Customers Similarity and Community Discovery on Twitter

Matthieu Blais
ECE Paris, Ecole d'Ingénieurs
matthieu.blais@edu.ece.fr

Abstract

With their millions of users, social networks became unavoidable for a company that wants to increase its market share and to prospect potential new customers. The concern of a good digital marketing strategy is to target a community of people who can have an interest in the sold product. The aim of this research is to develop a method for discovering interests of company's followers on Twitter and for figuring out their community. However, many users do not publish regularly tweets and due to their limited length, they are difficult to process. Therefore, a new strategy is reported for discovering the points of interest of each follower. Instead of analyzing tweets, this research propose to analyze the links between a company's follower and the other users they follow. Based on a Latent Semantic Indexing method, this research shows that the followers of a company have similar specific interests. Knowing this, the company can prospect new users to analyze if their interests match those of its followers' community.

Introduction

For a few years, Internet has totally changed the behaviors of the consumers. According to Council (2012), potential customers tend to self-diagnose their problems and form their opinions about a product online. This is the reason that all the companies should go through a digital transformation in the next few years (Westerman, Tannou et al. 2012) and they will have to adopt new strategies with these new tools to develop their business.

One of them is the social media. The propagation of information on these platforms is very fast and an easy to share message can be read by millions of potential customers (Akrimi and Khemakhem 2012). Social networks are much more efficient than the other traditional communication channels to make profits and to prospect new customers (Kaplan and Haenlein 2010). Companies must participate in Facebook or Twitter because with their millions of users, they become formidable tools for marketing and market research departments (Brennan and Croft 2012). Potential customers need to interact with companies (Council 2012) and social websites provide this opportunity, for building important meaningful relationships with them (Davis Mersey, Malthouse et al. 2010). On the other hand, one communication mistake can lead to costly consequences for the brand (Westerman, Tannou et al. 2012).

The main idea is to adjust the online marketing strategy in regard to the interests of customers (Vinerean, Cetina et al. 2013). To be successful in social media marketing, companies have to take time to recognize and highlight the interests of their customers. The consumers want to share their positive feelings and their enjoyment (Akrimi and Khemakhem 2012) so analyzing their interests will help companies to adapt their communication. It is important to notice that any re-tweeted tweet reaches an average of a thousand users (Kwak, Lee et al. 2010)! However, although the traditional knowledge discovery methods are part of our daily lives (John 1999), Hu and Liu (2012) claim that there is not enough of analyzer tools for social media. Indeed, the messages posted on social media websites such as Facebook and Twitter present a challenging style of text for language technology due to their short length and informal nature (Ritter, Clark et al. 2011).

Twitter is a social network where users can publish micro-blogs. Being able to discover the opinions and the interests of users on Twitter would be very useful for companies because this website provides valuable indicators of current trends, public reaction and interests (Khartabil 2013). However, it is challenging because the messages tend to be very short (Neubig and Duh 2013). Moreover, the language tends to be more casual with many orthographic errors, slang and abbreviations. The traditional text mining algorithms are not efficient enough for discovering the interest of users in generated social media content (Marujo, Ling et al.). The common methods consist of analyzing hashtags, sentiments (Chamlertwat, Bhattarakosol et al. 2012) and similarity of keywords in messages (Larsen and Song 2015). All these technics use Tweets, however, many users do not often tweet and only top 15% of Twitter users account for 85% of all tweets (Leetaru, Wang et al. 2013).

The purpose of this research is to overcome this problem. If the users do not publish Tweets, how can we process? The idea consists of analyzing the links between each follower and the other users they follow. The hypothesis is the topics of interest of followers of a company can be found by analyzing who they follow. Then, the objective is to analyze

the results to identify the similarities between all the company's followers and to show they form a community.

The remainder of this paper is summarized as follows. Section 2 reviews existing works about the topic extraction of Twitter data. Section 3 describes the method used to achieve the objective of this paper and the results and the analysis are presented in Section 4. Then, I discuss my work and the future studies before concluding in the last section.

Related Work

A method to identify a community is the recommendation system that can be used on Amazon for example (Linden, Smith et al. 2003). It provides an effective way of targeting for marketing. This system allows to have a real-time personalized profile for each customer. Any change in the user's data will make compelling recommendations. Then, we can use the common recommendations to identify a community.

Another method is to use text mining. Analyzing user's tweets one by one is not efficient with the actual text mining algorithms (Marujo, Ling et al.). The objective is to build a larger document for discovering interests of customers and a good opportunity is to use the user timeline. Indeed, the home pages of users open a whole range of possibilities in marketing research because the mining of correlations can be done simply by looking at the co-occurrences of terms in home pages (Adamic and Adar 2003). Instead of analyzing the tweet topics one by one, we should analyze all the home page and look for correlations between the keywords of each tweet.

