Study of Competition in the Textile Sector by Twitter Social Network Analysis

Ana Belén Gil-González
abg@usal.es
University of Salamanca, Spain

Angel Luis Blanco-Mateos
angeluisblanco@usal.es
University of Salamanca, Spain

Fernando De la Prieta
fer@usal.es
University of Salamanca, Spain

Ana de Luis-Reboredo
adeluis@usal.es
University of Salamanca, Spain

ABSTRACT

Social networks have become a place where brands or companies are advertised to increase their market share. The companies allocate resources to hire community managers that disseminate quality content and perform customer service tasks. This article presents the development of a tweet retrieval tool and user account information to generate simulation and a strategic analysis of a set of competitors in the Twitter social network during a period of three months. A collection of tweets related to four related textile brands has been obtained. The information retrieval process is done using Python to access the Twitter REST service whose results are indexed using Elasticsearch. Once the data is arranged, Kibana is used to create an interactive panel, with different visualizations, that helps to extract the information hidden in the data, proposing different alternatives, its advantages and its disadvantages.

.

KEYWORDS:

Information retrieval (IR), Information visualization, Competitive analysis, Textile industry, Social networks, Twitter, digital marketing.

INTRODUCTION

The use of social networks provides an expression and communication way to create and disseminate digital content in the social media strategy of any company. The use of social networks on Internet translates social structures and behaviors into the digital world. Social Networking Service (SNS) are software platforms that are presented through a user interface, to make content accessible. The services of social networks can be classified (Castañeda & Gutierrez, 2010) in (1) generalists, as a simple way to share information or (2) of specific purpose, where users share information oriented to a specific theme or format.

Twitter, since its birth in 2006, has become the ideal generalist social communication media for the dissemination of any message worldwide. The social network Twitter allows to send text messages of short length, with a maximum of 280 characters, which encourages the use of concisely and directly, basic elements of any good message in marketing support. Actually, it works in a similar way to a microblogging service. Each account has a kind of wall on which you can post these messages, called tweets. Users can follow other accounts by subscribing by becoming followers. Each account can publish its own messages or share in its wall the contents of other accounts, which is known as retweeting. These publications are viewed by all his followers. Twitter allows the sending of direct messages to other accounts, responses to a tweet, or the use of tags called hashtags that are preceded by the symbol #.

Its use by brands builds a direct channel of public brand communication that allows knowing at the time everything that is being talked about this brand. Well developed a communication strategy on Twitter will increase the visibility of the brand and the impact of its contents will forge the image of the brand and its online reputation. Having a good strategy on Twitter is therefore vital for companies, so it is interesting to make a comparative study of the practices developed by them and how they affect the behavior of their users. For this, a tool for the collection of tweets and support for the tasks of competence analysis and brand management is proposed. Four textile brands have been selected for the case study, competing to capture the largest possible market share of clients whose age ranges represent potential users of social networks: ZARA, H&M, MANGO and GAP. All of them are multinational companies and have a Twitter account, data on which the study will be carried out.

The rest of the article is organized as follows. Along next sections will be explained the experimental architecture of the framework for the recovery and visual representation of the data. Presented the architecture will be shown the main characteristics of the visual representation system of the data obtained for the case study. the article concludes with some reflections on the work proposed and future work.

RETRIEVAL ARCHITECTURE FOR TWEETS MANAGEMENT

Technological systems inside we live, make the information we manage and consume grow in highly complex structures. The Social Network Analysis (SNA) constitutes a theoretical and methodological advance that investigates a finite set of actors and the relationships that link them, likewise, social networks are considered social structures where communication processes and/or transaction between people take place.

Tweets have their own characteristics. First of all, they are short texts, with a maximum of 280 characters, and in many cases they use a form of language that departs from standardized language, including elements typical of Twitter. Therefore, the development of information retrieval systems on them may differ from traditional approaches.

The availability of this information through an architecture oriented to the recovery of Tweets allows to elaborate a tool that performs the collection of the data and its interactive visualization for its study. The possibility of making decisions in real time based on accurate data is crucial in the management of communication of a brand (Celeste et al., 2015).

The framework presents an application for Tweets retrieval based on the use of ELK stack technology (Elasticsearch + Kibana) (Fig. 1). Starting from information retrieval using Python to access Twitter REST service whose results are indexed using the Elasticsearch. This solution provides with a full-text search engine, with a REST interface and JSON documents without schema. For the introduction of information in Elasticsearch, a Python script has been used as a data access layer to read the Tweet messages and preprocess and format them to build the data set in which the analysis will be performed.

A collection of tweets is obtained between January 1, 2017 and April 9, 2017, related to all tweets from 4 companies in the textile sector and the mechanisms are enabled for their visualization in the user interface (UI) through the use of Kibana software, with its XPack plug-in, built on Elasticsearch.

