Department of Computer Science
Series of Publications A
Report A-2019-4

Understanding Social Media through Large Volume Measurements

Ossi Karkulahti

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, in Auditorium PII, Porthania building, on the 9th of October, 2019 at 12 o'clock.

University of Helsinki
Finland

Supervisor
Jussi Kangasharju, University of Helsinki, Finland

Pre-examiners
Mourad Oussalah, University of Oulu, Finland
Jari Veijalainen, University of Jyväskylä, Finland

Opponent
Yang Chen, Fudan University, China

Custos
Jussi Kangasharju, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645
ISBN 978-951-51-5508-5 (paperback)
ISBN 978-951-51-5509-2 (PDF)
Helsinki 2019
Unigrafia

Understanding Social Media through Large Volume Measurements

Ossi Karkulahti

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
ossi.karkulahti@helsinki.fi
https://cs.helsinki.fi/Ossi.Karkulahti/

PhD Thesis, Series of Publications A, Report A-2019-4
Helsinki, October 2019, 116 pages
ISSN 1238-8645
ISBN 978-951-51-5508-5 (paperback)
ISBN 978-951-51-5509-2 (PDF)

Abstract

The amount of user-generated web content has grown drastically in the past 15 years and many social media services are exceedingly popular nowadays. In this thesis we study social media content creation and consumption through large volume measurements of three prominent social media services, namely Twitter, YouTube, and Wikipedia. Common to the services is that they have millions of users, they are free to use, and the users of the services can both create and consume content.

The motivation behind this thesis is to examine how users create and consume social media content, investigate why social media services are as popular as they are, what drives people to contribute on them, and see if it is possible to model the conduct of the users. We study how various aspects of social media content be that for example its creation and consumption or its popularity can be measured, characterized, and linked to real world occurrences.

We have gathered more than 20 million tweets, metadata of more than 10 million YouTube videos and a complete six-year page view history of 19 different Wikipedia language editions. We show, for example, daily and hourly patterns for the content creation and consumption, content popularity distributions, characteristics of popular content, and user statistics.

We will also compare social media with traditional news services and show the interaction with social media, news, and stock prices. In addition, we combine natural language processing with social media analysis, and discover interesting correlations between news and social media content.

Moreover, we discuss the importance of correct measurement methods and show the effects of different sampling methods using YouTube measurements as an example.

Computing Reviews (2012) Categories and Subject Descriptors:

Human-centered computing → Collaborative and social computing → Collaborative and social computing systems and tools

General Terms:
Measurement, Human Factors

Additional Key Words and Phrases:
Social Media, Natural language processing, Wikipedia, Twitter, YouTube

Acknowledgements

I would like start by thanking my supervisor Professor Jussi Kangasharju for his spurring guidance and support. I am grateful to all my Collaborative Networking (CoNe) colleagues throughout the years. Thank you to my co-authors for all the publications we were able to produce together. For their valuable comments and time, I thank both pre-examiners.

I appreciate the support provided by Future Internet Graduate School (FIGS), Doctoral Programme in Computer Science (DoCS), and Helsinki Institute for Information Technology (HIIT). Thanks to the Department of Computer Science, Business Finland (formerly Tekes), and the Academy of Finland for funding my research.

Finally and above all, I extend my loving gratitude to Mengyuan, Olivia, and to the rest of my family.

Helsinki, September 2019
Ossi Karkulahti

Contents

Chapter 1

Introduction

The amount of user-generated web content has grown drastically in the past 15 years and social media services like Facebook, Instagram, Twitter, YouTube, and Wikipedia are exceedingly popular nowadays. Furthermore, the modern mobile devices are equipped with operating systems that are open for third-party software development and all more often with a high-speed Internet connection. Thus, we see that in future, even a larger part of the Internet content, be that blog entries, media, information, or entertainment, is created by users, instead of commercial or public entities. In social media, users can actively create and consume content and the threshold for publishing content is low. Users can connect with other users and also interact with and share their content. This is a development from early Web content that, in a way, resembled a bulletin board where mostly big companies and organizations posted content and the users passively consumed it. In general, social media users can be classified using three participation levels: the users who actively create and interact with content, the users who sometimes create and comment on others' content, and the users who passively consume content created by others.

In earlier research, social media has been measured in multiple ways. The popularity of content has been a common characteristic to many of the earlier social media measurement papers. Typically, the measurements have been conducted by collecting data, ranking the content based on e.g. number of views, then presenting the overall content popularity and seeing if it can be approximated with e.g. some form of power law distribution. Also, the content virality has been a subject of research interest that relates to the content popularity. Recently, sentiment analysis of the social media posts has been a rising topic. Another commonly researched topic has been the networks that the social media users form when connecting and interacting with each other. We argue that it is important to research and measure

content popularity on social media services in order to see how users create and access content and to find out if there are some patterns that could be useful in the future in terms of content distribution, replication, and storage. The focus of this thesis is on the content popularity and the content creation and consumption patterns. We want to examine how users create and consume social media content, investigate why social media services are as popular as they are, what drives people to contribute on them, and see if there is a way to model the conduct of the users. Also, we attempt to find out regional or cultural differences in the user behavior. Furthermore, we compare the user-generated content against the content provided by the commercial entities and analyze the similarities and differences. In addition, we want to see if the various aspects of the content be that e.g. the creation, consumption, or its popularity can be characterized and linked to real world occurrences.

We selected to measure three prominent social media services, Twitter, YouTube, and Wikipedia. Twitter is a social networking and a so-called microblogging service that currently has more than 300 million active users, 600 million visitors every month, and an overall reach of over a billion people [64]. Wikipedia is an online encyclopedia that by default anyone can edit. It has more than 40 million articles in over 200 languages and the service has more than 374 million visitors per month [62]. Our third service, is a video sharing service YouTube that according to the service itself has 1.9 billion logged-in users visiting every month [70] and in total users watch over a billion hours of video each day. The services were selected as they all offer different kind of user-generated content. Twitter offers short textual content ranging from casual status updates to news information and to public service announcements. YouTube has video content that varies from entertainment to educational and now even to live and paid content. Wikipedia on the other hand offers educational and informative content in form of text-based articles that can include figures and occasionally other graphics and audio.

An essential part of the research is the large volume data collection. We have gathered more than 20 million tweets during 2010-2013, metadata of more than 10 million YouTube videos and a complete six-year page view history of 19 different Wikipedia language editions. We believe that the amount of content will form a solid basis for our study. However, the data collection process was not completely straight-forward. The vast amount of content made us to choose our methodology carefully. While conducting our research and reviewing earlier work we saw that measuring large systems or services is challenging and typically measurements are performed

via sampling since analyzing the complete system is either prohibitively expensive or even impossible. Naturally, the way the sampling is performed has a strong effect on the measurement results and the conclusions that can be drawn from them. Ideally, the sampling should be done in a way that produces a random, representative sample of the total system, but in many cases technological limitations on the sampling may skew the process away from getting a representative sample. Using such a biased sample may yield incorrect conclusions about the properties of the system and further affect any derivative work which uses those results as its basis. Thus, while the priority of the thesis lies on understanding content creation and consumption, we also want to highlight the importance of using proper measurement techniques and show how different sampling methods can lead to different conclusions.

1.1 Research Questions

The main goal of the thesis is to understand social media content creation and consumption and in order to reach that goal we use the following research questions to guide our work:

- When and in which amount is content requested and created? We will want to examine what are the rates of content creation and consumption in Twitter, YouTube, and Wikipedia in order to produce e.g. hourly and daily patterns.
- What kind of content is the most created and consumed? We will want to know what are popular content types. We will use e.g. Twitter hashtags and keywords, Wikipedia classifications, and YouTube categories as the type identifiers.
- Are there cultural and regional differences? All the big social media services are very global and we will investigate if there are any differences in the user behavior between the various regions and cultures.
- Can the content consumption, creation, and popularity be modeled? Previously, e.g. the YouTube content popularity has been modeled extensively, but we want to investigate and see if we can apply further patterns to the content. This allows us e.g. to understand the user behavior, and, while out of the scope for this thesis, should be also useful in terms of content storing and distribution.
- How does the used data collection method impact the data and the results? Given the vast amount of the content and the corresponding

metadata, most of the earlier research on social media is based on some kind of sampling. We want to examine the effects of the selected sampling method on the results and conclusions. We both collect data and use earlier research results for the examination.

- How do social media services compare against traditional news services? We want to see what is the interplay between the news and social media services, for example, how people react to certain kind of news.

1.2 Research Contributions

In this thesis we present the following main contributions:

- We present metric measurements and analysis of three large social media services, namely Twitter, YouTube, and Wikipedia
- We compare social media with traditional news service
- We show how real-life events are portrayed on social media
- We combine natural language processing (NLP) with social media analysis, and discover interesting correlations between news and social media
- We show the importance of correct measurement method with YouTube as an example

1.3 Thesis Organization

In addition to the introductory chapter, this thesis contains an overview of the selected services, related work, and data collection process in the next chapter. In Chapter 3 we will focus on the consumption of Wikipedia content by examining in total 19 Wikipedia editions. Then, in Chapter 4, we survey Wikipedia activity against traditional news services and in Chapter 5 we track the interactions across business news, Wikipedia page views, and stock fluctuations. In Chapter 6 we will examine Twitter content creation e.g. in Madrid and Liverpool and in Chapter 7 we present a combined analysis of news and Twitter messages. Chapter 8 contains our work on YouTube, where we aim to show the importance of correct measurement methodology. Finally, Chapter 9 concludes the thesis.

1.4 List of Published Work

Chapters 8, 4, 7, and 5 of this thesis are based on the following publications:

I. Karkulahti, O and Kangasharju, J. "Youtube Revisited: On the Importance of Correct Measurement Methodology." Traffic Monitoring and Analysis. Springer International Publishing, 2015.

II. Du, M, Kangasharju, J., Karkulahti, O., Pivovarova, L., and Yangarber, R. "Combined analysis of news and Twitter messages." Proceedings of the Joint Workshop on NLP&LOD and SWAIE SemanticWeb, Linked Open Data and Information Extraction, 2013.

III. Karkulahti, O and Kangasharju, J. "Surveying Wikipedia activity: Collaboration, commercialism, and culture." Information Networking (ICOIN), 2012 International Conference on. IEEE, 2012.

IV. Karkulahti, O. & Pivovarova, L., Yangarber, R., and Kangasharju, J. "Tracking interactions across business news, social, and stock fluctuations." Advances in Information Retrieval. Springer International Publishing, 2016.

My contribution to the publications:

Articles I: I conducted most of the work alone. Analysis, conclusion, and writing were guided by professor Kangasharju.

Articles II & IV: I conducted social media data collection and data analysis alone. Overall data analysis, results, and conclusions were done together with other authors. I have not been part of the earlier development of PULS system.

Article III: The research work was jointly done with professor Kangasharju based on the data collected by Lasse Nordgren.

Chapter 2

Overview

In its simplest form social media can be seen to consist of all the social networks that operate over the Internet where users can produce and consume content. The birth of social media is attributed to the development of technologies that are loosely labeled under the term Web 2.0. It is common to many popular social media services that they only offer a platform or a tool and produce very little, if any, content by themselves. The term user-generated content (UGC) refers to the content created by Web users. The Organisation for Economic Cooperation and Development, OECD, provides three characteristic for UGC [49]. First, the content needs to *published*, be that on a public website or on a social networking site for a specific user group. Secondly, the content itself needs to be the product of *a certain amount of creative effort* in which the creating user adds his/hers own value to the content. The third characteristic for UGC is that it is produced *outside of professional routines and practices*. In their highly cited work [33], Kaplan and Haenlein couple Web 2.0 and UGC to describe social media as *"a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content"*.

In the same work, the authors also presented a model to classify social media services, which can be be seen in Table 2.1. The classification is composed of two dimension "Social presence / Media richness" and "Self-presentation / Self-disclosure" and three levels low, medium, and high. The first factor of the first dimension is social presence which is influenced by intimacy and immediacy of the communication in a service. Direct communication between two persons is more intimate than communication that is mediated and for example live video chat is more immediate than group messaging. Media richness, in turn, relates to the notion of information richness theory. In short, video with audio is a richer communication

Table 2.1: Social media service classification model [33].

		Social presence / Media richness		
		Low	**Medium**	**High**
Self-presentation / Self-disclosure	**High**	Blogs	Social networking sites	Virtual social worlds
	Low	Collaborative projects	Content communities	Virtual game worlds

medium than just audio since it allows more information and thus less ambiguity over the communication. In the second dimension, self-presentation stands for a concept which states that people usually want to give a positive and consistent impression of themselves for others and normally a person can give a deeper impression by the means of more self-closure and personal exposure. The three services that we focus on this thesis can be classified according to the model as follows: Twitter as a social networking site, YouTube as a content community, and Wikipedia as a collaborative project.

Interestingly, social media services seem to scale well to support and serve the needs of a wide range of users, e.g. casual users, celebrities, politicians, startups or huge multi-national organizations and companies. Normally, users produce content that other users can consume, share, and collaborate with. The users share their experiences, sorrows, and joys of life. Nowadays, it is common to see that even a big company and its brands are not presented by a single account, but instead accompanied with the individual accounts of its staff members. It is also not uncommon that companies want to attract or recruit popular users to promote or endorse their products. Thus, it is important to note that an increasing amount of content is produced by users that are paid or otherwise endorsed to do so. Also, in an increasing amount, the services providers want to differentiate their content from competitors, by having exclusive and original content. This kind of content monetization is making it harder to differentiate between UGC and professional content. Finally, it should be noted that, users can produce content solely for themselves, e.g. to store and archive family pictures and videos.

Most of the popular social media services are free to use and are funded by advertisements. However, the increasing amount of criticism and concern created by the intrusive information gathering to enable more and more accurate targeted advertisements and content have made the services to look for other sources for revenue. For instance, currently many popular third-party games on Facebook sell small items or other unlockables,

such as cosmetic in-game items, extra lives, levels, and characters. Naturally, Facebook will have its cut of the microtransaction fees. Twitter has a division that is providing an enterprise level analysis platform for a price and YouTube offers now also paid content and an ad-free subscription service. In contrast, Wikipedia's primary source of funding comes through donations.

2.1 Services

We have chosen three prominent social media services to be measured, namely Twitter, YouTube, and Wikipedia.

Twitter is a social networking and a so-called microblogging service that currently has more than 300 million active users, 130 million active daily users, 600 million visitors every month, and an overall reach of a billion people [64, 65]. The service was launched 2006. Addition to individual users, many commercial organizations, e.g. newspapers and TV stations, as well as public entities publish and promote their content through Twitter. In a similar vein, although the overwhelming majority of the users are regular people, the most popular ones, measured by the number of followers, are predominantly celebrities of some sort, that is, actors, athletes, musicians, politicians, etc. According to the company itself, 80 % of users are using mobile devices [64]. Due to its popularity, extensive research about its users and content has already been published, such as [39]. Twitter is an alluring research subject, not only because of its popularity and because it offers a free API access to data (with limits), but also as it is inherently light-weight, which makes it accessible to the different types of users. Until late 2017 the posts were limited to 140 characters, a legacy from the time when the system was thought to be SMS-based. The current limit is 280 characters. The character limit not only keeps the post short and simple, but also makes it acceptable and even preferable to post small messages but more often. On Twitter the relationships are by default directed, that is, user A can follow user B's posts without B following A's. The posts on Twitter are referred to as tweets and at the moment of this writing there are more than 500 million tweets created daily. We do agree that Twitter has a reputation of being a platform for people to utter what they had for breakfast or whatever trivial thing they are just doing. However, while at first this may sound meaningless, that is the kind of information that we first inquire for in many of you daily conversations. There seems, indeed, to be a yearn for social awareness. Besides, as our results will show, Twitter is used to share much more.

Wikipedia is the epitome of Internet-age collaboration and crowdsourcing. The online encyclopedia is free to use and, barring some time or otherwise sensitive articles, anyone can read, write, or edit any article. Currently, the service tops 40 million articles in over 200 languages and the service has more than 374 million visitors each month [62] who in total generate more than 16 billion page views [20]. Given its popularity, Wikipedia has become the standard, albeit sometimes criticized, reference of fact-finding in the common every-day use and has now found acceptance even in the educational and academic circles. The service launched in 2001 and in the fad-riddled world of Internet services, it has now been around more almost two decades and its popularity shows no signs of fading. Due to the popularity, extensive amount of research about Wikipedia has already been published starting from [67]. A lot of the early research concentrated on the credibility and accuracy of the Wikipedia articles. After that the focus shifted to establishing the motivating reasons for the editing and to measuring the content and its creation.

YouTube is a video sharing service which launched in 2005 and was acquired by Google in the following year. The service allows anyone to freely publish video content and according to the company itself, the service reaches over one billion people and e.g. in the United States more 18-49 year-olds than any cable TV network [70]. And while it may have be overtaken by Netflix in respect of overall traffic accounted for, and Chinese services such as Youku, Sohu, and iQiyi might soon, at least combined, have more users and views, YouTube is still without a doubt an exceedingly popular and widely recognized video service. In addition to the original website, YouTube is now available on smart phones, tablets, smart TVs, gaming consoled, etc. Mobile devices account for more than 70 % of all video views [70]. Financially, it is reported that YouTube's gross advertisement revenue might have been high as $5.6 billion in 2013 [19] and by 2019 it has paid out more than $2 billion to rightsholders after allowing content makers to claim money for their work in 2007 [70].

In Chapters 4, 5, and 7, we will compare the social media services with some traditional **news services**. In Chapter 4 we selected the online services of Helsingin Sanomat which is Finland's newspaper of record and the world-wide leading news provider BBC (British Broadcasting Corporation). In Chapters 5 and 7 we will use RSS feeds to collect data that contains news items from hundreds of news service providers.

2.2 Related Work

Research on social media services has been attracting increasing attention in various fields. Twitter has been founded to be a crucial source of information about public moods and opinions, for example, on topics of public concern such as political changes and elections [16], or revolutions [43]. Twitter is also found useful useful for monitoring of natural disasters and epidemics of infectious disease [41]. At the same time, Twitter is a problematic source since traditional NLP methods for information extraction, opinion mining, etc., are not directly applicable to very short texts, or texts using communication styles peculiar to social media [63].

Work similar to ours was reported in [60], that first used a fact extraction system to find events related to social unrest and cross-border criminal activity, and then tried to find additional information by using Twitter feeds. Authors of [5] trained a classifier to distinguish tweets that relate to real-word events from tweets that do not. They demonstrate that event-related tweets are quite rare; the majority of tweets do not contain events.

Kwak et al. [39] compared topics that attract major attention on Twitter with coverage in other sources, namely, Google Trends and CNN headlines. They have found that Twitter can be a source of breaking news as well. [72] used topic modelling to compare Twitter with the New York Times news site. They found business being among the top-10 topics on Twitter, but also mentioned that business-related tweets rarely express opinions.

Kruger et al. [37] manually analyzed 500 random tweets related to Adidas, and came to the conclusion that the company uses Twitter to promote their brand. The authors of [31] manually prepared a list of companies and brands belonging to different business sectors, and then collected tweets related to these companies and brands. They demonstrate that approximately 20% of tweets contain mentions of companies or brands, which means that Twitter is an important marketing medium, however, only 20% of the tweets that mention companies and brands express a sentiment about them.

Recently, the use of mobile devices as a "second screen" to augment TV consumption has been widely studied. According to Smith [56] already in 2012 over a half of adult cell phone owners in the US used their phones while they were watching TV. Schirra et al. [55] found that especially scenes with sadness/grief, character growth, and humor prompted users to tweet while watching TV. They also concluded that users who are watching alone are particularly motivated to tweet. Pittman and Tefertiller [52] collected

tweets using hashtags relating to TV shows and conluded that asynchronous streaming Netlix shows resulted in more engagement than the traditionally broadcasted TV shows. Giglietto and Selva [22] collected more than two million tweets relating to political TV shows and found that "the use of Twitter to express the viewers' personal opinions on the show is the most frequent in our sample". Leng et al. [15] examined the role of Twitter as a second screen during the end of the National Hockey League (NHL) regular season and playoffs in 2015. They presented temporal and spatial analysis of Twitter usage and concluded that "majority of these tweets are done using mobile devices, that the tweeting actively is heavy tailed, and roughly half of the tweets are retweets" and that "the usage patterns provide clear evidence that Twitter is used for real-time second screen usage". They also found that goals typically create seven to eight time spike compared to in-game baseline. Yu and Wang [71], in turn, collected tweets in real-time during five 2014 FIFA World Cup games and found that "tweets were used to express joy and anticipation and to express emotions" and "U.S. fans' fear and anger were common and reflected U.S. team's goals or losses".

Gabielkov et al. [21] studied news sharing and reading behavior of Twitter users highlighting "the ability of social media to cater to the myriad taste of a large audience". In their work Johnson and Yang surveyed 242 persons and concluded that social motives and information motives were the two largest factors for using Twitter [32].

Wikipedia has been studied a lot through out its 15-year existence. Its accuracy and reliability has been heavily scrutinized, most notably by [23]. The motivating factors for users to spend their time to contribute have been covered in many works. Nov [48] found that the strongest factors are fun and ideology and that there is a correlation between motivation level and contribution amount. Another survey [38], answered by around 100 PhD students, listed "to educate humanity/raise awareness", "feels like I'm making a difference", and "to give back to the Wikipedia community" as the main reasons to contribute. Rask [53] concluded that level of human development is a stronger factor for increased contribution than level of technological development. The motivation searching research has a lot in common with the similar work done in the area of open-source software development.

The findings from [30] state that most of the information to Wikipedia comes from online sources and Lih [42] points out that there is a linkage with edit-peaks of articles and news coverage on the topic.

Research done by Voss [67] showed that the number of articled per author follows the power-law. The research included a lot of statistics

which indicated that the growth of Wikipedia had been exponential from 2002 to 2005. According to [40], the traffic (page visits) in the English Wikipedia does not follows the power-law distribution after the 1000 most visited articles.

Kittur et al. [35] studied English Wikipedia during 2001-2006 and proposed that the influence of the biggest contributors had decreased during that time. Adler et al. [1] have researched how to measure the quality, and not only the quantity of the contributions to the Wikipedia.

Cha et al. [10] analyzed the video popularity of YouTube in 2006-2007. Their dataset consists of video metadata formed by crawling the indexed pages and getting videos belonging to certain categories. They had 1.7 million videos from Entertainment category and another 250,000 from Science category. Their results showed that the video popularity ranking of both categories exhibited power-law behavior "across more than two orders of magnitude" with "truncated tails" but "the exact popularity distribution seems category-dependent." The authors called for further research on the subject. The traces collected by the study have been a source for [66].

Cheng et al. [12] also measured and examined, among other things, the popularity of YouTube videos. They collected metadata for three million videos in 2007 and for further five million in 2008, using breadth-first search (BFS) starting with an initial video and asking its related videos and then their related videos until the fourth depth. Looking at video popularity they observed that: "though the plot has a long tail on the linear scale, it does not follow the well-known Zipf distribution." and found "that the Gamma and Weibull distributions both fit better than the Zipf, due to the heavy tail that they have". Since the authors were concerned that the BFS method would be biased towards more popular videos, they formed another dataset by collecting metadata of videos from the recently added list during a four-week time window. Comparing the two datasets they concluded that the videos from the recently added list exhibit popularity where: "There is a clear heavy tail" and "verifying that our BFS crawl does find non-popular videos just as well as it finds popular ones".

Szabo and Huberman [59] took a slightly different approach and wanted to see whether it is possible to predict content popularity. In the case of YouTube they measured the popularity and view counts of new videos for 30 days. Their data is from 2008 and consists of 7,146 videos selected daily from the recently added list. They chose the list over other alternatives in order to get "an unbiased sample". They concluded that the popularity of a YouTube video on the 30th day can be predicted with a 10 % relative error after 10 days.

In the research mentioned above, the data has been collected either by BFS crawling, or by selecting videos of a certain category or by picking most recent videos. We will show in the results section the problems that are associated with the methods and popularity distributions they produce.

Another method is used e.g. by Gill et al. [24] who analyzed the traffic between a university campus and YouTube servers. They concluded that "video references at our campus follow a Zipf-like distribution". They reasoned it to be partly because YouTube did not allow video downloading, meaning that a user had to issue another request to see the same video again. They also found out that on a longer time frame the most popular categories were Entertainment, Music, and Comedy. Che et al. [11] found that in 2013 the two most popular categories were Music and Entertainment while Gaming was the third most popular.

Bärtl [2] studied YouTube videos between 2006-2016 and observed an "overwhelming dominance of very few channels over the rest of content on YouTube" and also concluded that "findings from social media data can differ dramatically, depending on the data collection method, the time frame covered and the analytical approach".

Zink et al. [74] also measured the YouTube viewing and traffic patterns on a campus level and studied the effects of proxy caches to reduce traffic. Khan [34] surveyed YouTube users and analyzed the different motives for user to participate on the content and passively consuming it.

The authors of [50] analyzed the content history of YouTube through more than 76 million videos and were able to identify patterns in the content related to internal and external events.

On a more general level, the importance of a correct sampling method has been noted e.g. by Krishnamurthy et al. [36] who used three different data collection methods and analyzed their strengths and weaknesses in order to examine Twitter and improve the prior research, and by Stutzbach et al. [57] who introduced a technique for a more accurate and unbiased sampling for unstructured peer-to-peer networks.

2.3 Data Collection

There are a few commonly used ways to gather data from social media services. The best and the most comprehensive way would be to obtain a complete dataset of a service. This is, however, in many cases impossible because of commercial reasons or simply because data, or some parts of it, is inaccessible or too vast to process. Probably, the most typical way to collect data is to request it through an API (Application Programming

Interface) provided by the service in question. An API response can be metadata or actual content data. The metadata usually includes descriptive information and statistics about the content. For example, a YouTube API response is metadata rather than the actual video file, containing the video's characteristics like id, duration, category, and view count. The Twitter API returns both metadata and also the actual tweet text. The API's are usually rate-limited, meaning that only a certain number of queries can be made per a unit of time (usually a second or an hour). The third way, is to collect data 'somewhere from the middle'. For instance, in the past YouTube videos and their access could monitored e.g. at a university's campus network cache-level. However, the increasing use of HTTPS protocol is hindering this approach. The fourth way is to get the data from an external source.

Typically system measurements are performed via sampling since analyzing the complete system is either prohibitively expensive or even impossible. Sampling us usually done by periodically querying the API of a service using parameters that will yield the desired sample. Commonly, a query is formed based on only one or few hand-crafted test queries, then executed hundreds or thousands of times automatically where the frequency is normally dictated by the rate-limit and later the data is checked for its validity.

In our research, we used the APIs of Twitter and YouTube, whereas the Wikipedia data is compiled from an external page request log. Both YouTube and Twitter APIs are rate-limited. During our measurements, the APIs also evolved to restrict access to only authenticated requests, meaning that we needed to register for API keys. An API key is then included in each query and rate-limits are then applied on it. This influenced our data collection process quite heavily. Before the queries were authenticated, the rate-limits and other restrictions were based on the IP address where to query originates. That allowed us to utilize a computing cluster with 200 nodes so that the rate-limits and restrictions would apply to all nodes individually. In other words, compared to a single machine, the cluster offered us 200-fold access to the API. However, after the mandatory change to authenticated requests we could still use the cluster, but all the queries were tied to a single API key. Both services offer API keys free of charge, while commercial options with more favorable restrictions also exists.

