First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3597

Know your customers from Twitter contacts: automatic discrimination of peer contacts from news sources

Munar, Antoni[a]; Chiner, Esteban[a]
[a]GFT Group, Av Barón de Carcer 48, 46001, València, Spain.

Abstract

Know your customer is a core element of any customer relationship management system for mass service organizations. The emergence of social networking services has provided a radically new dimension, creating a more personalized, deeper, ubiquitous and almost real time relation with customers. At the same time, some of the more widespread social network platforms seem to be evolving not only as social networks between individuals but also as mass information distribution media. When knowing your customer through social networking services, it may be of interest to disambiguate which part of the customer context in the network relates to his peers from other sources. In this paper we present an algorithmic approach to disambiguate one aspect of such relation, as expressed in the nature of the contacts established in the social network: with peers or with organizations, news media or influencers. We focus in the case of Twitter where a simple supervised linear regression can provide a ranking score, effectively discriminating and ordering by closeness peer and other types of contacts (mass media or influencers). Such discrimination can serve as a preliminary step for deeper analysis or privacy protection of customer interaction and is suitable for implementation in automated Big Data systems.

Keywords: *Big Data; Social Mining; Twitter; Social Networks; Know Your Customer; KMeans clustering; Linear Regression.*

1. Introduction

Since the advent of mass production, mass services and mass marketing in the second half of the twentieth century, customer relationship management (CRM) systems have become a core component of the modern company in the race to improve customer satisfaction and retention in an increased competitive environment (Injazz 2003). Such technology applications have required an integration of front office systems (sales, marketing, customer service) and back office systems (financial, operations, etc..) with a set of well defined “customer touch points” (e-mail, direct mail, media marketing, etc..) carrying mostly a transactional (customer – company) or unidirectional (company – mass media) character.

The full emergence of social networking services in the early 2000s (Boyd 2008) supporting a series of globally wide spread social networks has added a radical new dimension. Information about customer preferences, customer-company customer-product and company-competitors is nowadays hidden inside of vast amounts of unstructured social data as blogs, posts in social networks, reviews in collaborative sites or opinions instantaneously expressed in micro-blogging sites, to name a few (Oberhofer 2015). Companies have been increasingly turning to the most widespread social networks (Facebook, Twitter, etc…) to engage with customers trying to integrate the anonymous social network information into their processes and operations (Heller 2011), to gain a 360º view of their customers, in what is known as “know your customer” (Bielski 2001). This integration has consisted mainly in engaging in conversations of different type (customer complains, brand promotion and company news), reputation alerts and sentiment analysis and community analysis of opinion leaders and influencers (Oberhofer 2015, Zhang 2015).

However, it has been stated that some of these Social Network Services (*e.g* the microblogging platform Twitter) act more as information networks that social networks (Myers 2014), being consistenly used by customers increasingly as a source of news of events outside the realm of direct acquitances and relatives (Mitchell 2015). These news can be originated and propagated either from peers or directly from media or companies with Social Network engagement.

In this paper we present a machine learning procedure allowing in a Social Network like Twitter, discriminate and rank by a perceived degree of “closeness” which contacts of given customers in the social network are most probably news sources (media or influencers) rather than peers from their most inmediate circle. Such discrimination can provide a first disambiguation of the community context of the interacting customers (“external” news media from peer circle) in a scalable and prompt way, avoiding biases in subsequent analysis. Also, such automated procedure can prove itself very appropriate as a first step of automated analysis of customer social network communities, where the computational complexity and amount of information needed for the analysis of the community graph (relations, sentiment analysis, natural language processing, theme detection) are greatly complicated by the presence of news media or extremely popular contacts, due to the great activity and number of subsequent contacts.

This paper is organaized as follows. Section 2 describes some relevant aspects of the type of data used from the Twitter public application program interface (API). Section 3 describes the developed KMeans algorithm together with the linear regression model. Section 4 presents the obtained results. Section 5 discusses the validation of both models. Finally, the conclusions summarizes the obtained results.

