Managing Big Data: GameTweets

Thijs Wiefferink
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
t.w.wiefferink@student.utwente.nl

Mark Meijerink
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
m.j.meijerink-1@student.utwente.nl

Alex Aalbertsberg
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
a.p.aalbertsberg@student.utwente.nl

ABSTRACT

The purpose of the project was to find out whether there is a correlation between the amount of tweets sent around a game's release date and the game's sales records. For this research, we used a data set provided by Twiqs, which contains over two billion Dutch tweets. We found good indications that there is a relation between tweeting about games and its sales records. However because of the inprecise sales records data it is currently not possible to statistically prove this.

Keywords

Big Data, MapReduce, Hadoop, Twitter, Games, Twiqs, Social media

1. INTRODUCTION

This paper has been written for the course Managing Big Data at the University of Twente. The goal of this paper is to explain the preparation, execution and results of the project we performed for this course. The project required us to perform operations on a large data set, and use the results for the purpose of answering a research question. Several data sets were made available to us on the CTIT cluster at the University of Twente.

We decided to use the Twitter data that is provided by Twiqs.nl, which contains a large part of the Dutch tweets from the period of December 2010 until November 2015. Twiqs aims to collect all tweets that are sent in The Netherlands or have a Dutch language. The tweets are filtered by language, only the tweets that are determined to be Dutch by an algorithm of Twiqs are included in the data set. Because of Twitter API limits not all tweets are collected, but estimations by Twiqs indicate that around 80% is collected.

We wanted to research the Twitter data in the gaming area, in order to see if there is any relation between the number of tweets about a certain game (especially around the release date) and the number of copies that a game sells. Therefore the research question is as follows: *Does the number of tweets about a game relate to the number of copies sold in Europe?*

For the number of copies that a game sells, we will use the top 20 sold games from the video game sales chart on the website called VGChartz [1] and filter this list to find the games that are released since 2011. The reason why we are filtering this list from the year 2011 onward is because the Twitter data also spans the period from 2011 until now. The filtered and ordered list on sales in Europe can be found in Table 1. This list will be used during this research.

[1]http://www.vgchartz.com/

To get some first impressions about the data, we have used the web interface of Twiqs to search through the Twitter data. We have searched for 'witcher' (which is supposed to catch the game 'The Witcher 3') in the time period of January 1st, 2015 until December 31st, 2015 to see if we could find the release date. Twiqs searched in 288480444 tweets with this term and time period, and found 12032 tweets. The results show a clear peak around the end of May, which corresponds to the release date of May 19th, 2015. We also searched for 'battlefield', 'fifa 15' and 'minecraft' to get some more examples of game data. Out of our examples we can conclude that we can find enough tweets to find something meaningful.

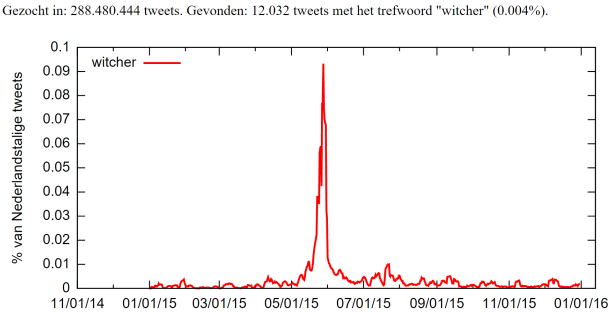

Figure 1: Search on Twiqs.nl for 'witcher' in the period of January 2015 until December 2015.

1.1 Need

The results of the research will confirm or deny that there is a relation between the amount of copies that were sold and the number of tweets about the game in the release year. This relation would be useful for game publishers, game developers, game review sites and customers. For example, if there exists such a relation, then the amount of tweets that are sent at the time of release could potentially be an indicator for the estimated amount of copies that will be sold for the game. This data could be useful for publishers and developers, as it will show them the amount of excitement and interest that is being displayed for their game upon release.

1.2 Task

With use of MapReduce we will search in all available tweets and count the number of tweets that are about a certain game. These counts will include time data so that grouping can be used on a certain scale to get a good graph for the number of tweets related to time. It will probably make sense to group data per day.

1.3 Preprocessing

The following operations will be performed on the tweet text and search terms before using them: convert to lower-

case; remove all punctuation; remove all whitespace characters. These operations will ensure that we need just a few variations of the search terms, and that the search results are a lot more consistent. Using the lowercase versions makes sense because we do not want to search for each casing, and changing to lowercase does not introduce any (or at least an unnoticeable number) false positives. Removing punctuation also helps, as official review sites will often use official names like *Call of Duty: Advanced Warfare*, but customers will likely not use punctuation in this way, so the special characters from these need to be removed so that they match the search terms. Removing special characters also ensures that we take Twitter hashtags into consideration while running the job. Removing whitespace is possible because having the words of a game name directly after each other does not make any existing word, so it is not likely to appear in a tweet that is not referring to the game.

