimes referred to as nowcasting as opposed to forecasting). Such nowcasting can be of interest when official reports about the variable of interest are private, delayed or costly. For example, Choi and Varian correlated search data with economic activity to predict current automobile sales, unemployment claims, travel destination planning, and consumer confidence [18, 19]. Much work has further been devoted to estimate influenza infections and cancer developments as people tend to search for specific diseases and related keywords when they get infected [30, 67, 46, 35, 22]. Goel et al. [36] extended this line of work to predict future consumer activity such as box office revenue for feature films, video game sales, and the Billboard song rankings by looking at how many people searched for the respective movies, games, or songs online weeks before their releases.

Popularity of online content has often been treated as a single variable (e.g., total number of views) instead of a time series. To forecast eventual popularity such as the total number of views either early popularity [83], or content-based features such as publisher, subjectivity, or occurrence of named entities [8] have been used. A different approach is taken by Radinsky et al. [70] who predict the top terms that will prominently appear in future news (such as prediction "oil" after observing "dollar drop"). More recent work treats popularity of queries and clicked URLs of search engines as time series and uses state space models adapted from physics for forecasting [71].

The only work on forecasting Wikipedia article popularity known to the author of this thesis restricts itself to the case of featured articles on the main page and accounts for daily and weekly cycles in viewing behavior [85]. These featured articles all exhibit very similar behavior during the time of main page exposure (particularly after outliers are explicitly removed prior to both modeling and evaluation).

Forecasting trending topics is different in the sense that we require a fully automatic system and that often there is little historical information available (unlike for many economic variables of interest). For example, few people were aware of "Costa Concordia" before the ship sunk in January, 2012. Similarly, to make predictions about "Emmy Awards 2012" one needs to understand the relationship of this event with previous ones such as earlier instances of the Emmy Awards (also illustrated in Figure 25 of Chapter 5). We assume that semantically similar topics share characteristics and behavior and therefore could improve forecasting accuracy. This assumption has not yet been explored in previous work (e.g., [71] which models single time series). In addition, our proposed approach can forecast popularity of arbitrary articles that exhibit very diverse viewing dynamics. This differs dramatically from the experimental conditions in [85] which were characterized by the authors as "nearly ideal". The diverse viewing dynamics include structural breaks, unexpected shifts in the corresponding time series, that were introduced in Section 2.2.6. One example for structural breaks in the context of trending topics was given in Figure 6 on page 10 that illustrates a 2000-fold increase in Wikipedia article views following Whitney Houston's death.

3.4 SOCIAL MULTIMEDIA APPLICATIONS

A first application using social media in the domain of social multimedia systems has been explored in [52]. This work uses the Flickr photo upload volume of specific topics to predict real-world behavior. Specifically, the photo upload volume is used as exogenous input to inform autoregressive nowcasting models (i.e., the model requires the Flickr upload volume at time t to produce a forecast for political election results at time t). This model is evaluated in a case study for a few monthly political election results and product sales. Further, the Flickr queries relevant to the forecast subject of interest are chosen manually (e.g., using "Hillary" instead of "Clinton" to avoid images by Bill Clinton for the 2008 Democratic Party presidential primaries).

Another application that explicitly focuses on trending topics identified in different online and social media channels uses them to dynamically form and extend the concept vocabulary of video concept detection systems [15]. They further find trending topics to be strongly correlated with changes in upload volume on YouTube and demonstrate a visual classification of these trends.

SocialTransfer is another system that uses trending topics obtained from a stream of Twitter posts for social trend aware video recommendation [74]. Learning new associations between videos based on current trends was found to be important to improve the performance of multimedia applications, such as video recommendation in terms of topical relevance and popularity. The popularity of videos is also used in [90] to drive the allocation of replication capacity to serve social video contents. This work analyzes real-world video propagation patterns and finds that newly generated and shared videos are the ones that tend to attract the most attention (called temporal locality). They further formulate the challenge to estimate the videos' popularities for video service allocation for which they use the number of microblog post that share or re-share the video. These insights are then incorporated into the design of a propagation-based social-aware replication framework.

Two other research prototypes that seek to enhance the multimedia consumption experience by extracting trending topics and events from user behavior on the Web are SocialSensor [26] and TrendMiner [77]. The former particularly emphasizes the real-time aspects of multimedia indexing and search over multiple social networks for the purpose of social recommendations and retrieval. TrendMiner focuses on real-time methods for cross-lingual mining and summarization of large-scale stream media and use cases in financial decision support and political analysis and monitoring.

The work presented in this theses differs from the related work described in various ways. In contrast to the work by Jin et al. [52], we present a fully automatic system that produces daily forecasts for example 14 days in advance. Further, the proposed system automatically identifies semantically similar topics and uses their historical popularity to inform forecasting. These forecasts can be then be utilized in various ways in social multimedia systems such as determining which concept models to train [15], which videos to replicate across a content delivery network [90], or which videos to show to a user based on future high-profile trends [74].

4 MULTI-CHANNEL ANALYSIS

This section introduces the dataset used for the analysis of trending topics across media channels and presents an analysis of their temporal characteristics as well as insights into the relationship between channels and topic categories.

4.1 METHODOLOGY

To facilitate a comprehensive analysis of trending topics across multiple channels, the top trends from three major online media channels were crawled on a daily basis. With the intent to capture people's communication needs, search patterns, and information demand, trends are retrieved from Twitter, Google, and Wikipedia, respectively (Section 4.1.1). Further, similar trends are clustered across time *and* channels (Section 4.1.2) to later be able to compare the different manifestations of a particular trend across multiple channels (Section 4.1.3). A conceptual overview of this process is also shown in Figure 14. Please note that the pipeline and data presented in [15] were reused for the purpose of the analysis in this chapter. After exploring expert and automatic descriptions of trending topics (Section 4.2), the lifetime of trends in the different channels is analyzed (Section 4.3) as well as determined whether trends start, peak, or end faster in a given channel than in others (Section 4.4). Lastly, the category distribution of online trends is characterized and it is investigated whether channels tend to particularly emphasize certain content (Section 4.5).

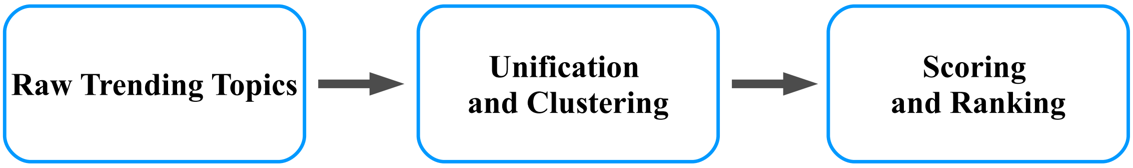


Figure 14: A conceptual overview of the process used to discover unique trending topics from noisy trends observed in online and social media channels.

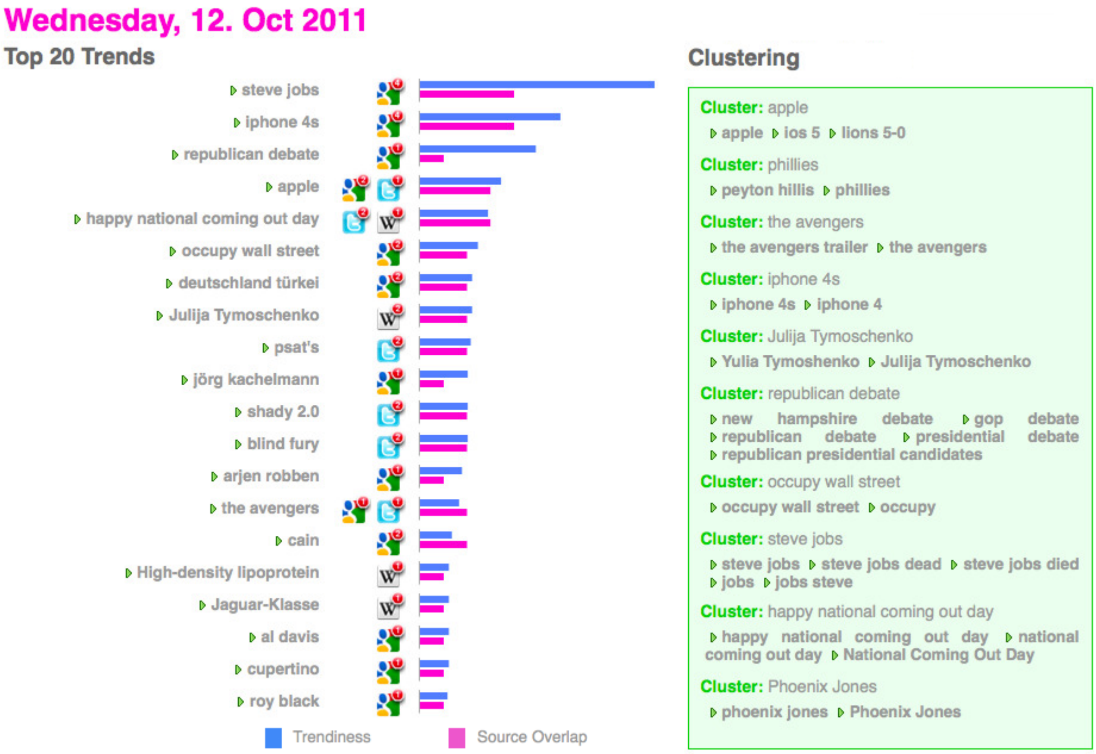
4.1.1 *Raw Trending Topic Sources*

As motivated above, top trends from Google, Twitter and Wikipedia are used as a starting point by retrieving ranked lists of popular terms from 10 different sources on a daily basis: five Google channels[1] (Search and News for USA and Germany as well as the Trends feed), three Twitter channels[2] (daily trends for USA and Germany as well as the Daily Trends stream), and two Wikipedia channels[3] (popular articles in the English or German language). For each of these feeds we retrieve 10-20 ranked topics resulting in 110 topics per day. In total, the dataset covers the observation period Sep. 2011 - Sep. 2012 and contains roughly 40,000 potentially overlapping topics. Example topics are shown in Figure 15 which illustrates all three steps. On the given day, people talked about and searched for Steve Jobs's death through "steve jobs", "steve jobs dead", "steve jobs died", "jobs", and "jobs steve" (see the "steve jobs" cluster on the right side). Further note that the raw trends also include different spellings such as in the case of "Yulia Tymoshenko" and "Julija Tymoschenko".

4.1.2 *Unification and Clustering of Trends*

To connect multiple instances of the same trending topic across time and media channels, a unification and clustering of the individual topics obtained in the previous step is performed. First, the individual topic strings are mapped to a corresponding Wikipedia URI by selecting the top-most Wikipedia result when performing a Google search for that topic (found to be more robust than more

[1] Retrieved from `http://www.google.com/trends/` on April 16, 2013.
[2] Retrieved from `https://dev.twitter.com/` on April 16, 2013.
[3] Retrieved from `http://dumps.wikimedia.org/other/pagecounts-raw/` on April 16, 2013.



Clustering

Cluster: apple
apple · ios 5 · lions 5-0

Cluster: phillies
peyton hillis · phillies

Cluster: the avengers
the avengers trailer · the avengers

Cluster: iphone 4s
iphone 4s · iphone 4

Cluster: Julija Tymoschenko
Yulia Tymoshenko · Julija Tymoschenko

Cluster: republican debate
new hampshire debate · gop debate · republican debate · presidential debate · republican presidential candidates

Cluster: occupy wall street
occupy wall street · occupy

Cluster: steve jobs
steve jobs · steve jobs dead · steve jobs died · jobs · jobs steve

Cluster: happy national coming out day
happy national coming out day · national coming out day · National Coming Out Day

Cluster: Phoenix Jones
phoenix jones · Phoenix Jones

Figure 15: Discovered trending topics obtained from raw trends observed on October 12, 2011 (shortly after Steve Job's death). The left side illustrates the ranking by trend score.

direct methods on Wikipedia). This unifies topics with different spellings or paraphrases. Then, two URIs are clustered together if their Levenshtein distance is below a certain threshold (set to 0.35 × word length). This process can be thought of as a greedy version of a hierarchical agglomerative clustering using the Levenshtein distance. The method allows to unify topics such as "steve jobs", "steve jobs dead", "steve jobs died", "jobs", and "jobs steve" into a single cluster (see Figure 15). Overall this results in 2986 clusters or individual trends. A cluster is now represented by its most prominent member and will be referred to as a *trending topic* for the rest of this work. Another resulting cluster for the "Olympics 2012" was depicted as a word cloud in Figure 2 on page 4 (with larger font size indicating higher term frequency). Note that while the automatic unification and clustering perform quite well occasionally there are errors such as the "apple" cluster at the top right in Figure 15 that also includes "lions 5-0" which refers to the Detroit Lions football team that started their 2011 season with a 5-0 record (their best since 1956). This error occurred because of the syntactic similarity of "lions 5-0" to "ios 5" (referring to the operating system by Apple).

4.1.3 *Ranking Trending Topics*

To reason about the popularity of these trending topics, they are assigned trend scores based on the following method: For each day and for each of our 10 feeds, the rank at which a topic appears is recorded. These ranks are combined using Borda count, obtaining a score for each day that is assigned to the topic's cluster (from the previous step). These daily scores are also illustrated in Figure 15 by the blue bar ("Trendiness") which essentially represents how important the trend was in its respective channels. A different way to judge the relevancy of a trend is "source overlap" (pink bar) which counts the number of channels the trend was featured in. This differentiates trends that have more global impact and are featured in nearly all channels (e.g., "steve jobs") from trends of smaller reach (e.g., "arjen robben").

To measure the impact of a trending topic (cluster) over its overall lifetime, we define its global trend score by the sum of its daily scores over the entire observation period. The top trends with respect to this global trend score are shown in Table 1. Please refer to the Appendix A on page 101 for the full list of the top 200 trends that were discovered from September 2011 until September 2012. The top 10 trends along with their trend scores were also shown in Figure 1 on page 2. Different patterns of user engagement could be observed such as the two peaks for "Olympics 2012", the gradual increase and steep fall for "Black Friday deals", and the recurring pattern of "Champions League".

4.2 CORRESPONDENCE TO REAL-WORLD EVENTS

The discovered trending topics show a clear correspondence to topics and events of significant interest in the real-world. More specifically, the trending topics in Table 1 shows sport events, product releases, (death of) celebrities, holidays, and

	Topic		Topic		Topic
1	olympics 2012	11	christmas	21	ufc
2	champions league	12	steve jobs	22	iphone
3	iphone 5	13	manhattan	23	happy new year
4	whitney houston	14	academy awards	24	kindle
5	mega millions numbers	15	formula 1	25	ncaa brackets
6	closer kate middleton	16	justin bieber	26	em 2012
7	facebook	17	joe paterno died	27	amanda knox
8	costa concordia	18	battlefield 3	28	earthquake
9	black friday deals	19	muammar gaddafi dead	29	mayweather vs ortiz
10	superbowl	20	ufc	30	santa tracker

Table 1: International top 30 trending topics during Sep. 2011 – Sep. 2012. There is a wide variety of trending topics including sport events, product releases, celebrity news, incidents, political movements, and entertainment. Please note that the US presidential election was in Nov. 2012 and is therefore not listed here.

incidents. However, a simple textual label for a trending topic might not provide a sufficient description in real-world applications. Particularly, users in marketing, journalism, and finance might be interested in a more detailed description of the trend and what it refers to specifically. In this section, we provide evidence that the trending topics in our dataset match real-world events in a timely manner allowing for manual expert descriptions of the trending topic. Then, we outline an automatic approach to generate such trend description and present promising results that suggest further investigation.

4.2.1 *Expert Descriptions*

To determine whether emerging trending topics correspond to and timely match actual real-world events, example trending topics with a clear temporal locality were chosen and annotated based on expert knowledge. In the following, example results for earthquakes and Champions League soccer matches will be presented.

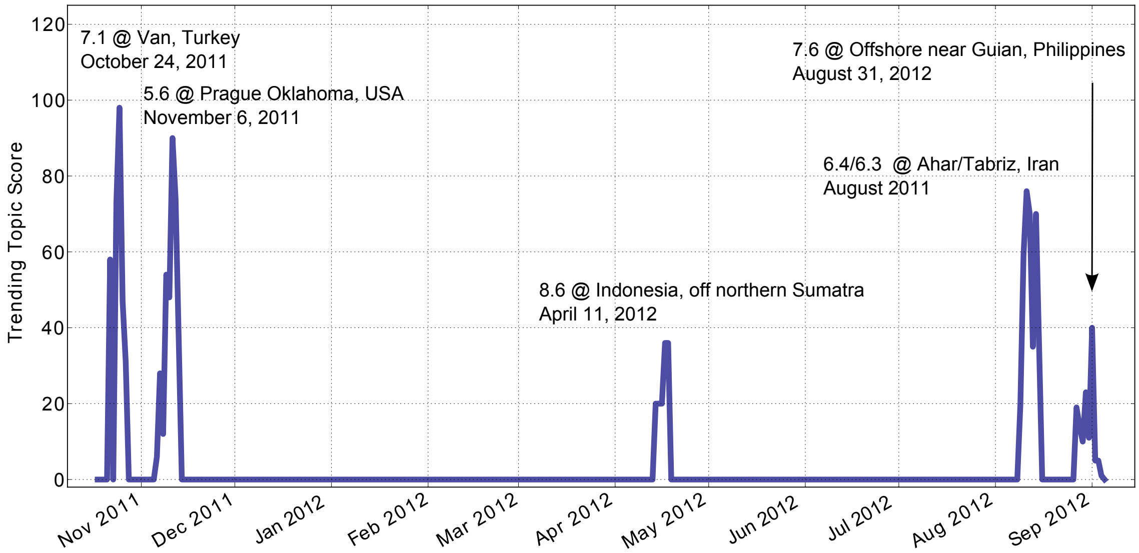


Figure 16: All spikes in attention of the trending topic "earthquake" co-occur with major earthquakes in the world. The corresponding earthquakes are listed with magnitude, location, and date.

All trends corresponding to earthquakes within the one year observation period were annotated using a public log of major earthquakes[4]. This manual investigation is illustrated in Figure 16 which indicates that whenever the trending topic "earthquake" experiences significant spikes in attention major earthquakes hit different locations on the globe. Please note that we report only the most prominent earthquake during these time periods. Of course, other locations were struck by earthquakes as well during these periods. Also, the same location might have experienced one or several aftershocks that are not listed here individually.

A second example is given in Figure 17 which depicts all spikes of the trending topics "Champions League". Please note that all spikes correspond to Champions League games days from the group phase starting in fall until the finals in May[5].

[4] Annotation selected from lists of earthquakes in 2011 and 2012 as published by Wikipedia at `http://en.wikipedia.org/w/index.php?title=Earthquakes_in_2011&oldid=540584243` and `http://en.wikipedia.org/w/index.php?title=Earthquakes_in_2012&oldid=548743916` (retrieved April 7, 2013).