Then, traditional methods of text mining can be applied. One approach for topic modeling is the generative probabilistic approach such as the Latent Dirichlet Allocation algorithm (Blei, Ng et al. 2003). It is used for large corpora or website but, it is less efficient for Twitter(Zhao, Jiang et al. 2011). Consequently, many LDA-like algorithm have been developed. Blei and Lafferty (2006) apply the LDA with an extension, the correlated topic model (CTM). Zhao, Jiang et al. (2011) also discuss another alternative for the classical LDA. They tried to adapt it to short messages and they show their Twitter-LDA algorithm got better results than the traditional algorithm. There are three standard steps of key phrase extraction: keyword ranking, candidate key phrase generation and key phrase ranking. For the last step, they used context-sensitive topical PageRank.

Khartabil (2013) developed the L-LDA model and also claims it has better performance than the classic LDA. Ritter, Clark et al. (2011) also used a LabeledLDA for classifying Named Entity.

Another approach is to use a non-probabilistic method like Latent Semantic Indexation (Řehůřek and Sojka 2010) or non-negative matrix factorization. It is a method to identify patterns between the terms and concepts contained in an unstructured collection of text.

Lee, Palsetia et al. (2011) use another method for text classification. They construct word vectors to classify the topics using a Naive Bates Multinomial classifier. Marujo, Ling et al. (built an indexer toolkit to extract a list of candidate keywords and train a decision tree in order to predict the correct keywords. Then, they use the Brown algorithm for clustering the keywords.

Methodology

Twitter is a social network where users can publish micro-blogs called tweets, messages with less than 140 characters, and follow other users. Following someone means taking a subscription to see all the tweets of this person. Unlike Facebook, users can usually follow everybody without being accepted by the other person. They can follow everything they want. Moreover, Twitter user accounts are public so it is easy to get data.

The objective of this research is to show that followers of a company have common interests and they form a community. The purpose is to study whom they follow. Figure 1 below illustrates the context of the study.

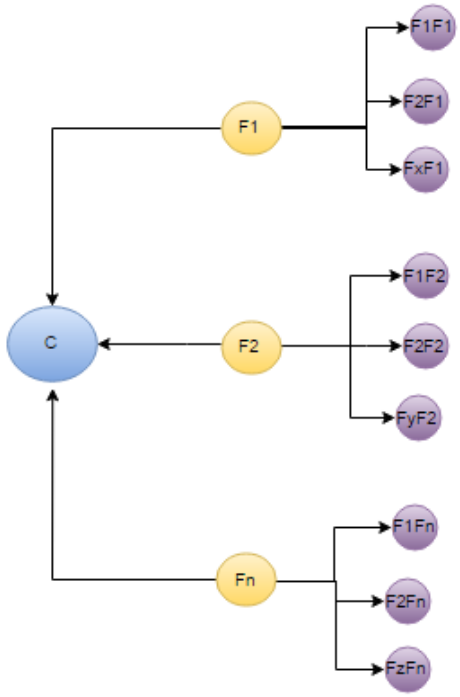


Figure 1: Context of the method

C is an active company on Twitter. It has *n* followers, F_1, F_2, .. F_n. These followers also follow other twitter accounts. F_1 follows *x* users, F_2 follows *y* users, etc. It is possible that F_1 and F_2 follow the same user, but this information is not used. A user can be described with three parameters: the Twitter

ID, unique for each account, a user profile containing all the public user information such a short description (140 characters), a website URL and a name and a user timeline containing all the tweets linked to the user (own tweets, re-tweets, messages addressed to him).

The idea is to discover the topics of interest of each F_n by analyzing the content of each of their F_zF_n. If F_n follow F_zF_n, it means they have an interest in F_zF_n (who they are and what they discuss in their tweets) (Akrimi and Khemakhem 2012). Therefore, both user profiles (who the user is) and user timeline (what the user discusses) can be exploited for analyzing the content of F_zF_n.

1. *Select x companies to study*
    a. *Get Twitter data for the x companies*
    b. *Extract keywords of each F_zF_n of each company*
    c. *Model Topics of interest*
    d. *Measure similarities*
2. *Comparison of topics of interest for company C_1 and other companies C_x*

Figure 2: General Algorithm

After extracting the keywords of the content of each F_zF_n account, it is possible to know the points of interests of each follower F_n of the company C. Then, the similarity in these interests can be measured between all of the F_n and we can deduce if the followers of this company have common points of interest or not.