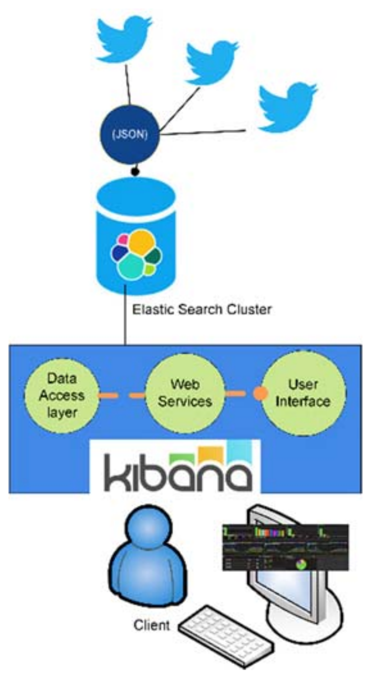

Fig. 1. Retrieval Architecture

Retrieval and Information indexing in JSON format

The social network Twitter allows us to retrieve the publications of specific users in JSON format with a series of well-defined attributes. This format does not define any way to deal with dates, locations or more complex objects. It will be needed to transform the information to the domain of our indexing system through the tool Elasticsearch[1].

Example of transformation of the data domain

```
{
  "mappings": {
    "tweet": {
      "properties": {
        "in_reply_to_screen_name" : {
          "type" : "keyword"
        },
        "coordinates": {
          "type": "geo_shape"
        },
        "created_at": {
          "format": "EEE MMM dd HH:mm:ss Z YYYY",
          "type": "date"
        }…,
      }
    }
  }
}
```

By separating the data obtained in a method that receives a user name and returns an array with the documents to be indexed, we can generate a method that needs: a list of users to be analyzed, the index where to store this set of users, the configuration for transforming data from one domain to another and a connection to our storage system. In this case the following method has been developed in Python:

Example of retrieval and indexing function.

```
def queryUsersToTwitterAndStoreToElastic(users,
  index_name, elastic_search, mapping_settings):

  doc_type="tweet"

  print('Descargando tweets de :' + str(users))

  # Borra el índice si existe
  if elastic_search.indices.exists(index_name):
    elastic_search.indices.delete(index_name)

  # Crear el índice con un mapeo de datos concreto
  elastic_search.indices.create(index=index_name,
```

[1] Elasticsearch Mapping, https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping.html, last accessed 2018/05/07

```
    body=mapping_settings)

  # Para cada usuario de la lista descarga e indexa los
  # tweets en elastic
  for screen_name in users:
    print 'Getting tweets from: ' +  screen_name
    alltweets = tw.get_all_tweets(screen_name)

    print "Indexando tweets en elasticsearch..."

    for tweet in alltweets:
      try:
        res = elastic_search.index(index=index_name,
                doc_type='tweet', body=tweet._json)
      except:
        print "Error indexando un tweet"

    print "Indexado de tweets de " + screen_name + " finalizado"
```

Selection of the test set

Defining a set of competing companies is the responsibility of the top management of any company. In this case, four brands have been chosen while competing for the largest possible market share of customers whose age ranges represent potential users of social networks: ZARA, H&M, MANGO and GAP. All of them are multinational companies and have a Twitter account.

The management of social networks is done by the figure of community manager, individuals or companies responsible for creating a community of users that improve the image of the company by creating relevant content for their followers and performing customer service tasks. The responsibility of those for these accounts is very high because they speak on behalf of the company that hires them.

It is the task of the computer area to create technological solutions that automate the extraction of information from the competition and allow the managers of social networks to discover strategies, virtues and defects, both their own and those of others. This work will allow to analyze the results and make adequate decisions.

The information shown hereafter, as a generated resource consists of a Twitter corpus that includes the period between January 1, 2017 and April 9, 2017, approximately 3 months. This timeframe is not chosen arbitrarily since in it ZARA publishes approximately 3000 Tweets. The following is an example of a series of analysis results to give an example of the usefulness and exploitation capacity of the information collected.

CREATING AN ENVIRONMENT DRIVEN BY INTERACTIVE REPRESENTATIONS

Knowing the Volume of Information over the Period Analysed

The proposed solution integrates with Kibana the development of interactive visualizations for the user, which allows its analysis and progressively by discovering the information hidden in the data. In this article we will follow an approach that will go from the general to the particular. However, the inverse option is also possible if for example we analysed a specific day and we were generalizing to compare it with previous days.

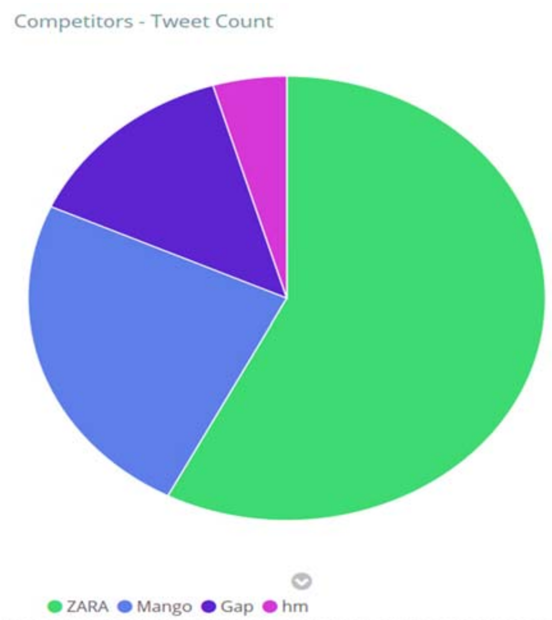

Fig. 2. Sector diagram for the number of tweets

It is important to emphasize that the information analysed depends on a period of time of 3 months and not on the total volume of information available to users. We can have an idea of the number of publications of each account with Figure 3. In addition, we can have a series of additional metrics such as the total number of Favorites, Retweets and the average of both thanks to Figure 4.

