Final Assignment: Social Web Application

The Social Web - Group 23

Felicia Hotie
VUNetID: 2507098
fhe900@student.vu.nl

Mark van der Laan
VUNetID: 2080796
mln241@student.vu.nl

Marc Went
VUNetID: 2507013
mwt680@student.vu.nl

Group Report

1. INTRODUCTION

A Social Web application is introduced in this document. This application has been designed and developed to gain insight in the duration of trending topics of two social media platforms and a historic social media source.
In the remainder of this document is the Social Web application - the prototype - further described. Information about the selected trending topics of the application is given in [Section 2]. [Section 3] includes a description of the data mining and analysis.

2. TRENDING TOPICS

The duration of trending topics on Reddit and Twitter have been compared with De Digitale Stad of twenty years ago. De Digitale Stad (1995) - abbreviated by DDS - was the first online community in the Netherlands and the first virtual city in the whole world[6].
A selection has been made for the trending topics that are included in the Social Web Application. The selected topics are three public holidays in the Netherlands, namely Saint Nicholas' Day, Christmas and New Year's Eve. These three holidays are among the most popular holidays in the Netherlands. Furthermore, these holidays occur in the same period of time, namely the month December[3].
Posts from the three social media sources have been retrieved for the three holidays by entering a number of search queries. The DDS newsgroup contains only posts in Dutch and therefore solely Dutch posts from the other two social media platforms have been retrieved. Therefore, the entered search queries consist of Dutch keywords. The following keywords have been used for retrieving posts about the three selected holidays: 'sint', 'sinterklaas', 'kerst' and 'nieuwjaar'.

3. DATA MINING AND ANALYSIS

The data mining and analysis from the three different social media sources are explained below.

3.1 De Digitale Stad

Because De Digitale Stad is a data set of twenty years ago, there is no API to search through the data. Nor is there one single file containing all the discussions or topics. DDS was built for people to learn about the new technology called "The Internet", one did not have an internet connection to the house, nor had a router to route the internet. Therefore, the city of Amsterdam set up a project to get people acquainted with "The Internet". Public terminals were set up where people could sign up, create their own House - we now know it as a homepage, go to the post office to get their mail - which we now know as email. One could also go to a bar and have a beer and a chat - this is known as chat rooms or IRC. The backup data of DDS was spread out over three servers, each server had their own role.
During previous research, several interesting folders were found containing newsgroup posts, chat rooms and emails. To gain valuable information about topics in the past a custom parser had to be written. To ease the parsing process, the focus shifted to only chat rooms called 'Bruine Kroeg' - which in turn is a reference to the many bruine kroegen in the city of Amsterdam. This data was easiest to parse because it was a single line per message, starting with 9 to 10 digits which is a UNIX timestamp. This is easily parsable and usable to check what date a topic was written. Furthermore, a 'grep' command was used to search through this data for the topic.
Next all the data containing a keyword is parsed that the Unix timestamp is converted to only contain the date, month, year information. This way we can look for messages per day and not per second. Next the number of messages per that day are calculated as well as the average number of messages over the total search period - let us call it 'avg' for now. Next we calculate 10% of the total number of days a topic was discussed - let us call it 'ten_perc'. Now an algorithm to calculate when the topic was longest trending had to be created. This algorithm was defined as followed: A topic starts to become trending when the number of posts on that day is equal or higher than 'avg'. Between the beginning and the end two days higher than 'avg' number of posts there can be up to 'ten_perc' number of days lower than 'avg' when this is exceeded we can say a topic is not trending anymore. The code is available

3.2 Twitter

To mine the Twitter feed we had to think out of the box. The biggest problem is that Twitter does not allow tweets to be searched nor mined through their API, but there is a

side road that is not supported. The code written for this is just a proof of concept, it does not work properly yet nor is it fast. The reason is that every person can search the history of Twitter using a keyword and optionally a start and end date. This is then displayed as a Twitter feed. The nice thing of the Web is that everything is readable and often written in JSON. When the users reach the bottom of the page a request is sent to Twitter to give the next 18 tweets, including a code for the next 18 tweets. To harness this search ability we wrote a script to parse this data. This proof of concept shows that 1800 tweets (or 100 requests to Twitter) take 2 minutes to load. A response of Twitter consists of a pre-HTML-formatted text containing 18 tweets, photos etc. One of the HTML tags within the JSON response contains a UNIX timestamp. This timestamp - similarly to DDS parsing - is converted to a timestamp containing only date, month, year. This converted UNIX timestamp is then counted. When done manually, 18.000 tweets were loaded. This would take the code more or less 20 to 30 minutes - if Twitter does not ban the script. So to be able to make this project work real-time, search results should be cached. This cache is first checked for matching data and returned if possible. Due to time constraints this was not possible to implement. The current implementation tries to gain as much data as possible within 15 seconds - this is due to a PHP script timeout. After these 15 second the data is returned as well as the last item code. This last item code can then be used in the next request as a starting point.