[5] Annotation from `http://www.uefa.com/uefachampionsleague/season=2012/index.html` (retrieved April 19, 2013).

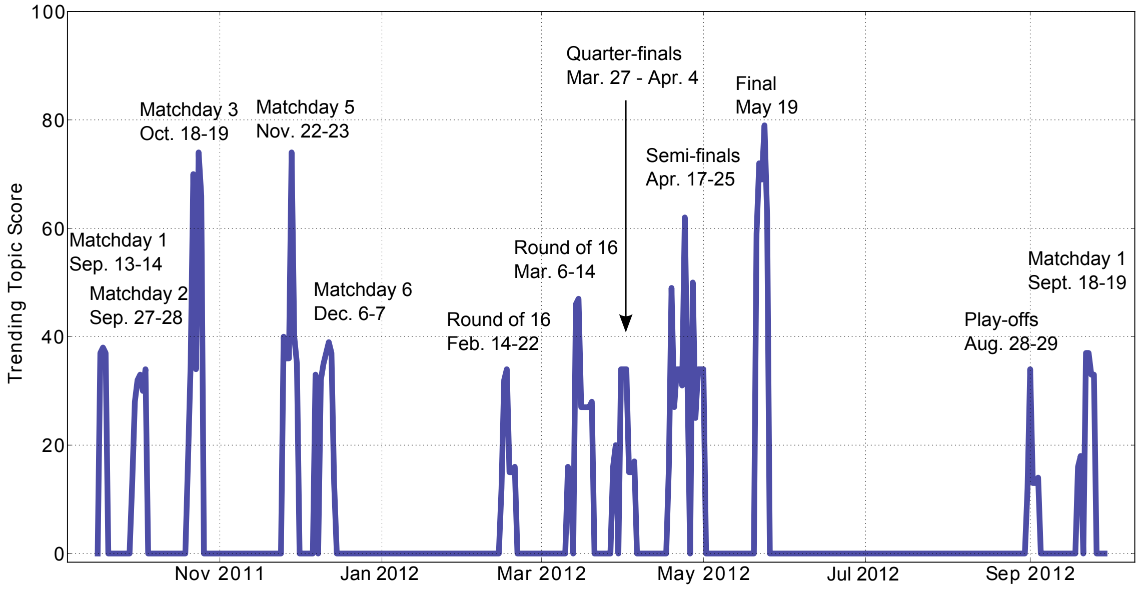


Figure 17: All spikes in attention of the trending topic "Champions League" co-occur with the game days of the European soccer tournament.

4.2.2 *Automatically Generating Descriptions*

The previous section illustrates that trending topics timely match real-world events and that therefore their occurrence can be explained and described manually by experts. However, manual annotation is cumbersome and may even be infeasible for the large number of trends that emerge every day. Therefore, automatic ways of explaining and describing trending topics are investigated in this section. First, given a trending topic, we need to retrieve more background information to "replace" the expert knowledge required for manual annotation, e.g. through explicit searches for related tweets, blog posts, or news articles. This information can then be displayed to the user alongside the trending topic. Note that is clearly suboptimal to list for example the headlines of all retrieved articles. First, one might find hundreds of articles on a given trending topic and second, many of the headlines are specifically designed not to maximize information throughput but to be "eyecatchers" that make you read the rest of the article. One example for this is "To Score Or Not To Score" [14], a New York Times article on the Super Bowl 2012, which was used in the case study below. Clearly, there is a need for scalable summarization of available information around trending topics that capture the interesting facts, dynamics, and changes of discussion.

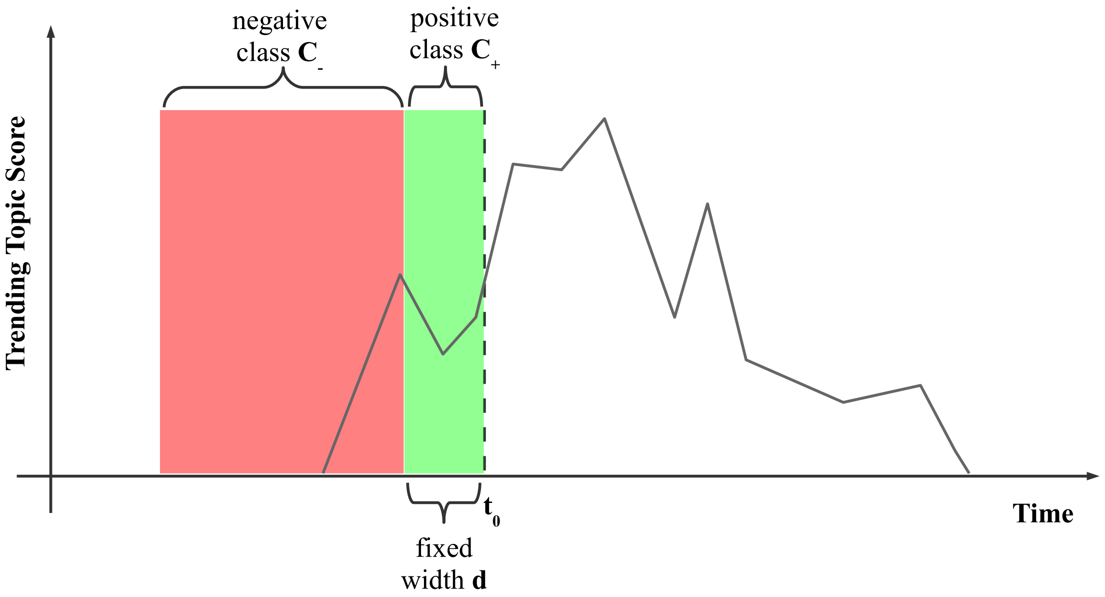


Figure 18: A conceptual overview of the two-class setup differentiating between old news and new news. Example results are given in Table 2 for the trending topics "Super Bowl" and "Joe Paterno".

This can be viewed as a selection problem, i.e. what are the most important and informative pieces among the retrieved background information? This problem is well known in the field of pattern recognition and statistical machine learning as feature selection where one is interested in selecting discriminative or significant features only. Therefore, the retrieved news articles are represented as bags of n-grams with their respective TF-IDF weights [7] and submitted to feature selection using Support Vector Machines (SVM) [23, 79] with L_1 loss [12] enforcing sparsity of selected features is performed.

There are different aspects one could be interested in describing for a given trending topics. First, what distinguishes this trending topic from other trending topics or general news coverage, and second, what distinguishes the current discussion on that trending topic from e.g. last week's discussion (of the same topic). The second aspect is much more interesting since it aims to capture what makes this topic important right now which is more likely to generate a useful description for the user. However, please note that the first aspect can easily be covered by a very similar approach. The conceptual idea to figure out what makes the current discussion about the trending topic special is illustrated in

Figure 18. At time t_0 the description for an emerging trend is generated. The news coverage regarding the given trend up until that point in time is split in two classes: The very recent news, for example of the last three days ($d = 3$), form the positive class C_+ (green), and the news coverage before this period form the negative class C_- (red). Now, a model such as SVM is employed that is able to compute which features, i.e. terms or sequence of terms (n-grams), discriminate between these two classes (C_+ and C_-). Such models typically assign weights to each feature such that the ones with the largest positive weights should represent what the recent news coverage is all about whereas the features with the largest negative weights represent major elements of the previous news coverage that are not discussed any more. In this case, an additional L_1 penalty is used to enforce sparsity on these weight vectors. Therefore, this approach automatically infers what distinguishes recent news from old news.

In the following, two case studies are described for which the discriminative terms are listed in Table 2. For these results the news articles came from The New York Times[6] representing professional news sources. However, experiments were also conducted using the social news source Reddit[7].

Example results of our case study for the trending topic "Super Bowl" are given in Table 2 (a) (based on 109 news articles resulting in 10169 features total). Here, the positive class C_+ consists of articles until the third day following the Super Bowl (February 5-8, 2012) whereas the negative class C_- contains all Super Bowl articles from the 30 days before the event. While headlines often conceal relevant information, this approach succeeds in automatically identifying major actors such as Steve Tisch (owner of the New York Giants), Ahmad Bradshaw (Super Bowl-winning touchdown), Derek Jeter (athlete compared to the Giants quarterback Eli Manning), Mario Manningham (Giants wide receiver whose catch is considered the key play of the game), and Tom Brady (New England Patriots quarterback) as well as related events such as Madonna's half time show performance, the popular Super Bowl ad by Skechers, and an unpopular one by General Electric. Related topics that are no longer part of the discussion include the regular season weeks, the playoffs leading up to the Super Bowl, the two teams that lost in the semi-finals against the Giants and Patriots (San Francisco 49ers and Baltimore Ravens), and Tim Tebow, who until a loss against the Patriots was hyped in the media.

[6] Retrieved from `http://www.nytimes.com/` on April 19, 2013.

[7] Retrieved from `http://www.reddit.com/` on April 19, 2013.

	Top Positive	Score		Top Negative	Score
1	electric	8.40	1	week	-8.17
2	madonna	7.29	2	playoff	-2.19
3	bradshaw	7.12	3	sports	-2.03
4	derek	6.80	4	49ers	-1.46
5	manningham	5.96	5	ravens	-1.18
6	felt	5.69	6	fumble	-0.62
7	skechers	5.36	7	coaching	-0.61
8	fared	5.34	8	tebow	-0.36
9	brady	4.55	9	defense	-0.17

(a) Discriminative terms for the trending topic "Super Bowl".

	Top Positive	Score		Top Negative	Score
1	died	35.49	1	new	-3.31
2	death	14.85	2	alabama	-2.76
3	points	9.70	3	brien	-2.25
4	department	6.78	4	houston	-1.38
5	life	4.98	5	athletic director	-1.21
6	free	2.99	6	family said	-0.70
7	twitter	2.88	7	treatments	-0.35
8	center	2.29			

(b) Discriminative terms for the trending topic "Joe Paterno".

Table 2: Top positive and negative terms discriminating between recent news and previous news coverage. The feature selection approach successfully identifies important parts of the event and news coverage. Please refer to Section 4.2.2 for more detail.

A second case study was conducted for the trending topic "Joe Paterno" who died on January 22, 2012 of complications of lung cancer. Similarly to the previous case study, the positive class consists of all related New York Times articles until three days after his death and the negative class consists of the thirty day period before that. The results are presented in Table 2 (b) based on 60 news articles from the three days following his death (resulting in 20589 features total). The proposed approach again succeeds in identifying major events, topics, and actors.

Joe Paterno, the former Penn State football head coach, died (by far the largest feature weight) at the Mount Nittany Medical Center and the following public memorial service was held at the Bryce Jordan Center (“center”). His donations helped to save Penn State’s classic department and he was one of the targets of the child sexual abuse scandal investigations by the U.S. Department of Education within the universities athletic department (“department”). Many articles further reflected on his life and accomplishments (“life”). Interestingly, Paterno’s death was prematurely reported on Twitter sparking a chain of erroneous reports [61] (“twitter”). Among the topics that were no longer part of the news coverage are Bill O’Brien (the football head coach hired after Paterno), Alabama (since the former Alabama coach Paul Bryant also died shortly after retiring), the athletic director Tim Curly who also was allegedly involved in the cover-up of the child molestation scandal, and reports that Paterno received chemotherapy treatments for his lung cancer. We believe these to be promising results that suggest further investigation as part of future work.

4.3 LIFETIME ANALYSIS ACROSS CHANNELS

Some trending topics such as “Champions League” occur multiple times within our one year observation period. To allow for an analysis of when channels begin and stop featuring particular topics, trending topic is divided into multiple sequences such that its daily trend score is non-zero for at least two out of three adjacent dates, i.e. “score gaps” of at most one day are compensated. An example how trending topics are split into multiple sequences is given in Figure 19. An analysis of the lifetime of trending topics in the different channels is performed for the top 200 trends (based on their global trend score) that were split into 516 sequences based on the described procedure (please refer to the Appendix A on page 101 for a ranked list of the top 200 trends). Table 3 summarizes the resulting number of topics and sequences for the different channels and their combinations. Note that the ten individual observed channels are mapped to their respective sources: Google, Twitter, and Wikipedia. The Google channel has the largest coverage of the trending topic sequences in our dataset (86.2%). About 9.3% of trending topic sequences occur in all three channels and between 11.0% and 33.7% occur in two channels. Of course, these numbers depend on how well

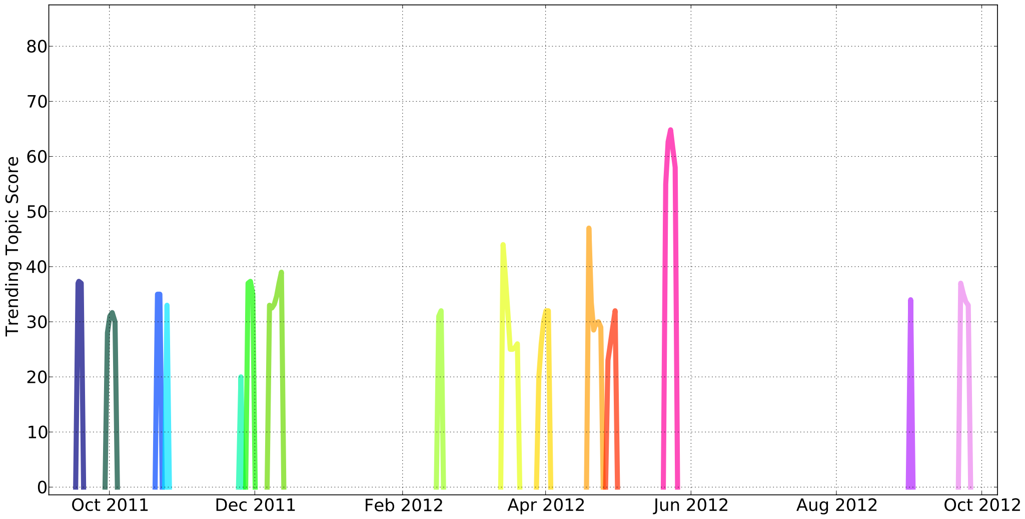


Figure 19: The trending topic "Champions League" is split into several trending topic sequences (encoded by color) for the analysis of temporal characteristics.

the clustering of trends (see Section 4.1.2) works. However, manual analysis of the results showed that the resulting clustering is meaningful in almost all cases (recall the example in Figure 15 on page 39). For brevity, trending topic sequences will simply be called trending topics for the remainder of this chapter.

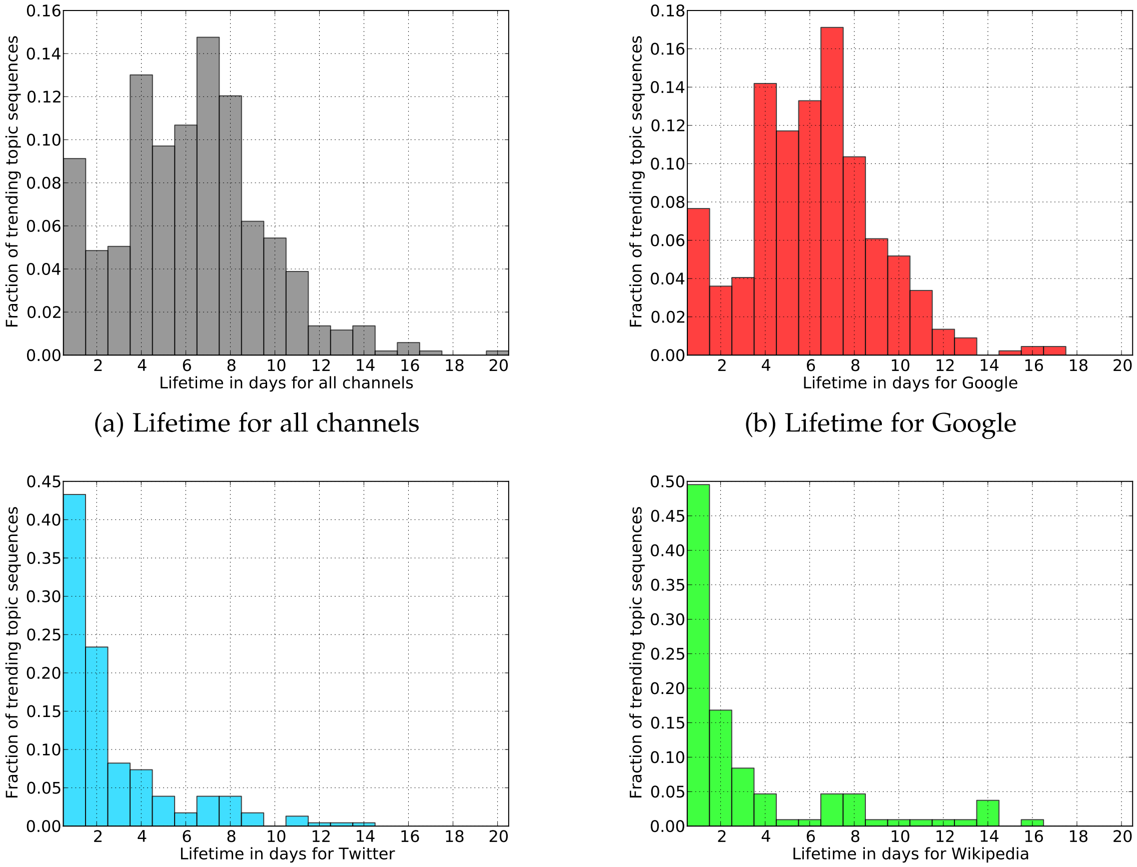
	#Topics	#Sequences
Total	200	516
Google	191	445
Twitter	118	232
Wikipedia	69	108
G & T	115	174
G & W	66	86
T & W	43	57
G & T & W	42	48

Table 3: Number of analyzed trending topics and sequences for Google (G), Twitter (T), and Wikipedia (W). A trending topic can subsume multiple sequences if it represents a reoccurring event (such as Champions League). The Google channel has the largest coverage of the trending topic sequences in our dataset (86.2%). About 9.3% of trending topic sequences occur in all three channels and between 11.0% and 33.7% occur in two channels.

We first deal with the question how long trending topics survive in general and whether there are differences in lifetime in the different media channels. Here, lifetime is defined as the number of consecutive days with positive trend scores. Histograms of the lifetime of trending topics are shown in Figure 20. It can be observed that the trending topics in the dataset rarely survive for more than fourteen days with most trending topics having a lifetime of less than nine days. Since Google covers a large share of top trends, the distribution for the channel looks very similar to the overall distribution. The lifetime of topics on Twitter is much shorter complying our expectations of the ephemerality of trends in this channel - about two thirds of the top trends only survive for one or two days. Interestingly, the distribution looks similar for Wikipedia. This poses the question when and why exactly people are turning to Wikipedia to satisfy their information needs. Our results suggest that Wikipedia article viewing is very event-driven rather than being a static encyclopedia that is occasionally used to look up certain topics.