The objective is to discover if the interests are usable to figure out a distinct community that can be used for marketing purposes or if they have common interests with many other users. The idea is to find if the followers of a company C_1 and a company C_2 have the same or distinct interests. For this, all the followers F_n of C_1 are supposed to be followers of C_2 as well and their similarities with the real followers of C_2 are measured. The interests of each F_zF_n are compared with those of C_2. In function of the result, we can deduce if the followers of C_1 and C_2 have common interests or not. This is repeated with *x* companies in different business areas. If the similarities between C_1 and C_n are always high, it means this method cannot be used for marketing purposes. In the following subsections, each step of the method is described more precisely.

Companies' selection

Four active companies on Twitter have been selected. The selection has been done on Twibs.com, a Twitter business portal, featuring many businesses who are on Twitter.

The criteria have been the number of followers and the business activity of the company. Only companies with less than 10,000 followers have been chosen. We assume users following these companies have an interest in their business rather than individuals who might follow only multinational brands. The second criteria has been to choose companies that have a different kinds of business to be able to observe a significant difference in the interests of each users.

Table 1 shows the characteristics of each company.

Twitter name	Location	Followers	Business Area
ShaunFKiely	Seattle	555	Point Of Sale
TruckingTools	California	3530	Truck
Andolasoft	San Jose	1612	Web & Mobile Apps
RivalBoxingGear	Montreal	3227	Boxing

Table 1: Companies selected

Data collection

The data collection is done via a Python program. Twitter provides REST APIs to programmatic access to read and write Twitter data. The python module Tweepy allows to use this API easily in a program.

The Twitter Id of all the followers F_n of the five selected companies have been downloaded. Then, for each of them, the users (F_zF_n) they follow have also been retrieved. The Twitter ids are downloaded for all F_zF_n followed by F_n and then, their user profile and their fifty last tweets of their timelines. However, Twitter limits the rate of requests that can be done by an application:

- 15 requests per 15 minutes window to get the ids of followers.
- 180 requests per 15 minutes window to get the user timeline (tweets and profile) of a user.

Therefore, the data collection takes a long time and it was impossible to collect all the data of all the followers of selected companies. For example, about a month is needed to collect all the data for one company with 2000 followers F_n who follows only 100 users F_zF_n each. Consequently, some criteria have been defined. First, to select followers F_n:

- The users have to be active on Twitter. If they did not have sent at least 50 tweets since the creation of their account, they are not considered active.
- The users have to follow a minimum of 25 accounts. It is for the same reason of inactivity. We cannot define the interests of users if they do not follow other users. Moreover, it would avoid the risk of studying a fake account.
- The users cannot follow more than 500 accounts. This limit is set to keep only users who have a real interest

in the company rather users who follow many accounts.

Second selection is done for downloading the accounts of F_zF_n:

- They have a complete user profile: a description in the profile because it is an important content to determine who the user is and a website.
- They are active. They publish at least 20 tweets per month.
- They have an English account. The other languages are not analyzed in this research.

Finally, the last step is to save data in a .csv file to analyze them. For each company C, their follower's ids F_n, the ids of the account followed by F_n, the description of the user profile and the last 50 tweets of F_zF_n users are stored.

Figure 3 summarizes this step:

1. *Get Followers of C*
2. *For each Followers F_n*
3. *If he is active and follows more than 25 and less than 500 users F_zF_n*
    a. *Get Users F_zF_n followed by F_n*
    b. *For each F_zF_n*
    c. *If he is active, he has a description, a website and his account is in English:*
        i. *Download user's description*
        ii. *Download last 50 tweets*
4. *Save data of the company in .csv file*

Figure 3: Data Collection Algorithm

At the end of this step, about 250Mo of data are stored so about 50Mo per companies (equal about 180 000 tweets in the .csv file)

Topic extraction

For this step, the objective is to extract the informative keywords from the downloaded data. Pronouns, common adjectives (good, best, etc.) and common verbs (like, go, etc.) are not informative keywords because they are not useful to determine a topic of interest. Moreover, the messages on social networks are not well structured. Abbreviations, hash-tags and links are often used but they are considered noisy words. Therefore, the first step was to clean the downloaded data using a Python program. Figure 4 shows the process followed for this:

1. *Remove links (url)*
2. *Tokenize the messages.*
3. *Remove stop words*
4. *Remove bigrams and unigrams*

Figure 4: Data cleaning algorithm

Getting tokens of a message consists of splitting the sentence in words and deleting the punctuation marks. Then, for each word, a stop word (a non-informative word), a unigram or a bigram, is deleted. To know if it is a stop word or not, the corpus of stop words in the natural language toolkit module is used. However, the list has been completed with some other words that are common on this social network like "Twitter", "Tweet", "account", "official", etc. At the end, the messages (descriptions and tweets) are transformed and they contain only informative words.