Fig. 3. Global metrics for the all accounts' tweets

It would be expected that the account with more retweets and more favorites was ZARA due to the notable difference in the number of publications. However, the tabular representation of these data does not confirm this hypothesis.

User Name	Tweet Count	Favorites	Retweets
hm	259	71,782	18,612
ZARA	3,240	13,326	2,668
Gap	767	4,103	1,131
Mango	1,359	3,144	1,104

Fig. 4. Interactive table of metrics for each brand ordered by the number of Retweets

Views Driven by the Most Used Hashtags

The social network Twitter owes part of its success to the creation of labels with the symbol of the pad (#) that categorize the publications by their most relevant words. The use of tags greatly facilitates the grouping of similar tweets. Each hashtag can also be understood as a strategy of positioning and differentiation with respect to the competition within the platform itself. If an account gets its followers to publish with these keywords will be generating free advertising, less intrusive and reach the target audience alone.

Cloud diagrams are a quick and simple solution that shows important concepts ordered by relevance depending on the size of the text, the color and other attributes of the textual format. However, these tag clouds, at this concrete application, have a series of problems: (i) they do not scale correctly; (ii) they do not show which user or users use them, if any other user is mentioned or in short how these labels are used; (iii) Filtering is strict. For example we see how ss17 and SS17 are represented differently.

These problems are solved by incorporating graph-based visualizations into the tool, as will be seen in section 4.

Fig. 5. Cloud Diagram with the most used tags by all accounts

Our representation is interactive, so we can select a label and see if it is used in conjunction with different ones. We will see how the entire control panel is updated to show only tweets that contain that hashtag selected. For example, we can select ss17 and the image represented in figure 6 would be shown.

Fig. 6. Cloud Diagram updated for the filter by hashtag #ss17

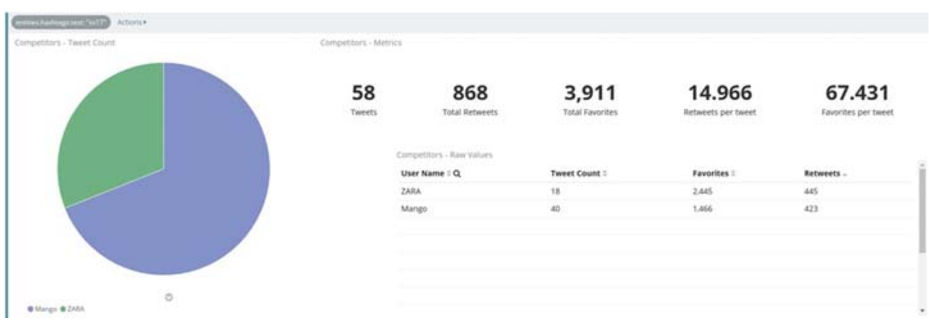

Fig. 7. Updated Dashboard for the filter by hashtag #ss17

Visualizations to measure the impact of competition

Measuring the impact of an account on social networks is more complicated than measuring the number of publications without having an associated tool that facilitates the task. In the previous section we have observed that the number of publications is not related to how their publications accept their followers. To analyse in greater depth the scope of each account, a series of visualizations governed by the date of publication are proposed.

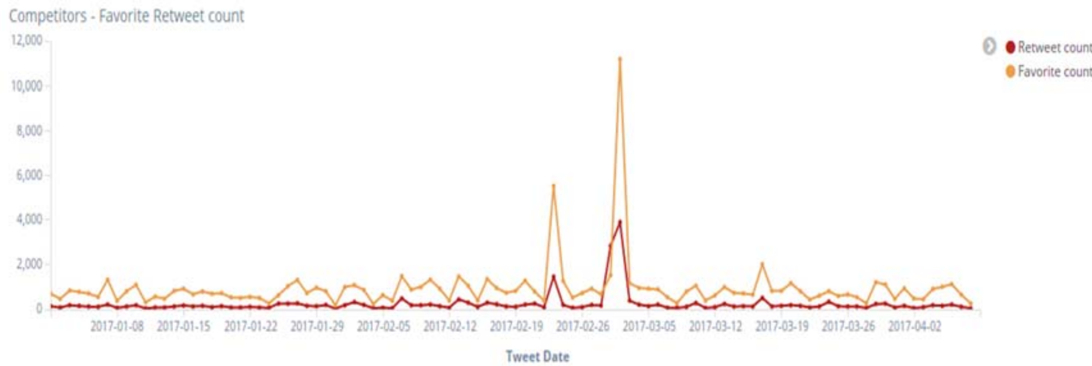

Fig. 8. Temporal evolution of the number of retweets and favourites of all accounts

First, we can create an interactive graph that shows the evolution of the number of retweets and favorites of all accounts. It is observed for these concrete data that both measures evolve pro-proportionally and that there are peaks of activity for specific instants.