In addition to the rate and access restrictions, both services limit the amount available data in other ways. Neither service provide a full-access to data. Twitter offers only data about tweets that are newer than 7 to 9 days and while the YouTube API allows historical travelling, only a small fraction of daily data is ever accessible. That is, we could make a query

where we ask for videos published between e.g. April 1-10, 2018, but the API would only return a certain number of 'pages' (iterations of the query) after which there was no way of getting more data for that specific frame of time.

Since we discovered that the access to data is limited quantitatively and time-wise, one of the very first questioned we needed to tackle was how should the data be gathered in order to get a representative sample of a system. We will present a detailed description of the data collection process for each service in the corresponding chapters. We also collected data from traditional news services. In short, we collected daily news text and related metadata from multiple news services using the RSS feeds that the news services provided.

Chapter 3

Wikipedia

In this chapter we focus on the consumption of Wikipedia content. We have chosen 15 Wikipedia editions (languages) with the most articles and hand-picked four more editions[1] due to research interest. We hypothesize that the selected editions should found a substantial basis for a comprehensive survey. While the Wikipedia traffic and page views have been examined before, to the best of our knowledge this is the first survey on this scale. As mentioned in Section 2.3 the source data for Wikipedia is the most comprehensive of the three services measured in this thesis and thus offers a good starting point to examine the social media services and the corresponding results.

Regarding to Wikipedia, our research points of interest are:

- When and in which amount are pages requested?
- Are there temporal patterns in the content consumption?
- What kind of articles are the most requested ones?
- Are there differences in e.g. article consumption patterns and content popularity between editions and are the differences culturally linked?

We analyzed the data using two different approaches. First, we formed various patterns based on time. The second approach is popularity-based. That is, we grouped the data e.g. by month and ranked the articles according to the number of times they were requested. The results are presented in two sections, but first we will describe the data.

[1] Arabic, Finnish, Korean, and Norwegian (bokmål)

Table 3.1: Dataset description.

Years	Editions
2008-2014	ar, ca, de, en, es, fi, fr, it, ja, ko, nl, no, pl, pt, ru, sv, uk, vi, zh

3.1 Description of Data

The basis of our survey is the extensive raw data from Wikipedia's affiliate statistic website[2]. We collected 15 TB of data which covers seven years (2008-2014). The data was collected between 2012 and 2015. The dataset overview is shown in Table 3.1. The raw data is divided to files by hour (24 files per day) and each file contains data in a row format where the first column indicates the edition in question, the second column is the title of the page, and the third column is the number of non-unique requests during the hour. So, for example the following line:

en Main_Page 242332

says that the main page of the English Wikipedia was requested over 240,000 times during the specific hour implied in the filename of the raw data file. In other words, the raw data tells us how many times each article was requested per hour. The raw data covers each article and every language edition of Wikipedia, but as mentioned, we will concentrated on 19 editions, which are: Arabic (ar), Catalan (ca), Chinese (zh), Dutch (nl), English (en), Finnish (fi), French (fr), German (de), Italian (it), Japanese (ja), Korean (ko), Norwegian (no), Polish (pl), Portuguese (pt), Russian (ru), Spanish (es), Swedish (sv), Ukrainian (uk), and Vietnamese (vi).

The raw data is not perfect, for some periods the data is missing and occasionally we were required to clean the data when we have deemed it to be erroneous. For example, we observed that one time the raw data implied that the English edition's monthly views would have grown by a factor of 10,000 from previous month, which we see unrealistic or as a result of some kind of denial-of-service type of an attack. Examining such situation would not match our scope which is to study how the data is consumed by the end-users. Please note that we use the terms 'page request' and 'view' quite interchangeably and that we have left out any page request which did not refer to an existing Wikipedia article.

As a reference, Table 3.2 shows the numbers of articles and edits for all the surveyed editions, as of Feb 2013. The edits are counted from the set up of the edition and they give a general idea how actively the edition has

[2] `http://dumps.wikimedia.org/other/pagecounts-raw/`

Table 3.2: Wikipedia statistics: number of articles and edits as of February 1, 2013 [68].

	N of Articles	N of Edits
ar	211K	12M
de	1.5M	119M
es	966K	68M
fr	1.4M	88M
ja	849K	47M
nl	1.2M	35M
pl	950K	35M
ru	966K	60M
uk	429K	12M
zh	671K	26M

	N of Articles	N of Edits
ca	394K	11M
en	4.2M	589M
fi	315K	13M
it	1M	61M
ko	230K	12M
no	368K	12M
pt	770K	35M
sv	770K	20M
vi	576K	10M

been modified. In addition, we used the resources provided by the excellent DBpedia community [14]. Their data of article types and categories is especially useful when we examine the most requested articles.

3.2 Time-based Results

In this section we will present result from our study that are all related to time, please note that all times are in UTC. We have divided the 19 editions into five groups:

A) Catalan, Spanish, French, Italian, and Portuguese

B) Japanese, Korean, Vietnamese, and Chinese

C) Finnish, Norwegian, and Swedish

D) Polish, Russian, and Ukrainian

E) Arabic, German, Dutch, and English

Group A includes the Romance language editions. Catalan is the smallest of the editions and the French edition has the most articles. French, Spanish, and Portuguese are all spoken in multiple countries and timezones, whereas Italian and Catalan are more local languages.

Group B comprises four Asian languages. Interestingly, while Chinese and Korean are officially only spoken in time zones UTC +9 and UTC +8 respectively, they both have large diasporas especially in North America,

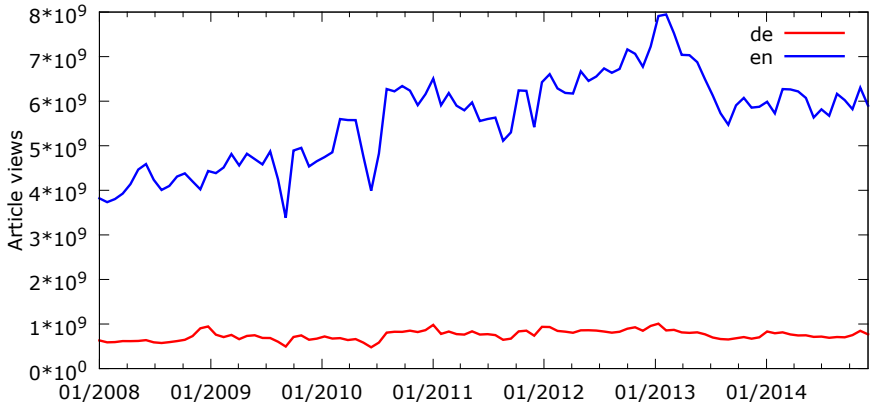


Figure 3.1: Wikipedia views per month for 2008-2014.

which a priori could influence the results. Also, the access to the Chinese edition has been periodically completely blocked and some articles are permanently not accessible in the mainland China.

Group C has languages from three Nordic countries that can be seen culturally very similar. In Group D are three editions from the Slavic language family. Finally, Group E consists of somewhat arbitrarily of four editions of which Arabic has speakers spanning over multiple time zones and English is the current global lingua franca and clearly the largest Wikipedia edition.

Let us start by looking at the big picture. Figure 3.1 plots the number of page request per month for the English and German edition throughout the surveyed period. As seen, the numbers for the English edition vastly exceeds the German one, which was the second biggest Wikipedia edition at the beginning of our survey. We can see that, in 2008, the mean for the English edition is roughly 4 billion monthly article views. A year later it is almost 1 billion more per month and in 2011 there were occasionally over 6 billion monthly requests. The highest peak comes in early 2013 where the requests peaked at close to 8 billion settling back down to the 6 billion at the end of 2014. Given that the English Wikipedia is overwhelmingly the largest one, it will be our primary interest and we try to be careful not to compare the English edition directly with the others as the size difference is obvious. The figure also reveals the nature of the raw data. The clearly visible sudden drops on the English edition's curve (e.g. Sep 2009) are caused by missing or lacking raw data. We wanted to show the drops as they were, for the sake of possible derivative work.

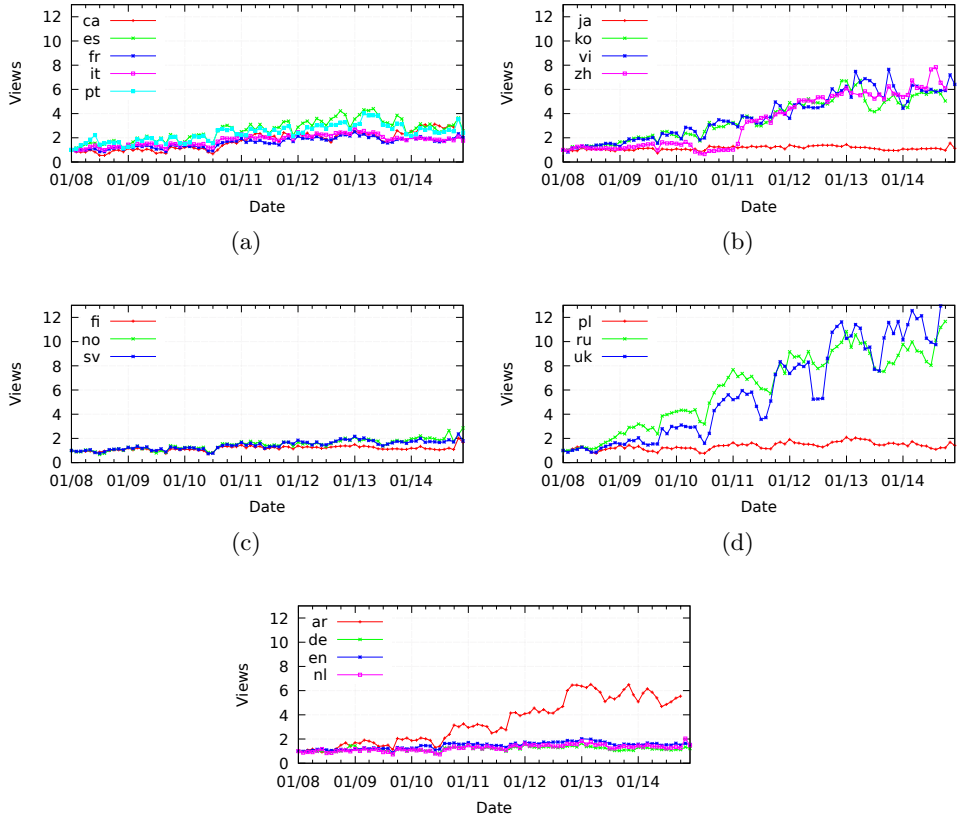


Figure 3.2: Page request evolution 2008-2014.

Figure 3.2 plots the page request evolution over the survey period for all the editions by monthly total views. Because of the difference of sizes of the editions, we use an index on the y-axis where 1 equals to the number of views on January 2008. The absolute page request numbers can be seen in Table 3.3. In Group A all the editions have doubled their monthly views and the Portuguese and Spanish exhibited occasionally four times as many views when compared to the beginning. In Group B the viewing for the Japanese edition has been relatively steady whereas the other editions have grown drastically, all having around six times as many views in 2014. The subplot for Group C shows that the Norwegian and Swedish editions have doubled in views whereas the growth has been more stale for the Finnish version. In Group D we see that the Polish version has periodically seen

Table 3.3: Number of articles and monthly requests in Jan 2008.

Edition	N of articles 2008/01	N of monthly request 2008/01
ar	68,000	15M
ca	98,000	9M
de	741,000	632M
en	2,100,000	3821M
es	318,000	300M
fi	149,000	40M
fr	605,000	296M
it	400,000	188M
ja	463,000	799M
ko	52,000	10M
nl	404,000	106M
no	150,000	21M
pl	457,000	226M
pt	346,000	112M
ru	230,000	70M
sv	268,000	47M
uk	91,000	4M
vi	29,000	7M
zh	163,000	39M

twice as much views when compared to the beginning of the period, but more strikingly the Russian and Ukrainian editions show multi-fold increase and interesting yearly cycle where page requests decrease sharply during the summer months. Finally, looking at Group E we see that the page request patterns for the Dutch, English, and German editions are similar, whereas the Arabic version has grown and is receiving approximately five times as many request at the end of the period than in the beginning.

3.2.1 Overview and Page Request Evolution

Overall, we have seen a rise in the number of page requests across the board. With the Chinese, Korean, Russian, Ukrainian, and Vietnamese editions, the number of page requests has grown multi-fold from 2008 to 2014. Not surprisingly, the growth of the Asian and Eastern European editions seems to concur with corresponding economical and technological development of the countries in question.

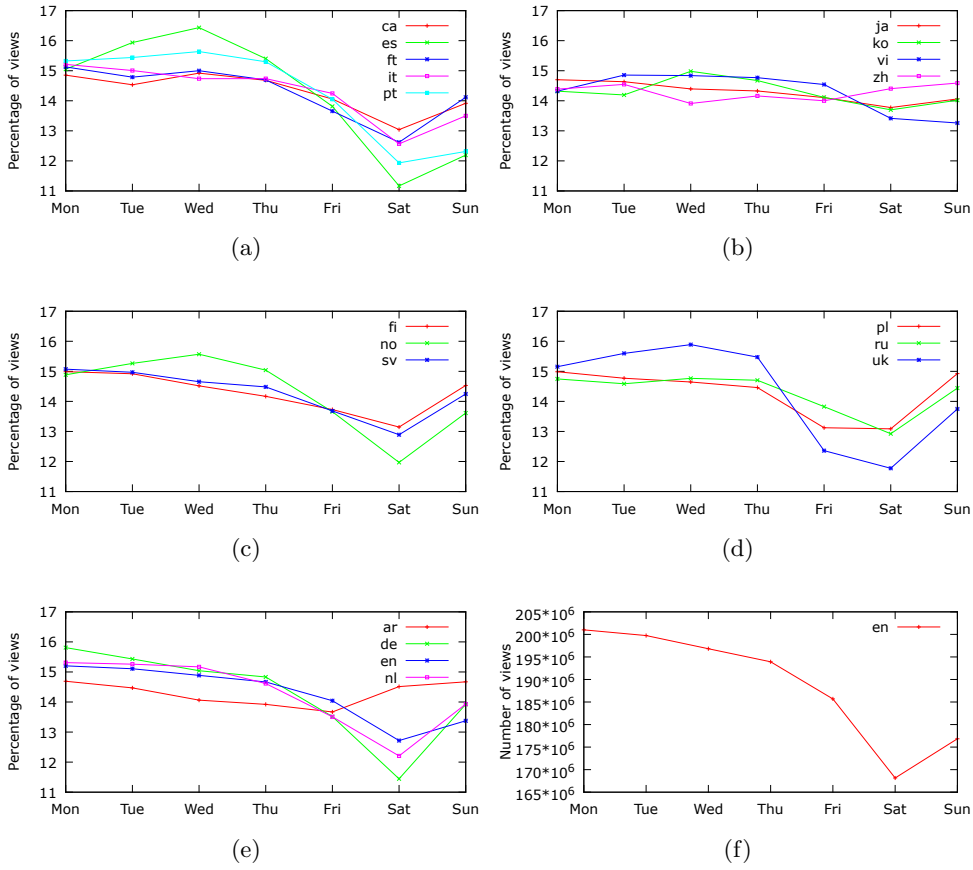


Figure 3.3: Daily Patterns.

3.2.2 Daily Pattern

Having seen the overview, let us move further. Figure 3.3 shows the daily distribution of the page requests for the different editions. In each subplot the value on y-axis stands for the percentage of page requests for the day indicated on the x-axis. In Group A the overall pattern is the same for the Catalan, French, and Italian editions where Saturday is the least active day. The Portuguese and Spanish versions exhibit even more distinct decrease in the number of page request during the weekends. On contrast, the page request distribute more evenly for all days when looking at the Asian language editions, where only the Vietnamese edition sees a drop during the weekend. The Finnish and Sweden editions show a slight drop in the request on Saturday, while it is more noticeable for the Norwegian

edition. Similarly, the Ukrainian version exhibit the drop more clearly than the other two languages in its group. Interestingly, the Arabic edition behaves like the Asian editions and the page requests increase from Friday to Sunday, which might be explained the way the day-offs fall in many Arabic speaking countries. Subfigure 3.3f plots the actual daily page request values for the English editions which range from the Mondays' over 200 million to the approximately 170 million occurring on Saturdays.

Of course, drawing too many conclusions just from the metric data is difficult, but there are noticeable the differences between the editions from the different regions and cultures. Overall, Saturday sees the least amount of page requests.

3.2.3 Hourly Pattern

Let us now look how page request behave on the hour-level. From Figure 3.4 we can see the hourly distributions for the editions where y-axis value stands for the percentage of page requests made during the hour marked on the x-axis. Overall, the most noticeable feature is the flatness of the English curve. With all other editions the page requests drop during 'night time' (remember that times are in UTC). This indicates that the viewing of the English Wikipedia occurs quite evenly across the different time zones.

In Group A the drop happens later with the Portuguese and Spanish version which we believe is caused by the influence of the South American users, whereas the French edition exhibits a pattern more fitting to European and African time zones. In Group B we see that the Vietnamese version interestingly zigzags from 02:00 to 17:00 for which we have no explanation. The editions of the C and D groups behave quite similarly with their respective group counterparts, which concurs with editions cultural and geographical similarity. In general, it is interesting to observe that e.g. with Groups C and D, pages are requested approximately evenly during the afternoon as in the evening, implying that Wikipedia is used during school and work hours.

The Subplot 3.4f shows the actual numerical values for the English edition and we see that mean of page requests range from a bit over 6 million to over 9 million depending on the time of the day. Most traffic is observed between 16-23 UTC.

3.3 Popularity-based Results

Thus far we have seen results that were all related to time, but in this section we are going to present results that are based on the popularity of

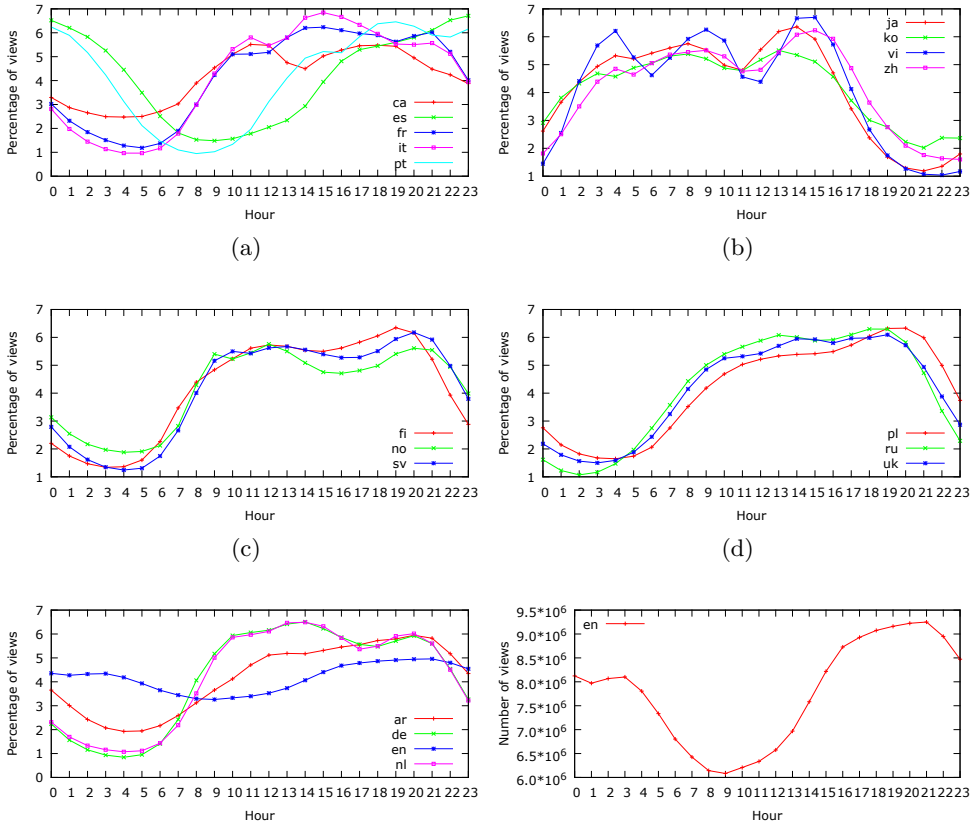


Figure 3.4: Hourly patterns.

the articles. While the format of this thesis is not suitable for presenting top lists as such, we will focus on modeling popularity on the edition level and aim to examine the top articles and also, where possible, discuss about possible reasons for popularity. We define a top article as an article that is among the edition's 1000 most popular. The popularity is based on the page requests. We will use the same grouping of the languages editions as with the time-based results.

3.3.1 Variation in the most Popular Articles

Figure 3.5 plots the fluctuation of top articles from month to month. That is, for each month for 2008-2014 we calculated the 1000 most requested articles and in the figure the y-axis shows the number of different articles

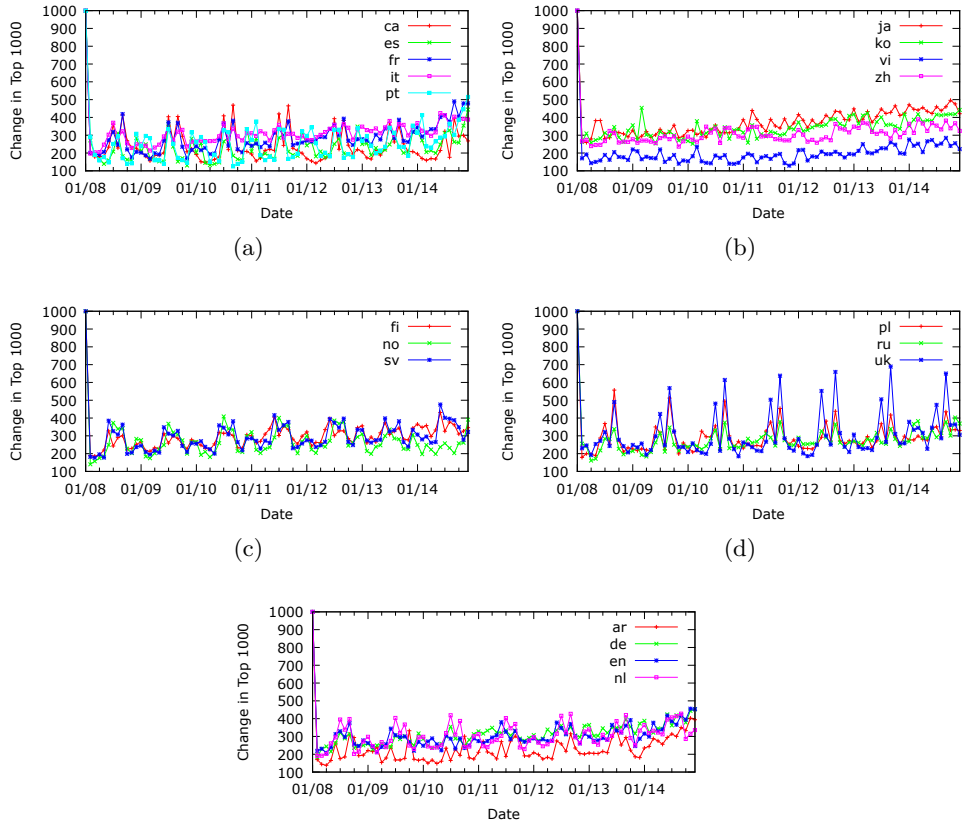


Figure 3.5: Article fluctuation in top 1000.

in the top 1000 most requested compared to the previous month. The 1000 is an arbitrary choice and difference between the articles around the 1000 rank can be only few views, but the metric is used as it allows us to compare the different editions.

The most striking feature of the plots is the double-peaking that happens mid-year with various editions. It is most clearly visible with the Ukrainian and Polish editions, but also is noticeable e.g. with all the Nordic editions. Taking the Ukrainian edition as an example, the curve indicates that every year during June-September up to 600 of the 1000 most popular article are different from the previous month, the change rate lowers for a month or two and then again more than half to top articles change again. We believe that it might relate to ending and beginning of school

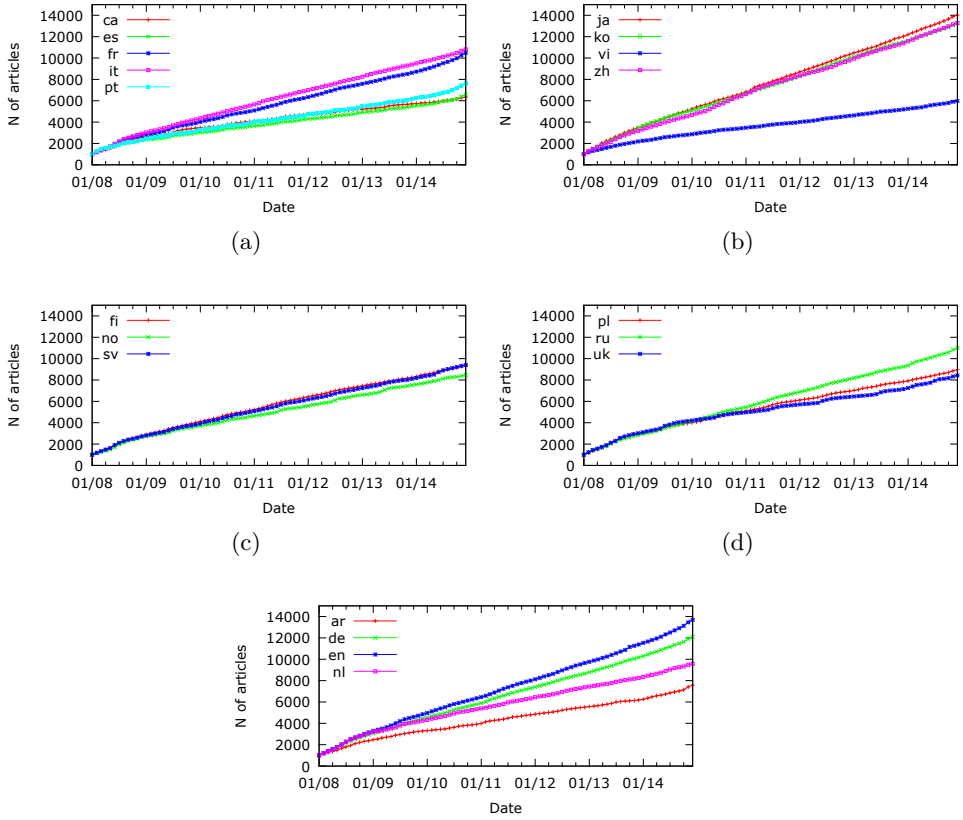


Figure 3.6: Total number of articles in top 1000.

and university years and their final and entrance exams, but at this point we have no evidence to support that nor can we offer any reason why it would concern only certain editions. As seen there are noticeable similarities in the patterns. All the curves in Group C have strong correlation with each other, as is the case with Group D. For reference, Table 3.7 shows a complete correlation coefficients matrix (Pearson).

Figure 3.6 shows how fast the total number of top articles increases over time. In other words, the line shows over time how many articles have been at least once in the monthly calculated list of 1000 most popular articles. The group A splits into two, the French and Italian editions behave almost identically both having in the end more than 10,000 top articles and the Catalan, Portuguese, and Spanish editions all have quite similar

lines and less variation in the top articles. The Vietnamese edition has significantly less top articles over time than the other Asian editions. The Nordic editions behave similarly, exhibiting all the same amount of variation in the top articles. With Group D we see that the Russian edition differs from the other two, having more variation in the top articles. The English edition show similar high variation as the Chinese, Japanese, and Korean versions, while the variation is slight lower for the German version. The Arabic version has the lowest number of top articles in the end of the group and the Dutch edition falls somewhere in the middle.