4.4 DELAY ANALYSIS ACROSS CHANNELS

For trending topics that occur in at least two of the three channels we record the day on which the trend starts, the day on which the trend peaks (through the trend score defined earlier), and the day on which the trend ends again. We then define the *delay* between two channels as the difference between the start/peak/end dates in channel X and the start/peak/end days in channel Y. Conceptually, this is also illustrated in Figure 21. Note that a positive delay means that the first channel X is slower, i.e. trends tend to start, peak, or end later in channel X than in channel Y. The mean delays for start, peak, and end are summarized in Table 4 and the full distribution over delays is shown in Figure 22. Interestingly, there are only marginal differences in starting delays (first of the three numbers) between the three channels with Twitter and Wikipedia being slightly faster than Google. These results are not very surprising considering that Osborne et al. found Twitter to be around two hours faster than Wikipedia [66] – a difference almost impossible to observe in our data of daily granularity. A much stronger effect is observed for peak and end delays (second and third number). Both Twitter and Wikipedia tend to peak more than two days before



(a) Lifetime for all channels

(b) Lifetime for Google

(c) Lifetime for Twitter

(d) Lifetime for Wikipedia

Figure 20: Lifetime in days for top 200 trending topics. The overall distribution is dominated by Google. Trends on Twitter and Wikipedia experience similar lifetimes and are much more ephemeral than those on Google.

Google. The picture is even clearer when looking at the end delays where Twitter and Wikipedia lead Google by three and four days respectively. Overall, these results add to the observed ephemerality of the Twitter and Wikipedia.

It might come as a surprise that Google seems to be much slower than Twitter and Wikipedia. Particularly, the end delay of up to four days was not anticipated in that magnitude. To verify these results we manually checked trending topics in this channels and found that some of the Google channels such as Google News seem to provide a summary instead of the current top news, i.e. aggregating the top stories of the last few days instead of only the present day. This can result in an artificial delay for the present analysis since trending topics in such

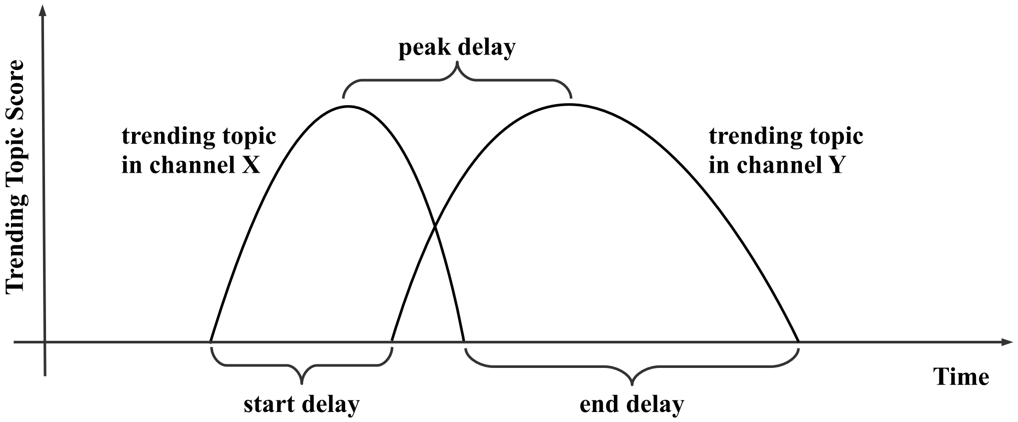
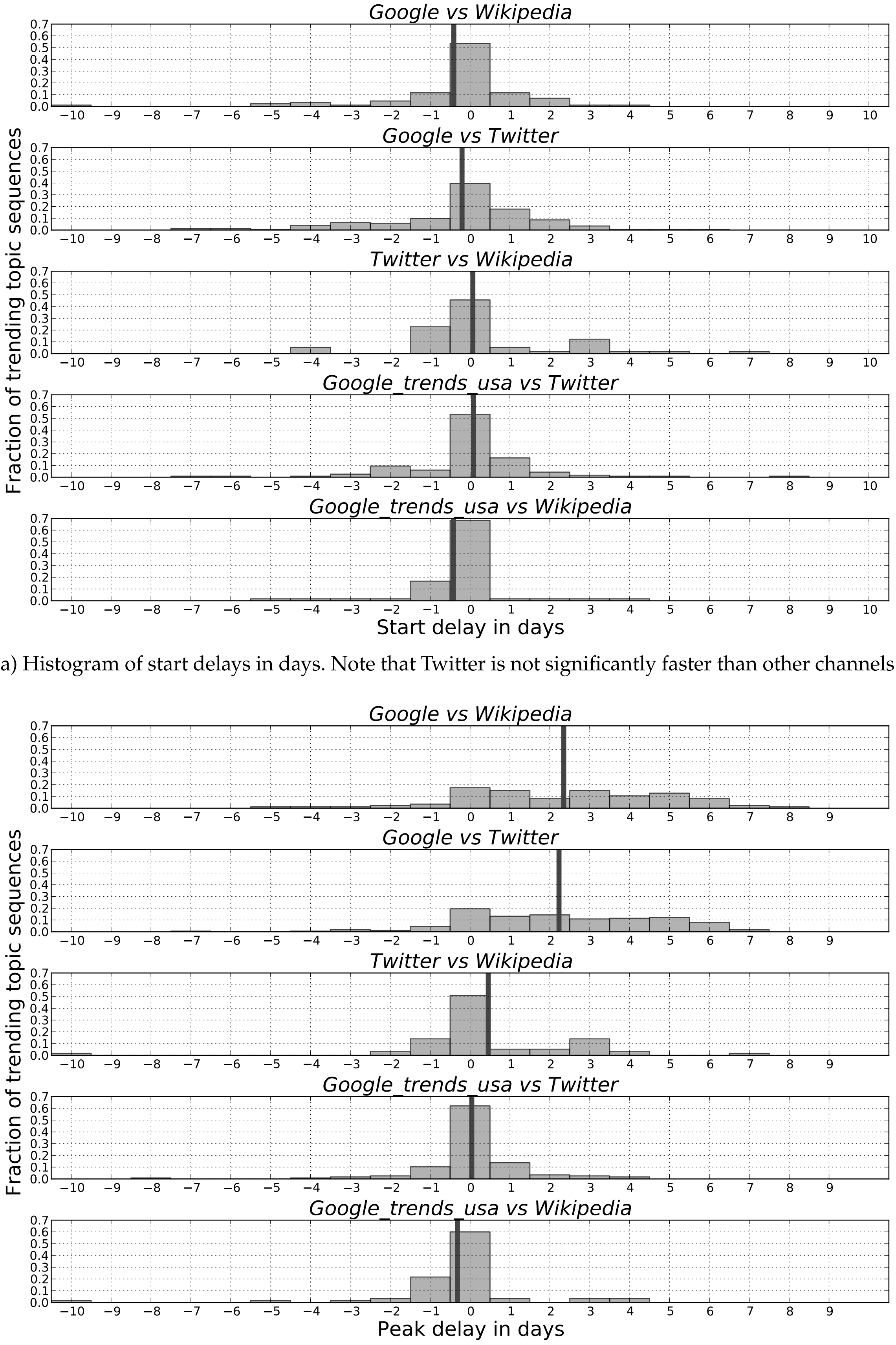


Figure 21: The delay between two channels X and Y is defined as the difference between the start/peak/end dates.

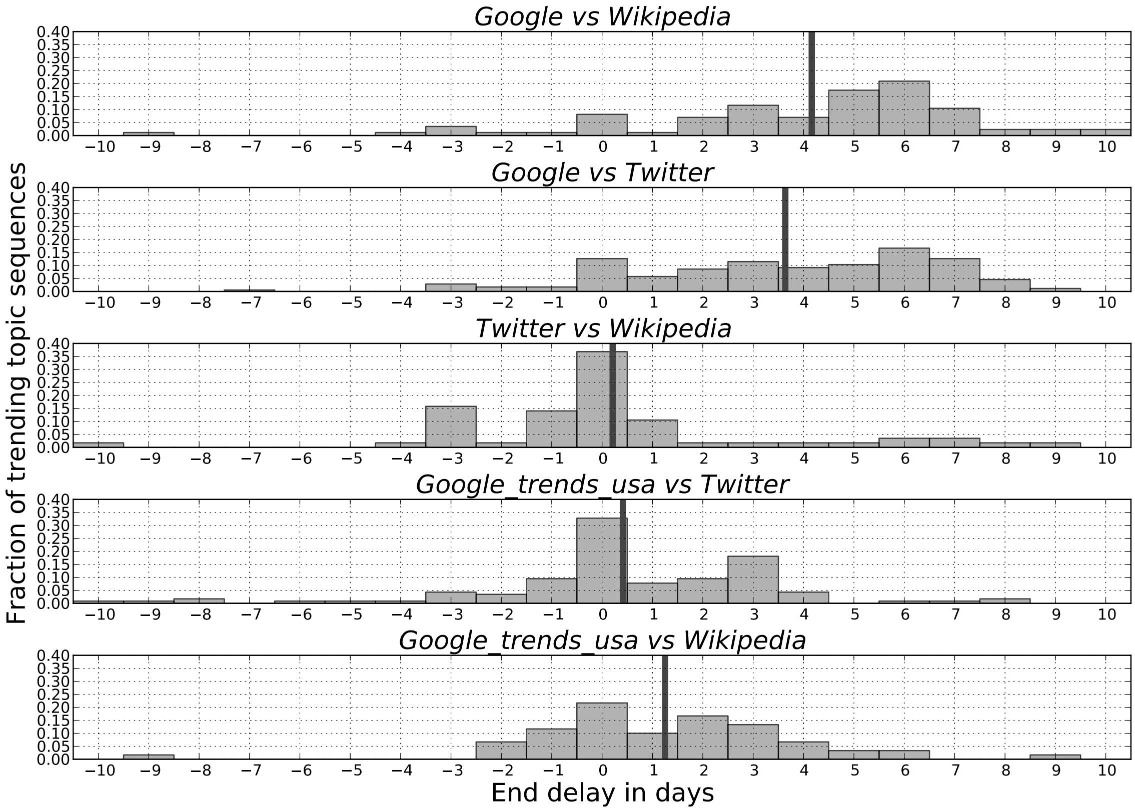
channels appear to live longer. This means that while people stop talking about certain topics on Twitter and stop accessing corresponding Wikipedia articles Google still includes these in their trend and news summaries. It would be very interesting to investigate how many people using these platforms are still interested in them. If this was the case, it would also raise the question why this information demand is not reflected in the Wikipedia access statistics. As such access statistics for Google are not publicly available it will be left to future work to answer this question. From this manual analysis, one of the Google channels, Google Trends, was not expected to be delayed or averaged in any way. To confirm this hypothesis, comparisons of this channel to Twitter and Wikipedia were added to the analysis. This channel peaks around the same time as Twitter and Wikipedia and trends tend to end around half a day after Twitter and more than one day after Wikipedia. This confirms that trending topics within this channel do not seem to be aggregated over several days as its behavior is much closer to ephemeral channels such as Twitter or Wikipedia.



(a) Histogram of start delays in days. Note that Twitter is not significantly faster than other channels.

(b) Histogram of peak delays in days. Note that Google peaks after Twitter and Wikipedia in most cases. However, Google Trends USA behaves similarly to Twitter and Wikipedia.

Figure 22: This figure is continued on the next page.



(c) Histogram of end delays in days. Note that trends in Google survive much longer than in Twitter and Wikipedia but that this behavior is weaker within Google Trends USA.

Figure 22: Histogram of delays in days between different media channels. The darker vertical bars represent the mean of the distributions (also summarized in Table 4). A positive delay means that the first channel is slower, for example the positive mean comparing Google to Wikipedia in the last plot (end delay) means that Google is *slower to end* a trend or equivalently that trends vanish earlier on Wikipedia than on Google.

	Twitter	**Wikipedia**
Google	-0.20 / 2.23 / 3.64	-0.41 / 2.35 / 4.16
Twitter	–	0.07 / 0.36 / 0.21
Google Trends	0.09 / 0.04 / 0.41	-0.42 / -0.32 / 1.25

Table 4: Mean delay in days between pairs of media channels (start/peak/end). Positive delay means that the "row channel" is slower than the "column channel". For example, the value 4.16 comparing Google to Wikipedia (third value/end delay) means that Google is *slower to end* a trend or equivalently that trends vanish about four days earlier on Wikipedia than on Google (see Figure 22 for the full delay distribution).

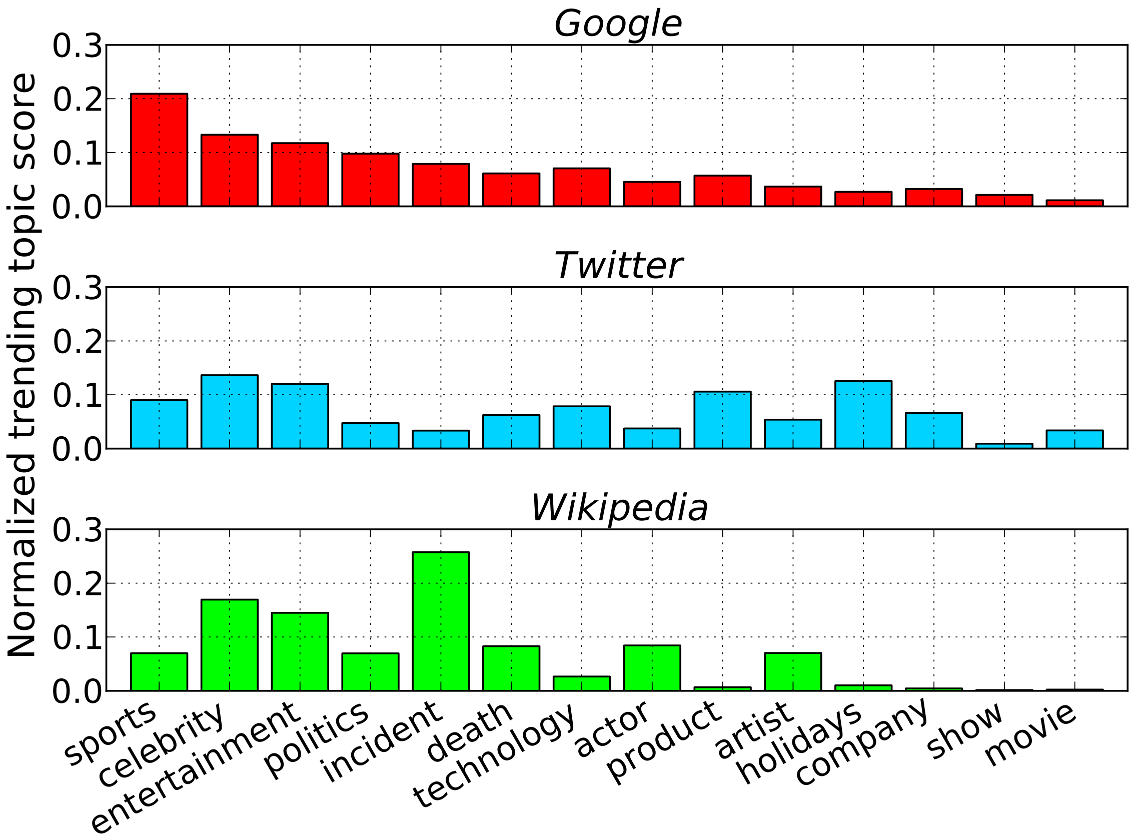


Figure 23: Normalized distribution of trending topic scores over trend categories in the individual channels. Note that channels tend to specialize in certain categories, e.g. Google for sports, Twitter for entertainment, products, technology, holidays, and Wikipedia for incidents, actors and artists.

4.5 TOPIC CATEGORY ANALYSIS ACROSS CHANNELS

Considering the different application domains of trending topics (recall Section 1.2) it is critical to understand why people turn to specific channels and whether these channels tend to emphasize certain kinds of topics. To shed light on what kind of topics are the most popular in the individual channels we manually annotated the 200 top trends with categories. The categories were chosen by examining the main themes of trends we found in our dataset. Please refer to Table 5 for more information about the individual categories, their descriptions, and examples. Note that a trending topic might be assigned to multiple cate-

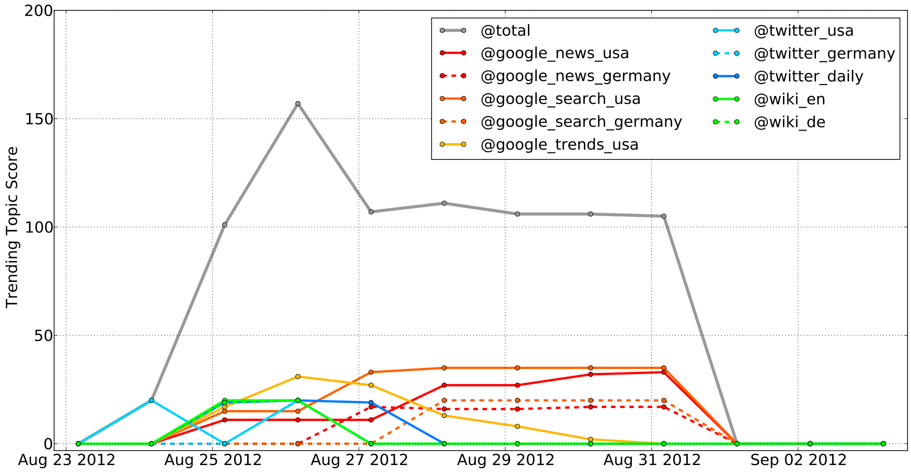


Figure 24: The event of Neil Armstrong's death on August 25, 2012 received much attention. Note that the peak of Wikipedia (green) is at the beginning of the trending topic period whereas news coverage continued for a couple more days. Twitter's first peak is caused by a confusion with Lance Armstrong who was banned from sport competitions on that day.

gories, i.e. the death of a popular music artist would be assigned to celebrity, entertainment, death, and artist to make an extreme example (similar to [1]). The engagement with respect to the different categories in a media channel is measured as follows: For each trend within the channel its score is assigned to all of its categories. Finally, we normalize the scores for each channel, e.g. to account for the dominance of Google for the scores overall. The resulting distribution over categories is displayed in Figure 23.