Subsequently, it is analysed which accounts have the greatest number of interactions with their publications each day by the brand. To do this, two graphs are made, Figure 9, which show the accounts that take the highest percentage of all the retweets of each day. It is observed that H&M being the account that less content spreads in the period studied, it wins the rest of the brands almost every day in both measures, which, as in the previous case, evolve evenly.

Fig. 9. Area Charts for the percentage of retweets and favorites for each account

Inspecting the peak of activity is a simple task and allows us to limit the period of the representations. Normally this information would not be enough and we would need to get to know the data of the publication. That is why raw data should also be able to be consulted. We see that the tweet that generates the peak of activity is from H&M and mentions an account that has several million followers.

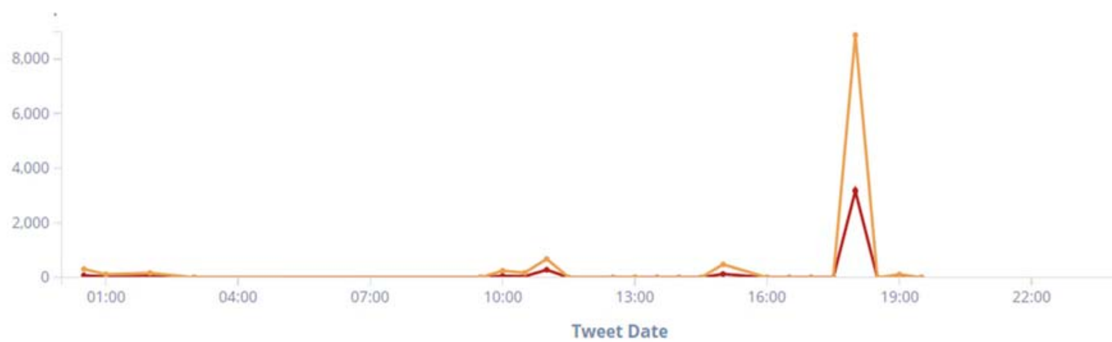

Fig. 10. Temporal evolution of the number of retweets and favourites of all accounts for March 2, day of activity peak

Raw Table Search

1 2 »

Time	_id	favorite_count	retweet_count	user.screen_name	text
March 2nd 2017, 18:00:05.000	AVtTsp7ITV9pj2LLGbil	8,881	3,176	hm	See all the great pics from our new #HMSpringIcons collection starring @theweeknd. Shop in store and online... https://t.co/mR9ld19bOy
March 2nd 2017, 11:06:13.000	AVtTsp7LTV9pj2LLGbin	418	225	hm	.@theweeknd leads the way to your ultimate spring wardrobe! #HMSpringIcons https://t.co/AN17F4f29X #HMMan https://t.co/N6Qzd780rX
March 2nd 2017, 15:01:16.000	AVtTsp7KTV9pj2LLGbim	255	81	hm	The collection, the front row, the models & the love! Relive the #HMStudio #SS17 show & shop the pieces now at... https://t.co/28fd8mLIs8
March 2nd 2017, 11:16:14.000	AVtTsR3TTV9pj2LLGaSY	264	59	ZARA	Woman | New Romantic collection. Check out the full editorial at https://t.co/xMlTz5Kvan #zarawoman #ss17 https://t.co/axopCRuJb5
March 2nd 2017, 10:08:51.000	AVtTsp7PTV9pj2LLGbip	248	58	hm	Indulge in #HMStudio highlights and shop all the looks at https://t.co/JDzYRhEF5Z. #SS17 #PFW https://t.co/no7BJ4Poh4

Fig. 11. Publications made on the day of activity peak by all accounts ordered by the number of retweets

KNOWLEDGE EXTRACTION THROUGH GRAPH-BASED REPRESENTATIONS

In summary, a graph is composed of a set of nodes or vertices and a set of edges that join them. The total number of nodes defines the size of the network, while the total number of edges defines the number of interactions between the elements of the network. On Twitter, users are shaped as nodes and their relationships as directed links graphing the relationships of friendship and followers. In addition, graphs and their clusters are visualized using the Multidimensional Scaling technique (MDS).

This technique consists of assigning spatial coordinates to the elements of a collection, trying to re-present the similarity of the different elements with the distance between them. Similar elements will be found next according to the resulting coordinates (David Easley & Jon Kleinberg, (2010).

Exploring a known hashtag

Imagine that we have discovered by chance, as explained in section *Views Driven by the Most Used Hashtags*, which the hashtag ss17 is used by different accounts and we need to know how each hashtag account uses.