Then, there were two options: using the descriptions or the tweets to determine the topics of interest. The description (Who F_zF_n is) is like a summary of the tweets (What F_zF_n discusses). Moreover, there is one description for about fifty tweets so the probability that the keywords of the descriptions are also keywords of the tweets is high. It would have been inefficient to use both of them at the same time. The choice has been done to study descriptions first because it should have less noisy keywords. However, after getting the results, a second analysis has been done using the tweets. The results of these two approaches are presented in Section 4.

The keywords of the sentences are used to build a dictionary and a corpus of word. The keywords have been grouped by document. A document is all the keywords of F_zF_n of one F_n. Therefore, if a company C has ten followers, $n = 10$, the dictionary has ten documents. The size of each document depends on the number of F_zF_n of each company's followers and the number of keywords extracted for them. The dictionary attributes a unique Id to all the words of a document and the corpus save the frequency of each of them. At the end of this step, we have n documents, a dictionary and a corpus for each company.

Topic modelling and similarity measurement

The objective of this step is to model the topics of interest of each document. For this, the Latent Semantic Indexation algorithm is used. Using the dictionary and the corpus, this algorithm builds an occurrence matrix and tries to reduce it, removing the less pertinent terms.

For this algorithm, the gensim python module is installed (Řehůřek & Sojka, 2010). In order to create the matrix, three main inputs are needed: the corpus, the dictionary and K, the number of topics (the latent variable). This variable is unknown so the first step is to start setting it to two, make the similarity measures and then, to try again with six, ten, fourteen, eighteen and twenty-two. The different results are presented in the Section 4.

After building the LSI model, the similarity between each document is measured. Each of them is transformed in a vector, using the dictionary and the LSI model. Then, the similarity matrix (cosine similarity) of the gensim module is

used to measure the similarity of a document with all the other documents. All the results are shown in section 4.

1. *Build LSI model with the corpus, the dictionary and the latent variable (number of topics)*
2. *For each document:*
   a. *Convert it in a vector using the dictionary and the LSI model*
   b. *Build the similarity matrix using LSI*
   c. *Measure the cosine similarity of the document with the other documents*

Figure 5: Topic modelling and similarity algorithm

Distinct community identification

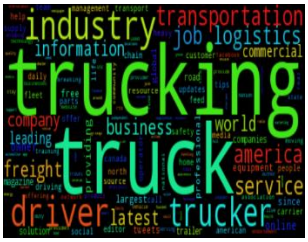
The objective is to validate the similarity between followers of the topic modelling step. Comparing the similarity between followers of different companies helps determine if the similarity measured previously allows the company to say if its followers who form a community or not. If they have common points of interests that are different from other companies, the company can use them to prospect other users. However, if all the followers of companies share the same interests, it means the followers of company do not form a distinct community.

For this, the same method of the previous step is used, but instead of building the model with the dictionary and the corpus of the company C_1, it is built with the dictionary and the corpus of another company. Then, the similarity of each document with the model of the other company is measured. We can compare if the similarity is about the same using the LSI model of another company.

Results

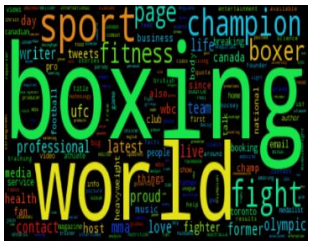
Wordcloud

First, a word-cloud is produced with the extracted keywords of each document F_n of each company. It is useful to visually see if there appears to be a difference in the keywords extracted. For displaying them, the WordCloud python module is used.

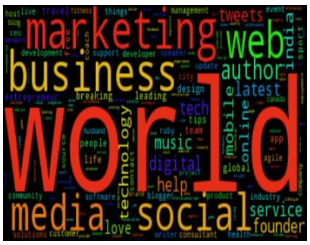


TruckingTools

ShaunFKiely



RivalBoxingGear



Andolasoft

Figure 6: Keywords of each company

We can notice a difference between all of them. The followers F_n of each company tend to have different kinds of points of interest. The more common words for TruckingTools are keywords related to truck. For RivalBoxingGear, words related to sport and boxing often come back. Andolasoft and ShaunFKiely seem to have some common words related to business like technology, marketing and social. In general, we can say, the followers of different companies tend to have different kinds of interest. At least, these first results give an interest to continue the research because we can expect to lately show followers of a company have same interests and they form a distinct community.