We can observe that the channels tend to specialize in certain topic categories. For example, the most popular category in Google is sports. A large share of the scores further is assigned to celebrity and entertainment categories. Google also has the highest relative share (15%) for politics. Twitter also features many trends in the celebrity and entertainment categories. Interestingly, it has the highest relative shares of trends related to products, companies and technology. One reason might be that a large fraction of Twitter are technology affine early adopters that like to share their thoughts on new products. Another interesting finding is that over 20% of the scores on Twitter are assigned to the holidays category. One could hypothesize that holiday related trends are big on Twitter because many people tag their posts and pictures with the same hashtags such as #christmas or #thanksgiving. Wikipedia clearly shows a specialization for categories that involve people and incidents such as disasters or the death of celebrities. Contrary

Category	Description	#Seq	Examples
sports	sports events, clubs, athletes	52	olympics 2012, champions league, bayern muenchen
celebrity	person with prominent profile	49	steve jobs, kim kardashian, michael jackson, neil armstrong
entertainment	entertainers, movies, TV shows	39	grammys, emmys, heidi klum
politics	politicans, parties, political events, movements	32	paul ryan, occupy, christian wulff
incident	an individual occurrence or event	27	costa concordia, hurricane isaac, virginia tech shooting
death	death of a celebrity	22	whitney houston, joe paterno died, neil armstrong
technology	product or event related to technology	20	iphone 5, ces, nasa curiosity
actor	actor in TV show or movie	18	lindsay lohan, michael clarke duncan, bill cosby
product	product or product release	15	ipad, windows 8, diablo 3
artist	music artist	15	justin bieber, miley cyrus, beyonce baby
holidays	day(s) of special significance	11	halloween, thanksgiving, valentines day
company	commercial business	10	apple, chick fil a, megaupload
show	TV show	7	x factor, wetten dass, the voice
movie	a motion picture	6	dark knight rises, hunger games, the avengers

Table 5: List of trending topic categories (ranked by their frequency in the data). Note that a trending topic can be assigned to multiple categories, i.e. the death of a popular music artist would be assigned to celebrity, entertainment, death, and artist.

to the intuition that Wikipedia is a slowly evolving channel which people use to read up on complicated topics, especially when also considering the temporal properties of the Wikipedia channel from the analysis above, it is conceivable that many users use Wikipedia during these trends and events to learn about or remind themselves about related topics. For example, when Neil Armstrong died on August 25, 2012 many people viewed the corresponding Wikipedia article e.g. to learn about who he was (American astronaut), what he was famous for (first person on the moon), and how old he was when he died (82) (see Figure 24). Note that this trending topic also mistakenly subsumes the lifetime ban from sport competitions of Lance Armstrong on August 24, 2012. However, the Lance Armstrong topic did not interfere significantly with the Neil Armstrong topic after the latter's death.

5 FORECASTING OF TRENDING TOPICS

As motivated in the introduction, anticipating changing information needs and shifts in user attention toward emerging trends is critical in many different application domains. However, forecasting trending topics is a very challenging problem since the corresponding time series usually exhibit highly irregular behavior (i.e., structural breaks) when the topics become "trending". In this chapter, first, three main classes of behavioral signals are distinguished and their implications for forecasting is discussed. Then, a novel forecasting approach is proposed that combines time series from automatically discovered semantically similar topics exploiting the observation that similar topics behave similarly.

The task in this chapter is forecasting of signals capturing online user behavior and interest in trending topics. Examples for such signals will be given below (Section 5.1) While the proposed approach (Section 5.2) will be able to compute such forecasts at any point in time, we are particularly interested in forecasting the behavior when the topic becomes trending. In contrast to other points in time where the topic receives the "usual amount" of attention, these are the more interesting and complex periods that feature an emerging interest in the given topic.

5.1 CHARACTERISTICS OF BEHAVIORAL SIGNALS

As the forecasting model should generalize to the behavior in multiple online media channels, first a manual analysis of such behavior patterns was conducted. It was found that patterns of people's attention are fairly similar across platforms.

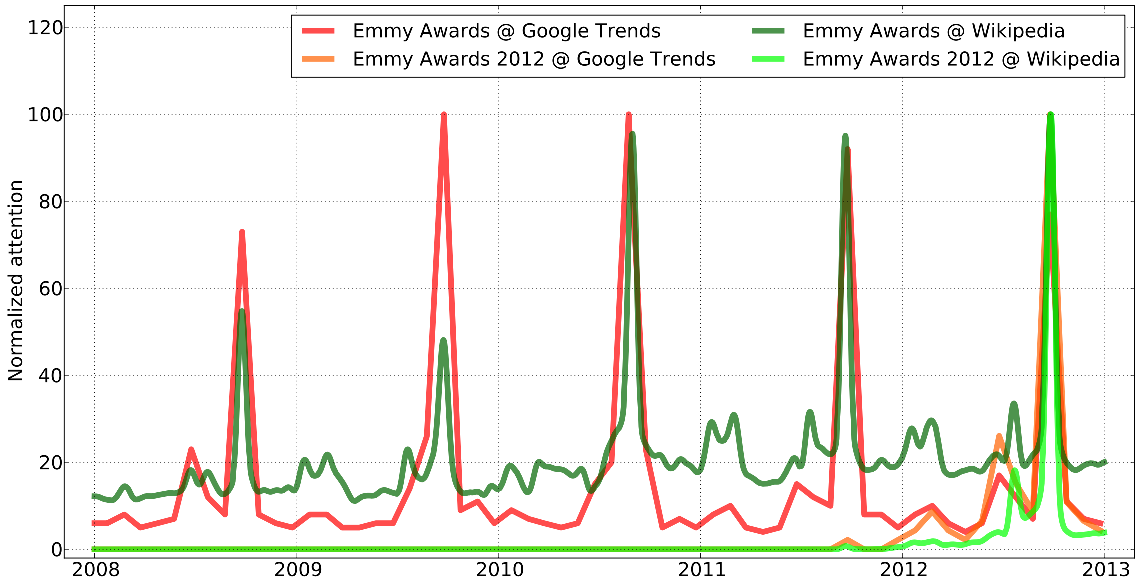
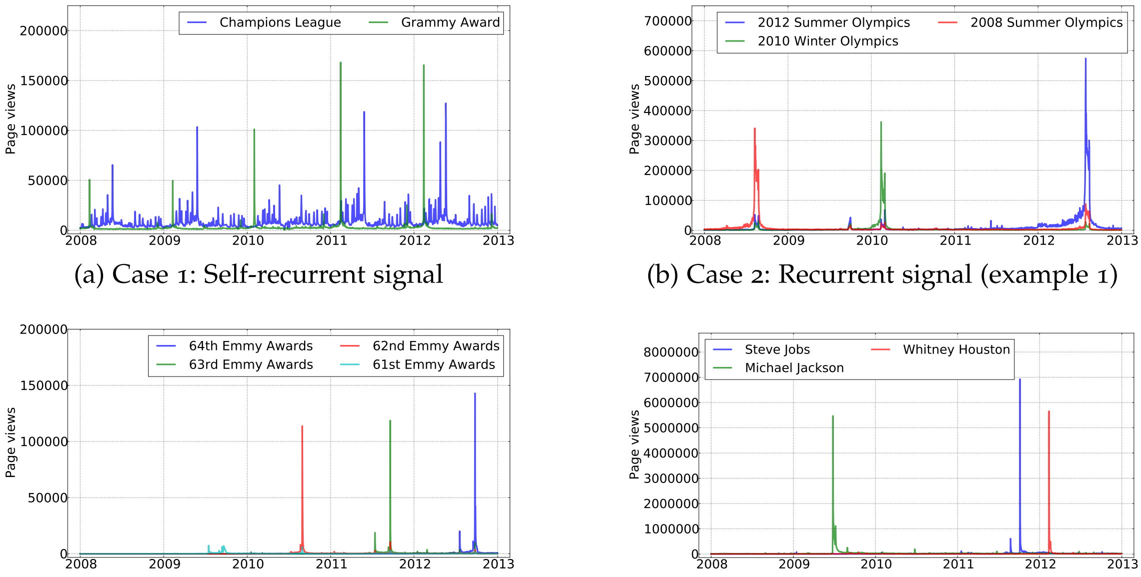


Figure 25: Normalized attention for "Emmy Awards" and "Emmy Awards 2012" on Google (via query volume) and on Wikipedia (via page views). All four signals exhibit similar behavior and share the last big peak in September 2012.

The normalized amount of attention toward the topic "Emmy Awards" and "Emmy Awards 2012" on Google Trends (search volume) and Wikipedia (page views) is shown in Figure 25. The two online media channels share the same high-level behavior, i.e. peaks tend to co-occur and are of roughly the same relative size. These similarities were observed across multiple topic categories and led to the decision to focus on Wikipedia page view data as it offers daily (even hourly) granularity instead of only monthly data that is available for example for Google. This finding is not surprising and supported by previous work which found bursts in attention on Google to correlate well with bursts on Wikipedia [72].

During this analysis, three main classes of signals were identified that have major implications for forecasting. Examples for all three classes are given in Figure 26.

CASE 1 - SELF-RECURRENT In the first case, the signal exhibits recurring patterns within the same signal. We call such signals *self-recurrent*. In the example (Figure 26 (a)), the self-recurring behavior is caused by the Champions League being a yearly soccer competition with very similar schedules each year (i.e., group phase starting in late fall and the finals in May) and the Grammy Award



(a) Case 1: Self-recurrent signal

(b) Case 2: Recurrent signal (example 1)

(c) Case 2: Recurrent signal (example 2)

(d) Case 3: Non-recurrent signal

Figure 26: There are three different classes of behavioral signals with large implications for forecasting: Self-recurrent, recurrent, and non-recurrent signals.

ceremony being a yearly event that people mostly are interested in during that time period.

case 2 - recurrent In the second case, signals do not exhibit recurring patterns themselves but these are contained within other related time series. Therefore, they are referred to as *recurrent* signals as recurrent patterns exist but not within the original signal itself. For example, Figure 26 (b) shows the interest in the 2012 Summer Olympics (blue). However, the corresponding time series never featured patterns as exhibited in the summer of 2012 because people were never very interested in the sports event before it actually happened. However, previous Olympics, the 2010 Winter Olympics and the 2008 Summer Olympics, feature similar behavior (e.g., the two peaks at the start and end of the event). Note that this is not a simple yearly seasonality as the Summer Olympics happen every four years. Furthermore, the Winter Olympics do not happen in Summer but in Winter (as the name already suggests). A second example (Figure 26 (c)) are the “64th Emmy Awards” in 2012. While in the previous case of the Grammy Award the time series itself contained recurrent patterns, now this behavior is spread over multiple time series, i.e. the interest in the Emmy Award 2011 is captured by the "63rd Emmy Award" and so on. Note that the yearly periodicity of the Emmy Awards is only one instance of recurring patterns for which peaks consistently recur after a nearly constant amount of time. For example, this might not necessarily be the case of boxing events which are scheduled individually and do not follow such a rigid pattern. In both examples, there is a clear structure to the recurrence of the event and regularity in naming/counting (e.g., by year or number of instance). However, this problem is generally not trivial as many different rules might apply as for example the Super Bowl using the Roman counting system (e.g., Super Bowl XLVI).

CASE 3 - NON-RECURRENT The third case captures *non-recurrent* signals that do not exhibit recurring behavior and for which there is no obvious related or preceding instance. The examples given in Figure 26 (d) show different celebrity deaths that received much attention from the general population. Obviously, people die only once so any form of self-recurrence is impossible.

IMPLICATIONS FOR FORECASTING The three introduced classes have a large impact on forecasting. The first case is the easiest since a model trained on past behavior should be able to capture the seasonality and structure and do reasonably well on forecasting. Autoregressive-moving-average models as introduced in Section 2.2.4 could be used (if the signals are stationary and do not exhibit structural breaks).

The second case is already much more challenging. In the case of self-recurrent signals, the signal itself can be used to build a model to forecast the future but in the absence of recurring patterns one is not able to do so and it is necessary to include other signals to inform forecasting, e.g. by using signals from a general set of other observed trends. Pushing this idea further, one might even be interested to extend this to a very large set of patterns such as the *entire* Wikipedia corpus. This should enable us to find very related signals such as previous instances of the same event for most trending topics. However, finding a particular pattern in a corpus of this size is clearly non-trivial as there are about nine billion patterns to choose from (assuming five years of daily page views on five million articles). If one is able to align these patterns correctly, they could be used as exogenous inputs in ARMA models. However, the proposed approach of Section 5.2 will provide a more natural way of solving this problem.

In the third case, it is much harder to find such related events as for example a particular person's death only occurs once. However, as the signal itself is usually insufficient to forecast user attention, the only way to proceed is to discover semantically similar instances. For example, the attention pattern toward Whitney Houston's death might look like the attention pattern that was paid to Michael Jackson's death, another very famous music artist. These two instances are semantically similar as both deal with the death of a famous music artist and celebrity. As the examples given in Figure 26 demonstrate semantically similar topics exhibit similar behavior. This observation will be illustrated in more detail in the following section. Therefore, the goal is to identify semantically similar topics exhibiting similar behavior that could be used for forecasting and then use this similar behavior to inform forecasting. This has an additional advantage of filtering the very large set of available patterns prior to more involved signal analysis.

5.2 SYSTEM OVERVIEW

A conceptual overview of our proposed system is presented in Figure 27. The very left shows the trending topic for the Summer Olympics 2012 along with its Wikipedia page view statistics in 2012. The task to be solved is to forecast the number of page views for a period of 14 days (yellow area) from the day indicated by the red line. This forecast start is triggered by the emergence of a corresponding trending topic in our observed channels (see Chapter 4). Note that the time series exhibits complex behavior such as the second, smaller peak at the end of the forecasting period which most likely corresponds to the closing ceremony event on this day. Based on the finding that semantically similar events can exhibit very similar behavior, the first step is to automatically discover related topics such as the previous Summer and Winter Olympics or FIFA/UEFA soccer championships (as illustrated by the second box). The second step in Figure 27 shows patterns of user engagement for the Summer Olympics 2008, the Winter Olympics 2010, and the UEFA Euro championship 2012 that were found to match the historic behavior of the 2012 Olympics best. These patterns show certain commonalities such as a second peak for closing ceremonies or final matches. The identified sequences from the previous step are then combined to a forecast

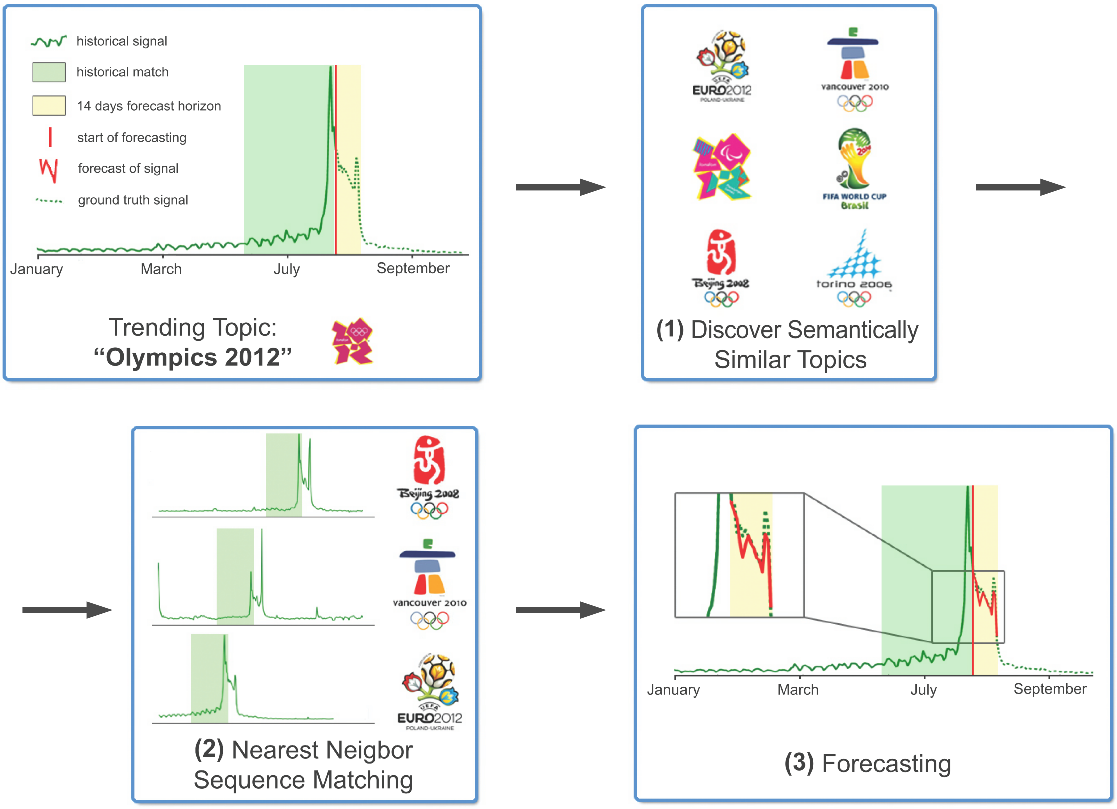


Figure 27: System overview of our proposed forecasting approach. For a given trending topic semantically similar topics are discovered (1) which are searched for patterns of similar behavior (2) which are then used to produce a forecast (3).

shown at the very right. In the following, these individual steps are explained in more detail.

Please note that we are neither able to nor do we attempt to predict incidents such as natural disasters or sudden deaths of celebrities in advance. However, even for unpredictable events like these, the patterns of user attention once this event has happened can be forecasted by taking previous instances of natural disasters or celebrity deaths into account.

5.3 DISCOVERING SEMANTICALLY SIMILAR TOPICS

The basis for our forecasting approach is the hypothesis that semantically similar topics exhibit similar behavior. Examples supporting this hypothesis were given in Figure 26 and Figure 27 which clearly show similar patterns of user engagement during the different events. For a given trending topic, we discover semantically related topics using DBpedia [6], a dataset containing structured information about several million named entities extracted from the Wikipedia project (recall Section 2.1.2). DBpedia provides rich semantic annotation such as category (via `dcterms:subject`) and type information (via `rdf:type`) and constitutes a natural choice as trending topics were already mapped to Wikipedia URIs. In the following, such annotations will be called properties for simplicity.