1.4 Further steps

A next step would be to determine the list of games from the Twitter data itself, instead of searching with a predefined list. Game names could be scanned for by checking for the words before/after the word 'game'. Then this game list could be used as input for the initial program to calculate the score for each game. Challenges for automatically determining the list of games would be finding the names, and keeping the list clean. This is hard because normally game names are not mentioned with a certain term in front or after it, so it is not trivial to find these.

1.5 Expected results

The result of this research will consist of comparisons between the amount of tweets that have been created about a game and its sales records. The results of the amount of tweets and the sales records will be visualized as a graph and will usually show a peak in the amount of tweets sent around the time of a game's release date. The answer to our research question depends on whether there is a correlation between the amount of tweets that have been sent and the amount of copies that were sold.

2. RELATED WORK

Before we started our project, we looked up several papers that preform similar projects in the past. The quality and importance of the found papers has been determined by checking the relevance and the credibility. The most relevant and important paper is discussed first in Section 2.1. The relevance is looked at from the technical perspective, for example in *Dealing with big data: The case of Twitter* [6] MapReduce is used, which also has been used for this research. We also looked at the relevance in terms of the topic the paper is about, for example how they used Twitter in comparison to the way it is used in this research. Finally the papers have been checked to see if the paper uses big data methods or more traditional methods to analyse the data used for the research.

2.1 Details

Dealing with big data: The case of Twitter [6] has been chosen as most important because it is describing the way that the Twitter data that we will be using for our research is gathered. The paper describes how the data has been filtered (by language for example), and how it has been processed. The data collected for the research of the paper contains more as two billion Dutch tweets, and is still expanding every day. The paper describes the collecting and storage and three case studies: "relating word frequency to real-life events, finding words related to a topic, and gathering information about conversations" [6]. The main result of the research is a website where users can search in a huge Twitter dataset, where the result is visualized with graphs and word clouds.

A second related paper is *Twitter based TV rating system* [2], this papers researches how social media data shows the popularity of a TV show. This data is used for advertisers to choose the right TV show for their product and to get the advertisement for a correct price. The system uses a normal MySQL database to store the data, in this case a MySQL database is sufficient because they only search for hashtags/search words related to a TV show. The system scrapes Wikipedia and IMDB for TV shows and then searches on Twitter with the names of the show and its cast. Then the tweets that are found are classified as the class "relevant" or "viewing", which mean related, but not viewing and viewing the show respectively.

The paper *MapReduce Functions to Analyze Sentiment Information from Social Big Data* [4] analyzes social media data with the Hadoop FileSystem and MapReduce to determine the sentiment of the data. The implemented system is capable of searching for certain keywords, and then determines the sentiment of the tweets. This helps to understand how a certain term is used, and can be used to do market research.

Then there is the paper *Analysis of Twitter Usage in London, Paris, and New York City* [1] This research has analysed Twitter usage in London, Paris and New York on usernames, gender, ethnicity and location. This way they determined what kind of users use Twitter in these cities, and found the high usage areas in these cities. The paper shows graphs and tables with the number of people of a certain ethnicity. They represent the results in tables and graphs.

The paper *Combined analysis of news and Twitter messages* [3] uses Twitter data to see how events of large companies show up on social media. The software tries to analyze the tweets of big companies as Facebook, Microsoft, Google, etc. and tries to gather useful and structured data from them. It for example combines news and Twitter messages to get information about an event, for example the announcement of a new Lumia smartphone by Nokia.

The last related paper is called *Modeling retweeting behavior as a game: comparison to empirical results* [5], it analyzes retweeting on Twitter. The paper considers a couple of different cases and tries to find out when retweeting occurs, and how/when the original tweeter reacts. They present their theories in tables and a graph.

3. MATERIALS AND METHODS

3.1 MapReduce

Collecting the tweets about the picked games starts with the MapReduce job that filters all tweets to the ones that match. To find matches a list of search words has been made for each game of the top 20. These search words are the full name of the game (`Grand Theft Auto V`), possibly a short notation of the game (`GTA V`), and possibly a different number notation (`GTA 5`). This would still not match a tweet that contains `GTA5` or `gtav`, so the search words and tweets have been processed. The search words and tweet text have been set to lowercase, and all characters that are not in `a-z, 0-9` have been removed. This means that a tweet text `GTA 5 is a very cool game!` would be processed to `gta5isaverycoolgame`. Then a contains check is ran for every search word of every game, which results in a

match at `gta5`. The filtering performed on the data ensures that all mentions of a game are caught, with exception of misspellings or very short notations. For example *Call of Duty: Ghosts* could be called *Ghosts*, but this word is too commonly used to add as search term, it would give too much false positives otherwise.