3.3 Reddit

Another feature of the application is the analysis of topic trending duration on Reddit. This feature contains the mining of posts & comments in Reddit and retrieving the corresponding dates. The posts and comments are selected based on a search query and time span. An API search query is executed by using a URL that requests the data on the Reddit site. This request uses the Reddit API to retrieve the data within a given time span. The time span is defined as a day, week, year or all time and can be found in the Reddit API documentation[12].
Requests return a JSON file that holds information about the found submissions. These submissions are stored in a tree-structured object. Every child contains one topic that is posted on Reddit. A child contains 44 data objects which stores information about, for example, the data in which it was posted, author, title of the submission, URL, number of comments, number of up-votes and down-votes.
In our application, the dates are retrieved from the submissions found with the search query and from all the comments of the corresponding post. Another JSON file is requested to parse the comments. Reddit users can reply on the posts and on other comments. This structure forms a tree that consists of the original post, comments on that post, replies to these comments and these reply can also have replies. To illustrate this structure, a diagram can be found in [Figure 1]. The frequencies of the retrieved dates are plotted in a diagram so the user can analyze the Reddit users' posting behavior. Python was used to implement this feature and currently it produces a graph that can be found in [Figure 2] and online at plot.ly [17]. The search query to produce this graph was 'christmas' and contains the posts found within the time span of a year. The online version is an interactive graph where the total amount of posts on a given day can be viewed more accurately. It is clearly visible that the topic becomes trending around the Christmas holidays and peaks at December 25th. Near the next holiday, which is new years eve, the activity starts to decrease significantly. A link to the code of this feature can be found in the references [16].

4. APPENDIX

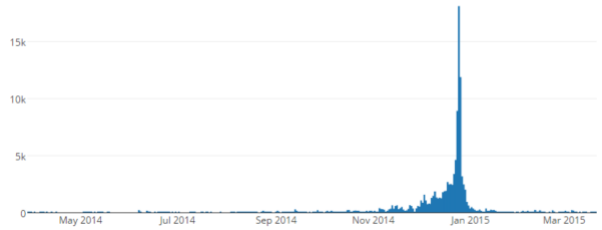


Figure 1: Graph of posts in Reddit using the search query: "Christmas"

Figure 2: Post structure in Reddit

Duration of trending topics on Twitter

1. INTRODUCTION

Twitter [1] is one of the most popular microblogging platforms [8] and has around 288 million active users per month [15]. The Social Web application - designed and developed by Group 23 - presents the duration of the trending topics on De Digitale Stad, Twitter and another social media platform Reddit. This application enables to compare the duration of trending topics on these three different social media sources. In this document an analysis is given about the trending topics and their duration on Twitter in comparison with De Digitale Stad from year 1995 [13].
In the remainder of this document the main theory behind the Twitter trending topic duration analysis is explained. The motivation for the application design is stated in [Section 3]. Furthermore, the limitations and possible scoping for the design are described in [Section 4]. Suggestions about how to evaluate the success of the Social Web application and how to improve and further develop the application design are given in [Section 5] and [6] respectively. Finally, a conclusion is given in [Section 7].

2. RATIONALE

One of the features of the Social Web application is the analysis of the trending topics duration on Twitter and comparing this data with the topics' duration on De Digitale Stad - abbreviated by DDS. The main idea behind this feature is that it is an analysis of the Twitter user posts in relation to the date. With this feature it is possible to answer the following question: 'What are the differences between the Twitter trending topics duration in comparison to the same topics duration 20 years ago?' The analysis of the Twitter data enables to compare the duration of the trending topics with the same trending topics' duration on another social media source. So, the behavior of the users of the two social media sources - from the past and present - can be evaluated.

3. MOTIVATION

In this section is the main motivation for the application design described. The target group of this application design are two different users. The added value of the application feature for these two users and the purpose of the application design are explained below.

3.1 Purpose

Each day are there around 500 million tweets sent by users on Twitter [15]. These users have different intentions and discuss different topics. Due to the growing popularity of Twitter, it is important to know for how long certain topics are trending among the Twitter users. There exists an interactive infographic - called 'What's up' - which enables to explore the trending topics on Twitter [4]. However, this infographic solely presents the Twitter trending topic data and gives only historical information about Twitter topics. Using the simple and easy-accessible Social Web application, it is possible to have an overview of the duration of certain trending topics on Twitter and other social media sources as well. Thus Twitter trending topics can be compared with the same topics discussed on other social media sources, such as De Digitale Stad. This way more insight is gained about trending topics duration on Twitter because it is possible to compare this duration with a social media platform of 20 years ago and other social media platforms.

3.2 Users

The analysis - of the trending topics duration on Twitter - is useful for two types of users. Firstly, the duration of trending topics on Twitter and DSS is relevant for Twitter users. The application design aims as an infotainment application for this particular user group. As a Twitter user it is interesting to know how long certain topics are trending on Twitter in comparison with 20 years ago. For example the following questions can be answered with the aid of the feature: 'How long was a certain topic popular on Twitter?' 'How long was a certain topic popular on the social media source of 20 years ago?' 'What are the differences between now and 20 years ago regarding the popularity of a certain topic on a social media source?' *Note: 'a certain topic' is in this current application design one of the three public holidays in the Netherlands, namely Saint Nicholas' Day, Christmas or New Year's Eve.* Secondly, it is easier for researchers to further analyze the duration of trending topics. The duration of trending topics is not limited to only Twitter data and therefore it is more interesting and valuable since more context is provided. Researchers can draw conclusions with regard to the user behavior on Twitter and DDS. For example, it can be concluded that most tweets about 'Sint' on Twitter are posted on December 5^{th} while the peak - for Sint as a topic - in DDS is on the 13^{th} of December. Another example is that the peak for Christmas tweets is on December 24^{th} and the Christmas peak on DDS is on 22^{nd} of December. This knowledge can be used to further analyze - e.g. why these findings occur - in future researches. Besides analyzing data, it is also possible for the researchers to forecast trending topics and their duration on Twitter because they gain more information and knowledge about the trending topics on Twitter. [2]

4. SCOPING

The application design has a certain scope and has a number of limitations. These limitations and scoping are with regard to the following: proof of concept, selected social media sources, selected trending topics and selected keywords. Below are these limitations and scoping for the application design discussed.