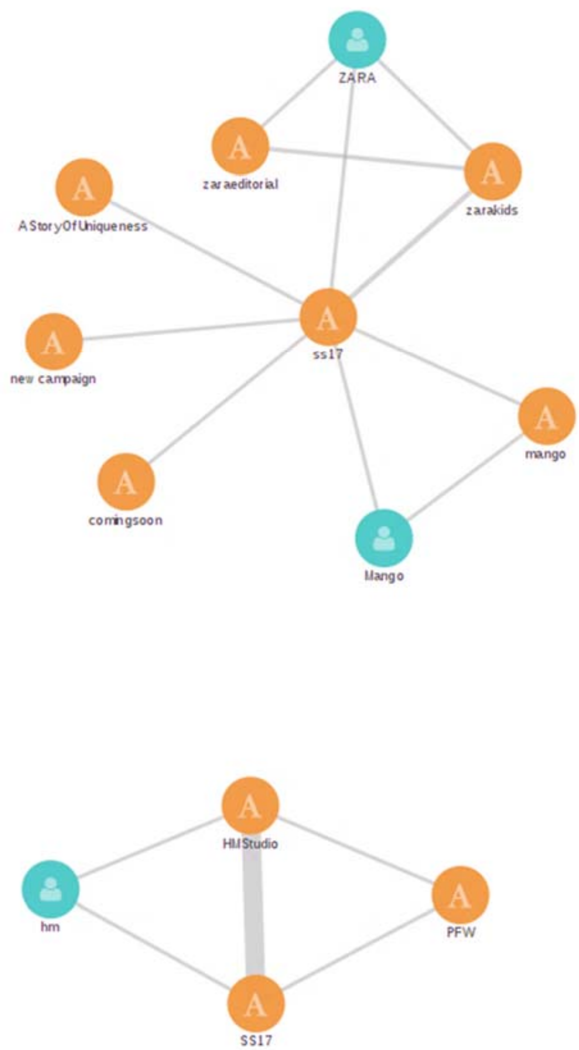

Fig. 12. Graph associated to the ss17 concept showing users that use the hashtag and other related hashtags

A representation based on graphs would allow us to discover how the publications of each account relate to their own and those of their competitors. The proposed solution involves making a weighted graph in which the nodes represent either a hashtag or a user that publishes.

The paths symbolize the relationships between nodes of the graph and the weight associated to each one of them will be the number of elements of the intersection of the set of documents that use a hashtag with the set of documents created by a user. That is, in this case the number of tweets that a user makes using a specific hashtag. We can see in Figure 12 the graph resulting from introducing the ss17 concept.

It is discovered that apart from ZARA and MANGO, the H&M account makes publications using the same hashtag but in capital letters. This data could well have gone unnoticed by a cloud diagram. We can check how the weights of the graph correspond to those of Figure 6.

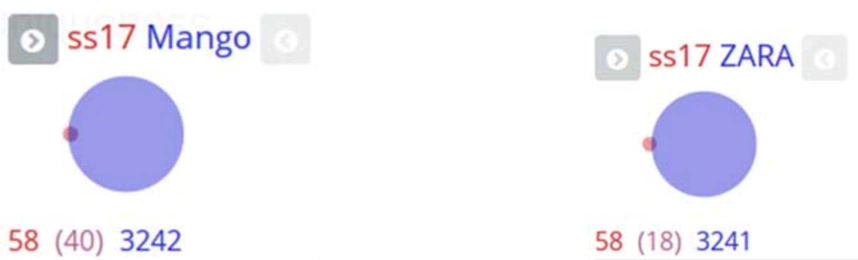

Fig. 13. Intersection of the set of tweets of the ss17 hashtag with the tweets of the users of MANGO and ZARA

Through the addition of different node types, the graph begins to grow and we begin to obtain more information. In this case, the relationship between users, mentions and answers is also shown.