If the mapping phase of the MapReduce job finds a match, then it outputs a line for the Reducer to process. The key of the line is build up like this: `<game>-<year>-<month>-<day>`. So this means it will output a text string as key, consisting of the game title, year, month and day. And output a `1.0` as double value. The reducer combines all lines that have the same key, adding their value together. This means that the result is a line for each day, that has the number of tweets for a certain game. With this data the number of tweets per game per day can be plotted.

3.2 Website

The results of the MapReduce task are processed by a Java program. This program collects all results from the different output files (one per machine in the cluster, 48 in this case) and groups them per game. The dates are parsed to a *milliseconds since 1970* time stamp to be used with the graphing library. Any gaps in the data are filled with zeros, since a gap means there are zero tweets about that game on that day. After the reading, grouping and repairing of the data it is printed to a PHP file, this file contains one line that assigns a JavaScript object to a variable. The object that it assigns has game names as keys, and as value it has an array. This array contains one array per day, which then contains a time stamp and count. This is the format that HighCharts, the chosen JavaScript graphing library, uses.

To compare the tweet counts to the number of copies sold the sales numbers of VGChartz have been used. The sale numbers are available for the ten weeks after a game releases on a certain platform (PlayStation 4, Microsoft Windows, etc.). These lines have been plotted in the same graph so that the correlation can be checked easily.

The website itself displays some basic HTML and loads the HighCharts library. Next, PHP includes the files with tweet count data and sold copies in a JavaScript block. Finally, JQuery is used to loop through the data and the HighCharts graphs are added to the DOM.

The implementation of the MapReduce job, Java programs and website can be found on GitHub[2].

4. RESULTS

With the materials and methods that have been described, we have executed a MapReduce job on the CTIT cluster. This job has counted the tweets that have been sent about a game on every day from December 2010 until November 2015. These tweet counts have been processed and their results have been made visible using highcharts, one of which can be seen below. We chose to use highcharts, because they contain some handy functions such as zooming in on a particular time period. The full results for all of the games in the top 20 can be found on the website that we have created. [3].

This image 2 represents the amount of tweets that were sent daily about the game Call of Duty: Ghosts. The release date for the game was November 5th, 2013.

The main result of this research was to check for a correlation between the amount of tweets that were sent about a certain game, and its position in the sales charts. For the purpose of checking for such a correlation, we decided to take the Twitter results and the sales charts, and create a list that contains both for all of the games. This list has been included in this report as 1.

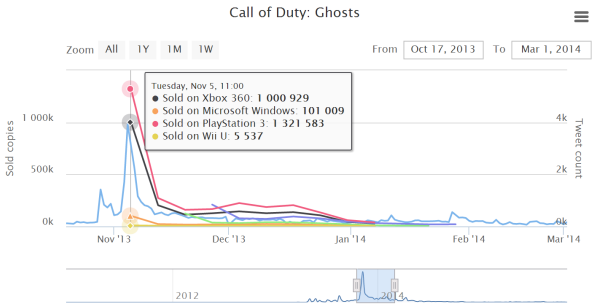

Figure 2: Graph for the game Call of Duty: Ghosts (zoomed in around the release date).

5. DISCUSSION

The MapReduce job took about 8 hours to complete on the cluster, so this meant that we could not perform this job often. To prevent us from needing to do this, we started by running our MapReduce code on smaller data sets, such as the number of tweets in one or two months. This allowed us to remove bugs from the code quickly, and test the data processing for the website before we ran the job on the complete data set.

There are several things that may have influenced the results, such as:

- We may not have been able to consider all tweets referring to specific games. Some tweets may only contain the name of the franchise, as people will not always spell out the game's full name. (FIFA 15 may become just FIFA, and Grand Theft Auto V may become just GTA).
- Unfortunately, the collection of tweets does not span the full 100% of Dutch tweets that were sent in the time period. The estimate is that the data set contains approximately 40% of all the tweets [6].
- The time period during which tweets have been collected spans from December 16th, 2010 until November 5th, 2015. Because some of the games were released either before or right before the end of this time period, we may not have all tweets that refer to this game around its release.
- Something else we have noticed, is that the number of tweets drops at around May 10th, 2014. This may be due to an error or an adaptation on Twiqs' side.
- Instead of only the number of tweets, also favorites and retweets of the tweet that mentions a game might be used. This might give a better estimation about the popularity of the game.