4.1 Proof of Concept

It was not possible to use the Twitter Search API[2] for searching and mining tweets from year 2014. The search results list is incomplete since it only returns the most recent and popular tweets. However, for the application design it is important that all the user tweets - from the specified date range - are returned. Therefore, an alternative code has been written as a proof of concept, but this proof of concept needs to be improved. Due to the time constraint it was not possible to improve the current implementation. In [Figure 3] are three visualizations of three selected trending topics

[1] `https://twitter.com/`

[2] `https://dev.twitter.com/rest/public/search`

shown. These line charts are the result of manual labor to demonstrate how the application design should look like after further improvements. The red lines indicate Tweets and the purple lines represent the user posts on DDS.

4.2 Social media sources

In the current application design is the duration of the Twitter trending topics only compared with the same topics on DDS and Reddit. The data on DDS is from 20 years ago and Reddit is live since June 2005 [7]. It could be possible to compare the Twitter data with more than two social media sources. However, the selected media sources in this current application design is sufficient for the Social Web application because of the time constraint. Comparing the Twitter data with two different social media sources - one from 20 years ago and a current social media platform - provides an adequate view of the differences between the duration of the trending topics for this application design.

4.3 Trending topics

Three Dutch public holidays - Saint Nicholas' Day, Christmas and New Year's Eve - are selected as trending topics in the application design, but more trending topics are discussed on Twitter. For example other public holidays and recurring events - e.g. April Fool's Day - are discussed as well on Twitter and DDS. In order to make it feasible to create the application design within the given time, a selection of trending topics has been made. It was out of the scope to include more topics in this application design.

4.4 Keywords

A number of keywords needed to be entered in order to retrieve tweets of the specified date range and topic. These keywords have been selected based on the three selected public holidays in the Netherlands. These selected keywords are: 'sint', 'sinterklaas', 'kerst' and 'nieuwjaar'. Only the tweets which include the exact phrase of these selected keywords are retrieved. However, there are search results retrieved which are not related to Saint Nicholas' Day when using the keyword 'sint'. These tweets contain the phrase 'sint', but are about other topics - e.g. Sint Antonius, Sint Maarten and Sint Petersburg. Furthermore, other keywords could have been used as well to retrieve Tweets about these topics. The retrieved data results list is not complete as a result of the selection of keywords in relation to the topics. The data does not include all the Dutch tweets regarding these topics and in some cases even contains tweets which are not related to the specified topic.

5. EVALUATION

The success of the application design can be evaluated in different ways. These evaluation methods can be divided into two groups namely the usability evaluation methods and the method which evaluates the Twitter results. In this section are three usability evaluation methods listed and explained. A method for evaluation and validation of the Twitter results is presented as well.

5.1 Usability evaluation methods

In order to have a successful application design, it is important that the application design is easy to learn, easy to use, easy to memorize, pleasant to use, efficient and supports users in making less errors. Therefore is the users' usability of the application design an important and relevant factor which measures the success of the application design. These evaluation methods are not only used to determine the success of the application design, but also to debug and improve the code. Feedback - including found problems and suggested improvements - is obtained when using the usability evaluation methods. Three different methods have been selected which can be used for evaluating the application design, namely the cognitive walkthrough, the think aloud method and the heuristic evaluation. These three methods are briefly described below.

5.1.1 Cognitive walkthrough

The cognitive walkthrough is focused on letting the evaluator step through a task scenario while the evaluator discusses each dialogue element and usability problems. Evaluator experts perform the evaluation on the application design. The cognitive walkthrough method simulates the problem solving process. A number of predefined tasks need to be individually executed step by step by the evaluator. Six standard questions - see [8.1] - have to be answered by the evaluator as well. Problems and issues are discovered when negative responses to the standard questions occur. Finally, the evaluators complete the evaluation by suggesting improvements for the application design. [9], [14]

5.1.2 Think aloud

A similar method is called the think aloud method. The user tells every spontaneously raised positive and/or negative thoughts while executing a certain task in the application design. A positive reaction could be for example 'I see the difference between the duration of the trending topics on Twitter and the duration of the trending topics on DDS'. A negative reaction example could be: 'what is the meaning of the red line'? While using this evaluation method, it is allowed to give the user a hint without disclosing the performance of tasks entirely. By using this method the thoughts of the users are evaluated and problems and other remarkable findings are revealed. [14]

5.1.3 Heuristic evaluation

The last usability evaluation method is the heuristic evaluation. At least five usability specialists need to carry out the heuristic evaluation individually. These specialists inspect the interface of the application design at least two times. The specialists need to get familiar with the application and interaction flow while going through the interface the first time. The second time, the specialists need to pay attention to the functionality and the features of the application design. Then, the application design need to evaluated with the aid of a number of heuristics. For each heuristic it needs to be checked whether the interface suffices and if there are violations of these heuristics. Each specialist completes the heuristic evaluation by listing their findings - such as occurrence of usability problems - of the application design. After the conducted evaluations by the specialists, the individual evaluations are compared with each other. Based on these usability heuristic evaluations, it can be concluded if the application design is successful or not and what needs to be improved. [9], [14]

5.2 Results evaluation method

A manual check can be performed to check whether the re-

sults in the analysis of Twitter trending topics are valid. It is possible to fill in a search query - e.g. Dutch tweets about 'kerst' in the month December - and validate these returned results. The amount of retrieved results in this search list needs to be compared with the amount of tweets in the application design. Next to the amount of retrieved tweets, the date of these tweets should be evaluated as well. This way, the accuracy and the validity of the Twitter results are evaluated.