2. Twitter Data

The public open Twitter API (REST APIs 2016) provides information about any Twitter account that has decided to make public the account. The Twitter Social Network Service can be considered as a directed graph (Myers 2014) where each account represents a node. The graph is directed because the relation (edges) between two nodes are asymmetrical. In Twitter terminology, Node B is said to be a friend of node A if A is subscribed to receive all the posts of B. On the contraire, node B is a follower of A, if B is subscribe to receive all the posts from A. For the account A the degree of the outgoing edges is called the number of “friends”. The degree of incoming edges is called the number of “followers”. For the purposes of this paper, for a given account only the number of followers and friends is retrieved. For model training and evaluation purposes, the twitter id (an arbritary combination of letters and numbers that can be used to access the timeline of posts of the account through the public Twitter web page) is also retrieved.

3. Model

The topological structure of the graph formed by the Twitter accounts follows a power law (Myers 2014). This means that some of the accounts have an exponential big number of followers (the celebrities or news media) while a very long tail of “normal” users will have a reduced number of both followers and friends. 95% of Twitter accounts have less than few hundreds of either friends or followers, with usually a bigger number of friends –*i.e.* accounts that they are following- than followers. For the 50% quantile, the number of friends is 39 while the number of followers is more than half: 26 (Myers 2014). The reduced number of contacts of normal people (compared with influencers) fits with the hypothesis that current people has limited resources to absorb and emit information, roughly of the same order, with a higher tendency to absorb than emit. Another factor is the reciprocity. If I follow you, and you are my real friend, most probable you will follow me as well. On the contraire, the accounts of celebrities, mass media and influences show a different structure. Usually an extremely number of followers, in occasions with comparable number of friends, if they choose to reprocicate or not. Another type is bots, spammers or marketers, with high number of friends, but very reduced number of followers. Despite the filter implement in Twitter (can not be a friend of more than 2000 accounts if more than 2200 accounts do not follow you) they may play a role. Because of its straightforward interpretation and availability the number of followers and friends, together with the reprocitiy –considered here as a categorical value- will be used as variables.

The discrimination of the contacts of a given account can modeled as a binary classification model: peer or mass media (or celebrity). For such classification, we study two different methods: KMeans clustering, which is an unsupervised method and linear regression, that is a supervised method requiring training. The value of the regressand variable of the linear model can also serve as an effective ranking score to order the contacts by perceived "closeness". Both methods can be easily implemented and parallelized in Big Data Systems. Unsupervised methods have the advantage that do not require training and are based only in some very general heuristic principles. Supervised models require training but can potentially lead to more precise results.

3.1 Unsupervised model: KMeans

KMeans (Kanungo 2002) is one of the more populars algorithms due to its simplicity and straightforward interpretation. KMeans tries to find clusters with centroids such that the mean square distance from each data point to its nearest centroid is minimized. As attributes we consider the difference of number of friends between the contacts of the prospect and the number of contacts of the prospect itself $d_{fr} = n_{fr}^{c} - n_{fr}^{p}$ together with the difference beteween the number of followers of the contact n_{fl}^{c}, and of the prospect $d_{fl} = n_{fl}^{c} - n_{fl}^{p}$. The prospect itself will have then $d_{fr} = 0$ and $d_{fl} = 0$. One meta-parameter of the KMeans algorithm is the *a-priori* number of clusters, that must be set ahead based on heuristic considerations. In our case, the "elbow method" which selects the number of clusters with the minimum total variance, is used (Kanungo 2002).

3.2 Linear regression

Linear regression for binary classification suffers from several shortcomings (Agresti 2011), namely heterocedascity of the residuals (wich causes difficulties interpreting the confidence intervals), unrealistic constant regression factors and the fact that the predicted regressand variable can be either bigger than 1 or negative, which is at odds with the interpretation of the regressand as a probability. However, they are sometimes used when what we want is simply to build a rank which values indicate some tendency and we are interested mainly in the extreme cases (contacts representing news media or celebrities are expected to be quite different from normal accounts due to the power law).