6. CONCLUSION

From the results, we are able to give an answer to our research question: Does the number of tweets about a game relate to the number of copies sold in Europe?

Certain games have a large variance in tweet counts around their release, after which the hype seems to die down quickly. This can for example be seen with the FIFA franchise. These games tend to release and then be played a

[2]https://github.com/NLthijs48/GameTweets
[3]http://wiefferink.me/GameTweets

Table 1: Number of tweets per game compared to number of copies sold

Game	Sold in Europe millions	Sold in world millions	Tweets total	Tweets 30 days around release	Tweets 7 days around release
Grand Theft Auto V	19.94	49.94	1114680	610750	479312
FIFA 15	11.27	18.03	270419	71405	30251
Call of Duty: Modern Warfare 3	11.09	30.47	834214	222893	154132
FIFA 14	10.84	16.88	493444	223642	79726
Call of Duty: Black Ops II	10.71	29.09	761844	136118	78677
FIFA 13	10.05	15.87	499469	210671	107023
Call of Duty: Ghosts	8.59	26.64	81476	29072	16613
FIFA 12	8.42	12.94	309791	110101	52854
Call of Duty: Advanced Warfare	7.44	20.7	72172	11704	6546
FIFA 16	7.43	11.71	65666	46821	22352
The Elder Scrolls V: Skyrim	7.37	18.7	1232093	48671	17285
Minecraft	6.8	18.76	1187771	9763	2697
Battlefield 3	6.41	17.23	287902	44576	17385
Call of Duty: Black Ops 3	5.59	15.07	235768	13990	7101
Battlefield 4	4.98	13.1	198555	29270	12834
Assassin's Creed IV: Black Flag	4.97	12.65	406618	40661	8444
Assassin's Creed III	4.88	12.96	276628	25344	11412
Assassin's Creed: Revelations	3.96	9.15	8662	2252	1124
Diablo III	3.93	10.01	49923	17000	10932
Far Cry 4	3.74	7.66	138114	10478	3907

lot by various players, then fall off before the release of the next year's FIFA game.

Other games tend to have a more constant player base. This can be seen with the game The Elder Scrolls V: Skyrim, where the majority of tweets are sent outside of the game's release date. Around 500 to 1000 tweets are sent about this game every day.

By looking at the graphs there is clearly a correlation between the number of tweets about a game and the sales numbers for some games, but definitely not for all of them. For example for Minecraft there is no clear relation, which is likely because of the small incremental updates the game has instead of full blown releases. The 1.0 update of Minecraft was more of a name change (removal of the *beta* tag) instead of a real release. However for games that do have a clear release date, the correlation shows quite clearly. For example the FIFA games show that there is a clear correlation, the curve of the tweet count matches that of the sales numbers quite well. Due to the lack of precise sales numbers (only a number for each week) we cannot statistically confirm this.

The sentiment of tweets might also have an impact on sales records. The tweets might also be about various flaws or bugs in a game, which will likely result in less copies being sold. Therefore, in order to give a better estimation of the number of copies sold depending on the amount of tweets, it might be useful to perform sentiment analysis on these tweets. Next to this more precise sales numbers of games would improve the results, it is likely that the used source has this data but does not publish data on the web.

7. APPENDICES

7.1 Results

The results can be found in Table 1.

8. ACKNOWLEDGEMENTS

We would like to thank the people at CTIT at the University of Twente for allowing us to use their cluster for our research. Additionally, we would like to thank Twiqs for collecting the tweet data that we used for our research. Finally, we would like to thank the teachers of the Managing Big Data course for their feedback on our project.

9. REFERENCES

[1] M. Adnan and P. Longley. Analysis of twitter usage in london, paris, and new york city. In *16th AGILE international conference on geographic information science, Leuven*, pages 1–7, 2013.

[2] A. D'Souza, R. Bathla, and N. Giri. Twitter based tv rating system. 2013.

[3] M. Du, J. Kangasharju, O. Karkulahti, L. Pivovarova, R. Yangarber, et al. Combined analysis of news and twitter messages. In *Proceedings of the Joint Workshop on NLP&LOD and SWAIE SemanticWeb, Linked Open Data and Information Extraction*, 2013.

[4] I. Ha, B. Back, and B. Ahn. Mapreduce functions to analyze sentiment information from social big data. *International Journal of Distributed Sensor Networks*, 501:417502, 2015.

[5] D. E. O'Leary. Modeling retweeting behavior as a game: comparison to empirical results. *International Journal of Human-Computer Studies*, 88:1–12, 2016.

[6] E. T. K. Sang and A. van den Bosch. Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal*, 3(121-134):2013, 2013.