There are a few interesting points in the results. First, while e.g. the Ukrainian edition showed the most variation when compared by month, its total top variation is actually just average when compared against the other editions, meaning that although the top articles change often, the pool of popular articles stays medium. We also see that Nordic editions behave almost identically indicating that, at least in some cases, the cultural and geographical locality can be seen in the behavioral habits of the users. Overall, seeing that the English edition has a lot variation in the popular articles is not surprising given its wide reach of users, but we also see that large and widely used editions like the Portuguese and Spanish have much less variation.

3.3.2 Characteristics of Popular Articles

Table 3.4 characterizes popular articles by four variables. The table is compiled so that again we calculated the 1000 most popular article for each month and then looked the popularity of the articles over time in more detail. Table 3.5 has an example. The article *Hard disk* in the English edition was the 50th most requested article in Jan 2008. One month later it was the 292 most requested. In Aug 2008 it was not among the top 1000. The example only shows the year 2008, but all the calculations were done for the whole period (2008-2014). Moving back to Table 3.4, the first column shows the total number of articles that were at least once in a month's 1000 most popular, the second column tells the percentage share of those articles that were always ranked among the top 1000, third column shows how much such articles cover of 1000, and the last column, in turn, gives the amount of the top articles that were in the top 1000 only one time. In other words, the table shows how stable is the set of the most popular articles in each edition.

The variation within the groups is noticeable. In Group A, the French edition has more than 10,000 top articles of which only 88, or 0.8 %, were every month within the 1000 most requested and 47 % of the top articles

Table 3.4: Characteristics of popular articles.

	Total N	Every time %	/1000 %	Only once %
ca	6345	2.3	14.7	39.9
es	6576	2.2	14.4	39.4
fr	10483	0.8	8.8	47.0
it	10812	0.9	9.6	42.3
pt	7622	1.4	10.9	41.6
ja	14021	0.5	6.8	46.9
ko	13279	0.5	6.9	39.7
vi	5972	2.3	13.8	30.6
zh	13283	0.5	7.3	45.0
fi	9376	1.5	14.5	43.0
no	8478	0.8	6.4	38.6
sv	9403	1.1	10.7	44.5
pl	8964	1.3	11.7	42.9
ru	10993	0.7	7.7	49.0
uk	8434	0.7	5.8	35.0
ar	7582	1.3	10.2	41.0
de	12088	1.2	14.0	55.0
en	13690	0.7	10.2	50.3
nl	9577	1.4	13.0	43.5

Table 3.5: Example of a top article.

2008	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Hard disk	50	292	172	119	671	679	573	-	-	-	746	-

only featured once in the monthly top 1000. In contrast, the Catalan and Spanish editions have more than 2 % the top of articles constantly in the top 1000 for the whole seven year period, meaning that approximately every seventh article in the top 1000 stayed there for the whole time. In Group B, the Vietnamese version differs from the others by featuring more ever-present top articles than the others. With the Finnish edition every seventh top article remained for the whole period, while with Norwegian version it happened only with every 16th article. The Ukrainian edition has one of the lowest number of ever-present articles, but has also only a small share of articles that featured in the top 1000 only once, which is consistent with findings from the previous section. The numbers for Group E tell that the English edition has few ever-present articles and half of the top articles were in top only once, while the German version has the highest rate of such articles.

3.3.3 Popularity Distributions

Let us know examine the editions' popularity distributions. Figure 3.7 shows how the editions' total views accumulate along with the popularity. The views are counted for the whole survey period and the articles are ranked based on the counts. The plots read so that e.g. looking at the Spanish edition in the Group A, we see that the most popular 10 % of the articles account for approximately 90 % of the editions total views. Interestingly, the Catalan edition differs from the other in Group A. Likewise, the Korean edition exhibits similar difference in Group B.

Lam and Riedl [40] examined the traffic (page visits) of the English Wikipedia and saw that it does not follows the power-law distribution after the 1000 most visited articles. They concluded that the page view distribution of the English Wikipedia follows a log-normal curve. Subfigure 3.7f shows our data plotted similarly and the seen distribution agrees with their finding, we notice especially the head and tail deviations.

Overall we see that a small part of the content create the vast majority of the page view requests. With all editions, the most popular 10 % of the articles create over 50 % of the total page views and with many editions the most popular 10 % are responsible of over 80 % of the views. In all cases most 90 % or more of the page views correspond to the most popular 50 % of articles. In general, the majority of Wikipedia traffic is caused by popular articles.

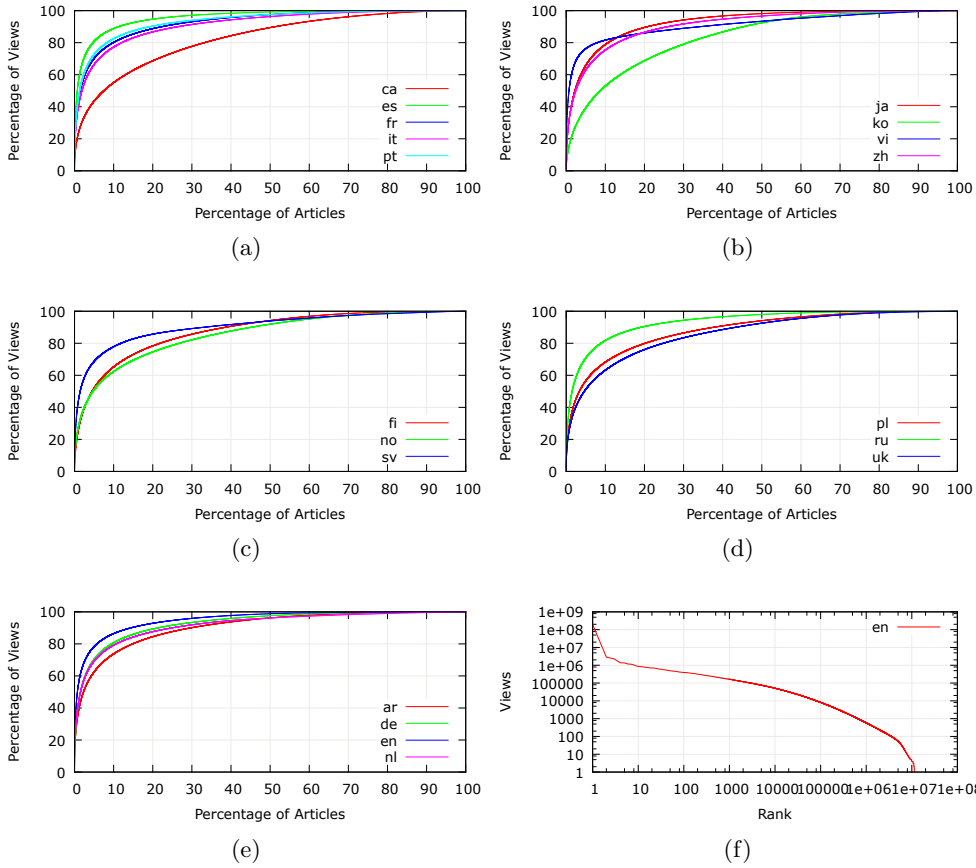


Figure 3.7: Popularity distributions.

3.3.4 Popular Article Types

Thus far, we have seen only numerical results. We will now briefly analyze the popular articles in the English edition by their type. We will use the meta-data information of the Wikipedia articles provided by the DBpedia community [14]. Unfortunately, the data was only available for the English language edition.

We used our previously calculated top article data and took the 1000 most popular articles of the whole survey period and matched them with the DBpedia data. We were able to get a type for 898 of the 1000 articles. As seen in Table 3.6 the single largest type is Person by a clear margin. As the Person type does not include the numbers for the types such as Musical Artist and Office Holder, it is safe to assume that at least every fifth top

Table 3.6: Top article types of English Wikipedia.

	Type	**N**
1	Person	184
2	Thing	166
3	Television Show	93
4	Country	84
5	Film	33
6	Musical Artist	27
7	Band	25
8	Office Holder	24
9	Disease	23
10	City	11

article is about a person. The other large type is Thing which seemed to be generic type for articles that were lacking better classification. Other popular article types are the entertainment related Televion Show, Film, and Band. Many articles about countries and cities are also be among the popular content.

3.4 Summary

In this chapter we presented a large measurement study of Wikipedia content consumption. We studied 19 editions and saw that the English language edition is by far the largest edition of Wikipedia and the edition had doubled the monthly views during the survey period. The number of monthly page request had increased drastically for many editions, the Ukrainian and Russian editions have seen an increase by the factor of ten. We were able to observe yearly, daily, and hourly patterns in the page requests. The editions did vary in many cases, but overall we could identify groups of editions based on similarity. The Nordic editions, for example, exhibit quite similar behavior in most aspect that we measured, which would indicate a link from the online behavior to the cultural similarities.

The popularity distributions revealed that generally a small part of the content causes most of the page request. With many editions the most popular 10 % of articles are responsible of over 80 % of the total page requests. We can generalize that majority of Wikipedia traffic is caused by popular articles. We also saw variation in the fluctuation of most popular articles between the editions, but overall 1-2 % of the top 1000 stayed the same for the whole survey period. We detected recurring patterns in the

most popular articles ranking changes. When we analyzed to most popular articles of the English Wikipedia we identified that the most popular article type is person, followed by articles relating to entertainment and places.

While in this chapter we focused on the consumption of the Wikipedia content, in the next chapter we will compare Wikipedia editing againts traditional news services and in Chapter 5 we will use Wikipedia as a source when we track interactions across business news, social, and stock fluctuations.

Table 3.7: Pearson correlation matrix for article fluctuation in top 1000.

	ar	**ca**	**de**	**en**	**es**	**fi**	**fr**	**it**	**ja**	**ko**	**nl**	**no**	**pl**	**pt**	**ru**	**sv**	**uk**	**vi**	**zh**
ar	1	0.72	0.86	0.9	0.75	0.84	0.85	0.87	0.83	0.82	0.78	0.73	0.78	0.69	0.86	0.81	0.66	0.87	0.84
ca	0.72	1	0.71	0.75	0.65	0.67	0.82	0.79	0.59	0.63	0.86	0.85	0.91	0.53	0.8	0.77	0.85	0.7	0.67
de	0.86	0.71	1	0.93	0.84	0.92	0.91	0.95	0.91	0.89	0.88	0.8	0.77	0.78	0.9	0.86	0.63	0.89	0.89
en	0.9	0.75	0.93	1	0.79	0.91	0.93	0.94	0.85	0.85	0.87	0.79	0.82	0.72	0.93	0.9	0.69	0.89	0.85
es	0.75	0.65	0.84	0.79	1	0.79	0.77	0.81	0.74	0.78	0.79	0.79	0.61	0.94	0.78	0.74	0.47	0.82	0.79
fi	0.84	0.67	0.92	0.91	0.79	1	0.86	0.91	0.84	0.83	0.88	0.81	0.76	0.71	0.89	0.93	0.62	0.87	0.85
fr	0.85	0.82	0.91	0.93	0.77	0.86	1	0.93	0.82	0.82	0.91	0.79	0.88	0.71	0.92	0.89	0.79	0.83	0.8
it	0.87	0.79	0.95	0.94	0.81	0.91	0.93	1	0.9	0.87	0.91	0.83	0.84	0.74	0.91	0.9	0.7	0.9	0.9
ja	0.83	0.59	0.91	0.85	0.74	0.84	0.82	0.9	1	0.88	0.77	0.65	0.66	0.71	0.8	0.75	0.57	0.86	0.87
ko	0.82	0.63	0.89	0.85	0.78	0.83	0.82	0.87	0.88	1	0.79	0.71	0.7	0.76	0.83	0.76	0.56	0.89	0.88
nl	0.78	0.86	0.88	0.87	0.79	0.88	0.91	0.91	0.77	0.79	1	0.89	0.88	0.7	0.88	0.92	0.79	0.83	0.81
no	0.73	0.85	0.8	0.79	0.79	0.81	0.79	0.83	0.65	0.71	0.89	1	0.8	0.69	0.8	0.84	0.65	0.74	0.73
pl	0.78	0.91	0.77	0.82	0.61	0.76	0.88	0.84	0.66	0.7	0.88	0.8	1	0.51	0.87	0.84	0.86	0.74	0.75
pt	0.69	0.53	0.78	0.72	0.94	0.71	0.71	0.74	0.71	0.76	0.7	0.69	0.51	1	0.7	0.64	0.38	0.77	0.71
ru	0.86	0.8	0.9	0.93	0.78	0.89	0.92	0.91	0.8	0.83	0.88	0.8	0.87	0.7	1	0.87	0.76	0.89	0.88
sv	0.81	0.77	0.86	0.9	0.74	0.93	0.89	0.9	0.75	0.76	0.92	0.84	0.84	0.64	0.87	1	0.73	0.81	0.78
uk	0.66	0.85	0.63	0.69	0.47	0.62	0.79	0.7	0.57	0.56	0.79	0.65	0.86	0.38	0.76	0.73	1	0.62	0.58
vi	0.87	0.7	0.89	0.89	0.82	0.87	0.83	0.9	0.86	0.89	0.83	0.74	0.74	0.77	0.89	0.81	0.62	1	0.92
zh	0.84	0.67	0.89	0.85	0.79	0.85	0.8	0.9	0.87	0.88	0.81	0.73	0.75	0.71	0.88	0.78	0.58	0.92	1

Chapter 4

Surveying Wikipedia Activity against Traditional News Services

This chapter is based on the Publication III (see Section 1.4)

In Chapter 3 we saw how Wikipedia content is consumed. In this chapter we focus on the users creating the content of Wikipedia. In particular, we try to identify patterns in how Wikipedia articles are created or edited and also compare different cultures and see how they affect Wikipedia editing activity. We also contrast Wikipedia article editing behavior with that of commercial news sites and notice several differences.

This chapter is organized as follows. Section 4.1 presents the methods of our data collection. Section 3.2 presents the results and discusses their implications. Finally, Section 4.3 concludes the chapter.

4.1 Description of Data

Table 4.1 shows a summary of the collected data. We selected 4 different Wikipedias for our study: Arabic, Finnish, Korean, and Swedish. The reasons for selecting these were that from these we were certain to be able to collect all the editing activity and they also represent a certain geographical and cultural spread. Finland and Sweden share many similarities in culture and society and we would expect them to exhibit similar kinds of activity. Same as Korean, the speakers of those languages are mostly concentrated on a single timezone, making editing activity easier to map in diurnal patterns.[1] Arabic is spoken over a very wide range of timezones

[1] Although there are a fair number of Korean speakers in North America, we did not see any clear evidence of large activity on their part in the Korean Wikipedia.

Table 4.1: Gathered data.

Service	Gathered data	Source	Period
Wikipedia	Recent edits in 4 languages	Portal page feed of each language	May 2009 - Apr 2010
HS	Latest news	Feed from portal page	Feb 2009 - Apr 2010
BBC	Latest international news	International news section feed	Feb 2009 - Apr 2010

and we wanted to see if that is visible in the levels of activity over the day. As examples of commercial news sites, we selected the BBC World News (BBC) and the leading Finnish daily newspaper Helsingin Sanomat (HS).

We downloaded RSS feeds from each of the 6 sources and used these feeds as sources for our data. For the commercial sites, these feeds contained the news items as they were published. For Wikipedia edits, the feeds contained information about what page had been edited and what changes had been made. We ran the data collection for approximately one year for each of the sources.

We gathered the data using simple scripts, which downloaded the aforementioned RSS feeds at regular intervals. The HS feed was taken directly from their homepage, the BBC feed from their international news section, and each Wikipedia feed from the corresponding portal page.[2] The intervals between two fetches of a feed varied from feed to feed. The goal was to balance between getting all of data and not getting too much of duplicates. Later, the downloaded feeds were parsed and pruned of duplicates[3] for analysis.

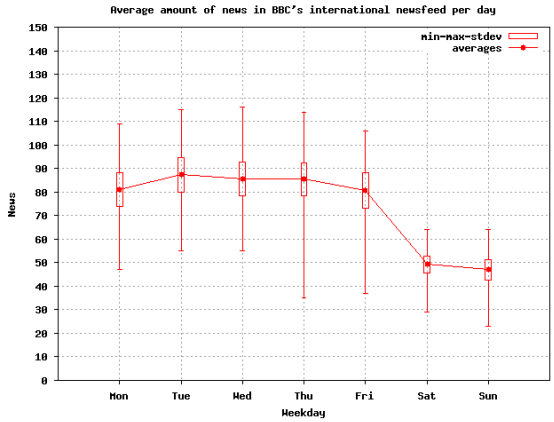
4.2 Results

This section is divided into three parts. First, we show results related to the commercial news services. This establishes a baseline against which we will then contrast the Wikipedia results, in order to compare Wikipedia with commercial services. The second part focuses on the Wikipedia results, and the third part examines the Finnish Wikipedia more closely by analyzing the users and changes in more detail.

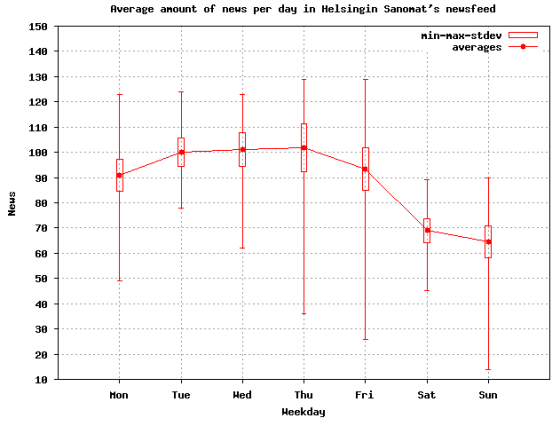
Note that all the times mentioned are given in local time, except for Arabic Wikipedia which is given in UTC. The local timezones are as follows:

- UTC for BBC
- UTC+1 for Swedish Wikipedia

[2] {ar, fi, ko, sv}.wikipedia.org
[3] Each item in the feed had a unique ID number to allow for pruning.



(a) BBC



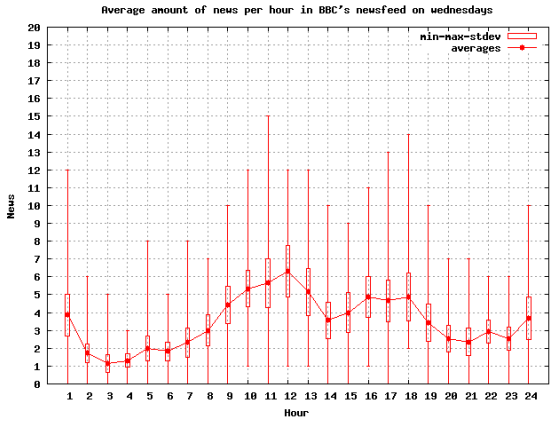
(b) Helsingin Sanomat

Figure 4.1: Daily averages for BBC and HS.

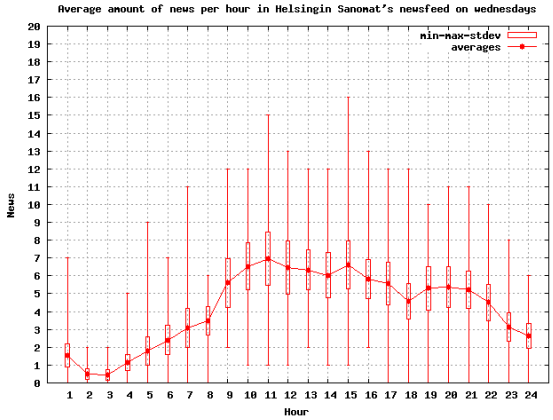
- UTC+2 for Finnish Wikipedia
- UTC+9 for Korean Wikipedia

4.2.1 Commercial News Services

We will start by presenting the weekly and daily activity distributions for the commercial news services. As shown in Table 4.1, we selected the international news section from BBC and download the corresponding RSS-feeds, each containing the latest headlines. The HS feeds included news from all categories.



(a) BBC



(b) Helsingin Sanomat

Figure 4.2: Wednesday averages for BBC and HS.

Figure 4.1 illustrates the daily averages for published news items for both BBC and HS. We have also calculated the minimum and maximum values and the standard deviation for the published items. As seen both services exhibit a similar and clear weekly pattern based on the working week; HS is averaging around 100 items per day, BBC a bit less. On weekends, the amount of published news items is significantly lower.

It is interesting to note that the national newspaper is more active in publishing news than an organization covering the whole world. However, we speculate that the reason is simply because the BBC international news feed only covers "large" events or events with international significance. Nonetheless, it would be interesting to study a wider range of news services

to see whether the number of published news items per day would be of a similar order of magnitude as the two we have chosen here.

Figure 4.2 plots the hourly averages for Wednesdays. As the level of activity was considerably lower on weekends, we examined a weekday more closely and arbitrarily picked Wednesday. Other weekdays exhibited similar results. This is an average of all Wednesdays over a period of more than a year. Both BBC and HS exhibit a clear diurnal pattern, with activity rising in the morning, leveling off for the afternoon, and slowing later in the evening. Interestingly, BBC shows a clear drop around 1–2pm local time, which we speculate could indicate a lunch break.

4.2.2 Wikipedias

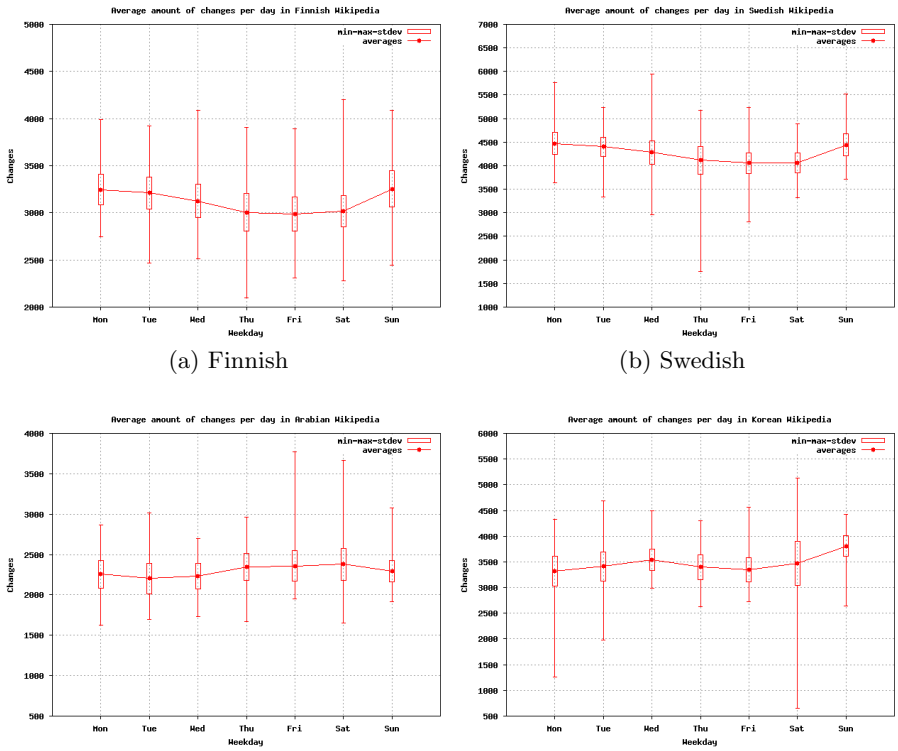
Figure 4.3 has the daily averages for all the four examined Wikipedias. As we can see, the changes made in the services distribute fairly equally over all days in all cases. The drop of activity on weekends that occurred with the commercial news services is not visible in the Wikipedias, quite the opposite, with Sundays typically seeing the highest average level of activity. Only the Arabic version has a slightly lower activity rate in Sundays, however, we should remember the fact that in Arabic countries the weekend falls on Friday-Saturday or in some countries on Thursday-Friday. Because of these differences between Arabic speaking countries, the lack of a clearly identifiable "weekend" is not surprising.

In terms of actual number of edits, the Swedish users are the most active among the four studied editions, followed by the Finns and Koreans at roughly equal level of activity, and finally Arabic users at a slightly lower level of activity.

The hourly averages for the four Wikipedias are presented in Figure 4.4. The activity levels follow natural diurnal rhythms. Interestingly, a great number of changes are made during working hours, which leads us to 2 different, but not mutually exclusive, conjectures about the people who edit Wikipedia. Either, the editors are people with "free" time during the day, e.g., students, or people actually edit Wikipedia during the working hours at work. Our methodology is not able to answer this question, nor are we aware of studies which would have looked at this question in more detail.

As expected, the editing activity in the Arabic Wikipedia is more spread out over the day, as a result of the large spread of timezones where native Arabic speakers live.

Korean Wikipedia editing activity is also closely tied to the diurnal pattern of Korea's timezone UTC+9. Even though there are a fair number



(a) Finnish (b) Swedish

(c) Arabic (d) Korean

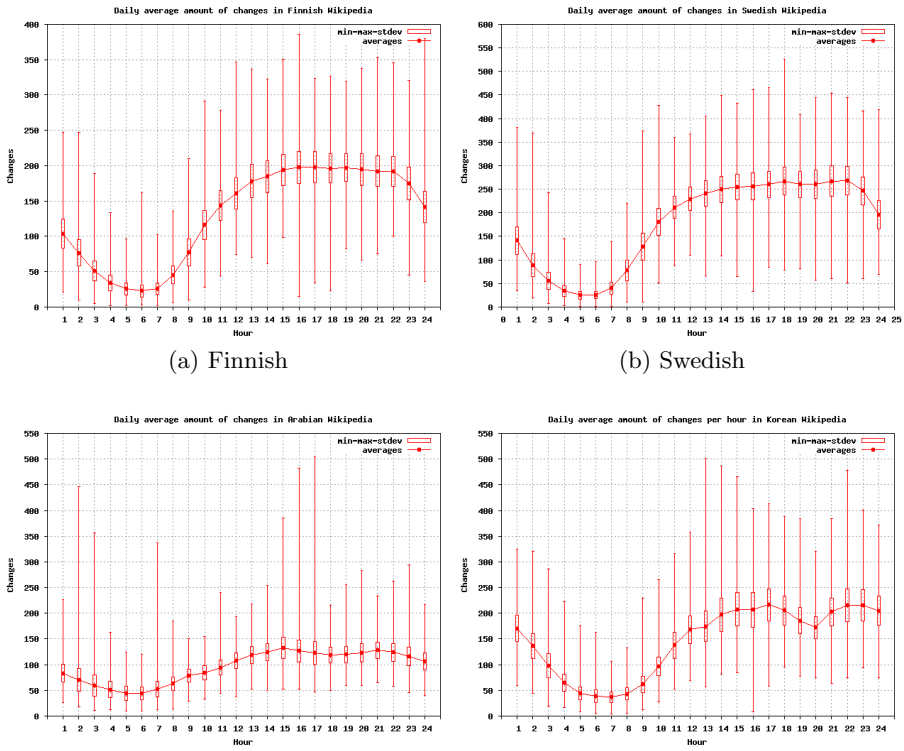
Figure 4.3: Daily averages for different Wikipedias.

of Korean speakers in North America, it seems that they do not contribute to Wikipedia in large numbers. If they did, their activity would be visible in the "Korean night", since the time difference is around 12 hours. However, the drop in Korean activity is similar to Sweden and Finland, leading us towards the conclusion that Korean Wikipedia is for the most part edited by people living in Korea.