6. FUTURE WORK

The current application design is a proof of concept and could have been further developed if time and efforts would permit. By caching the search results of Twitter, it is possible to process it in a real-time manner. Furthermore, adding more social media sources to the application design can lead to more representative and versatile data. For example adding Facebook or Google+ makes it more interesting for users to compare the duration of the Twitter trending topics. More social sources could also lead to an alternative analysis because of the large dataset. Next to adding more social media sources, it is also possible to add more public holidays - such as Liberation Day - and/or recurring events - e.g. April Fool's Day. This way, there is more available data. Also, more keywords can be used in order to retrieve the Twitter data. Other relevant keywords - regarding the selected topics - are for example: 'Pakjesavond', 'Kerstfeest' and 'oud en nieuw'. Besides the duration of the trending topics, other data can be added as well. For example the occurrence of certain words per topic, the sentiment of the Tweet - positive or negative - or if the Tweet includes a question etc. Moreover, tweets in other languages[3] can be included because Twitter is not only confined to Dutch tweets[4]. Especially for the public holidays Christmas and New Year's Eve it is interesting to know how long these trending topics are popular among non-Dutch Twitter users. Finally, because of the added data and data sources it would be a nice to use more visualizations - in addition to the line chart.

7. CONCLUSION

To summarize, one of the features of the Social Web application is the analysis of the duration of the trending topics on Twitter. The duration of the Twitter trending topics can be compared with the DDS duration of the same topics. Social media users and researchers can use this analysis of user posts - in relation with the date - to further analyze the differences between the two sources. This data analysis is rather interesting because Twitter is still used in 2015 while the DDS is from 20 years ago. The feature is still a proof of concept, but it can be further developed to have it processed automatically in the Social Web application. The application design can be improved and further developed if time and efforts would permit by adding more data, data sources, trending topics and tweets in different languages. This application design can then be evaluated by applying usability evaluation methods and validating the results of the application.

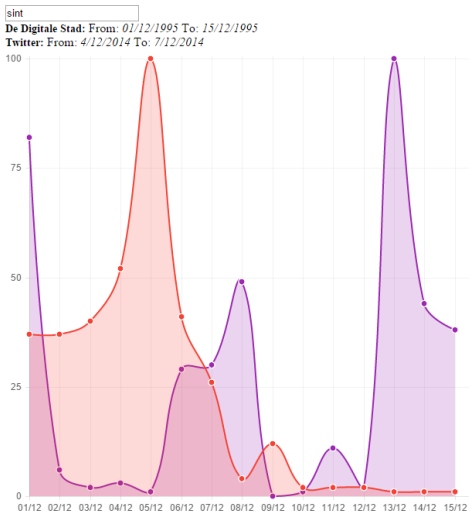
8. APPENDIX

[3] `https://mobile.twitter.com/trends`
[4] `http://www.twirus.nl`

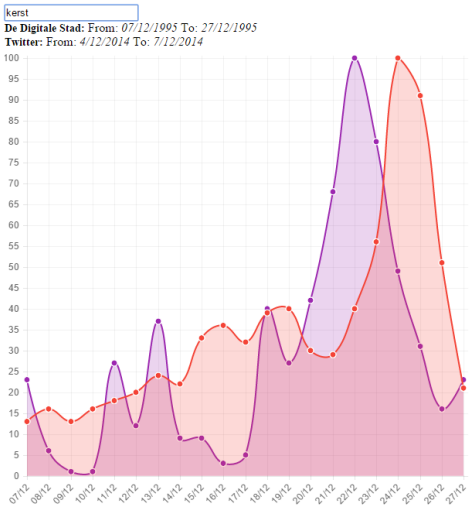
8.1 Cognitive walkthrough standard questions

1. Are the feasible and correct actions sufficiently evident to the user, and do the actions match with her/ his intention?
2. Will the user associate the correct action's description with what (s)he is trying to do?
3. Will the user receive feedback in the same place where (s)he has performed her/his action and in the same modality?
4. Does the user interpret the system's response correctly: does s/he know if s/he has made a right or wrong choice?
5. Does the user properly evaluate the results: is (s)he able to assess if (s)he got closer to her/his goal?
6. Does the user understand if the intention (s)he is trying to fulfill cannot be accomplished with the current state of the world: does (s)he find out alternative goals?