Similarity measure

In the step 4, the similarity between followers of a company using the cosine similarity has been measured. A histogram has been created to visualize the similarity. The first point was to set the value of the latent variable, and the number of topics of the corpus. On these charts, the value is 10. In the subsection "Number of topics, latent attribute K", we discuss the values that can take this variable but in order to discover community, it is better to take a value around 10.

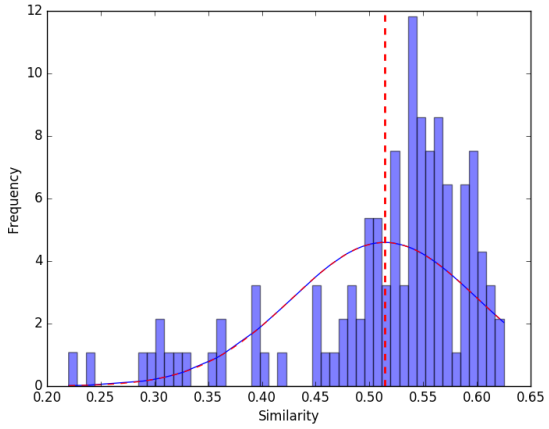


Figure 7: Andolasoft's follower similarities

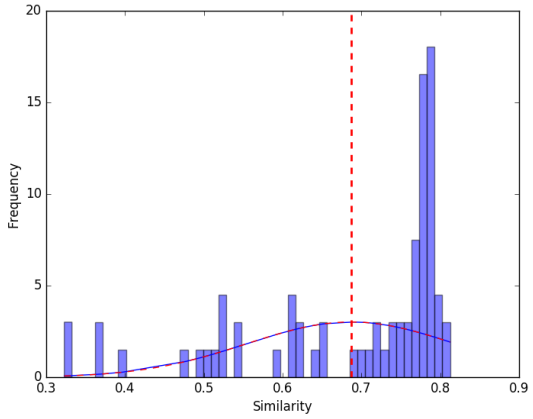


Figure 8: TruckingTools' follower similarities

The first chart shows the similarity between followers of Andolasoft. We can see a mean of 52% of similarity between each customer. It is not so high but it can be explained. On the word-cloud, the most frequent keywords tend to be general terms (marketing, business, social, and media) and not specific terms. Andolasoft is a company of Web and Mobile Apps but their applications can touch many different business areas. This could explain why the similarity is lower than on the second chart. Indeed, on figure 8, the mean of similarity for TruckingTools is really better, 69%. The words in the word-cloud were also more precise. All the companies have a similarity above 50% and the mean is 60%.

Community discovery

It can be shown that followers of company can have similarities so the objective is to exploit it. A way to do that is to compare the similarity of each follower with followers of other companies. The charts of the figures 9 and show clearly the similarity is lower when we compare followers of different companies.

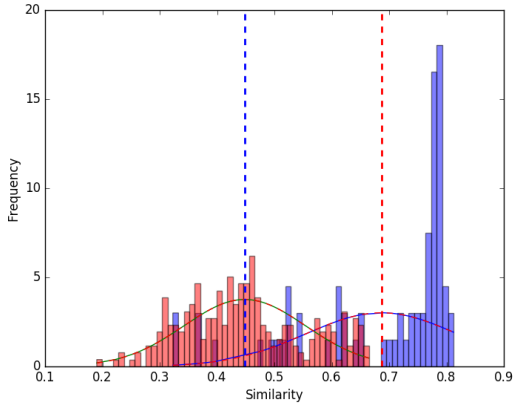


Figure 9: TruckingTools's follower similarities (blue) and similarities with other companies (red)

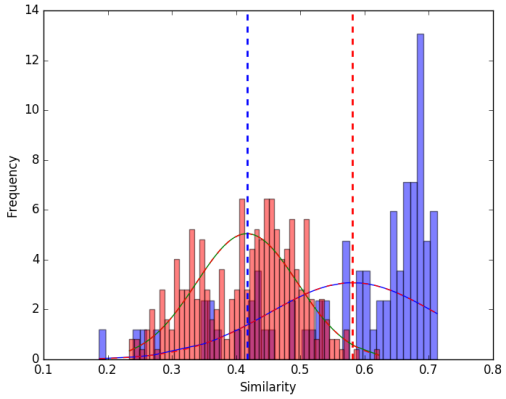


Figure 10: RivalBoxingGear follower similarities (blue) and similarities with other companies (red)