All in all, our results agree with daily distribution patterns observed in other user-generated formats discussed in [27]. This partially validates our methodology, and gives us confidence in the accuracy of our results.

4.2.3 Users and Changes

In this section we will show a few more observations that we have drawn related to users of the Finnish Wikipedia and the changes committed by them. User were identified and distinguished by their user names or IP-



(a) Finnish (b) Swedish

(c) Arabic (d) Korean

Figure 4.4: Hourly averages for different Wikipedias.

addresses. A single person could be using multiple usernames or IP addresses. We have no means of verifying either of these cases, however, as we are only interested in the overall level of activity, we deemed this to be a small problem.

From a sample of 1000 IP addresses we had, 97% of the addresses that were mapped to Finland by `whois`. This seems to indicate that Finnish Wikipedia is for the most part edited by people actually living in Finland and not by Finnish speakers living in other countries.

As shown in Figure 4.3a, Finnish Wikipedia has a lot of activity during the weekends. We also took a closer look at how the activity levels look like during major national holidays, such as Christmas and New Year. In the Finnish calendar, 25 and 26 December and 1 January are national holidays, and 24 December is widely observed as a full or partial holiday as well. Figure 4.5 plots the holiday season 2009, where a date shown on the x-axis is 00:00 of that day. During the main Christmas holidays 24

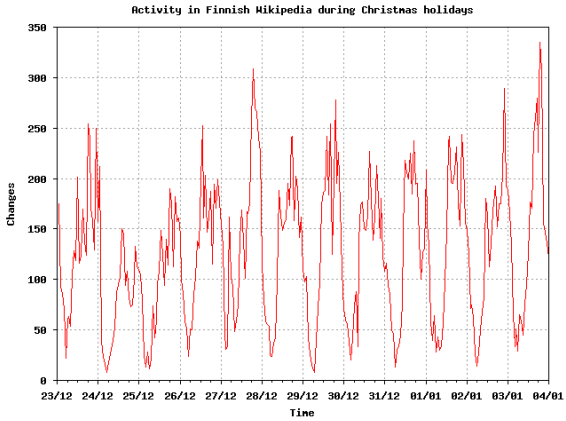


Figure 4.5: User activity during the holidays in the Finnish Wikipedia.

and 25 December, we see a clear drop of activity, whereas New Year on 31 December and 1 January show no special effect. However, in general the activity levels during the holidays are slightly higher than the long-term averages shown in Figure 4.4a.

Table 4.2 lists users categorized by activity into three groups; a user who has done more than 1000 changes to any articles during our study is a heavy user. The heavy users form only 0.2% of all users, however, they contributed almost 60% of all changes during the study. Medium users, 1% of all users, produce 13% of changes. The light and random users, who together represent almost 99% of users, did 29% of the changes. In other words, combining the two most active classes tells that less than 2% of users do 70% of the changes.

The facts that a lot of changes do occur during the working-hours and that 70% of the changes are done by a very small group are of users are somewhat in contradiction with findings of Kittur et al. [35] whose work showed a shift from admin-based editing to a more everyman's event. Our work would indicate that during the surveyed period at least the Finnish Wikipedia was still very much an admin-based Wikipedia.

Let us now examine the changes more closely. Figure 4.6 illustrates the number of changes made by the 500 most active users, strengthening the observation that the majority changes are made by a small group. Doing one change per day is not nearly enough to be ranked at the top of the list, for that one would have needed to commit tens of changes per day.

During our survey over 12 million changes were made into the Finnish Wikipedia. Figure 4.7 shows the 50,000 largest changes. A size of a change is noted by the accuracy of one character. As with the most active user

Table 4.2: Users groups and changes made.

Group	Number of changes	% of users	% of changes
Heavy	1000+	0.2	59
Medium	51-1000	1	13
Light	2-50	44	19
Random	1	54	10

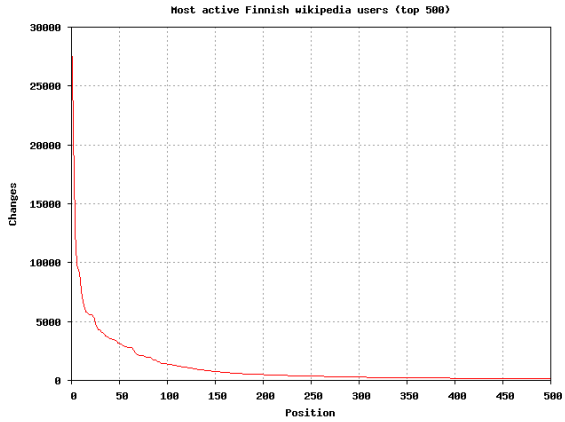


Figure 4.6: Most active users in the Finnish Wikipedia.

graph, the largest changes exhibit a power law distribution. A closer examination revealed that the largest changes were deletions of complete pages, that were then returned back to some previous version causing, in turn, another large change. We speculate that these are the results of either malicious defacing of pages or accidental deletions of content. On the other hand, over a half of the changes are less than 1000 character.

The results concur with the earlier discoveries by Voss [67], who in his work examined the edits made in the Danish, English, German, Hungarian, and Japanese Wikipedias.

4.3 Summary

In this chapter, we have studied activities in commercial news services and four different Wikipedias from around the world. We collected all the published news items and information about all the edits in the Wikipedias over a period of about 1 year (May 2009 - Apr 2010).

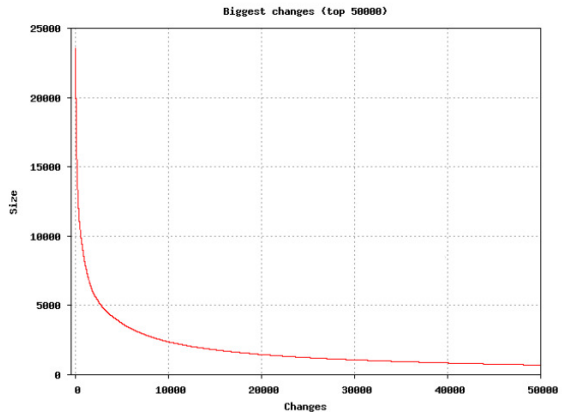


Figure 4.7: Biggest changes in the Finnish Wikipedia.

Our results show that the commercial news sites exhibit not only a clear diurnal pattern, but also a clear weekday-weekend pattern, with clearly lower levels of activity during weekends. Wikipedias, on the other hand, while showing a clear diurnal pattern, do not have a clear weekday-weekend pattern. Instead, the level of activity is relatively constant, with only a slight increase on Sundays. Our comparison across different language Wikipedias shows that they all follow a very similar pattern. Cultural and geographical differences in the Wikipedias we studied seemed to have very little effect on the level of activity. This leads us to speculate that the "trait" of editing Wikipedia is something to which individuals are drawn, not something specific to certain cultures. While we do not have any definite answer as to why this is so, one possible reason could be that the tendency to actively edit Wikipedia is an individual trait which transcends cultural barriers.

Chapter 5

Tracking Interactions across Business News, Wikipedia, and Stock Fluctuations

This chapter is based on the Publication IV (see Section 1.4)

In this chapter study the interplay among business news, social media, and stock prices. We believe that the combined analysis of information derived from news, social media and financial data can be of particular interest for specialists in various areas such as Web scientists, data journalists, and business analysts.

The nature of the complex relationships among traditional news, social media, and stock price fluctuations is the subject of active research. Recent studies in the area demonstrate that it is possible to find some correlation between stock prices and news, when the news are properly classified [61, 7]. A comprehensive overview of market data prediction from text can be found in [45]. In particular, [44] reported an increase in Wikipedia views for company pages and financial topics before stock market falls. Joint analysis of news and social media has been previously studied, *inter alia*, by [28, 60, 39]. The approach followed in these papers. has two interrelated goals: to find information complementary to what is found in the news, and to control the amount of data that needs to be downloaded from social media.

We use PULS[1] to extract events from news text. PULS is a framework for discovering, aggregating, vizualization and verification of events in various domains, including Epidemics Surveillance, Cross-Border Security and Business. We utilize the PULS system to collect on-line news articles from

[1]The Pattern Understanding and learning System: http://puls.cs.helsinki.fi

multiple sources and to identify the business entities mentioned in the news texts, e.g., companies and products, and the associated event types such as "product launch," "recall," "investment". Using these entities we then construct queries to get the corresponding social media content and its metadata, such as, Twitter posts, YouTube videos, or Wikipedia pages. We focus on analyzing the activity of users of social media in numerical terms, rather than on analyzing the content, polarity, sentiment, etc.

The main goals of this chapter is to combine NLP with social media analysis, and discover interesting correlations between news and social media.

5.1 Process Overview

Let's go over the processing steps. First, the system collects unstructured text from multiple news sources on the Web. PULS uses over a thousand websites which provide news feeds related to business (Reuters Business News, New York Times Business Day, etc.). Next, the NLP engine is used to discover, aggregate, and verify information obtained from the Web. The engine performs Information Extraction (IE), which is a key component of the platform that transforms facts found in plain text into a structured form.

An example event is shown in Figure 5.1. The text mentions a product recall event involving General Motors, in July 2014. For each event, the IE system extracts a set of *entities*: companies, industry sectors, products, location, date, and other attributes of the event. This structured information is stored in the database, for querying and broader analysis. Then PULS performs deeper semantic analysis and uses machine learning to infer some of the attributes of the events, providing richer information than general-purpose search engines.

Next, using the entities aggregated from the texts, the system builds queries for the social media sources, e.g. to search company and product names using Twitter API. The role of the social media component is to enable investigation of how companies and products mentioned in the news are portrayed on social media. Our system supports content analysis from different social media services. In this chapter, we focus on numerical measurement and analysis of the content. We count the number of Wikipedia views of the company and the number of its mentions in the news and then use time series correlation to demonstrate the correspondence between news and Wikipedia news. We also correlate these with upward vs. downward stock fluctuations.

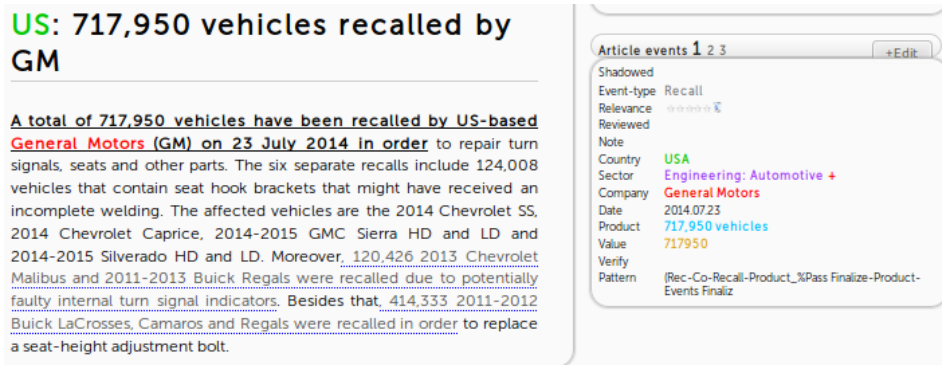


Figure 5.1: A news text and a product recall event produced by the PULS IE engine.

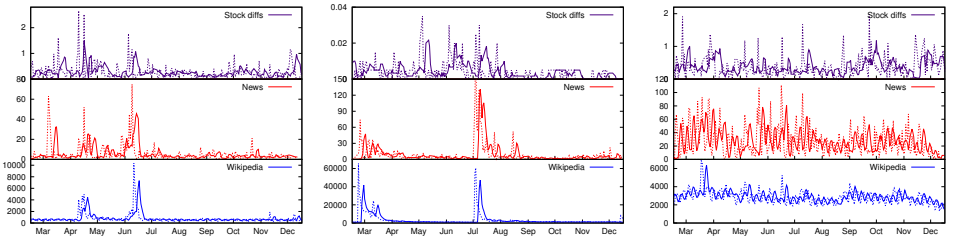
We have complete Wikipedia page request history for all editions, starting from early 2008, updated daily. We can instantaneously access the daily hit-count history for any Wikipedia article. Mapping a name of an entity to a Wikipedia article is not always trivial to do automatically, but the mapping appears to be easy in the vast majority of cases. Thus, we have used the Wikipedia data to explore and demonstrate visibility in social media in the results presented in the following section.

5.2 Results

In this section we demonstrate results that can be obtained using this kind of processing. We present two types of results: A. visual analysis of correspondence between Wikipedia views, news hits and stock prices, and B. time-series correlations between news hits and Wikipedia views.

In the first experiment we chose three companies: Alstom, Malaysia Airlines, and General Motors. We present the number of mentions in the news collected by PULS, the number of views of the company's English-language Wikipedia page, and stock data, using data from March to December 2014.

In each figure, the top plot shows the daily *difference* in stock price: the absolute value of the opening price on a given day minus price on the previous day, obtained from Yahoo! Finance. The middle plot shows the number of mentions of the company in news. The bottom plot shows the number of hits on the company's Wikipedia page. In each plot, the dashed line represents the daily values and the bold line is the value smoothed over three days.



(a) Alstom (b) Malaysia Airlines (c) General Motors

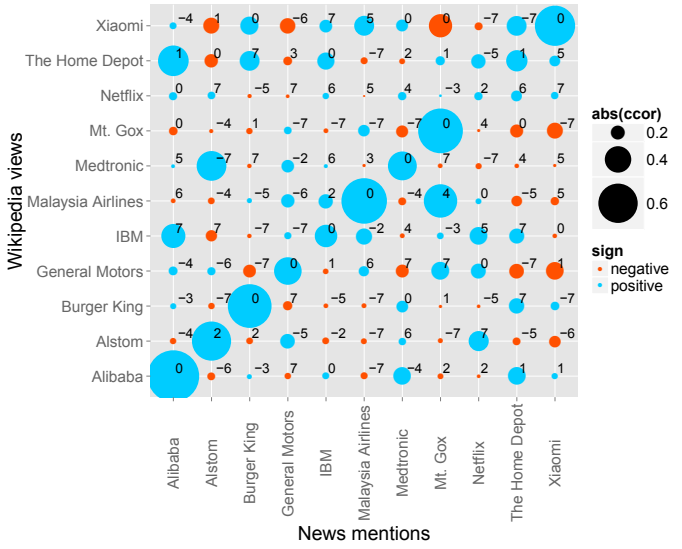
Figure 5.2: Daily differences in stock prices, number of mentions in PULS news and number of Wikipedia hits in 2014 for three companies.

Figure 5.2a plots the data for the French multinational Alstom. The company is primarily know for its train-, power-, and energy-related products and services. In the plot we can see a pattern where the stock price and news mentions seem to correlate rather closely. Wikipedia page hits show some correlation with the other plots. The news plot shows three major spikes, with two spikes in Wikipedia hits. The March peak corresponds to news about business events (investments), whereas the other peaks had a political aspect, which could trigger activity in social media; e.g., in June, the French government bought 20% of Alstom shares, which caused an active public discussion.

Malaysia Airlines suffered two severe incidents in 2014. On March 8, they lost one aircraft over the Indian Ocean, and on July 17 another was shot down in Eastern Ukraine. Strong correlation in the patterns between news mentions and Wikipedia hits is clearly visible in Figure 5.2b. The correlation with the stock price is less clear.

Figure 5.2c plots the data for General Motors, which was affected by numerous product recalls throughout the year. The company has been mentioned in the news and has been looked up on Wikipedia throughout the covered period. The stock price also oscillates over the entire year.

Although most of the local oscillations are due to normal fluctuations in the weekly flow of data on the Internet (with regular dips corresponding to the weekends), some broader-range correspondence is also discernible from the plots. Note, that the PULS IE system automatically assigns sentiment polarity to the news, classifying events as “positive” (e.g., investments, contracts, acquisitions) or “negative” (e.g., bankruptcies, layoffs, product recalls). This will form the basis for more detailed analysis of correlations with stock fluctuations in the future.



Circle width represents strength of correlation; color represents sign of correlation: blue is positive, red is negative; the numbers indicate the time lag (in days) at which the correlation with the greatest magnitude is obtained for the given pair: positive lag means that Wikipedia views follow news mentions.

Figure 5.3: Cross-correlation between Wikipedia views and mentions in PULS news for 11 companies.

In the second experiment, we choose eleven big companies from different industry sectors, namely Alibaba, Alstom, Burger King, General Motors, IBM, Malaysia Airlines, Medtronic, Mt. Gox, Netflix, The Home Depot, and Xiaomi. For each of these companies we collect two time series: daily news mentions and Wikipedia views during time period from March to December 2014. Then we calculate the cross-correlation between all possible pairs in these dataset, for a total of 121 cross-correlations[2]. We limit the lag between time series by seven days, based on the assumption that if there exists a connection between news and Wikipedia views it should be visible within a week.

The results of this experiment are presented in Figure 5.3, where the circle size represents correlation strength, the colors represents correlation sign: blue means positive correlation, red negative; the numbers mean the time lag at which the highest correlation for a given company pair was obtained: positive lag means that Wikipedia views followed news mentions, negative lag means that news followed Wikipedia views.

It can be seen from the figure that the largest correlations and the lowest lags can be found on the diagonal, i.e., between news mention for a company

[2]We use standard R ccp function to calculate cross-correlation.

and the number of views of the company Wikipedia page. Among the 11 companies there are two exceptions: The Home Depot and Netflix. For Netflix, news mentions and Wikipedia views do not seem to be strongly correlated with any time series. News about Alibaba show a surprising correlation with Wikipedia hits on Home Depot on the following day. At present we do not see a clear explanation for these phenomena; these can be accidental, or may indicate some hidden connections (they are both major on-line retailers).

The lag on the diagonal equals to zero in most cases, which means that in those cases the peaks occur on the same days. At a later time, we can investigate finer intervals (less than one day). We believe it would be interesting if a larger study confirmed that we can observe regular patterns in the correlations and the lags are stable,e.g., if a spike in the news regularly precedes a spike in the Wikipedia views, since that would confirm that these models can have predictive power.

5.3 Summary

We have presented a study of the interplay between company news, social media visibility, and stock prices. Information extracted from news by means of linguistic analysis was used to construct queries to various social media platforms. We expect that the presented framework and result would be useful for e.g. business analysts, marketing people, journalist, and researchers.

The results presented in Section 5.2 demonstrate the utility of collecting and comparing data from a variety of sources. We were able to discover interesting correlations between the mentions of a company in the news and the views of its page in Wikipedia. The correspondence with stock prices was less obvious. This could be improved by refining the forms of data presentation. For example, we have found that plotting (absolute) differences in stock prices may in some cases provide better insights than using raw stock prices.

Data could be improved by covering a wider range of data sources and social platforms, general-purpose (e.g., YouTube or Twitter) and business-specific ones (e.g., StockTwits). One could also analyze the social media content as well, e.g., to determine the sentiment of the tweets that mention a particular company. Covering multiple sources is important due to the different nature of the social media. Tweets are short Twitter posts, where usually a user shares her/his impression about an entity (company or product), or posts a related link. Wikipedia, on the other hand, is used for

obtaining more in-depth information about an entity. YouTube, in turn, is for both the consumption and creation of reviews, reports, and endorsements. This phase faces some technical limitations. For example, while Twitter data can be collected through the Twitter API in near-real time, the API returns posts only from recent history (7-10 days). This means that keyword extraction from the news and data collection from Twitter should be started immediately after the company or product appears in the news.

Further improvements could be achieved by building accurate statistical models on top of the collected data, and by exploring the correlations and possible cause-effect relations, etc. It should be possible to find particular event types (lay-offs, new products, lawsuits) that cause more reaction on social media and/or in stock prices than others. Likewise, it should be possible to develop predictive patterns of visibility on social media for companies and products, based on history or on typical behavior for a given industry sector.

Chapter 6

Twitter

This chapter will consist of results and observations drawn from our measurements related to the micro-blogging service Twitter. The motivation behind the survey is to understand better the reasons for users to create tweets, and see if the reasons correspond to real-life situations. We have collected more than 20 million tweets using two different methods. Our first data collection method was to gather tweets based on the location of users, that is, we collected tweets originating from such cities as Liverpool and Madrid. The second method is based on topical keywords, such as "H1N1" and "Olympics". From the data we have drawn hourly and daily patterns for the creation of tweets, user statistics, and timeseries for topical events. We also briefly examine the languages used in the tweets.

On contrast to the Wikipedia work presented on Chapter 3, in this chapter we are interested in the creation of the content. Also, with Wikipedia we had access to almost complete page view data. However, with Twitter we are limited to the data that the service offers through an API and we also had to take rate limitations into the consideration.

Our research points of interest are:

- Are there temporal patterns in the creation of tweets?
- What kind of users there are?
- How topical events are portrayed on Twitter?
- Are there geographical differences in the creation patterns or types of tweets?

The chapter is organized as follows. In the next section, we do introduce the data and the way it was collected. Thereafter, we start presenting results in Section 6.2. We also provide a short natural language analysis in

Section 6.5. Our summary is in Section 6.6 accompanied by a few words about the future work.

6.1 Description of Data

As mentioned, we executed the data collection by using two different methods. Common to both methods is that all the tweets were downloaded in RSS or JSON feeds which the Twitter API[1] provides and later parsed and pruned to eliminated duplicates. With the first method, we selected a bunch of cities, and collected tweets emitted from those cities. For example, we asked to get a RSS feed containing the one hundred latest tweets from Madrid and the enfolding 15 mile radius. We should note, that by no means we claim that the method collects comprehensively all the tweets from a specific location. That is up to the inner functionality of the Twitter service to which, of course, we have no influence. In addition, a user can set his/hers tweets to be private. However, we do stand by the validity of the method as the main target is to measure the fluctuation in the number tweets over time, not the absolute values. To this end, a solid level of tweets is better than being overwhelmed by the feeds, something that also affected our selection of the cities. We chose cities where the number of tweets for the location was reasonable. Having the data, we can form patterns and examine how real-life events are portrayed in the Twitter microcosm. Expectedly, the data will give us a worthy insight about users from certain areas and how they react to events and in which amount.

In more detail, the location of a Twitter user is indicated primarily by a geotag encompassed with the tweet, or secondarily and more commonly, by the location entered in the user's profile. This, of course, makes it possible that the RSS feeds include tweets from users who are not at the moment in the stated location, something that we accept as a weakness of the method. At the beginning, we requested the RSS feeds containing 100 tweets roughly every 50 seconds, thus, the maximum number of new tweets per minute that we could register was around 120 per city. However, we later tweaked the method so that we ended up with collecting approximately 300 tweets per minute for one city.

The motivation behind the method is to see accurately, even in the precision of a second, when people create tweets. Presumably, this would also give as indication to which real-life events cause the incentive to post a tweet and which do not. Furthermore, applying the method to various cities will reveal any cultural differences. While we collected tweets from more

[1] search.twitter.com/search.atom and later search.twitter.com/search.json

than ten cities, in this chapter we will concentrate mainly on presenting results and details from just two cities, namely, Liverpool, UK and Madrid, Spain. The two cities were selected as they differ, a priori, in language, working-hours, and culture. The results are presented in the next section.

Our second method was based on keywords. That is, we requested tweets based on certain keywords. For example, in the case of H1N1 virus, we formulated our requested so that we ended up with a RSS feed containing tweets that included at least one of the following keyword strings "swine flu", "swineflu", or "H1N1". Contrary to the first method, the tweets could have originated from anywhere. These feeds were requested repeatedly approximately every 20 minutes. The method will show how certain topics live and develop on Twitter. The approach, does give a clear and strong indication of which topics people are tweeting about. It is also a good way of seeing when some topics are starting to gain popularity and when the interest fades away. We also mixed the methods and in a few cases we performed keyword-based queries to data from a certain city which we had collected using the first method.

6.2 Comparison of Two Cities

Between January and June 2010 we gathered tweets from two cities, namely Liverpool, UK and Madrid, the capital of Spain, using our location-based data collection method. Overall, we collected 11 million tweets, 6 million from Madrid and the remaining 5 million from Liverpool. In this section will present daily and hourly patterns for the creation of tweets and statistics of the users. We have used data from 155 days to produce the Liverpool patterns and 136 in the case of Madrid. Furthermore, we will show a couple of occasions when tweeting seems to correlate with real-life events.

6.2.1 Daily Patterns

We will start by presenting the daily patterns for both cities. Figure 6.1 shows the average percentage of tweets created each of the week day in Liverpool and Madrid for the data collection period. As can be seen, in Liverpool the creation of tweets is spread quite evenly through-out the week with Sunday being the day when users are the most active, though only with a small margin. Interestingly, the Madrid users exhibit quite different behavior as the figure illustrates. In Madrid, the users are less active on the weekends than during the working week. The observation seems to indicate a cultural difference in the creation for tweets. To further confirm the difference we performed a chi-square test of homogeneity on the data

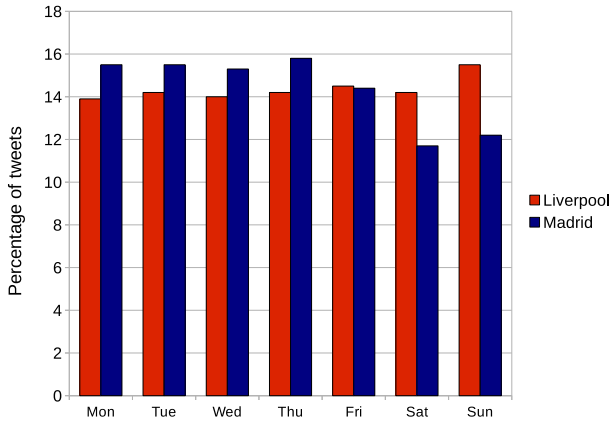


Figure 6.1: Daily patterns for tweets.

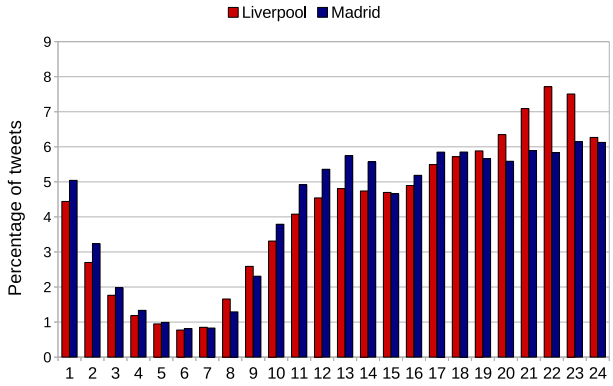


Figure 6.2: Hourly patterns.

(to counts, not proportions) and the resulting P-value <0.01 allowed us to reject the null hypothesis that the Liverpool and Madrid distributions would be from same group.

6.2.2 Hourly Patterns

Next, we will show the hour-by-hour distribution of the posted tweets. The patterns consists of averages for each hour calculated using every day of the week. Figure 6.2 depicts the hourly patterns of tweets from Liverpool and Madrid so that both are in local time. Again, the patterns do differ. With Liverpool, the pattern indicates that the users of Twitter are most productive in the evening hours, most notably during 20-23. Over 80 percent of

Table 6.1: Tweet statistics.

City	N	RTs	Links	Replies	Mentions	Plain
Liverpool	˜5M	0.07	0.16	0.42	0.16	0.33
Madrid	˜6M	0.11	0.32	0.33	0.22	0.26

tweets are posted during 12-01. The pattern resembles the ones that Guo et al. [27] presented in their work for blog article and blog picture postings. A closer analysis revealed that the pattern for all working days is similar and that, compared with Saturday and Sundays the main difference is that on weekends there is a little less activity on mornings, but activity more on the late evenings.