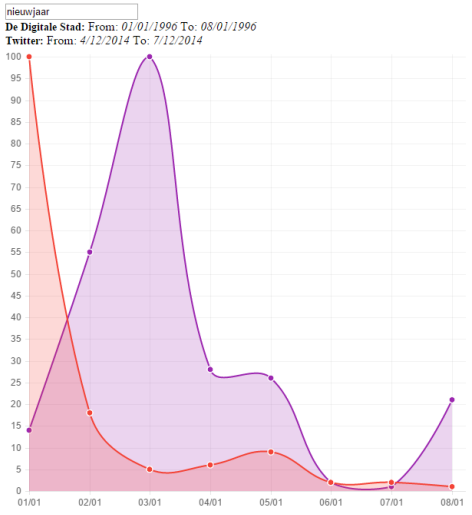
8.2 Images



(a) 'sint'



(b) 'kerst'



(c) 'nieuwjaar'

Figure 3: Topic relevance on Twitter (Red) and DDS (Purple)

Individual Report of Mark van der Laan

Trending topic in Reddit

1. INTRODUCTION

The selected feature of the application is the analysis of topic trending duration on Reddit. This feature contains the mining of posts & comments in Reddit and retrieving the corresponding dates. The algorithms in this features are described in [Section 3.3].
Social media has made an explosive growth. This set up an environment where many topics try to get the limited attention of internet users [5]. With social media, content is not only created by the general media, but also by the social media users themselves. In Reddit, this is done easily by uploading a link, which can be linked to an article, image or video. This generates a submission on the website that contains a title, link and a platform where users can comment and discuss the submission. Users can also create a text-based post where other users can comment and discuss the post as well.

2. RATIONALE

The ability to create content so easily results in having a large amount of information. This affects the ability of users to find the relevant information they want [1]. Understanding where the user pays attention to is important to gain insight into the development of problems in different cultures and opinions [18]. The ability to analyze trending topics with a given search query will result in a better understanding of user behavior and attention. In the application, it is possible to compare topics between Reddit, Twitter and De Digitale Stad. De Digitale Stad is dataset from 1995 and is interesting for comparison between past and present user behavior. Not all users from different social media platforms behave, in the same way. Therefore, it is interesting to analyze how topics trend between different platforms. In this report, the focus will be on the functionality of the Reddit trend analysis.
An analysis of the duration of topic trending on Reddit.com is particularly interesting due to the submission 'karma' system. This system enables Reddit users to up-vote or down-vote any submission on the website. These votes result in a ranking and determines whether a submission will be on the front page. The lower the ranking, the fewer amount of people that will see the submission. In addition to the submission voting system, Reddit also uses this technique for their comment feature. The ranking allows the most up-voted comments to be seen first and the comments with lowest ranks to be seen last. The ranking system in this feature also acts as a filter. It hides comments with many down-votes and is intended to hide irrelevant information such as spam. This 'karma' system makes the analysis of the topic trending more interesting because the higher a submission or comment has the more likely it is that it gets the attention of more users. The amount of attention it gets from users can be analyzed by looking at the amount of posts, but also the amount of votes.
In the current version of the application the feature mines Reddit submissions and comments on these submissions. From this data, the corresponding dates are retrieved and stored in a dictionary. This dictionary also holds the frequency of posts on a certain date. The submissions and comments are selected based on a search query and time span. A graph is plotted to compare the dates and frequency of posts on a given day, week or month. The analysis feature can help people to get a better understanding of trends, communication dynamics, agenda-setting and user behavior.

3. MOTIVATION

The main purpose of the application is to gain insight of topic trends and trend behavior between different social media platforms. The amount of posts can be seen on a given day and for a given topic. This information is useful for people who want to analyze the trending of topics and the relation between date and post amount. An analysis of multiple platforms and topics can be done, which is particularly interesting when any differences are visible. The range of users that would benefit from our application are ranging from marketers, data scientists, information scientists, computer scientists, social scientists and developers.
For marketers, it is useful because they can search a query related to products and analyze the effect of certain events like announcements, ads and conventions. When these events occur, an effect in the amount of posts about the related topic might be visible. From this data, further inspections can be done to see if a correlation between the date and popularity of the topic exists.
The application is also useful for data scientists, information scientists, computer scientists, social scientists. These scientists can also use the data to research whether certain correlations exist. They can use the data to test models that predict topic trends and gain insight of online social behavior. The application also allows an analysis of multiple sources, which can be used to research any differences between the sources. This is also useful for scientists and developers. It can help them analyze if any differences in the activity of the users using these different sources exist. They can create user profiles and test hypotheses related to user activity and attention. Using this information can help developers implement the characteristics of the favorite social media platform. The user profile and activity of Twitter, De Digitale Stad and Reddit users might be different due to the favoring of certain features that social media platform offer.
It is also worth noting that Reddit.com has more than 15 million unique visitors and approximately 150 million pageviews each month [11]. This makes an analysis of this social media platform more interesting due to the high traffic amount and structure of the website.

4. SCOPING

The scope of the application design is aimed at comparing a search query with two other social media sources. It aims at retrieving all the corresponding dates of related submissions and comments. It retrieves all comments of submissions that were found with the query. The query is used in the Reddit search API and retrieves all submissions that contain the query in the title or description. It is possible that certain comments might not directly be related to the given query. However, since the comments were found in a related submission, the assumption has been made that they are related. The nature of comments is to have a discussion about the submission. Therefore, these comment are taken