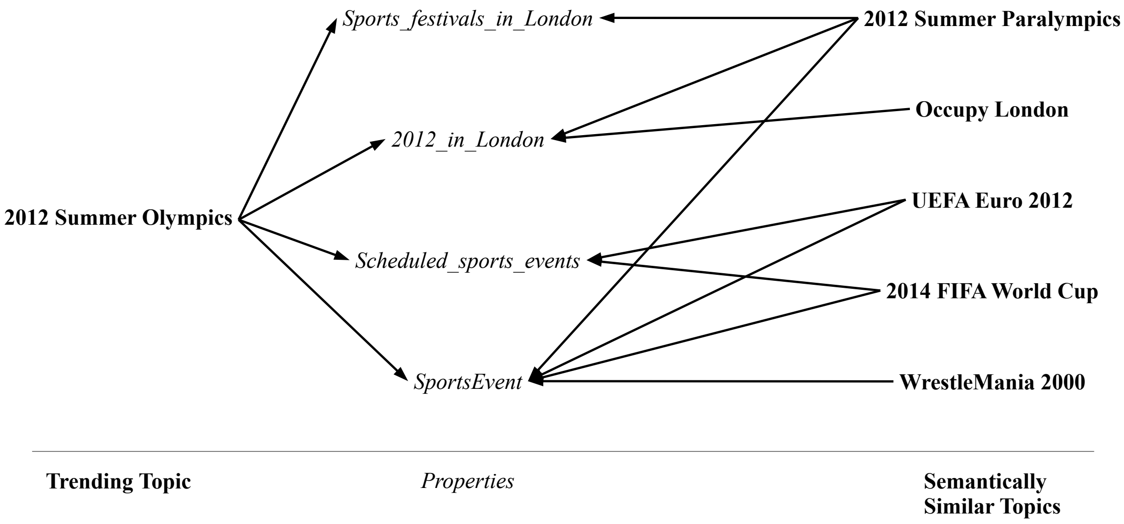


Figure 28: The 2012 Summer Olympics are annotated with certain properties on DBpedia which are linked to by other semantically similar topics as well.

Example properties for the Summer Olympics 2012 are given in Figure 28. The properties includes the type of event (sports), the location (London), and the year (2012). Essentially, topics and properties can be thought of as a large bipartite graph. In this graph, other nodes can link to the same properties as the trending topic Summer Olympics 2012. These nodes will be other topics that are semantically similar to the given trending topic. They can be discovered by traversing the graph and then be ranked by the number of properties that they share with the trending topic (through a SPARQL[1] query). For example, discovered semantically similar topics include the 2012 Summer Olympics (another Sports festival in London in 2012), the UEFA Euro 2012 and the 2014 FIFA World Cup (both scheduled sports events), WrestleMania 2012 (which is just a sports event), and Occupy London (no sports event but happening in London in 2012 as well).

Formally, a topic set $\mathcal{T}_{sim}$ is used which includes all discovered similar topics. For later comparisons, a topic set $\mathcal{T}_{self}$ is defined that only includes the trending topic itself, and $\mathcal{T}_{gen}$ which contains a wide variety of general topics (the top 200 trending topics as listed in Appendix A on page 101). In the following, we will use $\mathcal{T}$ as a placeholder for one of these topic sets. The elements will be referred to as "similar topics" also denoted by $sim_j \in \mathcal{T}$. For reference, the formal notation used in this chapter is summarized in Table 6.

5.4 NEAREST NEIGHBOR SEQUENCE MATCHING

Obviously not all time series corresponding to the discovered similar topics look exactly the same. Therefore, all these time series are searched for sequences that match historical behavior of the trending topic to be forecasted. For example, historical topics such as the 1896 Summer Olympics have gained very limited attention over the last years and are therefore unlikely to be representative for the large amount of engagement toward the 2012 Summer Olympics. The 2008 Summer Olympics would be a much better choice in that regard. To pick the right instances to inform our forecast, a short history window of the trending topic to be forecasted, i.e. the viewing statistics for the last two months, is compared to all partial sequences of the same length of similar topics denoted by sim_j in the topic set $\mathcal{T}$.

To capture this step in more formal terms, let $S_{\text{topic}}[t]$ be the time series for the given topic at time t. Further let us define

$$S^{t_0}_{\text{topic}}[t] := S_{\text{topic}}[t_0 + t]$$

[1] Retrieved from `http://www.w3.org/TR/rdf-sparql-query/` on April 20, 2013.

Notation	Explanation
$\mathcal{T}_{sim}$	Topic set containing all discovered similar topics
$\mathcal{T}_{self}$	Topic set only including the topic itselg
$\mathcal{T}_{gen}$	Topic set of general topics not based on semantic similarity
$\mathcal{T}$	Generic placeholder for a topic set
$\text{sim}_j \in \mathcal{T}$	Similar topic
t_0	Time of forecast
$S_{\text{topic}}[t]$	Time series for topic at time t
$S_{\text{topic}}^{t_0}[t]$	Shifted version of the time series that starts at t_0
$C(t_0)$	Sequence candidate set including all shifted time series
$S_{\text{sim}_j}^{t} \in C(t_0)$	Candidate sequence
$S_{\text{topic}}^{t_0}$	Time series of interest at point of time of forecast t_0
$N^k(S_{\text{topic}}^{t_0})$	Nearest neighbor set for time series of interest
$S_{\text{sim}_i}^{t_i} \in N^k(S_{\text{topic}}^{t_0})$	Nearest neighbor sequence (time series)
$d(\cdot,\cdot)$	Distance metric for time series
$\mathcal{F}(S_{\text{topic}}^{t_0})[\tau]$	Forecast for time series of interest τ days after t_0
$\alpha(\cdot,\cdot)$	Scaling function ensuring a smooth forecast continuation

Table 6: Summary of the formal notation used in this chapter to describe the forecasting approach.

as the shifted version of the time series (used for aligning multiple series below). In the following, a forecast of S_{topic} with a forecast horizon of h days starting at time t_0 is assumed.

Given a topic set $\mathcal{T}$ from the previous step, we define a sequence candidate set $C(t_0)$ that includes all possible shifted time series (denoted by $S_{\text{sim}_j}^{t}$):

$$C(t_0) = \{S_{\text{sim}_j}^{t} \mid \forall \text{sim}_j \in \mathcal{T},\ \forall t :\ t \leqslant t_0 - h\}.$$

The condition $t \leqslant t_0 - h$ ensures that we never use information more recent than $t_0 - h$, to allow for a forecast of h days without utilizing future information.

Given this candidate set, we now search for the k members $S_{\text{sim}_i}^{t_i}$ $(i = 1, \dots, k)$ that are the best matches for our time series of interest ($S_{\text{topic}}^{t_0}$). Note that these nearest neighbors are already optimally aligned to $S_{\text{topic}}^{t_0}$ through shifting the time series S_{sim_i} by a corresponding t_i. Further note that the same similar topic

sim_i can occur multiple times (e.g., for self-recurrent signals), i.e. $\mathrm{sim}_i = \mathrm{sim}_j$ is possible for $i \neq j$.

Formally, the nearest neighbor set becomes

$$N^k(S^{t_0}_{\mathrm{topic}}) = \{S^{t_1}_{\mathrm{sim}_1}, \ldots, S^{t_k}_{\mathrm{sim}_k}\}$$

where the $S^{t_i}_{\mathrm{sim}_i}$ are the k distinct elements that are smallest wrt. $d(S^{t_0}_{\mathrm{topic}}, S^{t_i}_{\mathrm{sim}_i})$ for all $S^{t_i}_{\mathrm{sim}_i} \in C(t_0)$. Therefore, the nearest neighbor set is always a subset of the candidate set $N^k(S^{t_0}_{\mathrm{topic}}) \subset C(t_0)$. Here, $d(\cdot,\cdot)$ is a distance metric between both time series which, in our case, only depends on a short history window of the time series. An interesting question is whether the metric should be scale invariant and in which form and to what degree. In the evaluation (Chapter 6), we compare three different distance metrics:

1. (squared) `euclidean` distance

   $$d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2,$$

2. euclidean distance on normalized sequences (referred to as `musigma`)

   $$x'_i = (x_i - \mu)/\sigma,$$

   where μ, σ are mean and standard deviation estimated from the respective time series, and

3. a fully scale invariant metric that was proposed in [94] called `y_invariant`,

   $$\min_{\gamma} d(x, \gamma \cdot y).$$

5.5 FORECASTING

Even the best matching sequences identified in the previous step might not be a perfect fit for the time series to be forecasted. Therefore, we rescale the matching sequences such that they all agree on the last observed value of $S^{t_0}_{\mathrm{topic}}$. This ensures that the forecast will be a continuous extension of past behavior. Now, the forecast $\mathcal{F}$ becomes the median over scaled versions of the sequences from the previous step ($N^k(S^{t_0}_{topic})$). The days to be forecasted are represented by $\tau \in [0, \dots, h-1]$ where again h is the forecasting horizon (usually $h = 14$ representing two weeks based on the lifetime analysis presented in Section 4.3). Our forecast can then be formally described as

$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^{t_0}_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau]\right), \text{ where} \tag{5.1}$$
$$\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) = S^{t_0}_{topic}[-1](S^{t'}_{topic'}[-1])^{-1}.$$

Function $\alpha(S^{t_0}_{topic}, S^{t'}_{topic'})$ adjusts the scale of nearest neighbor time series based on the last observed score such that the produced forecast is a smooth continuation of the previously observed scores. In practice, α is limited to a limited interval (e.g., $[0.33, 3.0]$) for robustness (to not emphasize possibly random fluctuations too much). The approach is further evaluated using the average instead of the median for forecasts. However, as presented in Chapter 6 the median proved to be more robust.

Summarizing, here is a high-level description of the full approach proposed in the previous sections:

Algorithm Nearest Neighbor Forecasting Using Semantically Similar Topics

Input: Trending topic `topic` and time when to start forecast t_0
Output: Forecast $\mathcal{F}(S^{t_0}_{topic})$

1. Obtain Wikipedia URI corresp. to trending topic `topic` (see Section 4.1).
2. Obtain semantically similar topics (topic set $\mathcal{T}_{sim}$) from DBpedia for Wikipedia URI (see Section 5.3)
3. Retrieve time series data for trending topic and all obtained semantically similar topics.
4. Use short history window before t_0 of the time series corresponding to `topic` to find the k best matching sequences (nearest neighbors) within all time series retrieved in the previous step (see Section 5.4)
5. Combine the k nearest neighbor sequences using Equation 5.1 to compute the forecast (see Section 5.5)

5.6 GENERALIZABILITY OF PROPOSED APPROACH

The main assumption of the proposed approach is the feasibility of measuring semantic similarity between signals (in addition to common signal-based similarity measures used in Section 5.4). Therefore, domain knowledge is required to be able to judge semantic similarity and to identify semantically similar signals within that domain. A related assumption is the availability of such similar signals, i.e. whether historical signal data is available for the identified semantically similar signals. This is necessary because we are essentially "transferring" certain patterns of past behavior of these signals to forecast the original signal.

In this work, we focus on trending topics in online media. Several online media channels satisfy the assumptions listed above including the ones analyzed in Chapter 4: Wikipedia, Twitter, and Google. In Wikipedia, semantic similarity is available through properties of DBpedia entities as described in Section 5.3. Google records the query volume for all search terms and further provides categories for these terms that could be used as a proxy for semantic similarity[2]. Similarly, on Twitter one could infer semantic similarity between tweets, trends, or users through text-based similarity measures (e.g. cosine similarity on TF-IDF scores [7]) or topic modeling approaches (e.g. Latent Dirichlet Allocation [13]). Furthermore, as illustrated in Figure 25 on page 60 and [72], other online media channels such as Google are found to exhibit very similar behavior to Wikipedia which is used an example in the context of this thesis. Therefore, the presented approach should immediately generalize to these channels.

Forecasting in the presence of structural breaks is a complex problem in several domains other than trends in online media. Our proposed method describes a general and systematic approach to forecasting in the presence of structural breaks and could be useful in all of these domains. One example is modeling stock prizes that can exhibit sudden changes in special situations. A classic example for structural breaks is the Volkswagen stock which temporarily tripled its value in October 2008 [37]. The corresponding time series is was already given in Figure 13 on page 28 which clearly shows unusual behavior in late 2008.

[2] See http://www.google.com/trends/ for details. Retrieved on April 10, 2013.

In this domain, semantic similarity between stocks could be based on whether the companies share certain attributes such as industry, market position, and board members, or whether they are or have been in similar situations (e.g., being acquired by another company). Together with the availability of historic stock price data this satisfies the assumptions of our proposed approach.

In the following chapter, we will empirically demonstrate forecasting performance of this approach for trending topics in a specific dataset that is superior to commonly used forecasting approaches introduced in Section 2.2.

6 EVALUATION

In this section, first the Wikipedia page views dataset is described before quantitatively evaluating the individual building blocks of the forecasting approach proposed in the previous section.

6.1 DATASET DESCRIPTION

Our proposed forecasting approach is evaluated on a large-scale dataset of page views on Wikipedia, an online collaborative encyclopedia that has become a mainstream information resource worldwide and is frequently used in academia [72, 92, 66, 85].

Reasons for this particular choice of social media channel were (a) the public availability of historical page views data necessary to build forecasting models (hourly view statistics for the last five years), (b) the size of the dataset allowing a comprehensive analysis of our proposed method across a wide range of topics (over 5 million articles), and (c) previous results that user behavior on Wikipedia (bursts in popularity of Wikipedia pages) is well correlated with external news events [72] and therefore support our proposed method's ability to generalize to different social media channels. Also recall Figure 25 which provided further evidence that the trend-related behavior is very similar across different channels. Please note that while we only provide results on Wikipedia our approach could be applied to any online and social media channel for which historic data is available (see Section 5.6).

The raw Wikipedia viewing statistics are published[1] by the Wikimedia foundation from where hourly view statistics starting from January 1, 2008 were obtained (2.8 TB compressed in total). These logs are aggregated to daily viewing statistics where URIs that have been viewed less then 25 times on that day are dropped.

[1] Retrieved from http://dumps.wikimedia.org/other/pagecounts-raw/ on April 10, 2013.

This does not introduce any bias since trending topics tend to accumulate view counts several orders of magnitude higher. For each day this results in approx. 2.5 million URIs attracting 870 million daily views. In total, the English and German Wikipedia feature more than five million articles that can be used for forecasting. During data preparation we noticed that Wikipedia did not record view counts for a few short periods of time during the last five years for which they are interpolated linearly.

Unlike related work based on query data from major search engines that use normalized viewing statistics (relative to the total number of queries on that day; e.g. [71]) this dataset allows us to work with absolute viewing statistics. The magnitude of these statistics can vary from 10^2 to 10^7 as illustrated by the log-plot in Figure 29. More examples for Wikipedia page view statistics have been given throughout the thesis such as in Figure 26 on page 61.

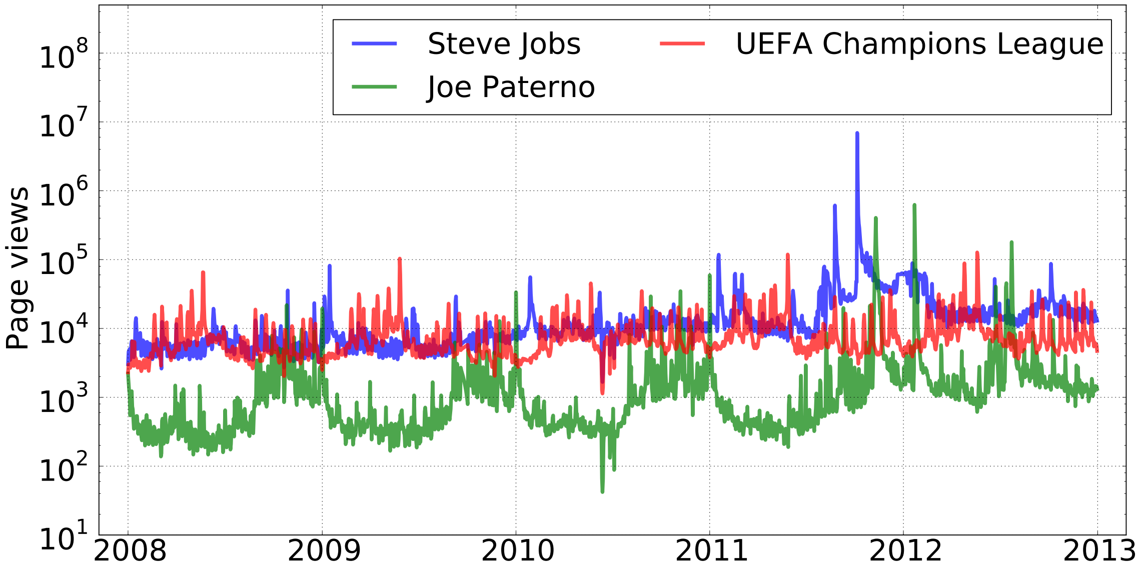


Figure 29: Example Wikipedia page view d statistics illustrating the large differences in user attention varying by up to five orders of magnitude.

6.2 PERFORMANCE METRICS

In the following, forecasts will be compared to the actual viewing statistics. The quality how well two time series match is captured by error metrics. A very common choice [48, 71] is root mean squared error, an *absolute* error metric, defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}\,,$$

where A_t is the actual value and F_t is the forecast value.

Note that unlike [71], in this theses, error is measured on the actual view counts instead of normalizing by the total views for each day as this glosses over large relative errors (i.e., even larger trends might only account for 10^{-5} of the daily views yielding very small error rates for virtually any forecast).

Absolute error metrics are suboptimal if the error is averaged over multiple instances of different magnitudes. This can be the case for trending topics, for example if one compares forecasting errors for the Michael Jackson article (reaching several millions daily views after his death) to lesser known persons such as Joe Paterno (usually having hundreds up to ten-thousands of daily views; recall Figure 29). Relative error metrics have the advantage of being easy to interpret and to be comparable across different time series in contrast to absolute metrics such as root-mean-square error. A standard choice for a *relative* error metric is mean absolute percentage error (MAPE), defined as

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|\,.$$

Both RMSE and MAPE will be used to judge and compare the performance of the proposed forecasting approach to different baselines that were introduced in Section 2.2.

6.3 EXPERIMENTS

We structure our experiments along the three main building blocks of our proposed approach to compare design choices for the individual methods independently (recall Figure 27 on page 64).