In Madrid, however, the users are nearly as actively creating tweets in the afternoon as they are in the evening, while there is a curious drop during 15-16. Again, we performed a chi-square test of homogeneity on the data and were able to reject the null hypothesis that the Liverpool and Madrid distributions would be from same group with P-value <0.01.

6.2.3 Statistics

Thus far, we have seen when and in which amount tweets are being created. However, now we will focus on the type of the tweets. Table 6.1 has the statistics of the tweets, but first a short rundown of the terms. RT in table means that the tweet is a retweet, which is somewhat analogous to forwarding an email. A retweet starts with RT. A reply, in turn, starts with @ mark and the username to whom to message is intended. A mention is indicated by using @ followed by a username in the body of the tweet and a tweet can include multiple mentions. Mentions let users associate other users to their tweets. We also counted the number of tweets that include at least one link. The figures in the table mean for example that 32 % of the tweets from Liverpool had at least a link on them and 16 % had at least a mention. It should be noted that a tweet can, for example, include a mention as well as a link.

Overall, roughly one third of all tweets from Liverpool and one fourth from Madrid were what we label as plain tweets. A plain tweet is not a retweet or reply nor it has links or mentions. One way of seeing a plain tweet is as a status update. However, having overall more than two thirds of non-plain messages means that tweeting is a heavily social experience, people are either sharing information with retweets and links or communication with other users by using replies and mentions. A noticeable difference, between

Table 6.2: Statistics of Liverpool users.

Group	N	N of tweets	% of users	% of tweets
Heavy	1046	1000+	1	55
Medium	7129	51-1000	7	23
Light	43889	2-50	41	7
Random	53317	1	51	15
	105 381			

Table 6.3: Statistics of Madrid users.

Group	N	N of tweets	% of users	% of tweets
Heavy	1163	1000+	0.8	53
Medium	9040	51-1000	6	24
Light	60371	2-50	42	8
Random	72559	1	51	15
	143 133			

the tweets from the two cities is that Madrid has double the percentage of tweets that have links. The high percentage of replies, especially in Liverpool, is also a noticeable stat. We will analyze the users from both cities more closely in the next couple of subsections.

User statistics

Tables 6.2 and 6.3 have user statistics for both cities. The figures include all tweet types. We have grouped the users according to their activity during the six-month period into four categories: heavy, medium, light, and random. A heavy user posts more than one thousand tweets, a medium user more than 50, but less than thousand, whereas a light user posts between 2-50 tweets. A random user corresponds to the one who creates only one tweet. We want to stress that the categorization is somewhat arbitrary and its main purpose is to serve as an instrument for us to examine the user behavior in the two locations.

Looking at the Liverpool user statistics, we see a couple of interesting things. First, there are 1046 heavy users, who make up approximately one % of all Liverpool users in our data, and they have produced 55 % of all tweets. The second noticeable thing is that over half of the users have not posted more than one tweet. Based on the figures, we can say that a small group of users produce most of the content, while most users produce very

Table 6.4: Most active users.

			% of all tweets				
Liverpool		Tweets	RTs	L	R	M	P
1	ebonyJCotter	41268	2	8	65	30	10
2	LiamHannah	24697	1	10	50	6	36
3	_charlottenberg	19857	1	3	60	8	30
4	CheshireJobsUK	16989	<1	71	0	0	28
5	leahoneill	16725	2	8	49	12	34
6	LiverpoolJobsUK	16279	<1	76	0	<1	23
7	kazuyanavy	15809	2	6	68	6	21
8	EpicDetector	15670	9	24	27	42	36
9	xSTEx	15327	1	10	75	5	12
10	SarahTheSkater	14007	<1	1	67	6	27
Madrid		Tweets	RTs	L	R	M	P
1	chibibun_bot	39348	0	<1	92	<1	7
2	Spainbot	38886	0	<1	94	<1	5
3	cosechadel66	30124	13	14	67	32	8
4	vuelosdesdeMAD	29964	<1	99	0	<1	<1
5	buscavuelos	29839	<1	99	<1	<1	<1
6	fmlopez48	28024	3	5	85	37	2
7	bonhamled	15804	6	67	20	9	12
8	LaTrinchera	15116	36	23	17	70	14
9	ClitterMonstaa	14964	11	3	19	18	59
10	tusanuncios	13425	0	99	0	<1	<1

little. The top 7-8 % of the most active users post almost 80 % of the tweets.

Having seen the differences between the users from the two cities in the daily and hourly patterns, it is intriguing to notice that the user statistics are, in turn, very similar. The relative numbers of the Madrid users are almost exactly the same as the users from Liverpool, even though there are nearly 40,000 more users in the Madrid dataset.

Most active users

As the heavy users have such a big part in the tweet production, let us now examine them more closely. The ten most active users of both cities during the measurement are listed in Table 6.4. The figures represent the

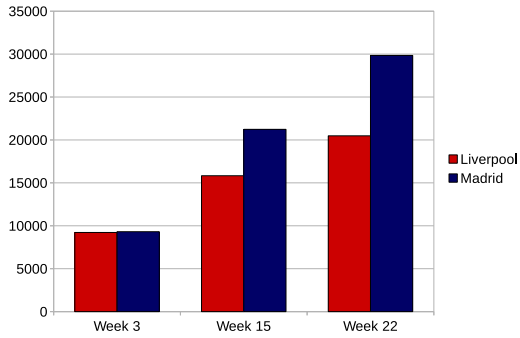


Figure 6.3: Surge in the number of users.

percentages of tweets with links (L), retweets (RTs), replies (R), mentions (M) and also the plain tweets (P). The usernames alone seem to reveal that we are facing a mixed group. Although, the many are regular people, there are a few exceptions. The fourth and fifth users of Liverpool are promoting available jobs and the two user with most tweets from Madrid are actually Japanese language bots, whereas the fifth and sixth are advertising flights and the tenth is generic announcement channel. We were unable to find a common factor for top user by inspecting tweet types posted by them as they do vary strongly among the users.

User evolution

While we conducted our survey, Twitter was reported to have a staggering growth-rate of 300,000 new users per day [26], a fact that did not go unnoticed in our measurements either, as we observed a significant rise in the number of users. Figure 6.3 shows the number of users for Weeks 3, 15, and 22 of 2010 that have posted at least one tweet during that week. In other words, we counted the number of active users for each of the three weeks. As seen, in Liverpool the number of active users has doubled in six months. In Madrid, the number of active users has tripled. Table 6.5 has further numerical details of the user evolution. We used to same grouping of users as in the previous section, but we only counted tweets from the week in question. When we compare weeks 3 and 22, we see that in both cities the random user group has become the biggest as expense of all other groups. The number of users that post only one tweet during a week has risen sharply and on Week 22 they made up almost have of all users in Liverpool and Madrid. Also, the figures show that the amount of tweets posted in a week has risen by around 70 % from Week 3 to Week 22 in both

Table 6.5: Evolution of users.

Group	Madrid			Liverpool		
	Wk 3	Wk 15	Wk 22	Wk 3	Wk 15	Wk 22
Heavy	0.06 %	0.05 %	0.03 %	0.08 %	0.07 %	0.04 %
Medium	9 %	5 %	4 %	9 %	6 %	6 %
Light	63 %	51 %	46 %	61 %	49 %	47 %
Random	28 %	43 %	49 %	31 %	45 %	48 %
N of Users	9312	21,235	29,845	9221	15,832	20,483
N of Tweets	198,958	292,826	340,424	184,890	233,058	310,475
Tweets/User	21.37	13.79	11.41	20.05	14.72	15.26

cities and the mean number of tweets per user has decreased. We believe this is caused by more casual people joining to try out the service as it has become more popular.

6.3 Events

One of the main motivation behind our research was to see how real-life events are portrayed on Twitter. Now, we will show two examples of real-life events and how they were seen on Twitter. Moreover, the examples will reveal us some reasons to why people are tweeting.

FC Barcelona - Real Madrid

The Spanish football year is typically distinguished by two meetings between heavy-weights FC Barcelona and Real Madrid. They are commonly considered as the most watched league football matches in the world, with a global audience of hundreds of millions [3]. The matches mix both pure sporting aspirations and political motivations in such lengths that any game between the two clubs is called el Clásico, the Classic. The latter meeting of the season 2009-2010 took place on April 10 2010, in Barcelona. The game was dubbed as the league title decider with the teams being tied on points before the game. Against that background, it is no surprise that the game, which FC Barcelona ultimately won 2-0, had a visible presence on Twitter.

Figure 6.4 indicates the tweet distributions of Barcelona and Madrid per minute during 20:00 (GMT), the time of the kick-off, and 22:15 that night. Tweets were collected separately from both cities. The three strongest peaks at 20:36, 21:15, and 21:50, identify the 1-0 goal, 2-0 goal, and end

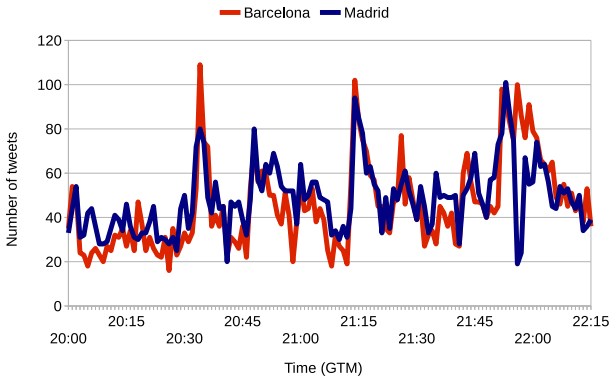


Figure 6.4: Barcelona and Madrid during el Clásico.

of the game, respectively. Interestingly, both curves show the same characteristics. Only the reaction during the half-time, 20:45-21:00 is more subdued in Barcelona. The results do hint that people are using Twitter to express both positive and negative thoughts, however, to validate this, a proper natural language analysis would be needed.

Brit Awards

The Brit Awards or the Brits, is an annual award show honoring mainly the Brittish pop music industry. The 2010 edition of the Brits was held on Tuesday, February 16. The event was broadcasted live on ITV1 starting 8 pm (GMT). As seen in Figure 6.5, the activity on Twitter in Liverpool during the Brit Awards is significantly higher than the average for that time slot. The average was calculated over other Tuesdays from the survey period. To verify the impact of the Brits Awards, we counted the words used in tweets that were posted from Liverpool during the show. The Table 6.6 lists a selected number of common words (more than 3 letters) and names that we subjectively see and have manually labeled as related to the Brit Awards. The R stands for the rank and N for the number of occurrences. The list has been made from all collected tweets published during 8-11 pm on the ceremony day. Clearly, the Brits has influenced a number of users to tweet about it, indicating that tweeting correlates with real-life events and creates a sort of social dimension to the television watching. In general terms, if such events are in the future identified in advance, this would give the social media service providers an opportunity to adapt their functionality accordingly and maybe work some kind of deal that would beneficial to all parties involved.

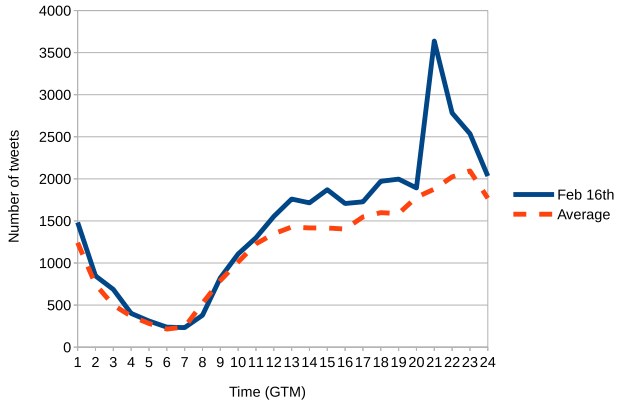


Figure 6.5: The impact of the Brit Awards.

Table 6.6: Common words during Brit Awards.

R	Word	N
2	brits	919
17	gaga	336
20	peter	286
36	robbie	285
36	cheryl	216
37	brit	213
45	lady	186
46	awards	183
56	liam	160
61	award	147
69	florence	139

R	Word	N
78	music	123
89	williams	112
106	cole	98
108	gallagher	94
120	alicia	88
129	kasabian	84
137	ladygaga	79
152	dizzee	70
164	allen	65
173	spice	62
200	oasis	51

6.4 Topical Situations

In their work [13] Cheong and Lee classified different happenings on Twitter as short, medium, and long-term topics. We have already seen examples of short-term topics, such as the Brit Awards and el Clásico football game. In this section we will produce results of topics that fit into the medium and long-term categories. To that end, we used our second method, that is based on keywords. In short, we wanted to see how some keywords are present in the tweets over time and the tweets can originate from anywhere. In other words, we wanted to collected as many tweets as we could with one or more corresponding keywords.

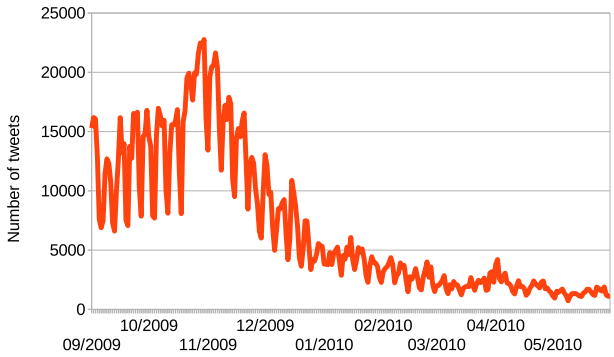


Figure 6.6: H1N1 on Twitter.

6.4.1 H1N1

The outbreak of the H1N1 virus caused worldwide preoccupation and extensive media coverage in 2009. Sensing that this would be an interesting topic to follow, we collected tweets that had at least one of the following keyword strings, "H1N1", "swineflu", or "swine flu" between September 2009 and May 2010. In total we collected 2.1 million tweets. Figure 6.6 illustrated that the virus and its development is undoubtedly visible on Twitter as well. The form of the curve correlates with the number of reported H1N1 cases presented for example in [8]. At the peak of the influenza, the chatter on the Twitter was at highest as well, and as soon as the spread of virus started to ease, the number of tweets began to decrease. The drops on the curve are on weekends, indicating that the users were tweeting about the virus much more actively during the working days. To analyze whether this kind of data can be used to predict the spread of diseases, is out of scope for our work, however the observations do prove that the social media is not a separate world and does reflect the real world events.

6.4.2 Winter Olympics

The 2010 Winter Olympics, was our other point of special focus. The event is interesting, as it is held in high regard in many countries, such as Canada, Germany, Russia, and in the Nordic countries, while being almost completely ignored in others. The Olympics were held on 12-28 February 2010 in Vancouver, Canada. Figure 6.7 shows the number of related tweets per day during a six-months time. The vertical lines indicated the start and end dates of the Olympics. Keywords that we used were, "vancouver", "olympic", and "olympics". In total we collected 2 million tweets. We see

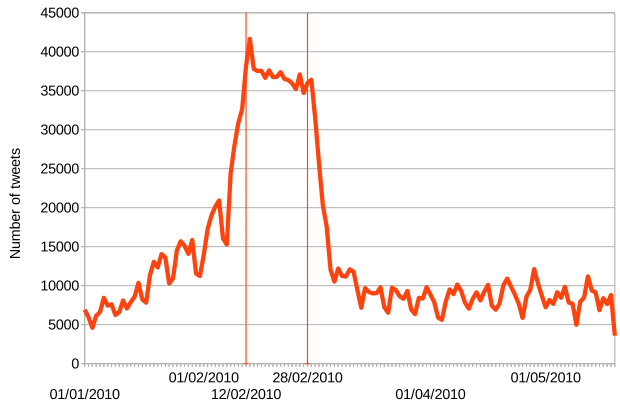


Figure 6.7: Winter Olympics on Twitter.

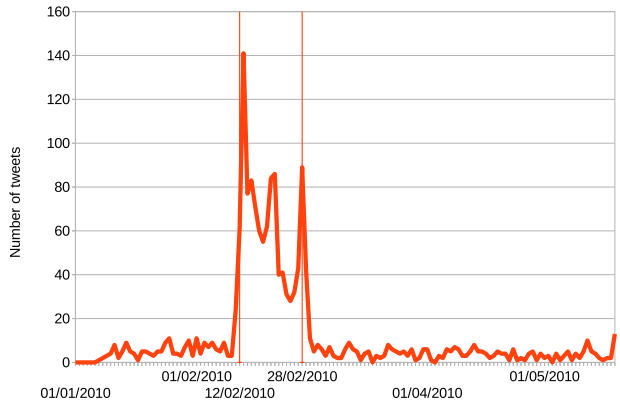


Figure 6.8: Winter Olympics on Twitter in Liverpool.

that there is a vast increase in the number of tweets before the start of the event and we can also observe that the number of tweets is decreasing as the event progresses. Again most of the activity occurs during the working days. It is evident that sporting events such as the Olympics generate the most discussion during event itself and that is also how it is seen on Twitter.

Winter Olympics in Liverpool tweets

In Liverpool the Winter Olympics was received with, at best, a lukewarm interest. Figure 6.8 shows the number of tweets from Liverpool with at least one of the keywords 'vancouver', 'olympic', and 'olympics' in them. As seen the distribution correlates with the event dates, however, even in

the busiest day, February 13, there are only 142 tweets, a meager 0.6 % of all the tweets collected that day. The lack of interest is not that surprising considering the fact that Great Britain won only one medal from the Winter Olympics, although a golden one.

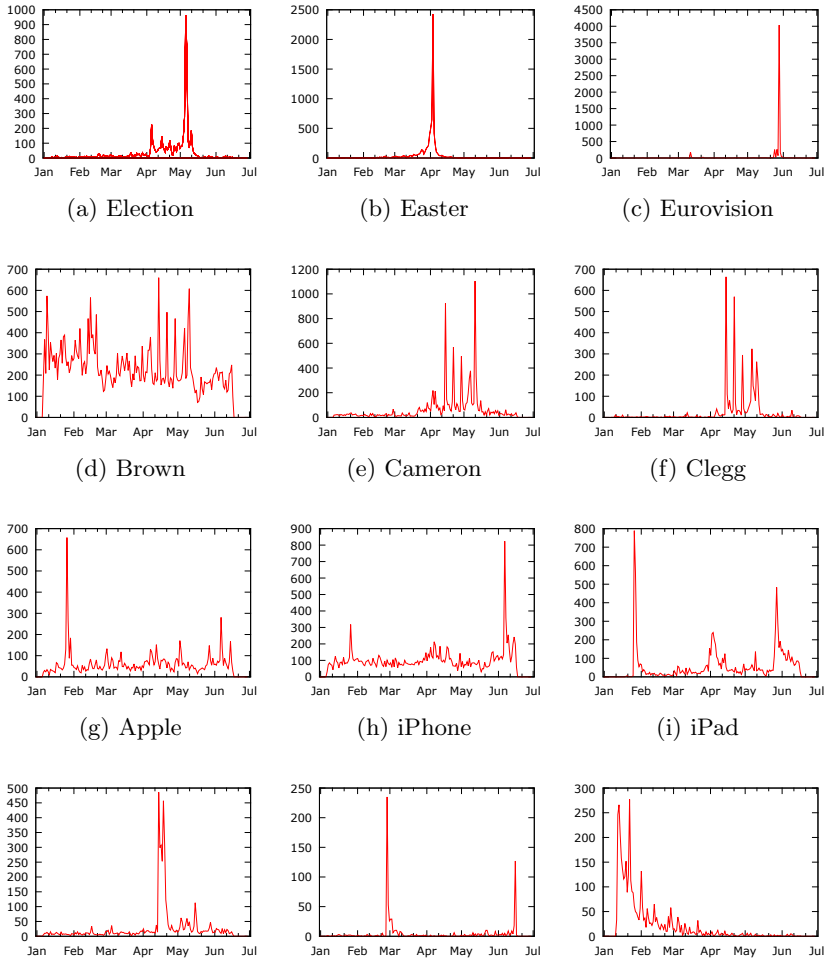
In the next subsection we will present more keyword-based results from the Liverpool tweets.

6.4.3 Liverpool Keywords

The previous examples showed that topical situation do present themselves on Twitter and offer interesting results. However, the problem is that in order to capture related tweets for an event or incident on Twitter one would need to a) to guess/anticipate the upcoming events and the related keywords or b) obtain a large catalog of tweets and do the analysis in retrospect. Given that gathering all the tweets that all users produce is not possible, at least for us, given the sheer numbers, we will settle on doing the analysis in a smaller scale using the data that we have from Liverpool. That is, we did not proactively collect tweets based on certain keywords, but instead used our location based data and calculated the occurrences of keywords from there.

Figure 6.9 shows a collection of keyword-based topics. Please note the changing y-axis in the plots. Each subfigure plots the number of occurrences of a particular keyword (not case-sensitive) given in the caption. The timeline is from beginning of the year 2010 to the end of June of the same year. Some, if not all, of the words are ambiguous by nature but as stated earlier we are interested in examining the differences in the fluctuation of the tweets and mapping it to corresponding real-life events.

In the first row we have keywords corresponding to three big events. With the 'election' keyword, Figure 6.9a, we wanted to capture the race and campaign of the UK general election. As seen, the number of tweets matching the keyword is relatively small until the peak around the election day May 6. However, quite interestingly the keywords formed of the last names of the three party-leaders and prime minister candidates show much more fluctuation. Curiously, the keyword 'cameron' referring to David Cameron, had the most occurrences in the election day, possible reflecting Cameron's eventual selection as the prime minister. The keywords, 'brown' and 'clegg' referring to Gordon Brown, the PM before the election, and Nick Clegg, leader of the Liberal Demoracts party, exhibit from April to May matching patterns which fit to the schedule of the televised debates among the party leaders. The debates were held on April 15, 22, and 29.



(a) Election (b) Easter (c) Eurovision

(d) Brown (e) Cameron (f) Clegg

(g) Apple (h) iPhone (i) iPad

(j) Ash and/or Volcano (k) Chile (l) Haiti

Figure 6.9: Keyword-based results from tweets originated from the Liverpool area.

The second event keyword shows that Easter celebration did generate a lot tweets, the Easter Sunday being clearly the most active day. The last plot on the top row is for the keyword 'eurovision' and, as seen, from all the events we examined the Eurovision song contest is the one with most collected tweets. Figure 6.9c plots the number of tweets having the keyword in them over time and from it we can see that on the contest's final, May 29, there are more than 4000 such tweets. Moreover, we can safely assume that there are a lot more tweets about the Eurovision song contest, although not having the keyword explicitly in them.

As the Apple brand is often considered being media-friendly, we wanted to see how it is portrayed on Twitter. The plots for keywords 'apple', 'iphone', and 'ipad' are on the third row. The announcement of the first iPad, on January 27, is clearly visible in both Figures 6.9i and 6.9g. The other peaks in Figure 6.9i coincide with dates when the device went on sale in North America and then in Europe, on April 3 and May 28, respectively. Interestingly, the highest peak for 'iphone' happened during the unveiling of iPhone 4, on June 7, but the corresponding peak for 'apple' in Figure 6.9g, is much more subdued than was the case with iPad.

On last row, Figures 6.9j-6.9l, we have three events related to natural disasters. The first one relates to the eruption of Eyjafjallajökull volcano in Iceland. An ash cloud formed by the eruption caused severe disruption to air traffic in Europe during April 15-23. Figure 6.9j shows that there were many tweets with either word 'ash' or 'volcano' posted that time. The last two keywords related to the earthquakes occurring in Chile on February 27 and in Haiti on January 12 2010. The earthquake in Haiti was especially devastating, causing more than 200,000 causalities according to the Haitian government. By comparing the number, it seems that, in our data, the catastrophes and tragedies do not count as big motivator to tweet when compared against the other events we have examined.

6.5 Content

In this section we analyze the languages used on Twitter in various locations and see what were some of the most interesting topics in Liverpool.

6.5.1 Language Percentages

In order to analyze the languages used on Twitter, we collected tweets from a multitude cities all around the world. The tweets were collected using our location-based method explained earlier in Section 6.1 and no keywords were used. We used NGramJ [46] to process a sample set from

Table 6.7: Language percentages.

City	Languages		
Amsterdam	Dutch	English	
	84	13	
Berlin	German	English	
	63	29	
Boston	English	-	
	97	-	
Copenhagen	Danish	English	
	21	64	
Helsinki	Finnish	English	
	14	68	
Liverpool	English	-	
	97	-	
Madrid	Spanish	Portuguese	English
	67	15	13
Paris	French	English	
	53	39	
Rio de Janeiro	Portuguese	English	
	81	12	
Santiago de Chile	Spanish	Portuguese	English
	74	15	5
St. Peterburg	Russian	English	
	84	8	
Stockholm	Swedish	English	
	64	27	

each city. The size of a sample was 100,000 tweets. Furthermore, a sample was pruned from retweets, usernames, links, and hashtags, so that the content is as much as possible pure text. Then the resulting text was given as an input to NGramJ to recognize the used languages. The results can be seen in Table 6.7. What is evident, is that the English language has a strong position in the tweets. English was detected in every city and was even more commonly used than the native languages in Helsinki and Copenhagen. However, in Amsterdam, a city with a fame as a cosmopolitan place, the English language was less prominent than in for example in Berlin and Paris.

6.5.2 Hashtags

The Twitter service does not allow many ways to annotate or categorize the content of a tweet. Fortunately, hashtags offer an originally unofficial

Table 6.8: Common hashtags used in Liverpool tweets.

R	Name		R	Name
1	#Jobs		6	#leadersdebate
2	#ff		7	#etsy
3	#ICD		8	#eurovision
4	#lfc		9	#bgt
5	#fb		10	#horror

mean to add some semantic information to the tweets. Use of # symbol in front of a keyword will indicate that the tweet is intended to be related to a certain topic, making it possible to group and search tweets by topics. For us, this provided a simple, yet effective method to analyze to what the content of a tweet relates to. Table 6.8 lists the most used hashtags in our Liverpool dataset that was collected by the location-based method. The common hashtag was jobs, which is utilized to mark open job positions. The second most used hashtag is ff, which stands for Follow Friday. The idea behind it is to recommend on Fridays other users some new users to follow on Twitter. ICD is the acronym of InCourts Daily, a Twitter feed that lists Crown Court activity. The fourth hashtag is for the discussion around the football club Liverpool FC. The number five is bit different, as it a sort of a technical command that allows people who have linked their Twitter and Facebook profiles to post tweets that will show on their Facebook page automatically. Etsy is hashtag for a popular e-commerce site and horror, at least partly, corresponds with one of their promoted campaigns. Bgt refers to the popular reality TV show Britain's Got Talent. All in all, according to the popular hashtags the content of tweets in Liverpool covered many topics such as employment, sports, commerce, politics, and entertainment.

6.6 Summary

In this chapter we analyzed Twitter content creation. We collected data using two different approaches, the first was based on the location of the users and the other used keywords. Examining the daily and hourly tweet patterns we have observed differences in the tweet creation between Twitter users in Madrid and Liverpool. The results also indicate that users are tweeting about current events, such as sporting events, elections, awards shows, and topical situations. One key observation is that the users are willing to express both their positive and negative thoughts.

During the measurement process we noticed that, both in Liverpool and Madrid, a small group of users produce most of the content, while most users produce very little. We saw that in our data a group of users that made up less than 10 percent of all users posted nearly 80 percent of all the collected tweets and that the majority of the user post rarely. We also noticed a sharp increase in both the number of users and the number of tweets during the measurement period.