into account of the total number of posts.
The current design of the application has certain limitations. Although the application is designed to analyze the trending of topics and general user attention and behavior, more data should be added to give more value to the results of the application. Sentiment analysis would give more insight in the sentiment of the topics and posts. Giving weights to the amount of votes could also make the results more dynamic. Currently, it is limited to retrieving dates and amount of posts. Also, the given time span of a query is limited to an hour, day, week, year and all-time as the Reddit API implemented it in this manner. However, the data mining feature can be modified to only store the date of a given time interval.
The Reddit feature is designed to discard all duplicate results by storing all the post identifiers and comparing new entries with existing identifiers. The current design also discards all posts that are placed by authors that contain 'bot' in the username. This is done due to the active bots on Reddit that users can use to add additional information in submissions. A limitation here is that real users with 'bot' in their username also get discarded. The decision for the current approach was made, due to the likelihood of actual bots having 'bot' in their username is higher than actual real users having this in their names. Using a list of all the actual bots in Reddit would result in more accurate results. Furthermore, the current design of the feature still parses all deleted posts. These posts still exist in Reddit and are also viewable by the users. They are often deleted either due to moderators, many down votes or deleted accounts. The current design of Reddit hides these deleted comments but offers a view button so the comment can still be viewed by users. These comments are only still viewable when other users have replied to this comment, else the comment is deleted from the website. This is done because else the replies on the deleted comment have no parent identifier and the syntax of those comments cannot be placed.
The syntax of the search query can also be a limitation when the query has multiple meanings. A good example of this would be the query: 'Apple'. This will most likely give results of submissions about both the brand Apple and the food.

5. EVALUATION

To evaluate the success of the application, a few steps can be taken. The success can be measured in terms of valid results, but also in terms of usability for the target group discussed in the motivation section.
To measure the validity of the results gathered by the Reddit analysis functionality a manual analysis can be performed on a few queries or a test feature can be built in the Reddit data miner. The metrics used to validate the results are the amount of parsed comments of a given submission and corresponding dates. A manual analysis can be done by selecting a few submissions that are retrieved using a search query. The given dates of the submissions should be compared with the resulting dates of the data miner. The amount of comments and the corresponding dates should be compared with the data mining results as well. This method can be automated and put in a testing algorithm in the Reddit data miner. This algorithm should compare the number of comments in a submission and the actual amount of comments parsed. The dates cannot be compared to this algorithm, since parsing the dates would be the same method. The dates can still be printed and manually compared to the dates of the replies in the submission.
A field study can be done to measure the usefulness of the application. Participants of the target group may test the application and can be interviewed to gather more information about what data they would need for certain goals. New features may result out of this study and an iterative process of design and prototyping will lead to a better version of the application that can be used for research purposes.

6. FUTURE WORK

The current state of the application is a prototype. The Reddit analysis feature works in Python and creates graphs in a Plot.ly account [10]. Currently, it shows only the amount of posts on a given date and the range of dates in which these posts occur. To further develop the application design several features would be useful to do more types of analysis. One feature that could be implemented would be counting the frequencies of words used in a given subsection or topic. This is interesting to compare between different platforms, topics and subsections. It would also give more insight into the user behavior and characteristics. Furthermore, certain new trends can be spotted when a word occurs often. For example, in the developer subsection of Reddit the frequency of the word 'JavaScript' occurs quite often. If data would be used from multiple time intervals, an analysis of time and different popular programming languages could be done.
Another improvement to the design would be to provide a link to the Reddit submissions, so that the user of the application could easily see in what context the topic is placed. This way it would also be possible to see how the Reddit users react to it, have a discussion and spot why certain topics are popular on a given date. Furthermore, it would be interesting to implement a version that does a sentiment analysis, so that the semantics of the trend would also be visible. Another feature that is interesting for the same research purposes is to add an analysis of votes. This way the user can get insight in what submission topics are up-voted or down-voted the most.

Individual Report of Marc Went

When is a topic defined as trending 20 years ago

1. INTRODUCTION

A trending topic, a term everyone knows from twitter. Twitter has everyday - if not every hour - a list of the Top 5 Trending topics. Many topics are trending for a day or two, but then new events happen and the interest of people shifts. This is an well known phenomenon when large amounts of information is thrown at people. What is more interesting in my opinion is what we can tell from history. Twenty years ago a project was started by the city of Amsterdam to get people acquainted with The Web, this project was called De Digitale Stad, several backup tapes were recovered and the data is now search-able. We were interested in the comparison of the current social web like Facebook and Twitter compared to social web of then like chat rooms, newsgroups etc. To search De Digitale Stad data we first had to find the location of this data, since the project was huge we could not search everything. We focused on a relative clean dataset called plein. This data - as mentioned in the group report - was a single line per message with the timestamp included. I had to find a way to define 'trending' - or rather relevance since the term trending would stretch too far - for 20 year old data. Currently 'trending' consists of large sum of messages about a topic in a small amount of time, but back then a user had to physically access a terminal to chat. Therefor, a topic could spread longer in time staying relevant. So for me the most interesting feature of our prototype would be how to determine when a topic is trending using data of 20 years ago.