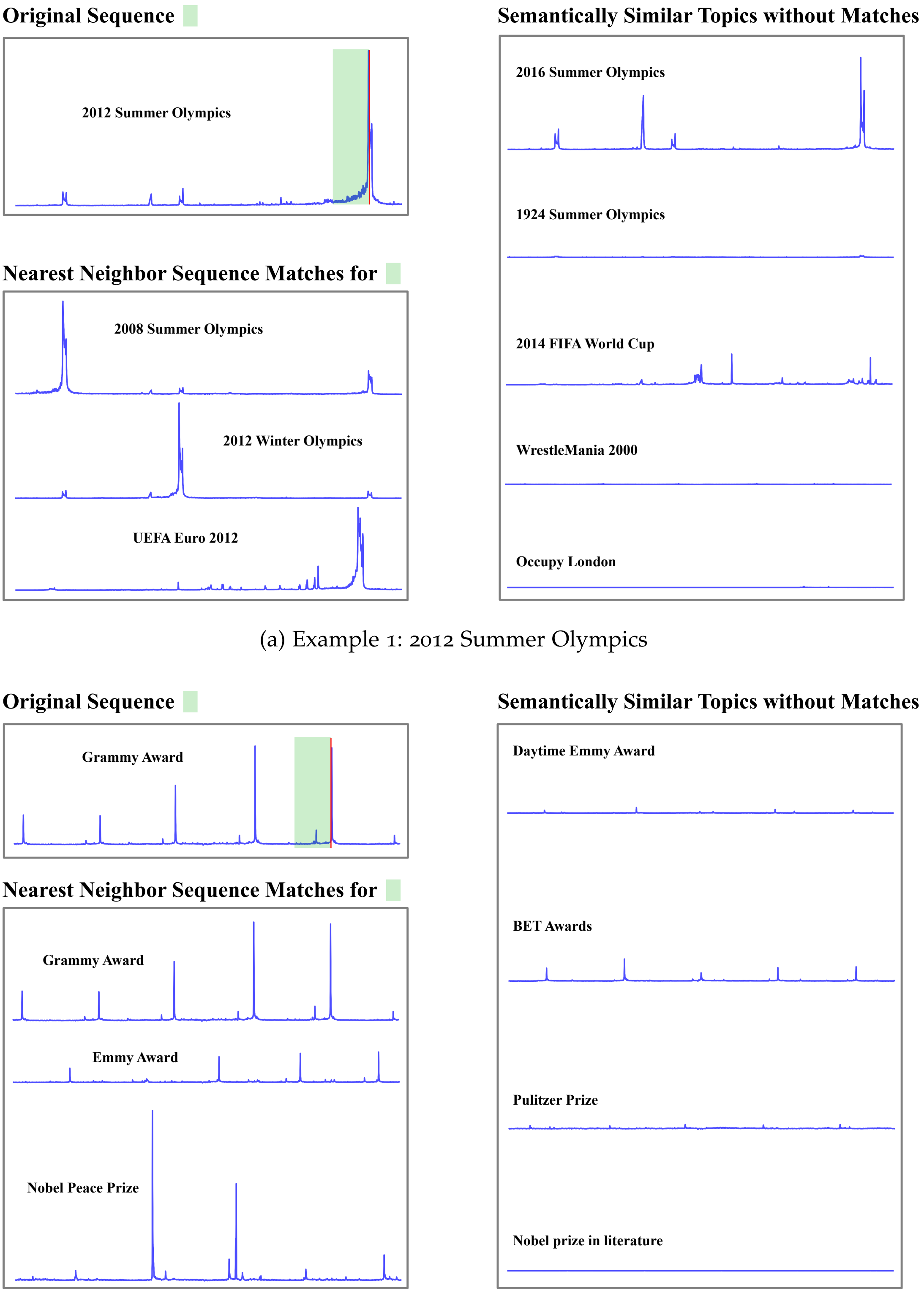
6.3.1 *Discovering Semantically Similar Topics*

The influence of discovering semantically similar topics is evaluated in two ways. First, qualitative results are presented by showing retrieved similar topics for a small sample of trending topics. Second, this part is evaluated indirectly by comparing forecast performance (i.e. through the third step) of using semantically similar topics from DBpedia to using a general set of topics (the top 200 trending topics).

Subsets of semantically similar topics for the top trends are shown in Table 7 (ignore bold print for now). Note that similar events or previous instances are successfully identified for events such as the Olympics, the Super Bowl, and UFC events for all of which we identify other major sport events. For award ceremony events such as the Academy awards and the Grammy Award the proposed approach is able to discover previous instances of the ceremony or other well known awards such as the Emmy Award, the Nobel Prize, and the Pulitzer Prize. Further, the approach is able to discover similar people. For Whitney Houston other primarily female music artists are discovered while for Justin Bieber the approach returns pop stars of similar age and fame. In the case of Steve Jobs other wealthy, leading, and influential figures, particularly from the tech industry, such as Mark Zuckerberg, Steve Wozniak, and Bill Gates, are identified. The case of Joe Paterno is particularly interesting because it fully automatically discovers people that died from the same cause, here, lung cancer (Joe Paterno, Jack Ruby, and Paul Newman). The nearest neighbor topics of Tim Tebow, a popular football player and NFL quarterback, include other NFL players with a similar celebrity status (Reggie Bush and Michael Oher), players of the same team and position (Peyton Manning), and other rookie quarterbacks of similar age (Colt McCoy and Cam Newton). Lastly, another trending topic category is computer games (products) in which case other games are discovered that attracted similar amounts of attention (Battlefield 3 and Diablo III). On average, about 95 semantically similar topics were retrieved per trending topic. While one could easily retrieve more similar topics this tends to decrease the precision.

Trending Topic	Nearest Neighbor Topics
2012 Summer Olympics	**2008 Summer Olympics, UEFA Euro 2012, 2010 Winter Olympics** :: 2016 Summer Olympics, 2014 FIFA World Cup, 2006 Winter Olympics
Whitney Houston	**Ciara, Shakira, Celine Dion, Brittany Murphy, Ozzy Osbourne** :: Alicia Keys, Paul McCartney, Janet Jackson
Steve Jobs	**Mark Zuckerberg, Rupert Murdoch, Steve Jobs** :: Steve Wozniak, Bill Gates, Oprah Winfrey
Super Bowl XLVI	**Super Bowl, Super Bowl XLV, Super Bowl XLIV** :: Super Bowl XLIII, 2012 Pro Bowl, UFC 119
Justin Bieber	**Selena Gomez, Kanye West, Justin Bieber** :: Katy Perry, Avril Lavigne, Justin Timberlake
84th Academy Awards	**83rd Academy Awards, 82nd Academy Awards** :: List of Academy Awards ceremonies, 81st Academy Awards
UFC 141	**UFC 126, UFC 129, UFC 124, UFC 132, UFC 127, UFC 117** :: UFC 138, UFC 139, UFC 137
Battlefield 3	**Mortal Kombat, FIFA 10, Call of Duty: Modern Warfare 2, Portal, Duke Nukem Forever** :: Call of Duty: Modern Warfare 3, Call of Duty 4: Modern Warfare, Pro Evolution Soccer 2011
Joe Paterno	**Terry Bradshaw, Joe Paterno, Jack Ruby, Paul Newman, Jerry Sandusky** :: Lane Kiffin, Donna Summer, Joe DiMaggio
Tim Tebow	**Reggie Bush, Michael Oher, Peyton Manning, Tim Tebow** :: Colt McCoy, Cam Newton
Diablo III	**Call of Duty: Modern Warfare 2, Call of Duty 4: Modern Warfare, Portal 2, Portal, StarCraft II: Wings of Liberty** :: World of Warcraft, Deus Ex, Rage
Grammy Award	**Grammy Award, Emmy Award, Nobel Peace Prize** :: Nobel Prize in Literature, BET Awards, Pulitzer Prize
54th Grammy Awards	**53rd Grammy Awards, 52nd Grammy Awards, 54th Grammy Awards** :: 51st Grammy Awards, 2012 Billboard Music Awards, 2012 MTV Europe Music Awards

Table 7: Selected trending topics along with their nearest neighbor topics using category and type information on DBpedia (step 1). The ones chosen by nearest neighbor sequence matching (step 2) are in bold print. In some cases the topic itself can be used for forecasting (appearing in both columns), e.g. if the time series contains recurring patterns (recall Figure 25).



(a) Example 1: 2012 Summer Olympics

(b) Example 2: Grammy Award

Figure 30: Qualitative results indication the necessity of Nearest Neighbor Sequence Matching. All signals are Wikipedia article views of corresponding articles drawn to the same scale. Note that several semantically similar topics exhibit different behavior or are in less demand but that the ones returned by Nearest Neighbor Sequence Matching match the original sequence much more closely.

6.3.2 *Nearest Neighbor Sequence Matching*

While the previous step succeeds in discovering many semantically similar topics those topics do not necessarily all exhibit similar behavior and might not be good choices for forecasting. Two examples are depicted in Figure 30 on the facing page. Note that all signals are Wikipedia article views of the corresponding articles drawn to the same scale (at which some articles appear to have generated very few views over the last years). The 2012 Summer Olympics were one of the largest sport events during the past decade (see Figure 30 (a)). Therefore, events that have not occurred yet are unlikely to have generated similar volumes and patterns of user attention (indicated by the green box). This is the case for the 2016 Summer Olympics and the 2014 FIFA World Cup. While the 1924 Summer Olympics might have generated a lot of attention during the 1920's, this time period is not covered by our dataset and over the past few years there only has been very little interest in the topic. Similarly, while Occupy London matches the 2012 Summer Olympics in the sense that it is also an event taking place in London in the same year, not nearly as many people were interested in the event. However, applying signal-based Nearest Neighbor Sequence Matching successfully identifies good matches such as the 2008 Summer Olympics, the 2010 Winter Olympics, and the UEFA Euro 2012 championship that all exhibit behavior similar to the 2012 Summer Olympics.

The second example in Figure 30 (b) shows the Grammy Award and several semantically similar topics along with their Wikipedia page view statistics. Again, several semantically similar events such as the Daytime Emmy Award, the BET Awards, the Pulitzer Prize, and the Nobel prize in literature fail to match the given trending topic in terms of behavioral patterns. However, Nearest Neighbor Sequence Matching is able to retrieve those award ceremonies that generate a similar level of attention as the Grammy Award. Since the signal corresponding to the Grammy Award is self-recurrent (see Section 5.1) good matches are also found within the same signal. Two other major award ceremonies that should be used for forecasting were found to be the Emmy Award and the Nobel Peace Prize.

There are several different ways to identify the best nearest neighbor candidates for forecasting and the most natural is by using a distance metric between time series (as introduced in Section 5.4). The main design choice when matching sequences from similar topics to a short history window of our time series is the choice of the distance metric and the length of the history window. It intuitively makes sense to match recent behavior more than perfectly matching the behavior two years ago. Therefore, the length of the history window was chosen to be 60 days. While small changes did not significantly effect the results it is important for longer history windows to emphasize recent scores over long past scores by weighting them differently.

The experimental setup for this part as well as forecasting (Section 6.3.3) looks as follows. Trending topics acquired in Chapter 4 are used as a trigger for forecasting, i.e. for each of the 516 sequences of the top 200 trending topics (recall Table 3 on page 48), a forecast with a horizon of 14 days is made, starting on the day they first emerge. This window of 14 days was chosen since this represents a reasonable maximum lifetime for most trending topics (recall Figure 20 on page 50). Different methods will be evaluated by beginning with a forecast of 14 days, then a 13 day forecast after one day and so on (illustrated by τ on the X axis in the error plots below). As the forecasting problems become easier error rates can be expected to decrease over time, i.e. from left to right.

As introduced in Section 5.4, the metrics `euclidean`, `musigma`, and `y_invariant` will be compared. To do so, the quality of the (first) nearest neighbor returned by this metric is measured by its similarity to the actual viewing statistics over the next 14 days (similar to the forecast setting but directly using the nearest neighbor as the forecast). The results are depicted in Figure 31 and show the average MAPE error (lower is better, see Section 6.2) between the nearest neighbor and the actual viewing statistics (ground truth). We also show oracle performance by picking the best match for the ground truth from all sequence candidates (based on all similar topics $\mathcal{T}_{\mathrm{sim}}$). First, note that oracle performance is well above 0% error as there might be no perfect match among the sequence candidates. Further, we observe that `y_invariant` can retrieve poor quality matches whereas the simple `euclidean` distance performs equally or better than the other two metrics. Apparently, the invariance of `musigma` and `y_invariant` does not help in this task. Therefore, we use the `euclidean` metric for all following experiments. Further note that the best matches have between 82 and 319% error illustrating the high complexity of the task.

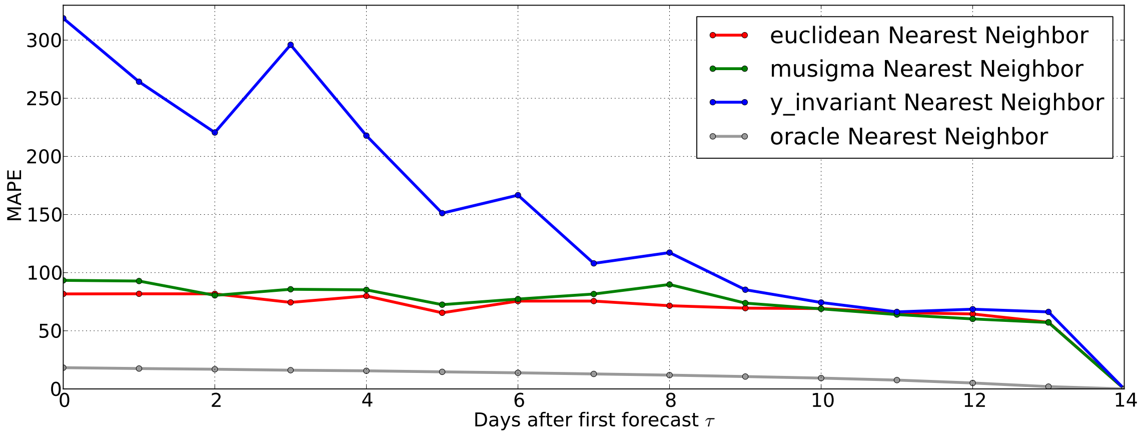


Figure 31: A comparison of nearest neighbor distance metrics. Euclidean distance performs at least as good as its more complex normalized/invariant counterparts. Further note that oracle performance is strictly greater than 0% error as there might be no perfect match among the sequence candidates.

6.3.3 *Forecasting*

As introduced above, forecasting performance is measured by forecasting the next 14 days for each of the 516 sequences of the top 200 trending topics at the point in time when they first emerge. The proposed approach is compared to several baselines that use a short history window of the time series itself (similar to [71]): a *naive* forecast (tomorrow's behavior is the same as today's) and a *linear trend* based on the last 14 days (as introduced in Section 2.2.3). We further compare to the *average trend* and *median trend* in our trending topics dataset as baselines that includes information from multiple time series. Note that this average and median trend are computed from μ/σ-normalized time series (i.e. we subtract the mean μ and divide by the standard deviation σ) since the average/median of actual view counts are actually very far from most trending topics and would perform poorly. To still be able to compute the error for actual view count prediction we then de-normalize the *average trend* and *median trend* baselines with the parameters of the time series to be predicted. Further, we compare against the performance of selected autoregressive models, namely AR(1), AR(2), ARMA(1,1), and AutoARIMA (please refer to Section 2.2.4 for a formal specification of these models). Note that while the following experiments were performed for different numbers of neighbors (for the nearest-neighbor-based methods) only the results for $k = 3$ are reported which performed best by a small margin.

	Method	**RMSEs in 1000**				
		$\tau = 0$	$\tau = 3$	$\tau = 5$	$\tau = 7$	$\tau = 9$
Baselines	naive	63.2	33.1	20.2	17.4	11.4
	linear trend	86.9	48.5	28.3	23.1	14.5
	average trend	49.3	25.9	22.0	19.9	18.3
	median trend	48.1	24.9	20.6	18.1	16.1
ARIMA	AR(1)	50.1	27.8	20.1	15.9	12.7
	AR(2)	75.1	31.7	22.6	16.0	13.4
	ARMA(1,1)	53.0	28.7	20.5	15.8	13.2
	AutoARIMA	58.9	30.7	26.9	19.5	16.7
Self	average	46.0	23.7	19.7	18.0	16.6
	average_scaled	44.6	21.9	17.6	15.5	13.8
	median	46.1	23.8	19.7	17.7	16.0
	median_scaled	44.9	22.3	18.1	15.5	14.4
Gen	average	45.7	22.9	19.2	16.1	14.1
	average_scaled	45.7	22.5	16.0	14.1	11.4
	median	41.4	21.2	17.6	15.4	12.9
	median_scaled	40.1	19.5	15.2	12.8	10.2
Sim	average	41.4	18.8	16.0	14.0	12.3
	average_scaled	39.6	17.1	13.7	11.6	10.0
	median	42.1	19.9	16.5	14.0	12.5
	median_scaled	41.0	17.9	14.2	11.5	9.8

Table 8: RMSE forecasting error for our baselines, selected autoregressive models, as well as methods using only the trending topic itself (*Self*), a general set of topics (*Gen*), or similar topics (*Sim*). The number of days after we forecast the rest of the 14 day period is represented by τ, e.g. $\tau = 5$ means that five days after the topic becomes trending we forecast the remaining nine days. Note that our proposed nearest neighbor approach outperforms baselines and autoregressive models in all cases and that using semantically similar topics outperforms a general set of topics which in turn outperforms using only the trending topic itself.

The RMSE forecasting errors are summarized in Table 8 which reports the results for several different instances of our proposed forecasting approach. *Self*, *Gen*, and *Sim* refer to the different topic sets $\mathcal{T}_{\mathrm{self}}$, $\mathcal{T}_{\mathrm{gen}}$, and $\mathcal{T}_{\mathrm{sim}}$ from which the nearest neighbor sequences are chosen (as described in the Section 5.3). On average, the RMSE of our best method is about 9k views closer to the actual viewing statistics than our best performing baseline, median trend, that already takes multiple time series into account (a relative improvement of over 17%). Compared to the worst performing baseline, linear trend, the proposed method is about 48k views closer on average (a relative improvement of over 54%). We can also observe that our proposed nearest neighbor approach outperforms autoregressive models in all cases which perform roughly on the same level as our baselines. Also, notice that the fairly sophisticated AutoARIMA model performs worse than its much simpler AR(1) counterpart – even though it aims to choose the best ARIMA model for the underlying data and was shown to perform well in several other forecasting competitions. This observation adds to our impression that autoregressive models (which assume stationarity; see Section 2.2.4) are not well suited to model trending topic time series with structural breaks. Further, it can be observed that taking the median tends to perform about as good or better than taking the average, using scaled nearest neighbor is better than unscaled nearest neighbors (i.e., not using the α function defined in Section 5.5), and that using semantically similar topics (*Sim*) is better than using a general set of topics (*Gen*) which in turn is better than restricting oneself to a single time series (*Self*).