Our quick natural language analysis revealed the strong position that the English language has on Twitter, especially in the Northern Europe. That popular hashtags in Liverpool covered many topics such as employment, sports, commerce, politics, and entertainment.

We saw in the results that for example the Apple product launches were clearly visible on Twitter. In the next chapter we will present a combined analysis of news and Twitter messages, in which we will also further examine the product launch visibility of numerous companies on Twitter.

Chapter 7

Combined Analysis of News and Twitter Messages

This chapter is based on the Publication II (see Section 1.4)

Our results from the previous chapter demonstrated that Twitter users do react to topical, news-worthy events. For instance, recall Figure 7.1, which plots the number of posts that contain keywords related to the 2009-2010 outbreak of H1N1 virus (swine flu). The curve matches almost perfectly with the peak of the outbreak and declines as the epidemic decayed. In this chapter we will show that the topicality can be extended to business events, such as new product releases, and some releases indeed generate a large number of posts.

We will argue that our practice is not applicable to the Twitter service exclusively, however we elected to survey Twitter for a number of reasons. It has a huge number of users and is used world-wide. Because tweets are limited in length, the amount of data to be collected is kept manageable and it also helps maintain the analysis process simple. However, the most important factor for us was its openness. By default all tweets are public and the service offers a relatively functional and free API for gathering data.

Until recently, a large part of research on social media has focused on analyzing and examining networks and graphs that emerge among users, references and links, and measuring patterns in creation and consumption of content. At present, more attention is being devoted to analyzing the vast volume of messages in the social media in terms of the *content* of the messages itself. Researchers in academia and industry are eager to mine the content for information that is not available from other sources, or before

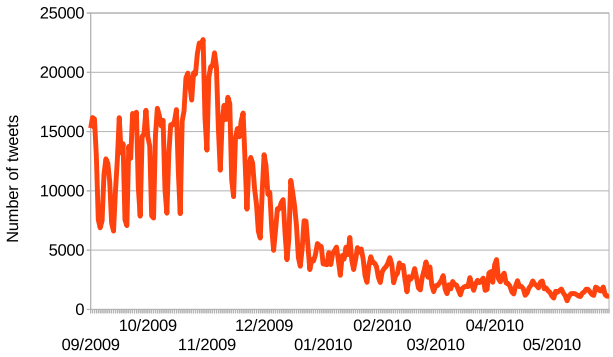


Figure 7.1: H1N1 on Twitter.

it becomes available from other sources (for example, see [4, 5], and other works of the authors).

However, our work is not aimed at event discovery in Twitter. Instead, we try to discover how events, which we find in other sources—e.g., in traditional media—are presented on Twitter. We assume that it is worthy to know not only what kind of events can be found in Twitter but also events that are not present in tweets. For example, continuing the previous example, we can note that apart from flu there are many other diseases that can be less represented or completely absent from tweets.

From the point of view of natural language processing (NLP), the immediate problem that arises is that the linguistic register and language usage that is typical for social media content—such as web logs, and especially the ultra-short messages, such as those on Twitter—is very different from the register and usage in "traditional," well-studied sources of on-line textual information, such as news feeds. Therefore, it has been observed that new approaches are needed if we are to succeed raising the quality of analysis of the content of social media messages to useful levels. This territory remains largely uncharted, though the need is quite urgent, since a better understanding of the content will enable developments in areas such as market research and advertisement, and will also help improve the social media services themselves.

In this chapter we examine how companies and products mentioned in the news are portrayed in message streams on the Twitter social networking service; in particular, we focus on media events related to the announcement or release of new products by companies. Our main research questions are: do interesting correlations exist between reports of a product release in the news and the volume of posts discussing the product on Twitter? Are

some types of products more likely to generate more posts than others? Do different types of products trigger the generation of different types of messages (e.g. retweets or tweets with links)?

One serious problem when conducting social media research is managing the data collection, and assuring that the system does not become overwhelmed with an enormous volume of data. We present a hybrid approach, where we first apply Information Extraction (IE) to messages found in news streams to narrow down scope of potentially relevant data that we will subsequently collect from Twitter. The volume of news is orders of magnitude smaller and more manageable than the volume of Twitter. In particular, extracting company and product names mentioned in the news will yield keywords that will match hot topics on Twitter. Although we may miss some important events on Twitter using this procedure, we reason that it is more tractable than continually keeping track of a large list of companies and products. An equally important factor is the fact that keeping lists of companies and products is not only impractical, but it is also insufficient, since new companies and novel products are introduced to the markets every day.

Our contributions and results include:

- we demonstrate how deeper NLP analysis can be used to help narrow down scope of messages to be retrieved from social-media message streams;
- we observe interesting correlations between events that are found in the two sources;
- we present some details about the content of tweets that correspond to news-worthy events: e.g., proportions retweeted messages and links, showing that sharing links is common when discussing certain products.

The remainder of the chapter is structured as follows. Section 7.1 describes the event extraction process, and covers the details of the data collection from Twitter. We discuss our results in Section 7.2, and Section 7.3 presents our summary and an outline of future work.

7.1 Description of Data

As in Chapter 5 we will use PULS system to extract events from news text. In Business scenario events typically include merges and acquisitions, investments, layoffs, nominations, etc. We focus on “New Product” events,

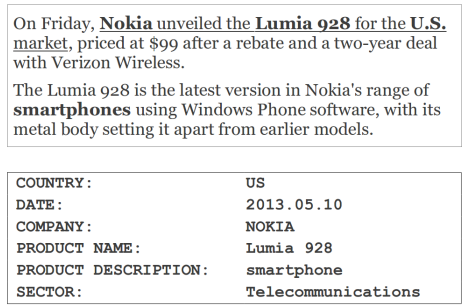


Figure 7.2: A news text and a "New Product" event, extracted from this document by IE system.

i.e., when a company launches a new product or service on the market. Figure 7.2 presents an example of a piece of text from a news article and an event structure extracted from this text. A product event describes a company name, a product name, a location, a date, and the industry sector to which the event is related. These slots are filled by a combination of rule-based and supervised-learning approaches [25, 69, 29].

For identifying the industry sectors to which the events relate, we use a classification system, currently containing 40 broad sectors, e.g., "Electronics," "Food," or "Transport." This classification system is similar to existing classification standards, such as the Global Industry Classification System (GICS),[1], or the Industry Classification Benchmark (ICB, http://www.icbenchmark.com/), with some simplifying modifications. The sector is assigned to the event using a Naive-Bayes classifier, which is trained on a manually-labeled set of news articles, approximately 200 for each sector, that we collected over several years.

We use the new-product events extracted by PULS to construct special queries to the Twitter API. One query contains a company name and a product name, which are the slots of a product event (see Figure 7.2). Every day we extract about 50 product events from news articles, and generate 50 corresponding queries to the Twitter API.

We then use the Twitter API and collect all tweets that include both the company and the product name. Below one can see an example tweet containing the company name *Audi* and the product name *A3*:

```
The new A3 from Audi looks great!
```

[1] http://www.msci.com/products/indices/sector/gics/

Time	Events	Tweets
Nov 2012–May 2013	1764	3,842,148

Table 7.1: Dataset description.

The Twitter API has some restrictions. While conducting our survey[2], we could make 150 requests per hour, asking for 100 tweets per request, yielding a maximum of 15,000 tweets per hour. We had at our disposal the University of Helsinki cluster consisting of approximately 200 machines, giving us the theoretical possibility to collect up to 3,000,000 tweets per hour.

While the company and product names are used as keywords in the Twitter query, other slots of the event are used for analyzing the results of the query. These slots, which include the industry sector, the country, the product description, and the date of the report, are used to label the tweets returned by the query. For example, we extract an event as in Figure 7.2 and get 2,000 tweets which contain both "Nokia" and "Lumia 928". Since the event is related to the industry sector "Telecommunications", we consider these 2,000 tweets are also related to "Telecommunications". Thus, we can group the returned tweets by industry sectors, country, etc., and analyze the flow of information.

The Twitter API lets us fetch tweets from seven previous days, and we kept collecting the tweets for each keyword for at least 3 days after its mention in the news. Thus, every keyword query has a time-line of roughly ten days around the news date.

The dataset is summarized in Table 7.1. We started the survey in November 2012 and the results include data collected through May 2013. In total, there are 1764 different events and in total close to 4 million tweets. In the final section of this chapter we will discuss how we plan to improve the data collection in the future.

7.2 Experiments and Results

7.2.1 Tweet Statistics Overview

First we present an overview of the tweet statistics. Table 7.2 summarizes the statistics, grouping the events based on the number of tweets they generated. The table also lists the total number of tweets, the percent of tweets that contain at least one hyper-link URL, and the percent of

[2]The access conditions have been recently changed

Number of tweets	Number of events	Links %	Retweets %	Unique tweets %
10k+	33	82	22	52
1k-10k	68	78	23	53
100-1k	109	79	24	61
10-100	258	84	18	73
1-10	249	85	12	85

Table 7.2: Overall statistics: number of tweets, links and retweets per event.

"retweets". A retweet is somewhat analogous to forwarding of an email. A retweets starts with "RT" abbreviation, making it easily distinguishable. Note that retweet can contain additional text compared to the original tweet, e.g., the retweeting user's personal opinion. The last column on the table represents the fraction of unique tweets; to count this number we subtracted from the total amount of tweets the number of tweets which were exactly identical. We pruned away the shortened link URLs from the tweet text when we calculated the uniqueness percentage, since the same URL can be shortened differently.

As can be seen from Table 7.2 there were 33 product events that generated more than 10,000 tweets. Strikingly, 82 percent of the tweets had a link. We checked a random sample through a subset of the tweets, and it seems that the single most common reason for the high number of links is that many websites today have a "share on Twitter" button, which allows a user to share a Web article with his/her followers by posting it on the user's Twitter page. The resulting tweet will have the article's original title, a generic description of the article (such as the one used in a RSS feed), and a link to the actual article. This can also be seen on the last column in Table 7.2, since the resulting tweets are always identical.

It is interesting to observe that the tweet uniqueness drops as the number of tweets increases. This would seem to indicate that the likelihood that an article is shared increases with the number of times it has already been shared. The same seems to hold for retweets as well. This corresponds to the observations found in literature: it was shown, [39], that if a particular tweet has been retweeted once, it is likely that it will be retweeted again. Similarly, tweets that contain a URL are more likely to be retweeted [58]. However, tweets related to business are rarely retweeted [72].

7.2.2 What is Tweeted most Frequently

The total number of distinct companies present in our data set is 1,140. The majority of these companies occur in one event only; for 50% of the companies less than 10 tweets have been returned. The list of most frequently tweeted companies is shown in Table 7.3. We show the number of events for a company in our dataset, the maximum number of tweets for any one event, and the total number of tweets for the company.

It can be seen from the table that only events related to well-known IT giants, (Facebook, Google, Microsoft), produce more than 100,000 tweets. Nokia, which is on the fourth position, produces 8 times fewer tweets than Google.[3]

Other companies in table are telecommunication and automotive companies, food and drink producers, cosmetics and clothing suppliers. By contrast airlines receive little attention, the news about opening new flight routes cause little response on Twitter. For example, the only tweet related to a new flight by Air Baltic between Riga and Olbia was found in a Twitter account which is specialized for the airline's news.

The list of the most frequently tweeted industry sectors is shown in Table 7.4. Note, that the business sectors are assigned to events, not to a particular company; for example, an event that describes Facebook launched "Home," an operating system for mobile phones, was assigned with the sector "Telecommunications Technologies", while an event that describes that Facebook launched Graph Search was assigned with sector "Media, Information Services".

As can be seen from Table 7.4, the sectors in our data are distributed approximately according to Zipf's law: the majority of tweets are related to a limited number of sectors, while the majority of sectors trigger little or no response on Twitter. For example, we do not find any tweets related to such sectors as "Construction" or "Minerals & Metals"; the "Agriculture" sector generated only 3 tweets.

Comparing Tables 7.4 and 7.3 we can observe that there is a dependency between the number of events related to a particular sector and the number

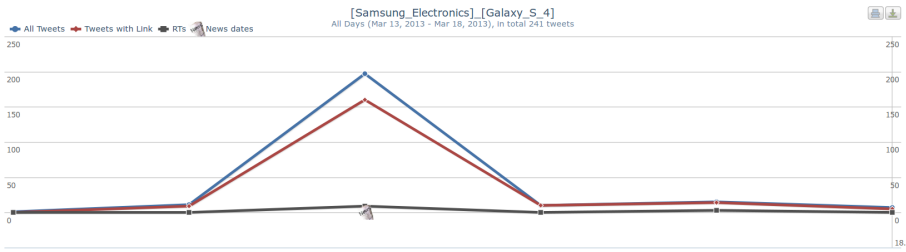
[3]We have found relatively few tweets related to Samsung Electronics, even though these events are about launching new smartphones and other gadgets, which seem to be very popular in Twitter. We believe that we did not find more tweets because the full name of the company—"Samsung Electronics"—is rarely used in the tweets, which tend to refer to it as "Samsung;" this type of synonymy will be taken into account in future work. The majority of tweets related to Samsung are links to news (see an example in Figure 7.3a); the text of these tweets are mostly identical (Figure 7.3b), which means that people do not type new information but only click the "tweet" button on the news page.

COMPANY	# events	max # tweets	total # tweets
Facebook	13	444188	1931445
Microsoft	18	440831	447104
Google	24	410986	877842
Nokia	8	52955	60655
Nintendo	2	46611	75275
Apple	8	19619	42243
Lamborghini	1	21951	21951
Adobe	3	16230	17801
Lego	2	15371	26001
Audi	9	13373	13829
Netflix	2	9880	14249
Casio	1	8970	8970
Amazon	5	8678	10079
Huawei	5	8559	8906
Sony	12	8081	12459
T-Mobile	2	7884	9043
Adidas	13	6487	9171
Acer	1	6099	8592
Volkswagen	2	4454	4454
Subaru	1	4397	4397
Macklemore	1	4301	4301
Zynga	2	4166	4170
Starbucks	1	3993	3993
Lenovo	2	3129	3129
Land Rover	3	2951	4619
Seat	1	2641	2641
Walmart	1	2575	2575
Samsung Electronics	24	2566	4578
Chevrolet	2	2517	2558
Coca-Cola	23	2432	5891
Deezer	1	2107	2107
Tesla Motors	1	2082	2082
Macef	1	2073	2073
Telefonica	6	2065	2090
Orange	7	1958	2532
H&M	2	1787	1787
Dacia	2	1650	1849
Intel	2	1649	1649
Dell	2	1074	2450
Lacoste	2	799	821

Table 7.3: Most frequently tweeted companies.

SECTOR	# events	max # tweets	total # tweets
Media, Information Services	109	444188	1534300
Telecommunications Technology	122	337776	531920
Information Technology	33	169086	182408
Consumer Goods	41	15371	29440
Drinks	94	3993	10312
Automotive Engineering	66	4454	10098
Transport	36	1714	9570
Cosmetics & Chemicals	113	3480	6194
Food	106	4369	5751
Energy	6	277	374
Finance	45	179	316
Textiles	10	166	290
Health	25	81	239

Table 7.4: Most frequently tweeted industry sectors.



(a) Number of tweets, links and retweets related to an event "Samsung Electronics launched Galaxy S4".

Timestamp (sort)	ID	Tweet	
2013-03-15T17:59:31	312624113040625665	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/98XgVNyfpp	campus42
2013-03-15T17:59:11	312624028839981057	Noticias TNO Venezuela - Samsung presenta el GALAXY S 4: Samsung Electronics anunció hoy la cua... http://t.co/5iL3luXEJf #noticias #tno	Grupotno
2013-03-15T17:56:59	312623476605341696	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/MzBzNoqzxa	World___News
2013-03-15T17:56:57	312623465184235521	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/aMFxz7ZmEC	SantinaMelgar
2013-03-15T17:56:54	312623455122124800	#TeamFollowBack How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. re... http://t.co/mLfYFHDsS4 #AutoFollowBack	Vermandita
2013-03-15T17:56:50	312623436721704961	#Tech How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-... http://t.co/FzsjeQjllh	zankrut
2013-03-15T17:56:50	312623435819917312	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed it... http://t.co/nvVT1PXlYW http://t.co/jczlrGv4c1	Riyajain1
2013-03-15T17:56:48	312623429033553921	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/nKRRFRzqRn	technewsplace
2013-03-15T17:56:46	312623421320216578	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/EgvCvVMNgR	drGalauuu
2013-03-15T17:56:44	312623412621213697	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line... http://t.co/GbSS116OHA	TomFlowers

(b) Tweets related to an event "Samsung Electronics launched Galaxy S4".

Figure 7.3: Samsung Electronics example.

of tweets related to this sector, whereas there seems to be no such relation between the number of events related to particular company and a number of tweets related to this company. For example, only one event involving

Acer appeared during the covered period—a launch of the "Iconia B1" tablet—but it drew more than 6,000 tweets.

The dependencies between the number of events and the number of tweets for companies and sectors are presented in Figures 7.4 and 7.5 respectively.

All events were taken from news written in English, but depending on the resulting keywords, the tweets that match the query could be in any language. Since we use the English names for companies and products there is an inherent bias toward countries that use languages with a Latin-based script. However, despite that were able to find many tweets for events that happen in countries that use non-Latin scripts, e.g., Russia or Japan. Two reasons for this may be that the larger companies operates globally, and that Twitter users tend to type company and product names in English even though they tweet in their own languages, see examples in Figure 7.6.

7.3 Summary

In this chapter we presented an combined framework, which allowed us to analyze the influence that business news have on tweets. We have demonstrated that the impact that new-product events have on Twitter depends more on the industry sector than on a particular company.

Our data, as it was shown before, include the event date and the timestamps for tweets. However, in the work this information has been overlooked in the analysis. Thus, in future work could focus more on the temporal dimension for example by adding more metrics, such as the time gap between the product launch and the peak of tweets.

Furthermore, we would like to see whether the impact created by a product launch based on the history could be predicted and to find out if there are some models to match that and the corresponding tweets. To solve this problem, the data collection process could be modified by monitoring several big companies for a longer time, in order to establish baselines. This would permit, first, to analyze the exact impact of a product launch on Twitter volume and, second, to measure an impact of corpus narrowing using information extraction.

Another aspect of the data, which would be interesting to further investigate, is location. As have been shown before, the business events include a country slot. Thus, tweets could be collected for certain location as we showed in Chapter 6 or geolocation techniques, [17, 6] could be utilized, to find the tweets' countries and to compare them with the countries found in news.

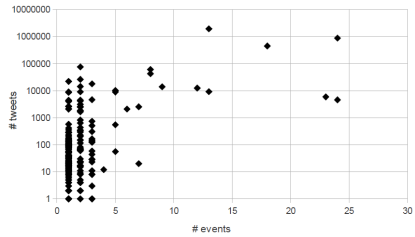


Figure 7.4: Number of events against total number of tweets for companies.

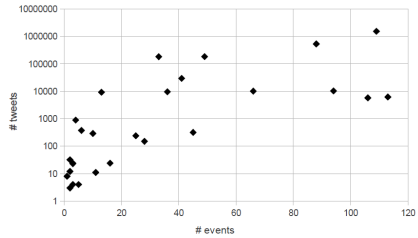


Figure 7.5: Number of events against total number of tweets for sectors.

2013-04-22T23:59:41	326485492076003328	みんなGoogle Glass好きなんやなぁ	mowsmow
2013-04-22T23:59:41	326485490150817793	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおおお！！！！！1」 こういう未来ですか？	maxonK
2013-04-22T23:59:40	326485486585655297	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおおお！！！！！1」 こういう未来ですか？	matoriv
2013-04-22T23:59:40	326485485310595074	エロがなければ映像ソフトの発展はなかったように エロがなければ革命的デバイスの発展も望めないのだ だからGoogle Glassがラッキースケベ共有のために使われるのは 極めて自然である	ragemax
2013-04-22T23:59:38	326485478796832768	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおおお！！！！！1」 こういう未来ですか？	Miyata_lori
2013-04-22T23:59:38	326485477521764352	RT @latercera: Google Glass utilizará pestañeos para sacar fotografías y los dedos para hacer zoom http://t.co/7mCluWBlWp	armagonttboy361
2013-04-22T23:59:37	326485475185537025	RT @ragemax: Google Glassが買えても きっと日本人ならろくでもない使い方しか思いつかないに違いない	uninosuke
2013-04-22T23:59:31	326485450359459840	Google Glass、どうせシャッター音消せないんだろ？意味ねえな	jiejumin
2013-04-22T23:59:31	326485448346198017	【速報】Google glassすごすぎｗｗｗｗｗｗｗｗｗ http://t.co/xKGlqeP9yg	asahirovip

Figure 7.6: Tweets related to an event “Google launched Google Glass”.

The query construction algorithm can be improved to find more tweets for compound company names, such as “Samsung Electronics.” This, however, cannot be done in a straightforward fashion: “Samsung” may likely refer to “Samsung Electronics”, though “Electronics” may refer to many different entities. Thus it is not possible simply search for all substrings of a company name, because such queries will produce too many false hits. We assume that special named entity recognition techniques, which have been developed for Twitter [54, 51], can be used to solve this problem. To improve coverage it is also possible to utilize automatic transliteration, which allows to map proper names from Latin to other scripts [47].

We have studied the most and least frequently tweeted companies and industry sectors. One could also focus to study the most frequently tweeted product types. Since every product found by IE system has a description (as presented in Figure 7.2), tweets can be grouped by product type. However, additional work is needed to merge such product types as, for example, "chocolate" and "chocolate candies". This could be achieved by using a Business concept ontology, which includes the long list of possible product types.

Chapter 8

YouTube

This chapter is based on the Publication I (see Section 1.4)

In the Chapters 3 and 6 we saw two service measurements. With Wikipedia we had access to complete, if not totally perfect, data of the page requests. With Twitter we needed to use the API provided by the service and take rate limitation in to the consideration and even then we saw that we can only obtain a small sample of the data. In this chapter we present measurement results of YouTube, but our main goal is to show the effects of four different sampling methods on YouTube.

Measuring large systems or services is challenging and typically measurements are performed via sampling since analyzing the complete system is either prohibitively expensive or even impossible. Naturally, the way the sampling is performed has a strong effect on the measurement results and the conclusions that can be drawn from them. Ideally, the sampling should be done in a way that produces a random, representative sample of the total system, but in many cases technological limitations on the sampling may skew the process away from getting a representative sample. Using such a biased sample may yield incorrect conclusions about the properties of the system and further affect any derivative work which uses those results as its basis.

YouTube is the largest and most popular video service on the Internet and has been an active focus in research for many years. Previously, YouTube's video popularity has been measured, for example, by crawling related videos [12], selecting videos belonging to certain categories [10], or by using a list of, e.g., the most recent videos [59] as the data-source. The problem with these methods is that, while the corresponding results of the measurements are valid as such, the methods lead to a biased sample, and thus, the results are not representative of YouTube in all respects. Since

other works may base their assumptions on the measured values, it is important that they indeed do represent the whole service and not a subset of it.

To demonstrate our case, we have collected four datasets, three by using methods from earlier research, and one by using a method that is based on random video IDs that has previously been used to estimate the number of videos on YouTube. We will show that, even though all data is obtained from the same source, via the YouTube API, there are noticeable discrepancies in the video popularity and other metrics depending on the method used.

The main contributions of the chapter are the following:

- We show measurement results of YouTube relating video popularity, age, length, and categories.
- We review prior YouTube measurements and data collection methodologies and show their differences.
- We compare existing methods for collecting YouTube video metadata.
- We demonstrate the differences in various metrics between the different sampling methods.

Our main goal is to highlight the importance of using proper sampling techniques and show how different sampling methods can lead to different conclusions. We do not aim to champion any of the methods, but instead try to raise attention to the way results are interpreted and accepted. That is, the authors of earlier work that we discuss and refer in this chapter have all been very clear to describe how they have conducted their research. In similar vein, the authors have been straightforward in expressing the limitations of their work and results. Thus, we see that the problem lies in the way the community accepts early results and rarely looks back. E.g. [10] is cited numerous times in reference to YouTube's video popularity, while, as we are going to show, the methods used in the paper are especially lacking in getting a representative sample and the popularity should only considered in the context of the way the data was collected.

We also argue that, while the wider consequences are out of our scope, the value of the result and the implications drawn from results span multiple research areas such as storage, replication, bandwidth and even wider disciplines such as marketing, user experience and user behavior.

The rest of the chapter is organized as follows. In Section 8.1 we discuss related work and review previous measurement methods that have been used on YouTube. Section 8.2 presents our data collection process. The

results are presented in Section 8.3 where we compare several key metrics obtained by the different methods and demonstrate their differences. In Section 8.4 we discuss about the validity of the method that is based on the random video IDs. Finally, Section 8.5 concludes the chapter.

8.1 Related Work

We already presented the related work in Section 2.2, but we will recap it here since it heavily ties with result presented in the following section.

Cha et al. [10] analyzed the video popularity of YouTube in 2006-2007. Their dataset consists of video metadata formed by crawling the indexed pages and getting videos belonging to certain categories. They had 1.7 million videos from Entertainment category and another 250,000 from Science category. Their results showed that the video popularity ranking of both categories exhibited power-law behavior "across more than two orders of magnitude" with "truncated tails" but "the exact popularity distribution seems category-dependent." The authors called for further research on the subject. The traces collected by the study have been a source for [66].

Cheng et al. [12] also measured and examined, among other things, the popularity of YouTube videos. They collected metadata for three million videos in 2007 and for further five million in 2008, using bread-first search (BFS) starting with initial video and asking its related videos and then their related videos until the fourth depth. Looking at video popularity they observed that: "though the plot has a long tail on the linear scale, it does not follow the well-known Zipf distribution." and found "that the Gamma and Weibull distributions both fit better than the Zipf, due to the heavy tail that they have".

Since the authors were concerned that the BFS method would be biased towards more popular videos, they formed another dataset by collecting metadata of videos from the recently added list for four weeks. Comparing the two datasets they concluded that also the videos from the recently added list exhibit popularity where: "There is a clear heavy tail" and "verifying that our BFS crawl does find non-popular videos just as well as it finds popular ones".

Szabo and Huberman [59] took a slightly different approach and wanted to see whether it is possible to predict content popularity. In the case of YouTube they measured the popularity and view counts of new videos for 30 days. Their data is from 2008 and consists of 7,146 videos selected daily from the recently added list. They chose the list over other alternatives in order to get "an unbiased sample". They concluded that the popularity of

a YouTube video on the 30th day can be predicted with a 10 % relative error after 10 days.

In the research mentioned above, the data has been collected either by BFS crawling, or by selecting videos of a certain category or by picking most recent videos. We will show in the results section the problems that are associated with the methods and popularity distributions they produce.

Another method is used e.g. by Gill et al. [24] who analyzed the traffic between a university campus and YouTube servers. They concluded that "video references at our campus follow a Zipf-like distribution". They reasoned it to be partly because YouTube did not allow video downloading, meaning that a user had to issue another request to see the same video again. They also found out that on a longer time frame the most popular categories were Entertainment, Music, and Comedy. Zink [74] et al. also measured the YouTube viewing and traffic patterns on a campus level and studied the effects of proxy caches to reduce traffic.

Brodersen et al. [9] studied the geographic popularity of videos and found that "about 50% of the videos have more than 70% of their views in a single region" and concluded that "videos exhibit strong geographic locality of interest". Given that all authors of the paper worked at Google, they were not limited by the API and chose randomly 20 million videos uploaded to the service between September 2010 and August 2011.