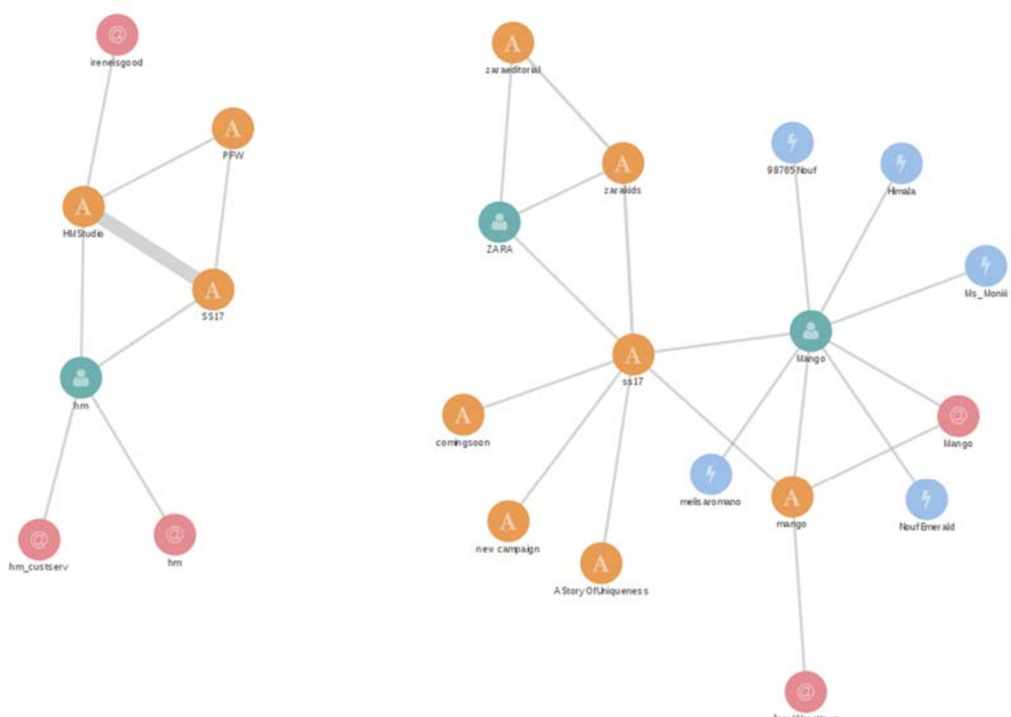

Fig. 14. Graph associated to the ss17 concept showing users that use the hashtag (person), other related hashtags (A), users mentioned when using the hashtag (@) and users to whom it is answered (lightning).

Exploration by users

The only knowledge that should be entered into the system are the names of desired users from whom you want to download tweets. Likewise, the concepts that must be entered if you wish to obtain information associated with them will be incorporated. In Figure 15, the hashtags used by ZARA are observed.

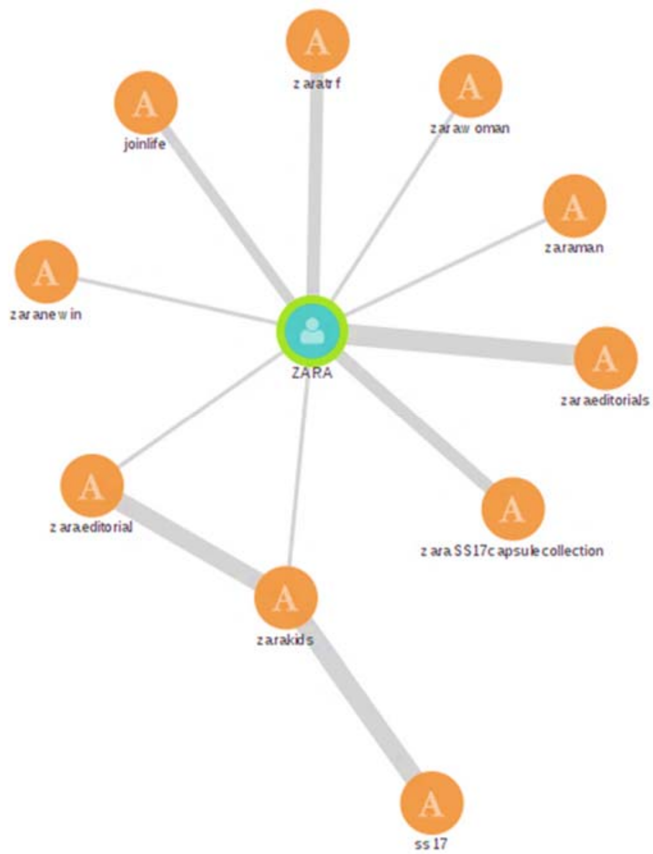

Fig. 15. Hashtags used by ZARA

To this graph we can add the mentions and users to whom ZARA responds. We will get a general idea of the interactions made by this account.