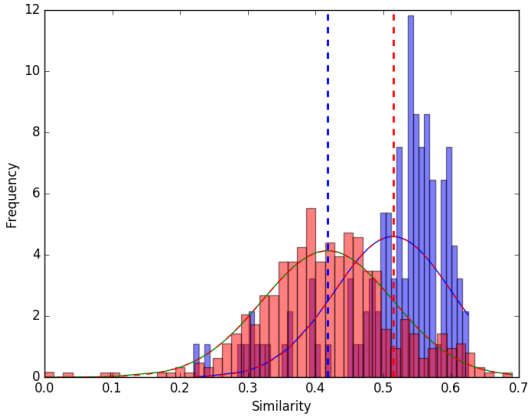


Figure 11: Andolasoft follower similarities (blue) and similarities with other companies (red)

The first chart shows the difference for TruckingTools. The similarity with the other followers of the other companies is 45% instead of 70% previously. On the second chart, we can see a difference lower with 42% instead of 58% for RivalBoxingGear. Almost all the other followers have a similarity less than the mean of similarity the company's followers. The last chart shows the result for Andolasoft. We can see a difference of only 10% but this is mainly due to the similarity of keywords observed previously.

Indeed, we can notice this on the figure 12. It shows the result of the measure of similarities with all the other company distinctly. We can see the purple histogram that seems to match very well with the Andolasoft. This histogram is the one of ShaunFKiely and as we already noticed, the word-clouds of these companies were similar. Both of them are company in a large domain of technologies so it is legitimate to find more similarities between them.

Both of them are company in a large domain of technologies so it is legitimate to find more similarities between them.

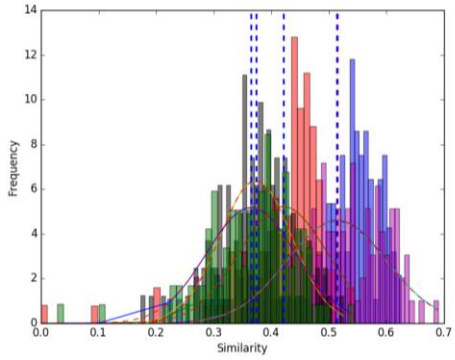


Figure 12: Andolasoft follower similarities (blue) and detailled similarities with other companies (red)

Number of topics, latent attribute K

The choice of the latent value K has been linked with the discovery of community. The real objective is not to get a similarity of 100% between followers because if the similarity is the same with the followers of all the other companies, it does not have an interest for company. The objective of this research is not to determine exactly which value of K we should choose but we tried to determine the best value for community discovery

We can notice that when K is low, the similarity is high. This trend can be observed with all the companies studied. On the figure 13, we can observe the effect of K on the similarity of RivalBoxingGear. If K=2, the similarity is 92%, if K=10, the similarity is 58% but if K=22, the similarity is 50%.

Now, if we compare the similarity of the followers of RivalBoxingGear and the other followers of other companies, we can see if K=2, the similarity is 80% whereas if K=22, it is 33%.

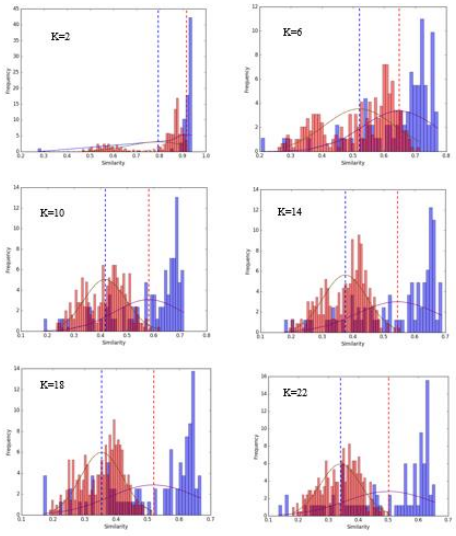


Figure 13: RivalBoxingGear similarities evolution in function of the value of the latent variable K

On figure 14, we can see the evolution of the similarity in function of K. The blue curve shows the similarity within the company, the green one shows the similarity with other companies and the red one is the difference between the two other curves. The same kind of chart is gotten for the other companies. This, is why the value K=10 seems to be the best value to balance between similarity between followers of company and dissimilarity with the followers of other company.

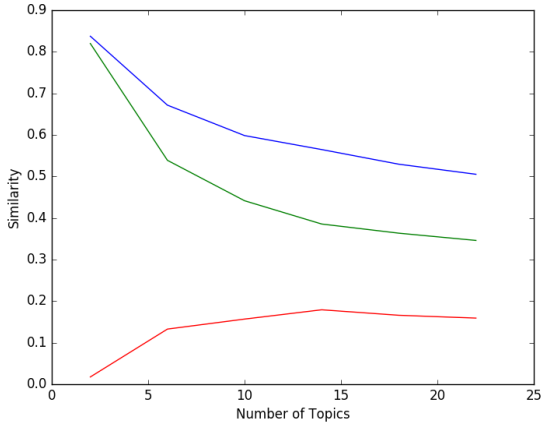


Figure 14: Evolution of the difference of similarities (red) between RivalBoxingGear's follower similarities (blue) and other companies' follower similarities (green) in function of K

User profile or user timeline

As already said, all the previous results have been gotten analyzing the user profile and their description. The objective now is to discover if we can have better results using the user timeline and the tweets. Here are the results for TruckingTools. First, the result of the word-cloud is similar. There are the same kinds of keyword.