The authors of [18] collected three datasets, one using top lists, another one consisting of videos that were known to be copyright protected, and third one using random lexical ontology based topics. They saw that popularity growth patterns varied depending on the used dataset. They also found that "that search and internal mechanisms, such as lists of related videos, are key mechanisms to attract users to the videos".

On a more general level, the importance of a correct sampling method has been noted e.g. by Krishnamurthy et al. [36] who used three different data collection methods and analyzed their strengths and weaknesses in order to examine Twitter and improve the prior research, and by Stutzbach et al. [57] who introduced a technique for a more accurate and unbiased sampling for unstructured peer-to-peer networks.

8.2 Data Collection

We have collected data using four different approaches. In the first approach, we started by periodically asking a list of the 50 most recently published videos using the YouTube API version 2 and later version 3. The list included information of the videos such as ID, view count, and

publish date. Having obtained the IDs of the videos, we later collected their view counts after 30 days. We had done similar surveys in 2009 and 2011 and we wanted to compare the results by doing the same procedure again in late 2013 and early 2014. We refer to this method as MR (Most Recent). The inherent problems of the MR method are that it is a slow way of collecting data and that videos for which data is collected are limited to similar age. The method is similar to one used in [59] and [12].

However, as it is not known in which manner videos end up on the MR list and thus it is not possible to know whether they constitute a representative sample, we simultaneously started collecting data using a different method in order to verify our results. In this approach, we generated random character strings and requested through the API a list of video IDs which include the string. Hence we call this method RS (Random Strings). In more detail, the method can be described as follows. We formed four characters long strings using random characters from 'a-Z', '0-9', '-', and '_'. As the YouTube video IDs are 11-character long strings generated with the same character set, we used the strings as keywords to request video IDs containing the random strings (4 characters were the shortest strings that returned matches consistently via the search). Resulting data also included video metadata such as duration, category, etc., and on average a random string yielded 6.9 video IDs. Besides randomness, the benefits of the method are that we were able to collect a very large number of video IDs with corresponding metadata and it provided a way to get a comprehensive sample of different-aged videos. Given that different strings might match to same ID, we further pruned out the duplicates.

Interestingly, for reasons unknown to us, with this method the YouTube API only returns video IDs that have at least one '-' in them, even though, in general, video IDs do not need to contain a '-'. The "-" was usually the fifth character of the ID. However, we argue that as the search strings are randomly generated (and the IDs are likely similarly generated, although this cannot be proven), statistically the sample obtained in this manner is equivalent to a random sample over all the videos; obviously this is a potential weakness of this method. Incidentally, Zhou et al. [73] provide a detailed description and discussion of the same method, with evidence to support that it indeed provides a random sample of the videos. However, their focus is on estimating the number of videos on YouTube and they do not investigate different metrics for the videos. They also mention a potential bias in other collection methodologies, such as BFS, but do not present any evidence of that. While we conjecture that the RS method provides a random sample, for the purposes of this chapter, i.e., to demonstrate the

differences between different sampling methods, it is not strictly necessary for the method to actually produce a random sample. We will talk more about the validity of the RS method in Section 8.4. A further limitation of this method is that it will not return videos with 0 views or deleted videos.

Our third method to collect data was to randomly select a video ID and then ask for its related videos and after that the related videos for all those videos up until to the fourth level. We set a limit of 50 related videos per one video, so theoretically one seed video could return up to 125,000 videos (50x50x50). The actual number of unique videos is naturally lower, due to overlap in the related videos. This can been seen as similar to breadth-first search and we shall refer to the method as BFS. As mentioned in Section 8.1 this method has been used earlier by [12]. This method is a fast way of obtain a large set of IDs, since the API allows getting the information of 50 videos with just one API request compared to the average of 6.9 obtained with the random strings. Because a video can be, and usually is, related to multiple videos, the method also needs pruning to remove duplicates.

Later we added a category-based method to complement our datasets. The goal was to collected a similar dataset that Cha et al. [10] had collected in their research. However, getting category-based metadata was not straightforward. When we queried the API to give metadata of videos belonging to a category, e.g. Music, the API only returned at best few hundred results per day. In addition, the list of metadata stayed stable for a long period, thus limiting the dataset to a very small size. Oddly, the amount of returned metadata increased as we added more fields to the requests. After a while, we found out that limiting the videos to a certain interval, based on their creation timestamp, actually greatly increased the amount of returned data. To comply with our original goal to measure popularity after 30 days we asked daily metadata of videos published 30 days ago and repeating this over multiple weeks. Using this modified approach we were able to obtain metadata of approximately 4000 videos per day per category. We chose to focus on eight categories: Music, Science and Technology, Pets & animals, Sports, People and Blogs, Entertainment, News and Politics, and Education. Similar to the BFS method you could get information of 50 videos with single API request. We refer to the category based method as CATS.

Table 8.1 shows an overview of the different datasets that we collected using the methods described above. In the following, we refer to the different datasets by their names and in some cases combine all three MR datasets into a single set, called MR. Similarly, the category-based datasets are combined into a single set that we named CATS.

Table 8.1: Description of datasets.

Set name	Method	Time period	N
MR-09	Most recent videos	summer 2009	9,405
MR-11	Most recent videos	summer 2011	8,766
MR-14	Most recent videos	late 2013 - early 2014	10,000
RS	Random id	early 2014	5M
BFS	BFS related videos	early 2014	5M
CATS	Category-based	summer 2015	649,629

8.3 Results

As described in the previous section, we have fours datasets collected using four different methods. Now we are going to show how the datasets differ according to different typical metrics that have been used in previous research on YouTube. We start with the video popularity ranking and then use number of views, age, length, and categories to further compare the datasets. Obviously, as the MR and CATS datasets are much smaller and the videos are by definition very recently uploaded (to the time when the dataset is collected), it does not allow one-to-one comparison with the other two methods in some metrics.

8.3.1 Popularity

Figure 8.1 plots the videos of RS and BFS datasets ranked based on the view count in log-log scale. Both datasets have 5 million videos. As can be obviously seen, there is a clear difference in the view count distributions provided by the two methods. The data collected using BFS method has a clear two-part distribution, with a quick-dropping tail. The RS data follows more closely a Zipf distribution, with a truncated tail. Across the board, the distribution of BFS data exhibits much higher popularity (higher view counts), being in parts four orders of magnitude higher (around the millionth most viewed video). Since RS represents a random sample, it can be argued that the BFS method provides videos which significantly over-estimate the actual view counts in YouTube. We suspect that when determining which videos to show as related videos, YouTube proposes videos that are more popular than average, and, thus, BFS datasets are prone to have inflated number of videos with high view counts.

A simple analysis reveals that the 10 most viewed videos in RS dataset account for 5 % of the total views, 100 most viewed for 17 %, 1000 for 43 %, and 10,000 (0.2 % of the total sample) for 74 %.

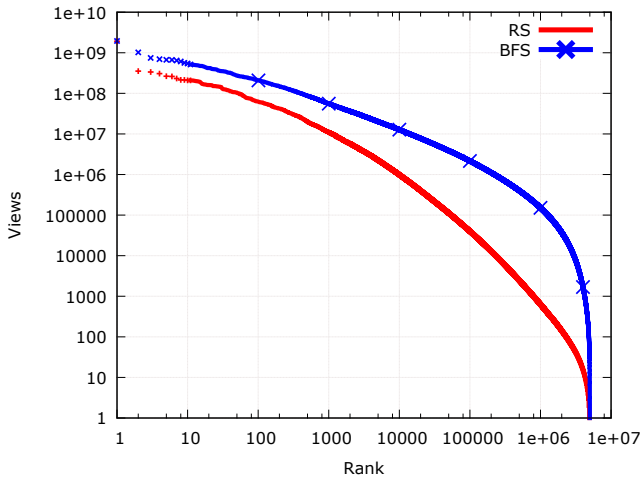


Figure 8.1: Popularity distribution by views.

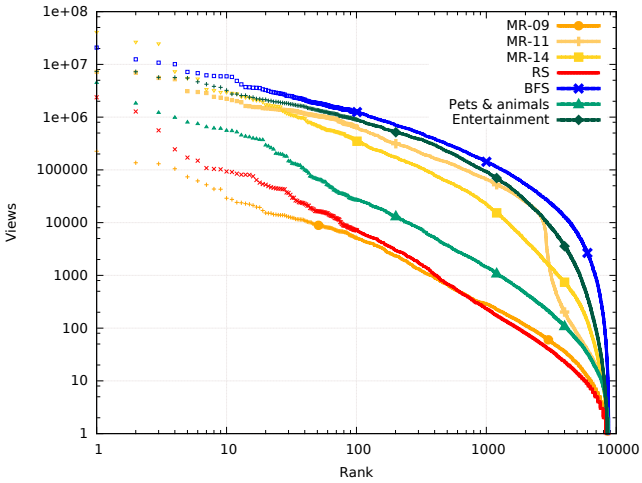


Figure 8.2: 30-day view count ranking comparison.

Popularity after 30 Days

Figure 8.2 shows the view counts of videos 30 days after their uploading, on a log-log scale, i.e., the plot captures the popularity of one month old videos, where as 8.1 plotted all videos. We show all three MR datasets separately. In order to keep the plot readable, we have included only two datasets from the CATS data, namely Entertainment and Pets & Animals. They were chosen as they were the most and least popular of all the collected categories

and the distributions of all the other categories would fall between the two. The x-axis is limited to 8766 which is the size of the MR-11 dataset (the smallest dataset in our study) to make the curves comparable. That is, we chose the 8766 videos of the MR-11 dataset and then randomly the same amount from other datasets. As can be seen, the datasets have noticeably different popularity distributions. In general, CATS, MR and BFS methods seem to overestimate the video popularity when compared to RS (Recall Figure 8.1 which shows the same result between BFS and RS across a larger dataset). Interestingly, the MR-09 and Pets & Animals shows relatively straight lines, close to that of RS, with a truncated tail, resembling the observations of Cha et al. [10], whereas the MR-11 would seem at least bipartite, pivoting around 12,000 views.

The view counts of MR-11, MR-14, Entertainment and BFS are orders of magnitude higher than those of RS. We suspect that this is because either a) new videos on the most recent list are more likely attract more views or b) being on the list will make the videos gain more views. The same conjecture applies also more or less to the related videos. Given that view counts of Pets & Animals are much higher than RS indicates that also the category-based data collection favors more popular videos.

8.3.2 View Count Accumulation

In this section we show how the view counts of videos from both RS and BFS datasets accumulate over a year. Unfortunately, the YouTube API does not offer historical view counts, that is it will only return the current view count of a video. However, the data is available through the YouTube website and can be obtained by emulating HTTP requests. Since, we did not want to overburden the service through this 'unofficial' channel, we chose to keep collected data sizes as modest.

Figure 8.3 plots the median values of the percentage of accumulated video views for 1 to 365 days from upload. The data size is 5000 videos for both RS and BFS. In more detail, we chose videos that were at least a year old, then for each of those videos we collected the daily view count values (1-365 days from upload) and calculated the accumulating view counts. So, e.g. as median value a video from RS dataset has 50 percent of its one-year views accumulated after 100 days, whereas at that point a BFS video has only be watched 30 percent of its one-year views. This indicates that videos from BFS dataset gather views more evenly throughout a year than the RS dataset videos, which could be caused by that YouTube is recommending them as 'watch next' and thus given them a steady number of views over time.

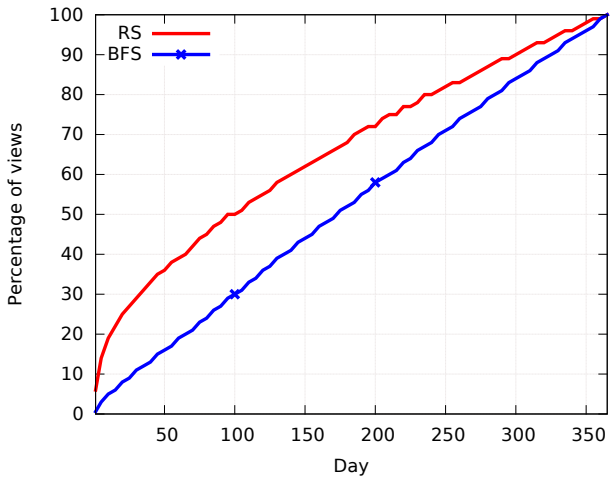


Figure 8.3: View count accumulation over a year.

8.3.3 Views

Table 8.2 lists the view count statistics for the datasets. It should be noted that the numbers for the MR and CATS datasets are not directly comparable with the others, since the datasets include mostly new videos and thus they have had a shorter time to accumulate views. As already stated, the BFS method favors more popular videos, which can be seen in the much higher mean and median values. In other words, in general, the videos of the BFS dataset are more viewed than those of RS. Figure 8.4 shows the different percentiles of the view counts. We can see that e.g. the 5th percentile of BFS is higher than the median of RS and across the board the BFS view counts are at least one order of magnitude higher than the RS ones. Figure 8.5 further illustrated this point by showing the median and the 5th and 95th percentiles of the RS and BFS datasets for eight years. For example, in the RS dataset the median value of 730-day-old videos is approximately 100 views. Looking at the percentiles we can see that there is overlap in the datasets, but the median of BFS is most of the time two orders of magnitude higher than the median of RS.

8.3.4 Age

Figure 8.6 illustrates the age distribution of the videos gathered by the RS and BFS methods. The MR data is left out as the age is already determined by the way the method works, limiting the data to new videos only. The plot is made by calculating the number of videos published on

Table 8.2: View count statistics of the datasets.

	N	Mean	Std. Dev	Median	Max
RS	5M	16,260	1,115,835	81	1,920,284,708
BFS	5M	260,019	2,595,870	19,217	1,950,573,461
MR	21K	68,553	1,205,992	461	111,762,034
CATS	650K	22,891	224,575	544	42,047,451

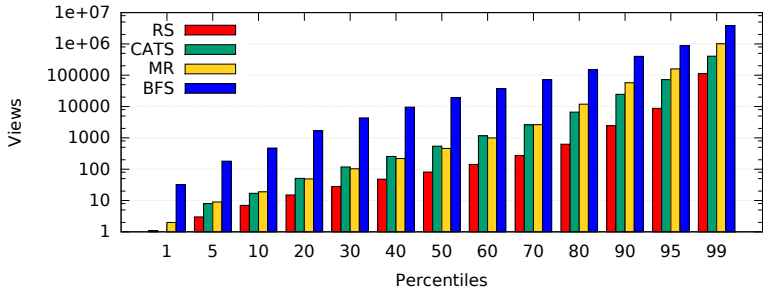


Figure 8.4: View count percentiles.

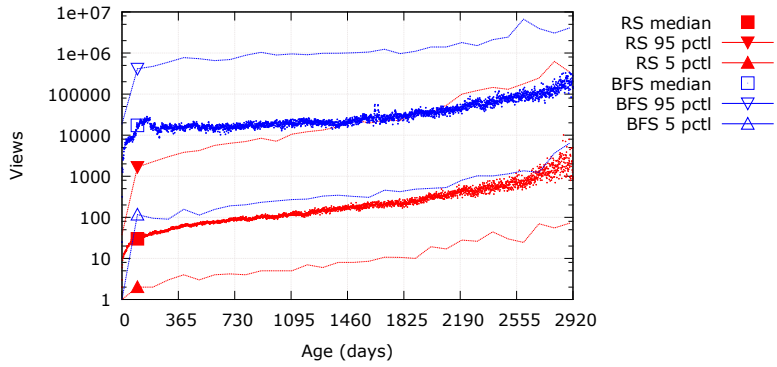
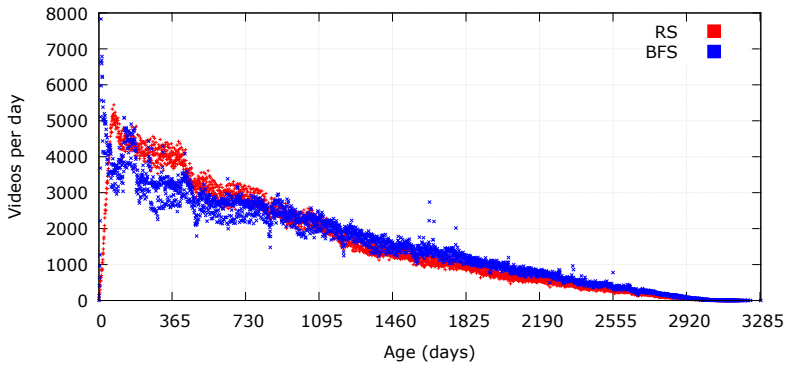


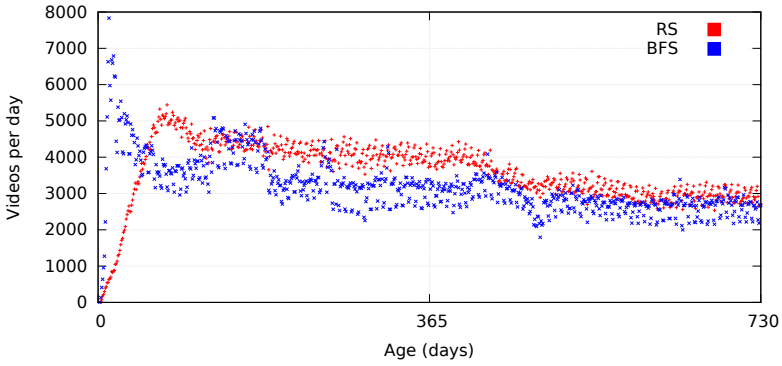
Figure 8.5: Median and 5th and 95th percentiles of RS and BFS.

each day. The BFS set has less videos that are newer than three years, when compared to the RS dataset. However, for very recent videos, the BFS dataset shows a considerable increase, reaching up to more than three times the number of videos with similar age in the RS set. It therefore appears that the selection of related videos is biased towards recent videos and implies that the BFS dataset has a disproportionate number of recent videos, when compared to the RS set.

The lack of new videos in the RS dataset is an artifact of the sampling method. This is because the method can only match existing videos and



(a) All videos

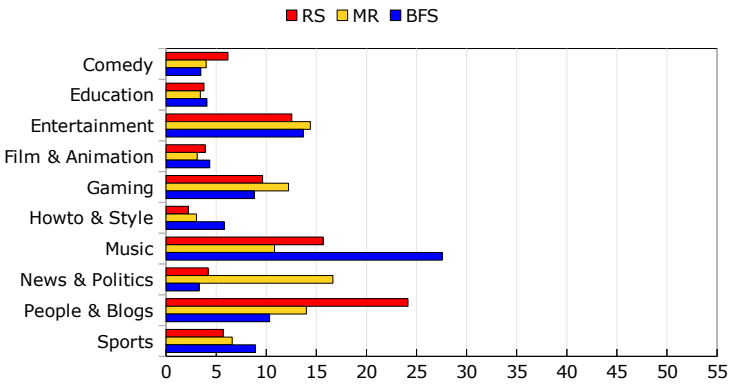


(b) Videos that are newer than two years

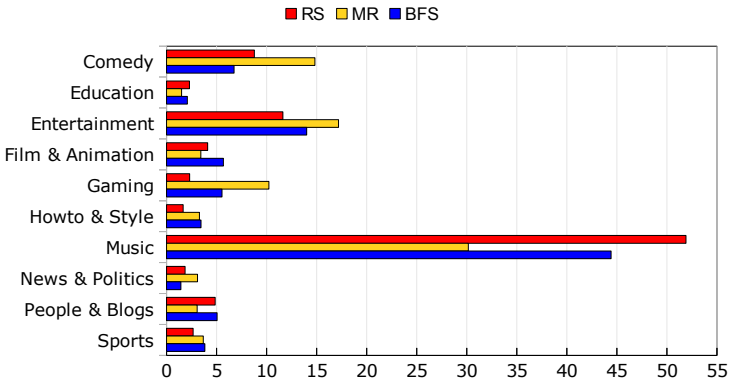
Figure 8.6: Video age distribution.

therefore videos that were uploaded after the data collection began have had a smaller probability of being selected, thus artificially reducing their number in the set. This effect can be eliminated simply by not counting the videos published during the data collection period.

On a more general note, looking at the RS data, we can see that the number of videos has grown rapidly, (even exponentially in some points), and continues to do so. Videos that are less than six months old make up 14 % of all video, less than one year 29 % and less than two years 53 %. In other words, majority of the YouTube content is newer than two years and 80 % newer than four years. Hence, the rate at which videos are uploaded to YouTube is still increasing and majority of videos have been published in the past two years.



(a) Percentage of videos



(b) Percentage of views

Figure 8.7: Video categories.

8.3.5 Categories

Figure 8.7a shows the fraction of videos in different categories in the different datasets, excluding CATS since the data is based on categories. The bars for MR combine all the three MR datasets MR-09, -11, and -14. Interestingly, the category with most videos is different in each dataset and the differences are significant. RS has most videos from the People & Blogs category, MR's biggest category is News & Politics, and Music is the largest category for BFS. When uploading a video, YouTube requires that the user sets a category for the video. If user does not explicitly define a category,

Table 8.3: Length statistics of the datasets.

	N	Mean	Std. Dev	Median	Min	Max
RS	5M	296	614	157	1	131,516
BFS	5M	512	1,181	247	1	800,492
MR	21K	545	1,535	190	1	45,122
CATS	650K	828	1,974	252	1	86,376

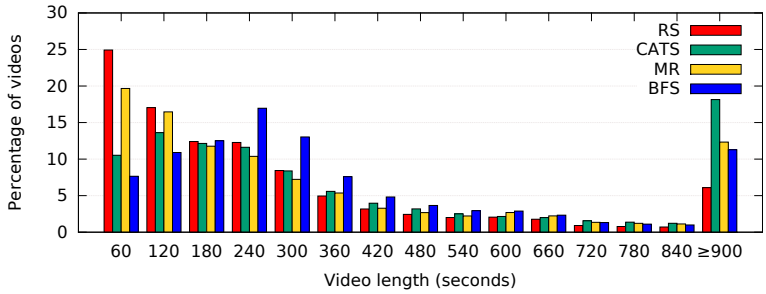
YouTube sets the video's category to the category of the last video that the user uploaded. If no prior upload exists, YouTube sets the video's category to People & Blogs, which is a very likely explanation why the RS dataset has the most videos in the People & Blogs category. Likewise, since MR takes the videos from the (curated) most recent list, it is not surprising that topical events dominate the list. For BFS, the high number of music videos is also not surprising since suggesting another music video as a related video to another music video seems intuitive.

However, even though the number of videos in different categories is very different for the three datasets, Figure 8.7b shows that the distribution of number of views across categories in the three datasets is very similar. Music is the most watched category for all three datasets, followed by Entertainment and then Comedy. Again, this highlights that the results from different methods may end up looking similar on some metrics, but not on others.

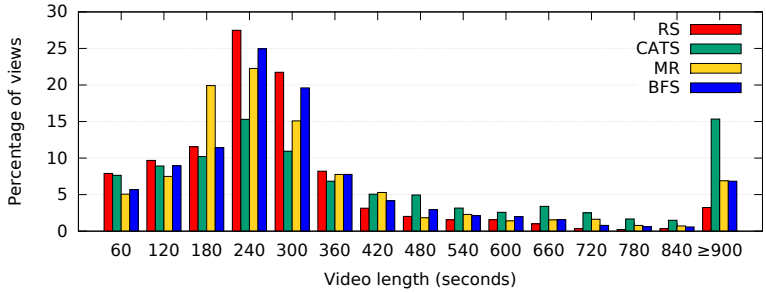
8.3.6 Length

YouTube used to cap the video duration to 10 minutes, but in July 2010 the it was extended to 15 minutes and a user can remove the limit completely by verifying the account. Table 8.3 shows the length statistics. The lengths are in seconds. We have manually checked that the maximum value for the BFS dataset is valid. The median video length is the highest for the videos of the CATS dataset, followed by BFS, MR and RS, CATS dataset has the also the highest mean and standard deviation values.

Figure 8.8a shows how the lengths of the videos in the datasets vary; the videos have been rounded to the next minute for plotting. Both RS and MR show that the most common length of a YouTube video is 60 seconds or less and that majority of video are less than three minutes long. The BFS and CATS data in turn indicates that most videos are between three and five minutes. This can be considered further evidence that BFS promotes certain types of videos forming a biased sample; as we already saw that BFS contains more music videos which are typically three to five minutes



(a) Percentage of videos



(b) Percentage of views

Figure 8.8: Video length.

long. Interestingly, MR and RS differ only in that MR has more videos over 15 minutes whereas RS has more videos of one minute or less.

However, Figure 8.8b shows videos between three and five minutes have the most views in all datasets, except CATS which has a high number of videos over 900 seconds. If this data were used to produce an estimate of how much traffic YouTube sees, RS, MR, and BFS would yield similar values, with MR being likely slightly below the others as it contains proportionally more videos of around 3 minutes, whereas CATS would have a higher traffic share of long videos.

Figure 8.9 show total duration of videos uploaded per day as a function of the age of the videos. This could also be used to obtain a rough estimate of total storage requirements of YouTube service. Again, BFS has longer video lengths. As the figure shows, the amount of data has risen almost exponentially for years. 40 % of the amount consists of less than one year old videos and 80 % of videos newer three years.

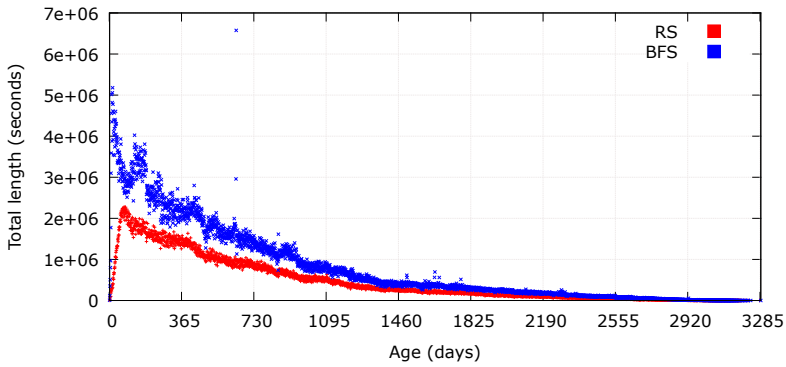


Figure 8.9: Total video length per day.

Table 8.4: Summary of methods.

Name	Used e.g. in	Characteristics
RS	[73]	Not fixed to any age or category, limited to certain video IDs
BFS	[12]	Fast, favors popular videos
MR	[59], [12]	Slow, limited to new videos, favors popular videos
CATS	[10]	API restrictions

8.3.7 Summary of Results and Methods

Table 8.4 summarizes the methods. When comparing the four methods among themselves, BFS tends to over-estimate most of the metrics we used and cannot therefore be considered a reliable way to represent the whole of YouTube videos. However, it is the fastest of the four for collecting a large dataset and seems to be able to capture popular videos which might have the most research interest. MR, on the other hand, is a very slow method, limited to new videos only, and it also tends towards over-estimation of the metrics. CATS in turn, based on our experience is restricted by the cumbersome API making the data collection limited e.g. in our case to only new videos. While we consider the RS method to be the most reliable, its weakness is that it is not very fast (recall that it returns on average 6.9 videos per query). Also, since almost all returned videos contain '-', there is potential for a bias in the returned videos, in case video IDs are not assigned randomly. We will discuss this in more detail next.