2. RATIONALE

The main theory of this application would be how to determine the relevance of a topic using only data. Take for example 'Sinterklaas' as a topic, from November the topic Sinterklaas starts to become relevant. In the middle of November he lands in the Netherlands and it's peak is on December 5^{th} on 'Pakjes avond' after which it stays relevant for a day or two and finally is replaced by Kerstmis. When talking about twitter relevance it is +- 2 or 3 days around December 5^{th}. But back in 1995 users would stay at home to celebrate with their family. So users would not go out to the city to login on a terminal to talk about 'Pakjes avond', they would do it a day in advance or a few days later. So it important to have a larger margin on the relevance of a topic, since the users would talk about it a few days after the event. So by creating a search command to look through all the chats on all the available 'pleinen' (city square) the dates on the topics could be extracted. First I looked through the data in excel, I had the extracted dates which I counted and plotted into a graph [Figure: 4]
This figure shows the raw search data plotted per day. It shows that on December 1st the topic became relevant and after December 15^{th} it died out. in between we see a big dip on December 5th which we can link to that people want to be with family. So an algorithm had to be written that would show the trending topic started Dec 1st and ended Dec 15^{th} (We humans can see is easily in the graph as trending or relevant)

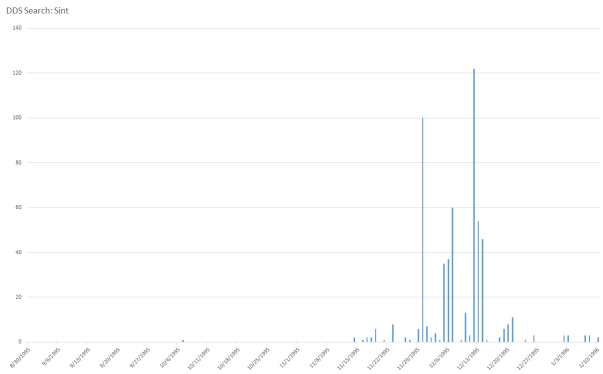


Figure 4: Raw DDS data plotted in Excel for 'sint'

3. MOTIVATION

3.1 Motivation for creating

My motivation to work with 20 year old data lies in the course: History of Digital Cultures (followed in January 2015 - by Gerard Alberts) Where in collaboration with the Amsterdam museum we found the Avatar generator that would create a random avatar for every user. The collaboration with Gerard Alberts and the Amsterdam Museum contact Tjarda de Haan continued after this course due to my interest in extracting more historic data. Because I study two Master programs simultaneously (Computer Science and Information Sciences) I wish to have my two theses overlap by subject. Excavating interesting information about De Digitale Stad would be a nice overlap when I can continue one thesis where the other ended. By doing research in advance I can ease my process later on. This is my motivation on including De Digitale Stad in as many courses as possible. Not only is this research very valuable for me, but for The Amsterdam Museum as well, currently they have a very small exhibition about DDS with an original Terminal, but this can be expanded by recreating De Digitale Stad again from the backup tapes. This research is therefore very valuable for the Amsterdam Museum since it gives a very nice insight into the history of human interactions in the beginning of the web.

3.2 Motivation for usage

The main usage of the system as a whole could be interesting to use to see how users interact differently on different social media. Reddit users post different kind, as well as different amount of messages compared to Twitter users. DDS users are totally different as well. The latter group of users could be of big interest for the Amsterdam Museum, since they have an exhibition of DDS. Visitors could interact with the system to see the differences between history and present use of 'social' media. For the system of trending-ness selection, it is different compared to the selection of trending topic on Twitter. The usage for this system would be more suitable when more web archeology is done, not necessarily DDS. It could help other developers, since the algorithm is thought about, but is not set in stone.

4. SCOPING

The scope of this research is quite limited, we know that there are many places on De Digitale Stad where users would vent their ideas and talk to one another. This data is avail-

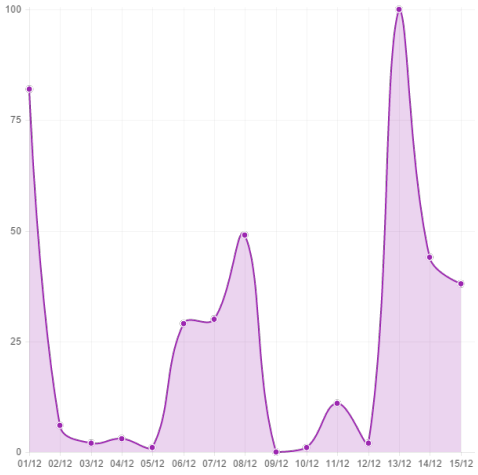
able from the backup tapes and can therefore be parsed. But because every interface had their own way of saving conversations, each interface will have to have a separate parser. To limit the amount of work to be done for this project (since it is a mere prototype) we choose to limit our view to the chat rooms on 'plein' (yes there were different types of chat rooms). Because of design decision the code only shows a limited part of all possible conversations on DDS.

5. EVALUATION

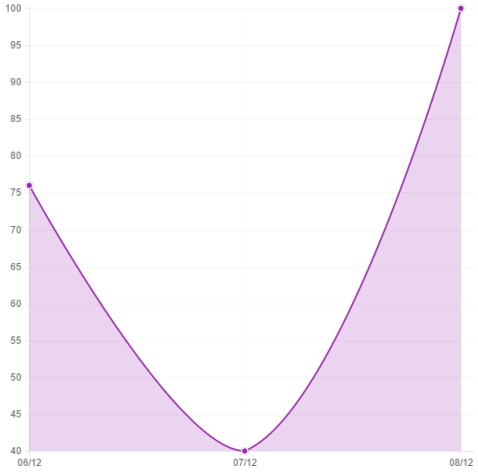
The success of this project is in my opinion really good, the code seems to work for the keywords we tested for (sint, sinterklaas, kerst, nieuwjaar) [19]. When the algorithm is executed we get the following results for the above mentioned keywords.
We can see in [Figure: 5] that, except 'sinterklaas' the terms span at least a week. Even though not all values are high the overall graph shows the relevance of a topic, where the first and last value have at least the average number of total posts. This data is very interesting for the Amsterdam Museum, especially if it is compared to twitter data, this data has peaks on the event's day, so December 5th for 'sint' and 'sinterklaas', December 24th depending on when Christmas is celebrated for 'kerst' and January 1st for nieuwjaar. One thing we had to keep in mind is to normalize the data, since there are many more users and therefore more posts on twitter than ever were on De Digitale Stad. The graphs would be disproportionate when the count we real values instead of relative values. So the graph shows 0 for the least number of posts - regardless if it is 0 or in the hundreds, and 100 for the highest number of posts. This way comparing the data is possible.