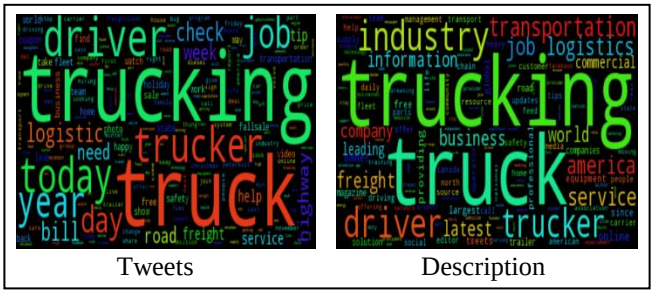


Figure 15: Word-clouds of tweets document and description document of TruckingTools

However, we can notice a significant change when we want to compare the similarities. Indeed, the similarity between followers of the company is 80% analyzing tweets instead of 70% analyzing descriptions.

The same trend has been observed with the four other companies. However, we can also notice that the results of the similarity with the followers of other companies is also higher: 74% instead of 45%. The difference is significant. Analyzing tweets allows to get a higher similarity but the community for the company is less distinct.

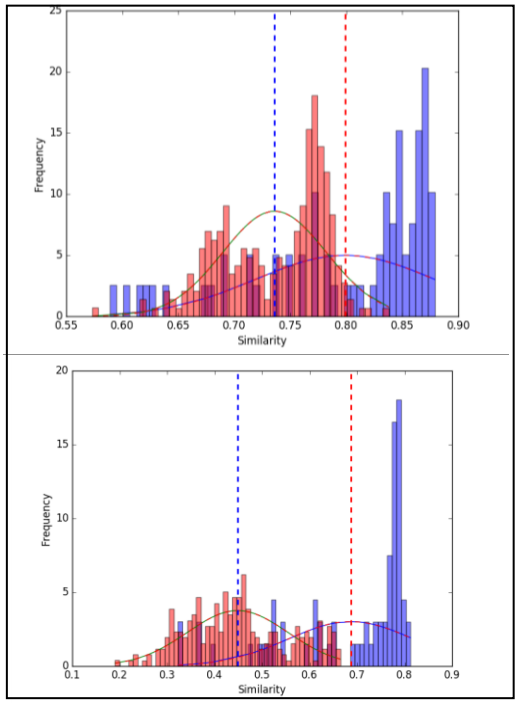


Figure 16: Comparison of the similarities results using tweets and description (TruckingTools)

Discussion

The first hypothesis was similarities in interest between followers of a company can be discovered by analyzing the users they follow on Twitter. The results tend to say we can do it. We can find a similarity of 60% between followers and even more if we change the latent variable in the algorithm. Moreover, the comparison with the followers of other companies shows we can notice a difference of 15% in their similarities. It allows us to figure out a distinct community for each company. The results show the similarities with other companies' followers tend to be below the mean of the similarity between the own company's followers. A company that wants to know if it can prospect users could use this result to decide if it is worth to prospect them. If the measures of similarity between these users and the company's community are above the mean value, the probability that they are interested is high!

However, we still can improve this. All users that are followed by company's followers are analyzed but some of them are not significant for extracting topics of interest. The users can follow organizations and companies but they can also follow their friends and their family. These kinds of user could be qualified to have noisy values. The link between the company's followers and these other users can be qualified of relationship link and it is different from an interest link. They can talk about everything (life, love, sport, books, etc.) and this can lead to similarity between followers of different companies. A future work would be to find a way to identify the links of following by relationship and to measure the similarities between followers without them. On the charts, we can also notice the similarity measures are often gathered in an interval of 10%. Analyzing who are the followers having a lower similarity could be informative and it determines if we can also remove these noisy values.

A question that appears during the research was to know if it is better to study user profile (description) or user timeline (tweets). The results are not significant enough to be affirmative on this point. After analyzing the data, some similarities between users were very low using the user profile whereas they were very high using tweets. After verification on Twitter, the similarity measured with the tweets seems more accurate but it needs to be proved. Two followers of RivalBoxingGear got a similarity of 35% using the description and then 75% using tweets but both of them were boxers and according to the users they follow, they seem to have many common points of interest. However, using tweets also improve the probability to have a noisy message and because of this, we cannot clearly distinguish the community of a company. If we can remove these noisy tweets, the similarity would be lower but the community would be more distinct. Another approach could be to remove all the most frequent words of each dictionary. The similarity between followers would probably be a bit lower and the similarity with the followers of other companies would be much lower as well.