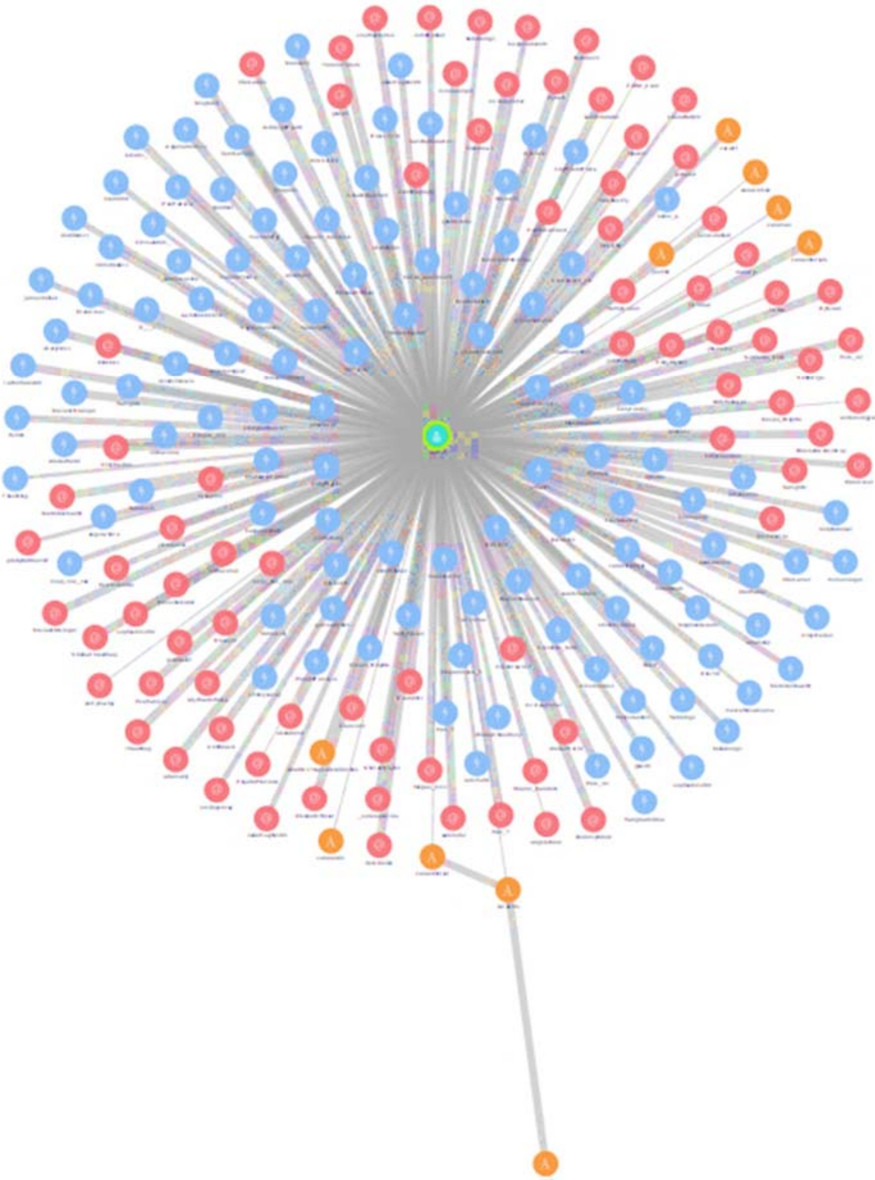

Fig. 16. ZARA Interaction Graph

In Figure 17 we observe the appearance of what are called communities in the network. A community is formed by a group of highly connected vertices with links scattered among the other groups of vertices. Each community is made up of a set of vertices with many connections to each other and each community is connected to the others through few links, generating the so-called bridges between communities (Newman, 2006). By sequentially adding each one of the studied accounts, we obtain that in the period of time analysed, the only connection between them is the hashtag #ss17 previously mentioned since each account generates a totally differentiated set to the rest. This gives us an idea of how segmented the market is and the loyalty of the clients to a specific brand of the four analysed.