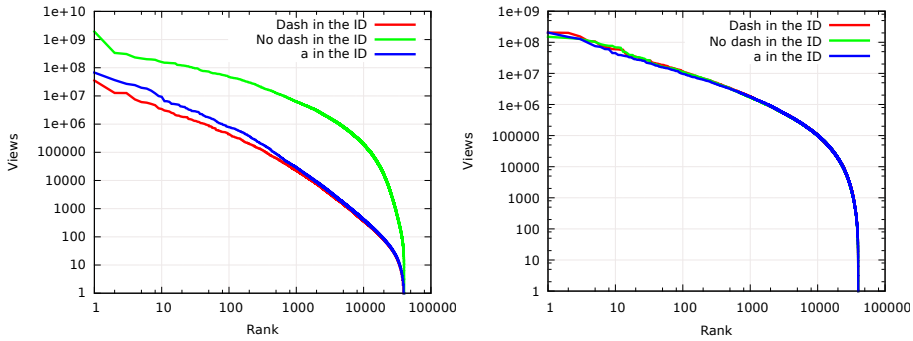
8.4 Discussion about the Validity of RS method

Like stated earlier, we started noticing that when we queried the API using RS method almost all the returned video IDs included a '-'. This was also marked by Zhou et al. [73] in their work, but they provided evidence to support that it indeed provides a random sample of the videos and that YouTube video IDs are generated uniformly from the ID space. Of course, calling the sample random is inherently dubious since clearly the videos are mostly selected from pool where the videos have at least one dash in their IDs. In our RS dataset, over 99 percent of the videos have dash in their IDs. However, what is left unclear is the role of the videos with IDs without the dash. We will now examine that.

Figure 8.10a show the popularity ranking plotted similarly to what we saw in Section 8.3.1. We have plotted three derived datasets from the original RS data. There were a little over 40,000 (out of 5M) video IDs without a dash in the RS data. We chose 40,000 of them as one dataset, second dataset consists of 40,000 video IDs with at least one dash in them, and to the third one we chose 40,000 videos with IDs that included at least one letter a. While the popularity distributions of 'dash' and 'a' datasets can be seen to behave similarly, we can observe a clear distinction with the dataset without dashes. The videos are in cases more than two orders of magnitude more popular than video in the other two datasets. This indicates that for some reason the API includes periodically more popular videos to the returned data, which seem to be recognizable by the lack of dash in the ID. Figure 8.10b plots the BFS data in similar fashion and as seen in the BFS data the videos without dashes in the IDs do not have a different popularity distribution. CATS data is not presented, but it behaves similar to BFS.

To further illustrate this point we have plotted in Figure 8.11 the percentage of video IDs with dash in the 100,000 most popular videos in the RS dataset. The figure plots percentage shares of videos IDs with a dash in them, based on the popularity ranking. To calculate the shares we grouped the videos into groups of one thousand videos based on their ranking. The figure shows us even more clearly that most popular videos in the RS dataset are the ones without a dash in the IDs.

Although the videos IDs without a dash only account for a fractional share of the whole dataset, they seem to represent a very large share of the most popular videos and if they are seen as an anomaly then they have clearly inflated to overall popularity of the RS dataset. A solution could be just simply prune them out of the dataset, but at this point we can not yet conclude that it would yield a 'better' dataset without any complications.



(a) RS (b) BFS

Figure 8.10: The popularity depending on the video ID.

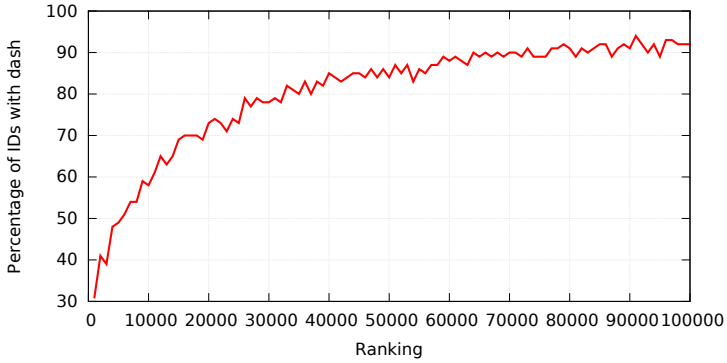


Figure 8.11: Role of dash in video popularity.

More research is needed on this aspect. However, we can still argue that when comparing against the other data collection methods, the RS does produce at least the most varied set of data, even if it cannot really fully be considered as a random sample.

8.5 Summary

In this chapter we have argued that data collection methodology can have a significant impact on what kinds of results can be obtained from measurements. We have used YouTube as an example and considered four different data collection methods, three from existing research and one adapted from previous work. By comparing the datasets obtained with the different methods, we have shown that they differ, sometimes greatly, in many of

the key metrics used in past research on YouTube. Even a large sample is not immune to the bias introduced by a particular sampling method, as the results of the BFS dataset demonstrate.

The random string sampling method behind the RS dataset has not been used to measure different metrics on YouTube whereas MR and BFS have been used in previous research to characterize YouTube. Given the large difference between RS and the others on several key metrics, it is natural to raise questions about the general applicability of previously obtained results on YouTube done via MR, CATS, or BFS methods. As we have shown that depending on the metric and the collection methodology, results may differ either qualitatively, quantitatively, or both, or they might not differ from the RS dataset. While we have strong reasons to believe that the RS method can produces a representative sample of YouTube, we cannot exclude a potential bias in it as we have shown in Section 8.4.

In essence, our results demonstrate that there is a need to understand the strengths and weaknesses of the different sampling methods in order to understand their impact on the measurement results. We believe that on the whole, a more critical approach to measurement methodologies is required in order to ensure that the measurements capture the essence of the measured system, to the extent that it is feasible.

Chapter 9

Conclusion

In this thesis we have studied social media content creation and consumption through large volume measurements. In Chapters 3, 6, and 8 we presented measurements of three large social media services: Wikipedia, Twitter, and YouTube. In Chapters 4 and 5 we surveyed Wikipedia activity against traditional news services and tracked the interactions across business news, Wikipedia page views, and stock fluctuations. In Chapter 7, we presented a combined analysis of news and Twitter messages.

The motivation behind our research was to try to see how users create and access content and find out if there are patterns in creation and consumption of the content. We wanted to investigate why social media sites are as popular as they are, what drives people to contribute on them, and if it is possible to model the conduct of users. We also wanted to find out any regional or cultural differences in the user behavior. Furthermore, we compared the creation and consumption patters of user-created content and commercial content.

We studied 19 Wikipedia editions and saw that the English language edition is by far the largest edition and its monthly views had doubled during the survey period. We also saw that the number of monthly page request had also increased for many editions, the Ukrainian and Russian editions have seen an increase by the factor of ten. The editions did vary in many cases, but overall we could identify groups of editions based on similarity. The Nordic editions, for example, exhibit quite similar behavior in most aspect that we measured, which would indicate a link from the online behavior to the cultural similarities. In addition, we were able to observe daily and hourly patterns in the page requests. The popularity distributions revealed that generally a small part of the content causes most of the page requests. With many editions the most popular 10 % of articles are responsible of over 80 % of the total page requests. In general

terms, majority of Wikipedia traffic is caused by popular articles. We also saw variation in the top article fluctuations, but overall 1-2 % of the top 1000 articles stayed the same for the whole survey period. When we analyzed to most popular articles of the English Wikipedia we identified that the popular article type is of a person, followed by articles relating to entertainment and places.

When we compared the activity of commercial news services with four different Wikipedia editions around the world, we saw that there were clear difference in creation patterns between commercial news sites and Wikipedia editing. Commercial sites followed a very clear diurnal pattern and a 5-day working week, while the diurnal patterns were more spread out with the Wikipedia editions. In the studied Wikipedia editions cultural and geographical differences seemed to have very little effect on the level of activity. While we did not have a definite answer, we reasoned that it could be so because the tendency to actively edit Wikipedia is an individual trait which transcends cultural barriers. We also studied of the interplay between company news, social media visibility, and stock prices. We were able to discover interesting correlations between the mentions of a company in the news and the views of its Wikipedia page. The correspondence with stock prices was less obvious.

We also found cultural and regional differences in the creation patterns on Twitter and based on the results, we can say that a small group of users produce most of the content, while most users produced very little. The results also indicated that users are tweeting about current events, such as sporting events, awards shows, and topical situations and we observed that the user are willing to express both their positive and negative thoughts. We noticed during the measurement process that the number of Twitter user had risen considerably.

We observed that product launches are clearly visible on Twitter when examined the product launch visibility of numerous companies. We observed interesting correlations between the business news and posted tweets and found out some details about the content of tweets that correspond to news-worthy events: e.g., proportions retweeted messages and links, showing that sharing link is common when discussing certain products. We demonstrated that the impact that new-product events have on Twitter depends more on the industry sector than on a particular company.

Throughout the thesis we also wanted to underline the importance of proper data collection methodology. We used YouTube as an example and considered four different data collection methods, three from existing research and one adapted from previous work. By comparing the datasets

obtained with the different methods, we showed that they differ, sometimes greatly, in many of the key metrics used in past research on YouTube. Even a large sample is not immune to the bias introduced by a particular sampling method. In essence, our results demonstrated that there is a need to understand the strengths and weaknesses of the different measurement methodologies in order to understand their impact on the measurement results.

9.1 Discussion and Future

An obvious next step would be to periodically repeat the measurements presented in this thesis and see if the new results match with our findings. We strongly believe that all our measurements are reproducible, although, as we saw, the staggering growth of content can make it difficult to get a representative sample. We hope that the research community would appreciate more follow-up work as we argue that, especially in the field of social media measurement, too often the first major paper about a certain service is adopted as the source and cited years later without knowing if the results and conclusion are still valid.

Overall, we have seen that it is possible to form patterns for the social media content creation and consumption. Also we have seen that users group based on their activity and that there are cultural and geographical differences in the user behavior. We observed that only a small part of the total content can cause the most of the traffic and that a small group of user can create vast amount of content. In addition, it seems most popular content has attributes pertaining to certain type of an event or content. These are all important factors in terms of content distribution, replication, and storage. While it was out of scope for this thesis, and a lot of prior research exists, we still see a lot of research potential in trying to predict and anticipate social media content popularity.

Also, we would like to point that during the writing process of this thesis, all of the three services that we concentrated on: Twitter, Wikipedia, and YouTube, decided to move to actively encrypt all the end-user traffic. Of course, by its nature traffic encryption can be seen as paradigm changing factor in the Internet traffic as it will make it difficult to e.g. cache content based on its popularity. However, we believe that it is still too early to say how the big picture will unfold and it would seem likely that the findings of this thesis would still apply in future content distribution.

We want to end by discussing about something that we were considering quite frequently during this research process. That is, how well different

models, be that inferential or descriptive, apply to social media content, especially when considering popularity? For example let us consider the popularity of an online video. At first it would seem simple enough to assume that the overall video popularity could at least be approximated e.g. as we saw in Chapter 8 with a Zipf distribution with a truncated tail. However, things get more complicated when we start asking questions such as: "What makes a video popular?", "Is there something inherent in the video's content or are there external driving forces that affect the popularity?", and "Do Recommender Systems just make popular content more popular?". Also, what are the roles of the service provider and other parties like the ISPs? We know that a social media service provider can inflate the view counts of content simply by displaying it more frequently to the users e.g. as recommended or promoted content. This of course, backfires if it discourages and displeases the users. However, it can be argued that it is in the service provider's interest to offer users videos with a minimal effort, be that measured e.g. by traffic or computational costs. That is, a popular service likely has users that just want to use the service in order to spend some time, without having any specific video (or type) in mind. Then it would be beneficial for the service provider to offer such user a video that, of course, the user would want to watch, but is also a video that is efficient to deliver. Furthermore, we argue that it is in the provider's interest to create popularity, as, we believe, many people would be interest to see a video which has, let us say e.g., more than 100 million or even billion views. For instance, was there something inherently popular in Psy's Gangnam style music video which at the time of writing has accumulated more than three billion views? Surely, a well-established artist is right to assume that posting a new music video will gather a lot of views, same goes to a tweet from a Twitter user with a lot of existing followers. However, at any point a piece of social media content might go viral, for basically any reason. The question is that how fitting it is then to use the statistical metrics and models with social media content? Or what is the right data collection methodology? One might argue that only a very small fraction of the content are externally influenced, but at the same time that seems to be the content that will generate the most traffic.

References

[1] B.T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, pages 1–10. ACM, 2008.

[2] Mathias Bärtl. Youtube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence*, 24(1):16–32, 2018.

[3] BBC. Barca & real renew el clasico rivalry . `http://news.bbc.co.uk/sport2/hi/football/7773758.stm`.

[4] Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. Identifying content for planned events across social media sites. In *The fifth ACM international conference on Web search and data mining*, Seattle, Washington, 2012.

[5] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on Twitter. In *International Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

[6] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on Twitter. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Atlanta, Georgia, 2013.

[7] Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew Richardson. Which news moves stock prices? A textual analysis. Technical report, National Bureau of Economic Research, 2013.

[8] L. Brammer, L. Blanton, S. Epperson, D. Mustaquim, A. Bishop, K. Kniss, R. Dhara, M. Nowell, L. Kamimoto, and L. Finelli. Surveillance for Influenza during the 2009 Influenza A (H1N1) Pandemic–United States, April 2009–March 2010. *Clinical Infectious Diseases*, 52(suppl 1):S27, 2011.

[9] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

[10] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, 2007.

[11] Xianhui Che, Barry Ip, and Ling Lin. A survey of current youtube video characteristics. *IEEE MultiMedia*, 22(2):56–63, 2015.

[12] Xu Cheng, Jiangchuan Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *Multimedia, IEEE Transactions on*, 15(5):1184–1194, Aug 2013.

[13] Marc Cheong and Vincent Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining*, pages 1–8, New York, NY, USA, 2009. ACM.

[14] DBpedia. `http://wiki.dbpedia.org/`.

[15] Daniel de Leng, Mattias Tiger, Mathias Almquist, Viktor Almquist, and Niklas Carlsson. A second screen journey to the cup: Twitter dynamics during the stanley cup playoffs. In *2018 Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–8. IEEE, 2018.

[16] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*. ACM, 2010.

[17] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. Carmen: a Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, 2013.

[18] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 745–754. ACM, 2011.

[19] Forbes. Google's youtube ad revenues may hit $5.6 billion in 2013. http://www.forbes.com/sites/timworstall/2013/12/12/googles-youtube-ad-revenues-may-hit-5-6-billion-in-2013/.

[20] Wikimedia Foundation. Wikimedia statistics https://stats.wikimedia.org/v2/.

[21] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? *ACM SIGMETRICS Performance Evaluation Review*, 44(1):179–192, 2016.

[22] Fabio Giglietto and Donatella Selva. Second screen and participation: A content analysis on a full season dataset of tweets. *Journal of Communication*, 64(2):260–277, 2014.

[23] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[24] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28. ACM, 2007.

[25] Ralph Grishman, Silja Huttunen, and Roman Yangarber. Event extraction for infectious disease outbreaks. In *Proc. 2nd Human Language Technology Conf. (HLT 2002)*, San Diego, CA, 2002.

[26] Guardian. Twitter has 105m registered users, 600m searches per day.. and more numbers from chirp. https://www.theguardian.com/technology/blog/2010/apr/14/twitter-users-chirp-details.

[27] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. Analyzing patterns of user content generation in online social networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 369–378, New York, NY, USA, 2009. ACM.

[28] Weiwei Guo, Hao Li, Heng Ji, and Mona T Diab. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of ACL-2013*, 2013.

[29] Silja Huttunen, Arto Vihavainen, Mian Du, and Roman Yangarber. Predicting relevance of event extraction for the end user. In *Multisource, Multilingual Information Extraction and Summarization*. 2013.

[30] I. Huvila. Where does the information come from? Information source use patterns in Wikipedia. *Information Research*, 15(3), 2010.

[31] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11), 2009.

[32] Philip R Johnson and S Yang. Uses and gratifications of twitter: An examination of user motives and satisfaction of twitter use. In *Communication Technology Division of the annual convention of the Association for Education in Journalism and Mass Communication in Boston, MA*, 2009.

[33] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[34] M Laeeq Khan. Social media engagement: What motivates user participation and consumption on youtube? *Computers in Human Behavior*, 66:236–247, 2017.

[35] A. Kittur, E. Chi, B.A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, 1(2), 2006.

[36] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24. ACM, 2008.

[37] Nina Krüger, Stefan Stieglitz, and Tobias Potthoff. Brand communication in Twitter—a case study on Adidas. In *PACIS 2012 Proceedings*, 2012.

[38] Stacey Kuznetsov. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2):1, 2006.

[39] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[40] S.T.K. Lam and J. Riedl. Is Wikipedia growing a longer tail? In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 105–114. ACM, 2009.

[41] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of NAACL-HLT*, 2013.

[42] Andrew Lih. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *In Proceedings of the 5th International Symposium on Online Journalism*, pages 16–17, 2004.

[43] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Danah Boyd. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5, 2011.

[44] Helen Susannah Moat, Chester Curme, H.Eugene Stanley, and Tobias Preis. Anticipating Stock Market Movements with Google and Wikipedia. In Davron Matrasulov and H. Eugene Stanley, editors, *Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale*, pages 47–59. 2014.

[45] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[46] NGramJ. `http://ngramj.sourceforge.net/`.

[47] Javad Nouri, Lidia Pivovarova, and Roman Yangarber. MDL-based models for transliteration generation. In *SLSP 2013: International Conference on Statistical Language and Speech Processing*, Tarragona, Spain, 2013.

[48] Oded Nov. What motivates wikipedians? *Commun. ACM*, 50(11):60–64, 2007.

[49] OECD. *Participative Web and User-Created Content*. 2007.

[50] John Paolillo, Sharad Ghule, and Brian Harper. A network view of social media platform history: Social structure, dynamics and content on youtube. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[51] Jakub Piskorski and Maud Ehrmann. On named entity recognition in targeted Twitter streams in Polish. In *The 4th Biennial International*

Workshop on Balto-Slavic Natural Language Processing : ACL 2013, 2013.

[52] Matthew Pittman and Alec C Tefertiller. With or without you: Connected viewing and co-viewing twitter activity for traditional appointment and asynchronous broadcast television models. *First Monday*, 20(7), 2015.

[53] M. Rask. The reach and richness of wikipedia: Is wikinomics only for rich countries. *First Monday*, 13(6-2), 2008.

[54] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[55] Steven Schirra, Huan Sun, and Frank Bentley. Together alone: motivations for live-tweeting a television series. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2441–2450. ACM, 2014.

[56] A Smith and J Boyles. The rise of the 'connected viewer'. pew research center's internet/american life project, 2012.

[57] Daniel Stutzbach, Reza Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2):377–390, 2009.

[58] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010.

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

[60] Hristo Tanev, Maud Ehrmann, Jakub Piskorski, and Vanni Zavarella. Enhancing event descriptions through Twitter mining. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[61] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 2007.

[62] TIME. Wikipedia turns 15 today. http://time.com/money/4178111/wikipedia-turns-15-today/.

[63] Mika Timonen, Paula Silvonen, and Melissa Kasari. Classification of short documents to categorize consumer opinions. In *Proccedings of 7th International Conference on Advanced Data Mining and Applications*, 2011.

[64] Twitter. Company facts. https://about.twitter.com/company/.

[65] Twitter. Q4 and fiscal year 2018 letter to shareholders. https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Shareholder-Letter.pdf.

[66] Vytautas Valancius, Nikolaos Laoutaris, Laurent Massoulié, Christophe Diot, and Pablo Rodriguez. Greening the internet with nano data centers. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 37–48. ACM, 2009.

[67] Jakob Voss. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics : 10th*. ISSI, July 2005.

[68] Wikimedia. Meta-wiki. list of wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias.

[69] Roman Yangarber. Counter-training in discovery of semantic patterns. In *Proc. ACL-2003*, Sapporo, Japan, 2003.

[70] Youtube. Youtube statistics. https://www.youtube.com/yt/press/.

[71] Yang Yu and Xiao Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Computers in Human Behavior*, 48:392–400, 2015.

[72] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer, 2011.

[73] Jia Zhou, Yanhua Li, Vijay Kumar Adhikari, and Zhi-Li Zhang. Counting youtube videos via random prefix sampling. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 371–380. ACM, 2011.

[74] Michael Zink, Kyoungwon Suh, Yu Gu, and Jim Kurose. Characteristics of youtube network traffic at a campus network–measurements, models, and implications. *Computer Networks*, 53(4):501–514, 2009.

TIETOJENKÄSITTELYTIETEEN OSASTO
PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports are available on the e-thesis site of the University of Helsinki.

A-2013-1 M. Timonen: Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion. 53+62 pp. (Ph.D. Thesis)

A-2013-2 H. Wettig: Probabilistic, Information-Theoretic Models for Etymological Alignment. 130+62 pp. (Ph.D. Thesis)

A-2013-3 T. Ruokolainen: A Model-Driven Approach to Service Ecosystem Engineering. 232 pp. (Ph.D. Thesis)

A-2013-4 A. Hyttinen: Discovering Causal Relations in the Presence of Latent Confounders. 107+138 pp. (Ph.D. Thesis)

A-2013-5 S. Eloranta: Dynamic Aspects of Knowledge Bases. 123 pp. (Ph.D. Thesis)

A-2013-6 M. Apiola: Creativity-Supporting Learning Environments: Two Case Studies on Teaching Programming. 62+83 pp. (Ph.D. Thesis)

A-2013-7 T. Polishchuk: Enabling Multipath and Multicast Data Transmission in Legacy and Future Interenet. 72+51 pp. (Ph.D. Thesis)

A-2013-8 P. Luosto: Normalized Maximum Likelihood Methods for Clustering and Density Estimation. 67+67 pp. (Ph.D. Thesis)

A-2013-9 L. Eronen: Computational Methods for Augmenting Association-based Gene Mapping. 84+93 pp. (Ph.D. Thesis)

A-2013-10 D. Entner: Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond. 79+113 pp. (Ph.D. Thesis)

A-2013-11 E. Galbrun: Methods for Redescription Mining. 72+77 pp. (Ph.D. Thesis)

A-2013-12 M. Pervilä: Data Center Energy Retrofits. 52+46 pp. (Ph.D. Thesis)

A-2013-13 P. Pohjalainen: Self-Organizing Software Architectures. 114+71 pp. (Ph.D. Thesis)

A-2014-1 J. Korhonen: Graph and Hypergraph Decompositions for Exact Algorithms. 62+66 pp. (Ph.D. Thesis)

A-2014-2 J. Paalasmaa: Monitoring Sleep with Force Sensor Measurement. 59+47 pp. (Ph.D. Thesis)

A-2014-3 L. Langohr: Methods for Finding Interesting Nodes in Weighted Graphs. 70+54 pp. (Ph.D. Thesis)

A-2014-4 S. Bhattacharya: Continuous Context Inference on Mobile Platforms. 94+67 pp. (Ph.D. Thesis)

A-2014-5 E. Lagerspetz: Collaborative Mobile Energy Awareness. 60+46 pp. (Ph.D. Thesis)

A-2015-1 L. Wang: Content, Topology and Cooperation in In-network Caching. 190 pp. (Ph.D. Thesis)

A-2015-2 T. Niinimäki: Approximation Strategies for Structure Learning in Bayesian Networks. 64+93 pp. (Ph.D. Thesis)

A-2015-3 D. Kempa: Efficient Construction of Fundamental Data Structures in Large-Scale Text Indexing. 68+88 pp. (Ph.D. Thesis)

A-2015-4 K. Zhao: Understanding Urban Human Mobility for Network Applications. 62+46 pp. (Ph.D. Thesis)

A-2015-5 A. Laaksonen: Algorithms for Melody Search and Transcription. 36+54 pp. (Ph.D. Thesis)

A-2015-6 Y. Ding: Collaborative Traffic Offloading for Mobile Systems. 223 pp. (Ph.D. Thesis)

A-2015-7 F. Fagerholm: Software Developer Experience: Case Studies in Lean-Agile and Open Source Environments. 118+68 pp. (Ph.D. Thesis)

A-2016-1 T. Ahonen: Cover Song Identification using Compression-based Distance Measures. 122+25 pp. (Ph.D. Thesis)

A-2016-2 O. Gross: World Associations as a Language Model for Generative and Creative Tasks. 60+10+54 pp. (Ph.D. Thesis)

A-2016-3 J. Määttä: Model Selection Methods for Linear Regression and Phylogenetic Reconstruction. 44+73 pp. (Ph.D. Thesis)

A-2016-4 J. Toivanen: Methods and Models in Linguistic and Musical Computational Creativity. 56+8+79 pp. (Ph.D. Thesis)

A-2016-5 K. Athukorala: Information Search as Adaptive Interaction. 122 pp. (Ph.D. Thesis)

A-2016-6 J.-K. Kangas: Combinatorial Algorithms with Applications in Learning Graphical Models. 66+90 pp. (Ph.D. Thesis)

A-2017-1 Y. Zou: On Model Selection for Bayesian Networks and Sparse Logistic Regression. 58+61 pp. (Ph.D. Thesis)

A-2017-2 Y.-T. Hsieh: Exploring Hand-Based Haptic Interfaces for Mobile Interaction Design. 79+120 pp. (Ph.D. Thesis)

A-2017-3 D. Valenzuela: Algorithms and Data Structures for Sequence Analysis in the Pan-Genomic Era. 74+78 pp. (Ph.D. Thesis)

A-2017-4 A. Hellas: Retention in Introductory Programming. 68+88 pp. (Ph.D. Thesis)

A-2017-5 M. Du: Natural Language Processing System for Business Intelligence. 78+72 pp. (Ph.D. Thesis)

A-2017-6 A. Kuosmanen: Third-Generation RNA-Sequencing Analysis: Graph Alignment and Transcript Assembly with Long Reads. 64+69 pp. (Ph.D. Thesis)

A-2018-1 M. Nelimarkka: Performative Hybrid Interaction: Understanding Planned Events across Collocated and Mediated Interaction Spheres. 64+82 pp. (Ph.D. Thesis)

A-2018-2 E. Peltonen: Crowdsensed Mobile Data Analytics. 100+91 pp. (Ph.D. Thesis)

A-2018-3 O. Barral: Implicit Interaction with Textual Information using Physiological Signals. 72+145 pp. (Ph.D. Thesis)

A-2018-4 I. Kosunen: Exploring the Dynamics of the Biocybernetic Loop in Physiological Computing. 91+161 pp. (Ph.D. Thesis)

A-2018-5 J. Berg: Solving Optimization Problems via Maximum Satisfiability: Encodings and Re-Encodings. 86+102 pp. (Ph.D. Thesis)

A-2018-6 J. Pyykkö: Online Personalization in Exploratory Search. 101+63 pp. (Ph.D. Thesis)

A-2018-7 L. Pivovarova: Classification and Clustering in Media Monitoring: from Knowledge Engineering to Deep Learning. 78+56 pp. (Ph.D. Thesis)

A-2019-1 K. Salo: Modular Audio Platform for Youth Engagement in a Museum Context. 97+78 pp. (Ph.D. Thesis)

A-2019-2 A. Koski: On the Provisioning of Mission Critical Information Systems based on Public Tenders. 96+79 pp. (Ph.D. Thesis)

A-2019-3 A. Kantosalo: Human-Computer Co-Creativity - Designing, Evaluating and Modelling Computational Collaborators for Poetry Writing. 74+86 pp. (Ph.D. Thesis)