6. FUTURE WORK

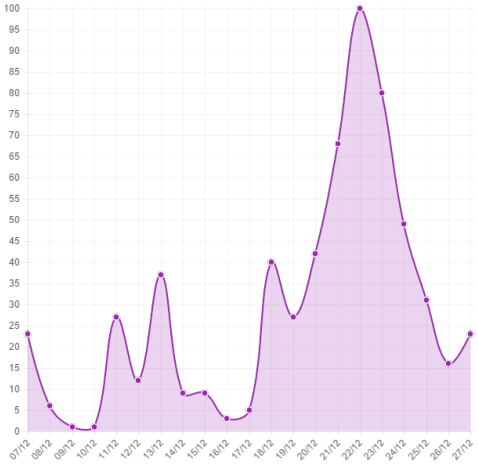
Regarding the project as a whole I would love to implement real-time search through Twitter and Reddit. These social media are great to compare statistics against, but not only between Twitter and Reddit themselves (to see the difference in people on those networks) but also between Twitter, Reddit and De Digitale Stad. But to search properly through De Digitale Stad, the search area should expand from only 'pleinen' to newsgroups and homepages as well. But creating parsers that are able to parse this data is a lot harder and therefore more time consuming. We currently have two proof of concepts ready for Twitter and for Reddit. They are not implemented into the code yet because they are not polished enough to give data necessary. Implementing Reddit is currently close to done, but there are some minor issues regarding speed. So a similar solution should be applied in the retrieval of Twitter data as well as Reddit data.



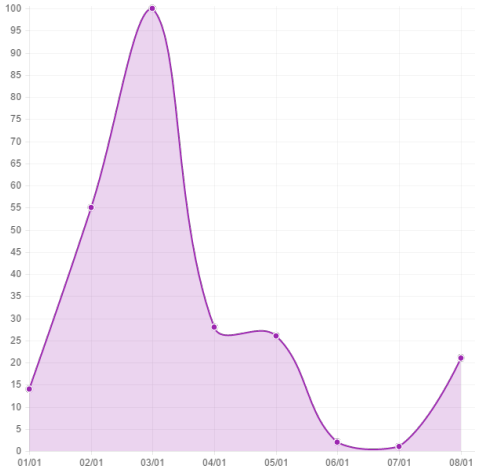
(a) 'sint'



(b) 'sinterklaas'



(c) 'kerst'



(d) 'nieuwjaar'

Figure 5: Graphs of data from De Digitale Stad

References

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

[2] T. Althoff, D. Borth, J. Hees, and A. Dengel. Analysis and forecasting of trending topics in online media streams. pages 907–916, 2013.

[3] I. am Amsterdam. Public holidays. `http://www.iamsterdam.com/en/visiting/plan-your-trip/practical-info/public-holidays`, 2015.

[4] L. Dugan. Explore the popularity and lifespan of twitter's trending topics [infographic]. `http://www.adweek.com/socialtimes/explore-the-popularity-and-lifespan-of-twitters-trending-topics-infographic/456645`, 2011.

[5] J. Falkinger. Limited attention as a scarce resource in information-rich economies*. *The Economic Journal*, 118(532):1596–1620, 2008.

[6] H. Gersonius. Re:dds. `http://hart.amsterdammuseum.nl/71509/nl/re-dds`, 2015.

[7] R. inc. Reddit. `http://www.reddit.com/about/`, 2015.

[8] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. pages 56–65, 2007.

[9] M. Matera, F. Rizzo, and G. T. Carughi. Web usability: Principles and evaluation methods. pages 143–180, 2006.

[10] Plot.ly. Plot.ly. `http://plot.ly`, 2015.

[11] Reddit. Reddit traffic. `http://www.reddit.com/r/AskReddit/about/traffic`, 2015.

[12] Reddit.com. Reddit api. `https://www.reddit.com/dev/api`, 2015.

[13] D. W. Society. Dds. `https://waag.org/en/project/digital-city-dds`, 2015.

[14] D. Stone, C. Jarrett, M. Woodroffe, and S. Minocha. *User interface design and evaluation*. Morgan Kaufmann, 2005.

[15] I. Twitter. Twitter. `https://about.twitter.com/company`, 2015.

[16] M. van der Laan. Link to reddit feature made in python. `https://drive.google.com/file/d/0BzqijByeqBn3UmFrUmVtU2NFM1k/view?usp=sharing`, 2015.

[17] M. van der Laan. Plot.ly christmas trend graph. `https://plot.ly/~mrkdoob/29`, 2015.

[18] C. Wang and B. A. Huberman. Long trend dynamics in social media. *EPJ Data Science*, 1(1):1–8, 2012.

[19] M. Went. Final assignment prototype - the social web. `https://mwent.info/vu/social/final/`, 2015.