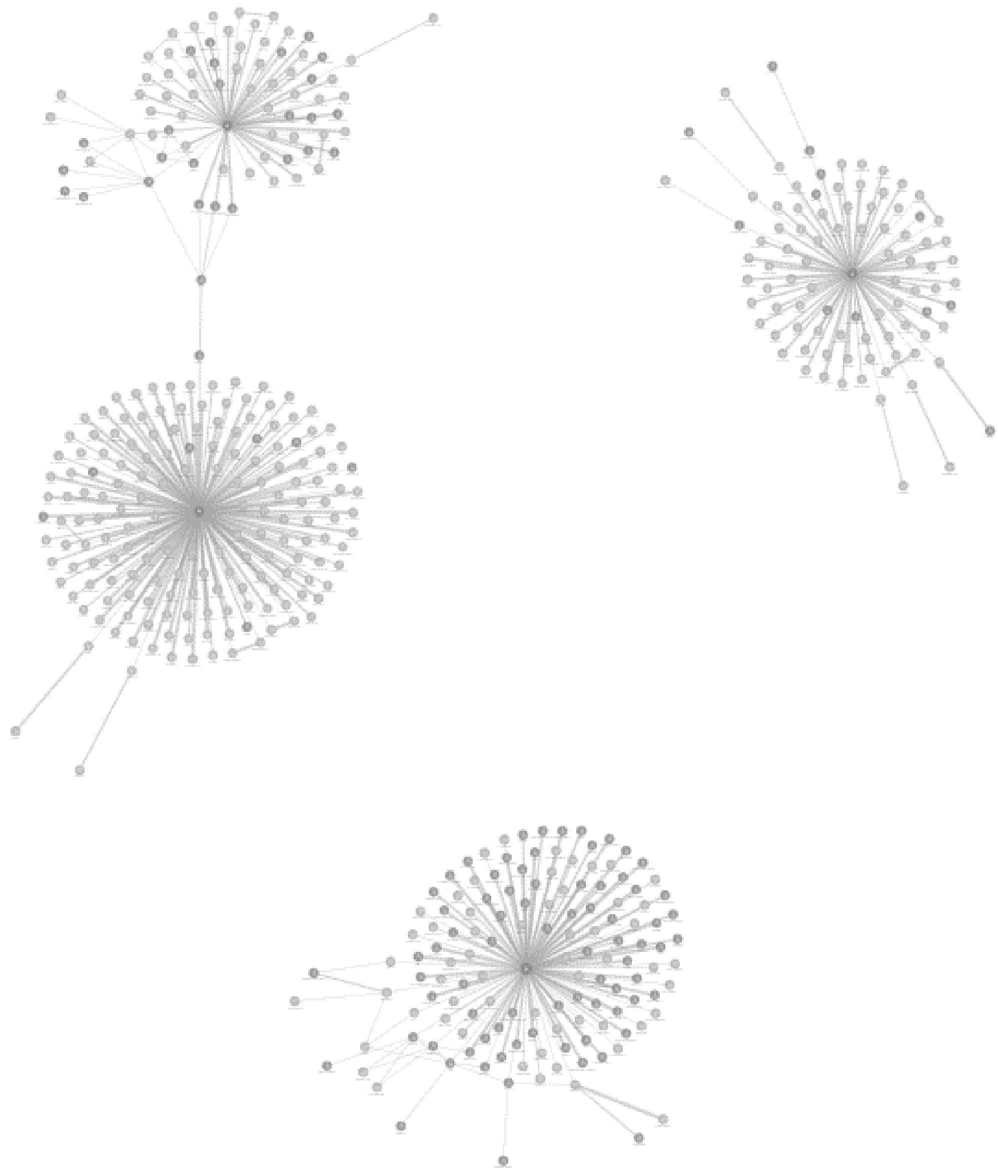

Fig. 17. Graph of connection between the set of competitors

CONCLUSIONS

Twitter is a social network that is mostly public, which allows brands to use two fundamental uses. The first consists in converting Twitter into the official means of communication and in a second place it allows these brands to perform social listening through said network. Brands use Twitter as a means to disseminate valuable content by making consumers share it by making it viral and making those brands gain visibility. This allows brands to generate communities as a sum of their followers, who become ambassadors of these brands at the same time helping their promotion. Only the content that followers consider of value are shared, so that it becomes a channel of support for marketing campaigns about the brand or about their services or products.

The presented tool allows the retrieval of tweets from one or several brands while comparing and analyzing them in common time spaces. Also facilitates to follow a specific hashtag and its temporal evolution. The visualization interface helps to discover the information hidden in the data. Interactive elements representations allow us to work from the general to the particular, choose a user, a period of time, a hashtag or any other filter that is convenient for us. However, to make possible to discover behaviors inside the information it is needed to propose different alternatives that answer at specific moments or any possible problem.

Representation by graphs is very useful when looking for relationships between the different elements and their attributes quickly, in case of not having information about the indexed data. The case study has shown us from this visualization that the four analyzed brands generate four distinct communities perfectly segmented.

Twitter allows brands to establish a dialogue with their consumers beyond the buyer-seller role. The case study allowed us to observe that in the case of ZARA, fundamentally, it uses Twitter as a customer service, generating a huge amount of tweets. For this reason measuring only the volume generated in the number of tweets is insufficient, as the vast majority are irrelevant and serve this customer service so they do not interest the rest of the followers, who do not share them. This is not the case with the H&M brand account, which with 7% of the tweets made by ZARA is capable of having a much greater impact because it searches for more marketing functions based on the virality phenomenon.

As future lines of work, the presented prototype allows an important improvement field. The most immediate ones will be to work on new representations that allow the possibility of filtering irrelevant publications for community managers in a simple way. The idea is to add more metrics that allow analysing and comparing a brand and its competitors, including new measures on the number of followers and their relevance in the community with the mentioned users, whether the messages are an answer or any other method that was necessary.

In addition, the tool will be prepared to expand the number of downloaded data and automate the unloading of certain labels automatically when they are used by a competitor company.

Acknowledgments: This work has been partially funded through the INNHOME project: Predictive Innovation for the Improvement of the Competitiveness of SMEs (RTC-2016-5315-6), Ministry of Economy and Competitiveness (Challenges Collaboration 2016) and Federal Funds.

REFERENCES

Castañeda, L. & Gutiérrez, I. (2010). Redes sociales y otros tejidos online para conectar personas. Aprendizaje con Redes Sociales. Tejidos educativos en los nuevos entornos. Sevilla: MAD Eduforma.

Celeste Lyn Paul, Randall Rohrer, Bohdan Nebesh, (2015) "A "Design First" Approach to Visualization Innovation", Computer Graphics and Applications IEEE, vol. 35, pp. 12-18, ISSN 0272-1716.

David Easley & Jon Kleinberg. (2010). Networks, Crowds, And Markets: Reasoning About A Highly Connected World by Cambridge University Press, 2010, 744 pp. ISBN: 0521 195330

Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103(23), 8577–8582. http://doi.org/10.1073/pnas.0601602103.

Artículo recibido: 20/11/2017
Artículo aceptado: 08/05/2018

Editor in Chief: Prof. Dr. Luis Camilo Ortigueira-Sánchez
Guess Editor for this paper: Prof. Dr. Antonio Navarro García