Conclusion

The objective of this research was to find a way to prospect users of Twitter who do not publish Tweets. The results show that we can find similarities between followers of company on Twitter by analyzing who they follow. They also show the followers form a distinct community, a group of people with the same interests and the community is different in function of the business area of companies.

The results can be used for companies that want to increase their market share. For each user, they can find if they have the same points of interests of their followers. Even if the user does not tweet, the analysis of the links with the other users provide information enough to detect points of interest and then to decide if the company should prospect him or not.

Acknowledgement

I thank Dr. Insu Song, the supervisor of my research for his helps and advice provided during all my work and also the James Cook University in Singapore that welcomed me all along my research. I also warmly thanks Christian Larsen for his previous work “Social Network Data Mining for Customer Relationship Management Application” that has been useful in my work.

References

Adamic, L. A. and E. Adar (2003). "Friends and neighbors on the web." Social networks **25**(3): 211-230.

Akrimi, Y. and R. Khemakhem (2012). "What Drive Consumers to Spread the Word in Social Media." Journal ofMarketing Research & Case Studies: 1-14.

Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. Proceedings of the 23rd international conference on Machine learning, ACM.

Blei, D. M., A. Y. Ng and M. I. Jordan (2003). "Latent dirichlet allocation." the Journal of machine Learning research **3**: 993-1022.

Brennan, R. and R. Croft (2012). "The use of social media in B2B marketing and branding: An exploratory study." Journal of Customer Behaviour **11**(2): 101-115.

Chamlertwat, W., P. Bhattarakosol, T. Rungkasiri and C. Haruechaiyasak (2012). "Discovering Consumer Insight from Twitter via Sentiment Analysis." J. UCS **18**(8): 973-992.

Council, M. L. (2012). "The digital evolution in B2B marketing." The Corporate Executive Board Company www. mlc. executiveboard. com (assessed November 2014).

Davis Mersey, R., E. C. Malthouse and B. J. Calder (2010). "Engagement with online media." Journal of Media Business Studies **7**(2): 39-56.

Hu, X. and H. Liu (2012). Text analytics in social media. Mining text data, Springer**:** 385-414.

John, G. H. (1999). "Behind-the-scenes data mining: a report on the KDD-98 panel." ACM SIGKDD Explorations Newsletter **1**(1): 6-8.

Kaplan, A. M. and M. Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media." Business horizons **53**(1): 59-68.

Khartabil, D. (2013). "Data Mining and Visualisation of Twitter Using Topic Modelling."

Kwak, H., C. Lee, H. Park and S. Moon (2010). What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web, ACM.

Larsen, C. and I. Song (2015). Social Network Data Mining for Customer Relationship Management Applications.

Lee, K., D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary (2011). Twitter trending topic classification. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE.

Leetaru, K., S. Wang, G. Cao, A. Padmanabhan and E. Shook (2013). "Mapping the global Twitter heartbeat: The geography of Twitter." First Monday **18**(5).

Linden, G., B. Smith and J. York (2003). "Amazon. com recommendations: Item-to-item collaborative filtering." Internet Computing, IEEE **7**(1): 76-80.

Marujo, L., W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. de Matos, J. P. Neto and J. Carbonell "Automatic Keyword Extraction on Twitter." Volume 2: Short Papers: 637.

Neubig, G. and K. Duh (2013). How Much Is Said in a Tweet? A Multilingual, Information-theoretic Perspective. AAAI Spring Symposium: Analyzing Microtext.

Řehůřek, R. and P. Sojka (2010). "Software framework for topic modelling with large corpora."

Ritter, A., S. Clark and O. Etzioni (2011). Named entity recognition in tweets: an experimental study. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Vinerean, S., I. Cetina, L. Dumitrescu and M. Tichindelean (2013). "The effects of social media marketing on online consumer behavior." International Journal of Business and Management **8**(14): p66.

Westerman, G., M. Tannou, D. Bonnet, P. Ferraris and A. McAfee (2012). "The Digital Advantage: How digital leaders outperform their peers in every industry." MITSloan Management and Capgemini Consulting, MA: 2-23.

Zhao, W. X., J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim and X. Li (2011). Topical keyphrase extraction from twitter. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics.

Zhao, W. X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan and X. Li (2011). Comparing twitter and traditional media using topic models. Advances in Information Retrieval, Springer**:** 338-349.