However, RMSE error has the disadvantage that it is dominated by the most popular trending topics with the largest view counts. Therefore, we choose the MAPE measure for the remaining analysis as this relative error metric is comparable across trending topics of varying popularity. Additionally, because MAPE errors can become disproportionally large (e.g. forecasting 1000 views when it is actually only 100 results in a 900% relative error), we drop obvious outliers (5%) and report the average error of the remaining sequences. Results for the baselines and the best performing methods from Table 8 will be shown (to be able to visually distinguish them in the error plot). Again, the methods are evaluated by beginning with a forecast of 14 days, then a 13 day forecast after one day and so on as illustrated by the X axis in Figure 32. We can observe that the proposed approach including median and scaling clearly outperforms all baselines as well as other instances of our framework. Our proposed method achieves a mean average percentage error (MAPE) of 45-19% (for $\tau = 0$ until $\tau = 13$). This corresponds to a relative improvement over our baselines of 20-90%

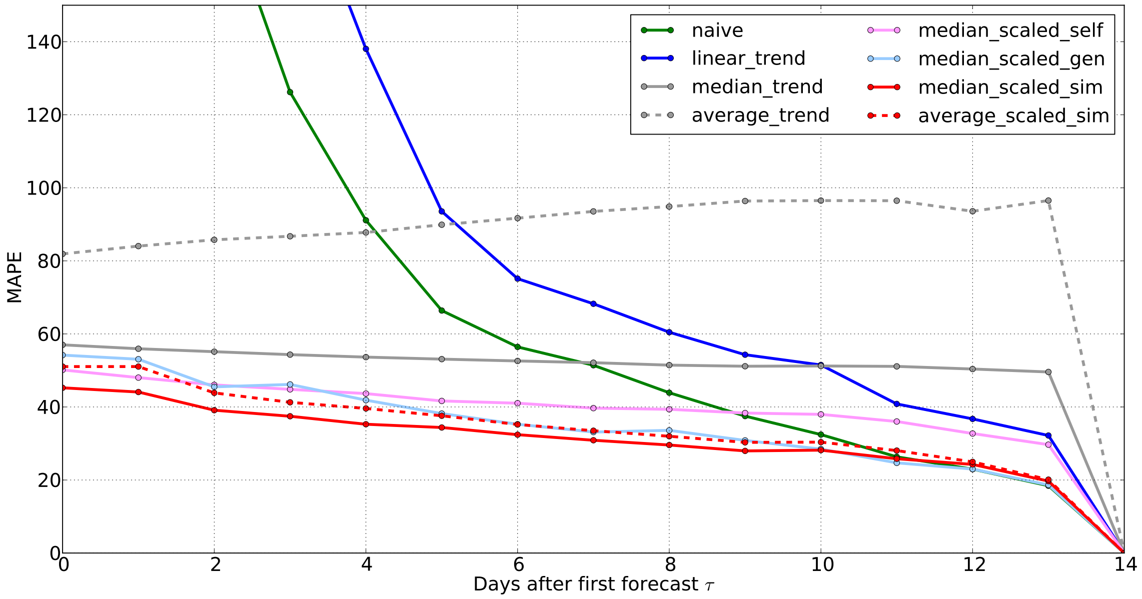


Figure 32: MAPE forecast error moving through a 14 day window, e.g. the error at 4 depicts the forecasting error averaged over the following 10 days. Simple baseline methods perform poorly for long-term forecasts. The proposed nearest neighbor approach outperforms all baselines and using semantically similar signals (red curve) performs significantly better than using general set of signals or just the signal itself.

with again median trend performing best and linear trend performing worst among the baselines.

6.3.4 *Qualitative Forecasting Results*

To get a sense for the behavior of different forecasting approaches it is important to analyze qualitative results as well. Example forecasts are given in Table 9 for four different trending topics: "Battlefield 3" (BF3), "The Hunger Games" (THG), "UEFA Champions League" (UEFA), and "Ultimate Fighting Championship" (UFC). The first column shows the day at which the forecast is computed starting at day 0, the day the topic becomes trending until day 12 where only the last two days have to be forecasted (corresponding to τ in Table 8 on page 82). All other columns depict the corresponding forecasts at these points in time (also indicated by the vertical red line). Note that these forecasts do change over time. Forecasts of three different models are shown. The proposed method `median_scaled_sim` (red) is compared to its variants only using the trending topic itself (*self*; pink) or using a set of general topics (*gen*; blue). The ground truth behavior is given by the green curve.

The "Battlefield 3" (BF3) example illustrates that it is often insufficient to only use information from a single time series (*Self*) or to choose from a general set of topics (*Gen*). The `median_scaled_self` forecast features a very fast drop-off since the signals itself does not contain any similarly high peaks. On the other hand, `median_scaled_gen` overestimates the engagement toward the topic as many of the trends in the general set exhibited such behavior. However, the variant using semantically similar topics leads to a much more accurate forecast.

In "The Hunger Games" (THG) example, the first forecasts all miss the upcoming peak because such behavior has not been encountered before (day 0). However, on the following day this mistake is recognized and new forecasts generated. Note that the approach using semantically similar topics is the only one that is able to capture the interesting dynamics leading to the second peak.

An example illustrating recurring behavior is given by the trending topic "UEFA Champions League" (UEFA). All three forecast models use patterns from the same time series as the recurring patterns can be re-used for forecasting (note that the blue and pink curve are hidden behind the red one). In this example, the proposed nearest neighbor approach succeeds in forecasting such recurring behavior that occurs with a certain regularity, i.e. whenever the Champions League matches are scheduled (see Section 4.2).

The "Ultimate Fighting Championship" (UFC) example illustrates that sudden, large peaks can be forecasted by the proposed technique before they happen. While `median_scaled_self` and `median_scaled_gen` obviously fail to capture the trending behavior, the variant that includes similar topics forecasts the amount of attention toward the UFC event with remarkable accuracy. There have been over one hundred UFC events that all exhibited very similar behavior, i.e. having a small peak around a month before the actual event (as seen in Table 7 these events are also discovered by the proposed approach). Because of this regularity, our approach succeeds in anticipating the Wikipedia page views very accurately. Therefore, this is another great example for the predictive power that comes from exploiting semantically similar events.

In general, it can be observed that nearest neighbor forecasting can lead to very accurate forecasts and that in many cases, only using information from a single time series or a general set of topics is not sufficient. Exploiting that semantically similar topics exhibit similar behavior improves forecasting in most cases (particularly, the very challenging non-recurrent cases; see Section 5.1).

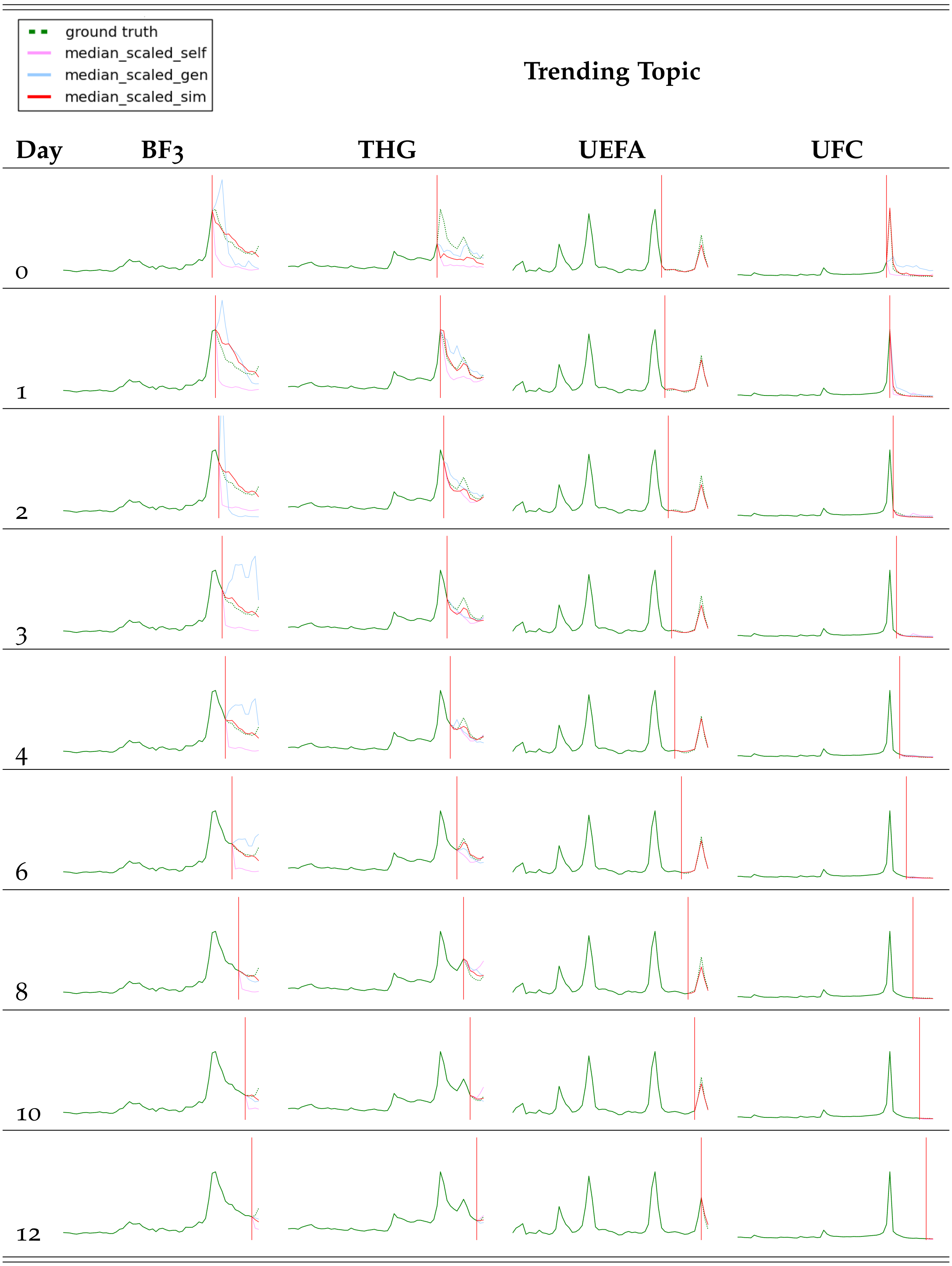


Table 9: Visualization of example trending topic forecasts for “Battlefield 3” (BF3), “The Hunger Games” (THG), “UEFA Champions League” (UEFA), and “Ultimate Fighting Championship” (UFC). Each column depicts multiple forecasts at different points in time (as indicated by the vertical red line). The individual graphs plot Wikipedia article views (Y axis) against time (X axis).

7 CONCLUSION

The final chapter of this thesis consists of a summary and a future outlook. First, major challenges related to trending topics in online media are summarized along with the main contributions, results and most important insights from this work. The second part then suggests further directions of improvement and outlines possible future work in the investigated area of research.

7.1 SUMMARY

Every day, petabytes of data capturing human behavior are created from online interactions such as social networks or web search engines. Much effort has been devoted to process these large amounts of data as well as aggregate and visualize certain parts of it. However, since this data contains valuable information about what people are interested in and how they feel about certain topics (and in this sense mirror society), there is an increasing interest in analyzing these datasets more thoroughly. Understanding trends in online media is a key part in many application domains such as economics, health monitoring, journalism, finance, marketing, and social multimedia systems. In these applications, it is important to understand various characteristics of different online media channels such as what kind of trends they feature and what their temporal behavior looks like. Some of them even require knowledge about trends in the future, i.e. forecasting the amount of people's attention and user engagement toward given topics in the very moment they emerge.

These two core challenges are addressed in this thesis by providing a multichannel analysis of trending topics based on an observation period of one year as well as proposing a novel forecasting approach for trending topics. Evidence was presented that semantically similar topics exhibit similar behavior, i.e. that there are common behavioral patterns that are shared among related topics.

Our proposed forecasting approach exploits this observation by automatically identifying semantically similar topics and utilizing historic behavior patterns to produce forecasts for trending topics. This forecasting model was evaluated empirically using real-world user behavior of a large-scale Wikipedia dataset and demonstrated superior performance over baselines methods and autoregressive models.

Summarizing, this thesis provided the following main insights into trending topics in online media:

1. Many more high-impact trending topics are featured on Google (86%) than on Twitter (45%) or Wikipedia (21%) based on the top 200 trends.
2. A significant difference in when trends start could not be observed across the different online media channels (possibly due to daily granularity of the data).
3. Trends on Twitter and Wikipedia are significantly more ephemeral than on Google ending between three and four days before they do so on Google.
4. All observed media channels tend to specialize in specific categories which has direct implications for building trend-aware information systems.
5. Semantically similar topics exhibit similar behavior, i.e. the attention toward a topic follows certain patterns common among similar topics.
6. It is possible to automatically discover semantically similar topics with high precision, e.g. by using semantic category information on DBpedia.
7. Using historic behavioral information of semantically similar topics can significantly improve forecasting performance compared to baselines, autoregressive models, and other nearest neighbor techniques.
8. Forecasts by the proposed approach are about 9-48k views closer to the actual viewing statistics than all baseline methods, and achieve a mean average percentage error (MAPE) of 45-19% for time periods of up to 14 days, a relative improvement over baselines of 20-90%.

7.2 FUTURE WORK

In the context of this thesis, several different research directions could be pursued to build upon the presented results. First, the multi-channel analysis could be extended to datasets with trends on a sub-day, e.g. hourly, time scale. This would allow for insights whenever time lags are significantly smaller than one day. Given the surprising result that trends in Wikipedia arise as quickly as for example in Twitter, we hypothesized that people tend to consult corresponding Wikipedia articles in early phases of the trend to learn about the subject at hand. As a first step of investigation, edit histories of Wikipedia articles could be analyzed to check what information is already included at such early points in time after real-world events. The multi-channel analysis could further be applied to other social media channels such as Facebook, Google+, or Pinterest to investigate their temporal behavior and emphasis on particular topic categories. Given that this work focused on the top trends that are often of national or even global interest it would be interesting to see whether the observed characteristics also hold for more localized trends such as city- and/or language-specific trends.

Today, Wikipedia articles are viewed many more times than a few years ago. This simple fact, a global trend in viewing statistics, could be incorporated in the proposed forecasting approach (Chapter 5). Invariance on the level of the distance metric between signals (Section 5.4) or forecasting (Section 5.5) would be viable options to do so. However, we found that there exists a trade-off between invariance and forecast robustness, at least for the distance metrics evaluated in this thesis (Section 6.3.2). Another promising direction is to explicitly detect and exploit seasonality in the signal or to include features other than the raw signals (such as the number of shared semantic categories). Formally, this could be achieved through a more general "learning to rank" framework [17]. A different approach would be to add a meta-learning step to choose the best model for a given signal (e.g., as in [71]). This could help answer when to restrict oneself to the present signal and when to include other similar signals (recall Figure 25).

In application domains such as journalism one might be more interested in the general shape of the forecast than the exact values. To optimize for this, error measures based on dynamic time warping (DTW) [10] could be used. Alternatively, one could predict "temporal clusters", i.e. transforming the time series forecasting problem into a classification problem (e.g. as in [94]).

Another promising direction would be to apply the presented insights into trending topics to improve video concept detection systems. Borth et al. [15] have found trending topics to be strongly correlated with changes in upload volume on YouTube. We believe that video concept detectors will benefit from contextual information about current trends that could significantly reduce current error rates. For example, the higher-than-usual upload volume of football videos during the Super Bowl can be incorporated in non-uniform concept priors that change over time, which would lead to prioritizing the concept "american football" over "soccer" or "golf" whenever the concept detectors are uncertain. Furthermore, trending topic forecasts could guide the allocation of computational resources to train concept models for trends on the fly.

BIBLIOGRAPHY

[1] Eytan Adar, Daniel S Weld, Brian N Bershad, and Steven S Gribble. Why we search: visualizing and predicting user behavior. In *Proc. of International Conference on World Wide Web*, 2007.

[2] Toni Ahlqvist, A. Bäck, M. Halonen, and S Heinonen. *Social media roadmaps: exploring the futures triggered by social media*. VTT Technical Research Centre of Finland, 2008. Available from `http://www.vtt.fi/inf/pdf/tiedotteet/2008/T2454.pdf`.

[3] James Allan. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer, 2002.

[4] Jesse Alpert and Nissan Hajaj. We knew the web was big... `http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html`, July 25, 2008. Web Search Infrastructure Team. Official Google Blog. Retrieved April 5, 2013.

[5] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[6] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[7] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[8] Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: Forecasting popularity. In *AAAI Publications, Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[9] Steven Van Belleghem. *The Conversation Company: Boost Your Business Through Culture, People and Social Media*. Kogan Page, 2012.

[10] Donald Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *AAAI 1994 workshop on Knowledge Discovery in Databases*, pages 359–370. AAAI, 1994.

[11] Mark A. Beyer and Douglas Laney. The importance of 'big data': A definition. `http://www.gartner.com/DisplayDocument?ref=clientFriendlyUrl&id=2057415.`, June 2012. Gartner.

[12] Jinbo Bi, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. Dimensionality reduction via sparse support vector machines. *The Journal of Machine Learning Research*, 3:1229–1243, 2003.

[13] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[14] Sam Borden. To score or not to score. `http://www.nytimes.com/2012/02/07/sports/football/super-bowl-46-after-giants-surreal-touchdown-debates-on-the-strategy.html`, 2012. New York Times. February 7, 2012. Retrieved April 8, 2013.

[15] Damian Borth, Adrian Ulges, and Thomas M Breuel. Dynamic vocabularies for web-based concept detection by trend discovery. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 977–980. ACM, 2012.

[16] George E P Box and Gwilym M Jenkins. *Time series analysis: forecasting and control*. San Francisco: Holden-Day, 1976.

[17] Christopher JC Burges, Robert Ragno, and Quoc Viet Le. Learning to rank with nonsmooth cost functionss. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 19, pages 193–200. The MIT Press, 2007.

[18] Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. `http://research.google.com/archive/papers/initialclaimsUS.pdf`, 2009. Google, Inc.

[19] Hyunyoung Choi and Hal Varian. Predicting the Present with Google Trends. *Economic Record*, 88(S1):2–9, 2012.

[20] Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman. Corpora for topic detection and tracking. *Topic detection and tracking*, pages 33–66, 2002.

[21] Michael P Clements and David F Hendry. Forecasting annual uk inflation using an econometric model over 1875–1991. *Frontiers of Economics and Globalization*, 3:3–39, 2009. Forecasting in the presence of structural breaks and model uncertainty.

[22] Crystale Purvis Cooper, Kenneth P Mallon, Steven Leadbetter, Lori A Pollack, and Lucy A Peipins. Cancer internet search activity on a major search engine, united states 2001-2003. *Journal of medical Internet research*, 7(3), 2005.

[23] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[24] Richard A Davis, Thomas CM Lee, and Gabriel A Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[25] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. Find me the right content! diversity-based sampling of social media spaces for topic-centric search. In *Proceedings of the 5th international AAAI Conference on Weblogs and Social Media*, 2011.

[26] Sotiris Diplaris, Symeon Papadopoulos, Ioannis Kompatsiaris, Ayse Goker, Andrew Macfarlane, Jochen Spangenberg, Hakim Hacid, Linas Maknavicius, and Matthias Klusch. SocialSensor: sensing user generated input for improved media discovery and experience. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 243–246. ACM, 2012.

[27] Jurgen A Doornik. Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data. 2009.

[28] eMarketer. Is Social Media Marketing at a Saturation Point? http://www.emarketer.com/Article/Social-Media-Marketing-Saturation-Point/1009273, August 17, 2012. eMarketer Report. Retrieved April 18, 2013.

[29] Marcelo Espinoza, Caroline Joye, Ronnie Belmans, and Bart DeMoor. Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. *Power Systems, IEEE Transactions on*, 20(3):1622–1630, 2005.

[30] Gunther Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association, 2006.

[31] User "Factoryjoe". Figure: Maslow's hierarchy of needs. `http://en.wikipedia.org/wiki/File:Maslow%27s_Hierarchy_of_Needs.svg`, June 18, 2009. Wikimedia Commons. Retrieved April 5, 2013.

[32] Fernando Fernández-Rodríguez, Simón Sosvilla-Rivero, and Julián Andrada-Félix. Exchange-rate forecasts with simultaneous nearest-neighbour methods: Evidence from the ems. *International Journal of Forecasting*, 15(4):383–392, 1999.

[33] Fernando Fernández-Rodríguez, Simón Sosvilla-Rivero, and Julián Andrada-Félix. Nearest-neighbour predictions in foreign exchange markets. 2002. FEDEA Working Paper.

[34] Sam Fiorella. The Addiction and Cost of Social Media. `http://www.huffingtonpost.com/sam-fiorella/social-media-addiction_b_2749102.html`, February 24, 2013. Huffington Post. Retrieved April 18, 2013.

[35] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[36] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[37] Steve Goldstein. Volkswagen shares skyrocket on Porsche move. `http://articles.marketwatch.com/2008-10-27/news/30797796_1_porsche-automobil-holding-vw-shares-volkswagen-shares`, October 27, 2008. Market Watch, The Wall Street Journal.

[38] Soren Gordhamer. 5 ways social media is changing our daily lives. http://mashable.com/2009/10/16/social-media-changing-lives/, October 16, 2009. Mashable.

[39] Carrie Grimes. Our new search index: Caffeine. http://googleblog.blogspot.de/2010/06/our-new-search-index-caffeine.html, June 8, 2010. Web Search Infrastructure Team. Official Google Blog. Retrieved April 5, 2013.

[40] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

[41] Anthony Ha. Denver Post Highlights Social Media Coverage (And Storify) In Its Pulitzer Win. http://techcrunch.com/2013/04/15/denver-post-pulitzer/, April 15, 2013. TechCrunch. Retrieved April 16, 2013.

[42] Kandace Harris. Using social networking sites as student engagement tools. *Diverse Issues in Higher Education*, 25(18):4, 2008.

[43] Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1991.

[44] Tom Heyden. Harlem Shake: Tracking a meme over a month. http://www.bbc.co.uk/news/magazine-21624109, March 1, 2013. BBC News. Retrieved April 16, 2013.

[45] Nielsen Holdings. State of the Media: The Social Media Report 2012. http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html, April 12, 2012. Nielsen Holdings Report. Retrieved April 18, 2013.

[46] Anette Hulth, Gustaf Rydevik, and Annika Linde. Web queries as a source for syndromic surveillance. *PloS one*, 4(2), 2009. Public Library of Science.

[47] Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 2008.

[48] Rob J Hyndman, Anne B Koehler, Ralph D Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3):439–454, 2002.

[49] Nat Ives and Rupal Parekh. Marketers Jump on Super Bowl Blackout With Real-Time Twitter Campaigns: Social-Media Teams at Oreo, Audi, Tide and VW React Swiftly. `http://adage.com/article/special-report-super-bowl/marketers-jump-super-bowl-blackout-twitter/239575/`, February 3, 2013. Ad Age media news.

[50] Alejandro Jaimes. A human-centered perspective on multimedia data science: tutorial overview. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1537–1538. ACM, 2012.

[51] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[52] Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. The wisdom of social multimedia: using Flickr for prediction and forecast. In *Proceedings of the international conference on Multimedia*, pages 1235–1244. ACM, 2010.

[53] Donald E Knuth. *The TEXbook*. Addison-Wesley, 1986.

[54] Eric Kuhn. Google unveils top political searches of 2009. `http://politicalticker.blogs.cnn.com/2009/12/18/google-unveils-top-political-searches-of-2009/`, December 18, 2009. CNN Politics. Retrieved April 5, 2013.

[55] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P Gummadi. Geographic dissection of the twitter network. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[56] Andrew Lipsman. comScore Releases November 2009 U.S. Search Engine Rankings. `http://www.comscore.com/Press_Events/Press_Releases/2009/12/comScore_Releases_November_2009_U.S._Search_Engine_Rankings`, December 16, 2009. comScore. Retrieved April 5, 2013.

[57] Ingrid Lunden. Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City. http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/, July 30, 2012. TechCrunch.

[58] Patrick Lynch. NASA Finds 2011 Ninth Warmest Year on Record. http://www.giss.nasa.gov/research/news/20120119/, January 19, 2012. NASA Headquarters release No. 12-020. Retrieved April 16, 2013.

[59] Abraham H Maslow. A theory of human motivation. *Psychological Review*, 50:370–396, 1943.

[60] Lisa Mason. Impact of social media on society: 5 times social changed the world. http://socialmediasun.com/impact-of-social-media-on-society/, July 4, 2012. Social Media Sun.

[61] Mike McCarthy. Websites apologize for prematurely reporting Paterno's death. http://usatoday30.usatoday.com/sports/college/football/story/2012-01-22/joe-paterno-websites-report-death-premature/52744180/1, February 1, 2012. USA Today. Retrieved April 19, 2013.

[62] Samantha Murphy. 9 Brands That Thought Fast on Social Media During the Super Bowl. http://mashable.com/2013/02/04/brands-super-bowl-social-media/, February 4, 2013. Mashable.

[63] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[64] John Neter, William Wasserman, Michael H Kutner, et al. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[65] Barbara Ortutay. Live action: Twitter grabs Super Bowl spotlight. http://bigstory.ap.org/article/live-action-twitter-grabs-super-bowl-spotlight, February 4, 2013. The Big Story, Associated Press.

[66] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. In *SIGIR 2012 Workshop on Time-aware Information Access*, 2012.

[67] Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[68] Pulitzer Price Jury. The 2013 Pulitzer Prize Winners: Breaking News Reporting. `http://www.pulitzer.org/citation/2013-Breaking-News-Reporting`, April 15, 2013. Pulitzer. Retrieved April 16, 2013.

[69] Erik Qualman. *Socialnomics: How social media transforms the way we live and do business*. Wiley, 2012.

[70] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Predicting the news of tomorrow using patterns in web search queries. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08.*, volume 1, pages 363–367. IEEE, 2008.

[71] Kira Radinsky, Krysta Svore, Susan Dumais, Jaime Teevan, Alex Bocharov, and Eric Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st international conference on World Wide Web*, pages 599–608. ACM, 2012.

[72] Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Traffic in Social Media I: Paths Through Information Networks. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 452–458. IEEE, 2010.

[73] Drew Rattray. Michael Jackson Dies and Takes the Internet with Him. `http://blog.tmcnet.com/design-vs-functionality/2009/06/michael-jackson-dies-and-takes-the-internet-with-him.html`, June 26, 2009. TMCnet Bloggers. Retrieved April 4, 2013.

[74] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. SocialTransfer: Cross-Domain Transfer Learning from Social Streams for Media Applications. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 649–658. ACM, 2012.

[75] Pamela Rutledge. Social networks: What maslow misses. `http://www.psychologytoday.com/blog/positively-media/201111/social-networks-what-maslow-misses-0`, November 2011. Psychology Today.

[76] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM, 2010.

[77] Sina Samangooei, Daniel Preotiuc-Pietro, Trevor Cohn, Mahesan Niranjan, and Nicholas Gibbins. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *AAAI Publications, Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[78] Dan Schawbel. 5 Reasons Why Your Online Presence Will Replace Your Resume in 10 years. `http://www.forbes.com/sites/danschawbel/2011/02/21/5-reasons-why-your-online-presence-will-replace-your-resume-in-10-years/`, February 21, 2011. Forbes. Retrieved April 18, 2013.

[79] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization and beyond*. MIT Press, 2002.

[80] Erick Schonfeld. Google Processing 20,000 Terabytes A Day, And Growing. `http://techcrunch.com/2008/01/09/google-processing-20000-terabytes-a-day-and-growing/`, January 9, 2008. TechCrunch. Retrieved April 5, 2013.

[81] Matthew Sparkes. Twitter and Facebook 'addicts' suffer withdrawal symptoms. `http://www.telegraph.co.uk/technology/social-media/9986950/Twitter-and-Facebook-addicts-suffer-withdrawal-symptoms.html`, April 11, 2013. The Telegraph. Retrieved April 18, 2013.

[82] Cisco Systems. The Zettabyte Era. `http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html`, May 2012. White Paper.

[83] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[84] Twitter Search Team. The Engineering Behind Twitter's New Search Experience. `http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html`, May 31, 2011. Twitter Engineering Blog. Twitter. Retrieved April 4, 2013.

[85] Marijn ten Thij, Yana Volkovich, David Laniado, and Andreas Kaltenbrunner. Modeling and predicting page-view dynamics on Wikipedia. *arXiv preprint arXiv:1212.5943*, 2012.

[86] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[87] Twitaholic.com. The Twitaholic.com Top 100 Twitterholics based on Followers. `http://twitaholic.com/`, April 4, 2013. Retrieved April 4, 2013.

[88] Twitter. Twitter turns six. `http://blog.twitter.com/2012/03/twitter-turns-six.html`, March 21, 2012. Twitter Blog. Retrieved April 4, 2013.

[89] Jose Antonio Vargas. On Wikipedia, Debating 2008 Hopefuls' Every Facet. September 17, 2007. The Washington Post.

[90] Zhi Wang, Lifeng Sun, Xiangwen Chen, Wenwu Zhu, Jiangchuan Liu, Minghua Chen, and Shiqiang Yang. Propagation-Based Social-Aware Replication for Social Video Contents. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[91] Jianshu Weng and Bu-Sung Lee. Event Detection in Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[92] Stewart Whiting and Omar Alonso. Hashtags as Milestones in Time. In *SIGIR 2012 Workshop on Time-aware Information Access*, 2012.

[93] Peter Whittle. *Hypothesis testing in time series analysis*. Almqvist & Wiksell, 1951.

[94] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

A

APPENDIX

The following table lists the top 200 trending topics of the dataset described in Section 4.1 on page 37. The listed score is the global trend score defined in Section 4.1.3 on page 40 and the topic categories were defined in Table 5 on page 56.

Table 10: Top 200 trending topics during Sep. 2011 – Sep. 2012 with their global trend score and topic category annotation.

	Topic	Score	Topic Categories		Topic	Score	Topic Categories
1	olympics 2012	4378	sports	101	emmys	458	entertainment
2	Joint-stock company	1960	other	102	scarlett johansson nude pictures	452	celebrity, actor, incident
3	champions league	1874	sports	103	bundesvision	452	entertainment
4	iphone 5	1852	product, technology	104	bon jovi	452	artist, celebrity
5	whitney houston	1841	celebrity, artist, actor, entertainment, death	105	jeremy lin	447	sports, celebrity
6	Ramona Leiß	1819	celebrity	106	andy griffith	445	actor, artist, entertainment
7	mega millions numbers	1809	other	107	nfl draft	445	sports
8	closer kate middleton	1679	celebrity, entertainment	108	davy jones dead	442	artist, death
9	Spezial:Suche	1653	other	109	the voice	441	entertainment, show
10	facebook	1550	company, product	110	x factor	441	entertainment, show
11	dfb pokal	1485	sports	111	megaupload	441	politics, company
12	costa concordia	1409	incident	112	let it snow	438	other
13	black friday deals	1380	other	113	windows 8	437	technology, product
14	superbowl	1370	sports	114	sikh	432	politics
15	christmas	1345	holidays	115	lebron james	421	sports, celebrity

continued . . .

. . . continued

	Topic	Score	Topic Categories		Topic	Score	Topic Categories
16	steve jobs	1342	celebrity, technology	116	paul ryan	411	politics
17	schlag den raab	1285	entertainment	117	medaillen-spiegel 2012 london	410	sports
18	Manhattan	1281	incident	118	tupac coachella	409	entertainment, incident
19	Daniel Lopes	1221	celebrity, artist, entertainment	119	ann curry	405	celebrity, entertainment
20	academy awards	1204	entertainment	120	olympics 2012 opening ceremony	401	sports
21	formel 1	1145	sports	121	batman	400	movie, entertainment
22	justin bieber	1096	celebrity, artist, entertainment	122	junior seau	399	celebrity, sports, death
23	joe paterno died	1050	incident, celebrity, sports, death	123	dortmund bayern	398	sports
24	battlefield 3	1041	technology, product	124	chisora	389	sports
25	muammar gaddafi dead	1021	politics, death	125	gauck	388	politics, celebrity
26	Wikipedia:Hauptseite	1017	other	126	rose bowl	386	sports
27	ufc	1000	sports	127	sandusky	383	sports, incident, politics
28	iphone	1000	technology, product	128	groundhog day	382	holidays
29	happy new year	885	holidays	129	george zimmerman	382	politics, incident, death
30	kindle	877	technology, product	130	jim henson	379	celebrity, entertainment
31	ncaa brackets	859	sports	131	voice of germany	379	entertainment, show
32	em 2012	826	sports	132	chelsea	377	sports
33	amanda knox	817	politics	133	usain bolt	376	sports, celebrity
34	earthquake	802	incident	134	susanne lothar	376	celebrity, actor, death
35	mayweather vs ortiz	790	sports	135	media markt online	374	company
36	santa tracker	788	entertainment, holidays	136	dwight howard	374	sports, celebrity
37	thanksgiving	787	holidays	137	skyrim	371	technology, product
38	apple	761	technology, company	138	iowa caucus	371	politics
39	bayern muenchen	750	sports	139	deutschland griechenland	368	sports
40	tim tebow	743	sports, celebrity	140	sage stallone	367	celebrity, actor, death

continued . . .

. . . continued

	Topic	Score	Topic Categories		Topic	Score	Topic Categories
41	nasa curiosity	740	technology, incident	141	satellite falling	365	incident
42	ios5	735	technology, product	142	golden globes	363	entertainment
43	halloween	731	holidays, entertainment	143	google+	363	technology, company, product
44	solar eclipse	730	incident	144	empire state	362	incident
45	occupy	724	politics	145	demi moore	362	actor
46	bundesliga	712	sports	146	instagram	357	technology, company, product
47	grammys	711	entertainment	147	powerball	355	other
48	neil armstrong	708	celebrity, death	148	beyonce baby	350	celebrity, artist, actor
49	michael clarke duncan	704	actor, celebrity, death	149	easter	346	holidays
50	grass gedicht	692	politics, celebrity, incident	150	bayern	346	sports
51	titanic papst	660	entertainment, politics, incident	151	daylight savings time	341	incident
52	clint eastwood speech	655	politics	152	breaking dawn	341	movie, entertainment
53	diablo 3	652	technology, product	153	heidi klum	337	celebrity, entertainment
54	direct tv	642	technology, company, product	154	memorial day	335	holidays
55	us open	641	sports	155	tony scott	334	celebrity, death
56	schlecker schliessungen	638	incident, politics, company	156	zerg rush	334	other
57	masters 2012	634	sports	157	chelsea barcelona	333	sports
58	christian wulff	633	celebrity, politics	158	piratenpartei	333	politics
59	dark knight rises	631	movie, entertainment	159	the avengers	331	movie, entertainment
60	michael phelps	630	sports, celebrity	160	virginia tech shooting	330	incident
61	esc 2012	628	entertainment	161	ramadan	330	holidays
62	unwetter	621	politics	162	wimbledon	329	sports
63	kony	621	celebrity, politics	163	lunar eclipse	327	incident
64	supertalent	612	entertainment, show	164	chick fil a	323	company
65	mitt romney	610	politics	165	prince harry	323	celebrity, entertainment
66	trayvon martin case	609	politics	166	miami cannibal	322	incident
67	kim kardashian	606	celebrity, actor	167	lolo jones	322	sports
68	klitschko	598	sports	168	chris brown	317	celebrity, actor, artist

continued . . .

. . . continued

	Topic	Score	Topic Categories		Topic	Score	Topic Categories
69	colorado fires	590	incident	169	eli manning	317	celebrity, sports
70	marco simoncelli	588	sports, celebrity, death	170	heesters	316	actor, artist, celebrity, death
71	acta	587	politics	171	2012	316	movie, entertainment
72	pacquiao vs marquez	578	sports	172	daytona 500	312	sports
73	morgan freeman	576	celebrity, actor	173	brigitte nielsen	310	actor, entertainment
74	sopa	564	politics	174	ces	309	entertainment, technology, politics
75	wetten dass	558	entertainment, show	175	bachelor	307	show, entertainment
76	hurricane isaac	551	incident, politics	176	dschungelcamp	305	entertainment, show
77	babak rafati	541	sports	177	space shuttle	296	technology
78	wetter	538	politics	178	london 2012	294	sports
79	kim jong il	533	politics, celebrity, death	179	mayweather	293	sports
80	troy davis	529	politics, death	180	niedecken	291	artist
81	dns ok	525	other	181	rodney king	287	incident, politics, death
82	ipad	513	technology, product	182	deutschland italien	286	sports
83	lindsay lohan	507	celebrity, actor, artist	183	tarifverhandlungen oeffentlicher dienst	286	politics
84	hunger games	505	movie, entertainment	184	kentucky derby	281	sports
85	medal count	496	sports	185	Lost Generation (poem)	280	entertainment
86	ios6	495	technology, product	186	Bundesstrafgericht	280	politics
87	higgs boson	492	technology, incident	187	gideon sundback	278	other
88	valentine's day	489	holidays	188	kate	278	celebrity, entertainment
89	dan wheldon	488	sports, death, celebrity	189	guild wars 2	272	entertainment, product
90	bettina wulff	485	politics, celebrity	190	michael jackson	268	artist, celebrity, death
91	rick santorum	484	politics	191	4th of july	267	holidays
92	james holmes	477	death, incident	192	gedanken	266	other
93	bill cosby	471	celebrity, actor, artist	193	reno air crash	265	incident

continued . . .

. . . continued

	Topic	Score	Topic Categories		Topic	Score	Topic Categories
94	handball em	470	sports	194	kristen stewart	265	artist, celebrity
95	silvia seidel	467	actor, celebrity, death	195	aurora shooting	265	incident, death
96	ernest borgnine	467	actor, death	196	primark berlin	261	company
97	tebow	466	celebrity, sports	197	bundesliga live stream	260	sports
98	miley cyrus	465	celebrity, actor, artist	198	ray allen	258	celebrity, sports
99	mario balotelli	463	sports, celebrity	199	gamescom	255	entertainment, technology
100	world series	462	sports	200	san diego fireworks	249	incident