In our model the rank variable y that can take values of 0 (peer contacts) or 1 (news media or celebrity) is modeled as a linear combination of: the categorical variable *reciprocity* with takes the value 1 when contact and prospect are followers and friends of each other (0 otherwise) and an followers-friends term composed of the product of the difference of number of friends between the contacts and the prospect $d_{fr} = n_{fr}^{c} - n_{fr}^{p}$ multiplied by the difference of the number of followers between the contacts and the prospect $d_{fl} = n_{fl}^{c} - n_{fl}^{p}$. This interaction term between variables tries to capture the reinforced effect of similar number of friends and followers.

$$y = C + B \cdot rec + A \cdot d_{fl} \cdot d_{fr}$$

4. Model Fit

4.1 KMeans Clustering

25 individuals with Twitter accounts were selected and asked to classify if they contacts were people that they personally know or celebrities or sources of news. Mass media or firm accounts are usually easy to identify because of the explicit name and content of the posts. Another 10 accounts were selected for model validation purposes.

Figure 1 shows the result of clustering for the contacts of a single individuals. The within cluster variance plot shows that the optimal number of clusters is three. The cluster containing the prospect (cluster No 1) is composed of contacts with almost the same number of friends and followers, and its clearly separated from two other clusters, where its components have a very high number of friends or followers. Therefore, by choosing this clusters one could discriminate between peer contacts and news or other contacts. The high variance in the number of followers inside the cluster of the prospect is due to the fact that the cluster comprises not only peer contacts.

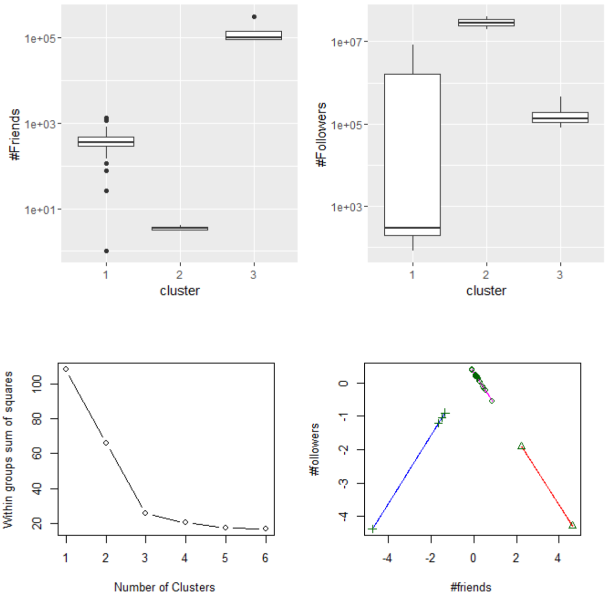

Figure 1. Top left: box plot of the number of firiends for the three different clusters. Top Right: box plot of the number of followers for the three different clusters. Bottom left: sum of within clusters variance as a function of the number of clusters. Botton right: cluster found the Kmeans algorithm. Circles correspond to cluster No 1, crosses to cluster No 2 and triangles to cluster No 3.

4.2 Linear Regression Model

Table 1 shows the results of the fit to the linear regression model with a set of contacts of different individuals. A set of 25 individuals with several tenths of contacts each where asked to indicate whether the contacts were peers or contacts from which they obtain information out of their inner circle. All the parameters with the exception of the intercept are statistical significant. As it could be expected, the reciprocity is a very strong factor in discriminating between peers, accounting for the most part of the discrimination. The followers-friends plays a residual role usefull to discriminate in the case when, *e.g.* a celebrity or organization systematically reciprocicates contacts. Figure 2 shows fit residuals

Parameter	Estimate	Std. Error	t-score	p-value
Intercept C	0.044	0.039	1.124	0.266
Reciprocity B	0.84	0.05	16.777	<2e-16
Followers-Friends A	0.15	0.024	6.147	1.13e-07

Table 1. Estimated values for the linear regression model.

that as expected for a linear regression classification present non-gaussian tails. As in the case of the clustering, contacts classified as peers show reduced and roughly equal number of friends and followers with small variance. Other regression classifiers like logistic regression do not yield statistical significant values for the coefficients due to the huge range of values that the variable followers-friends can take for one of the classes.

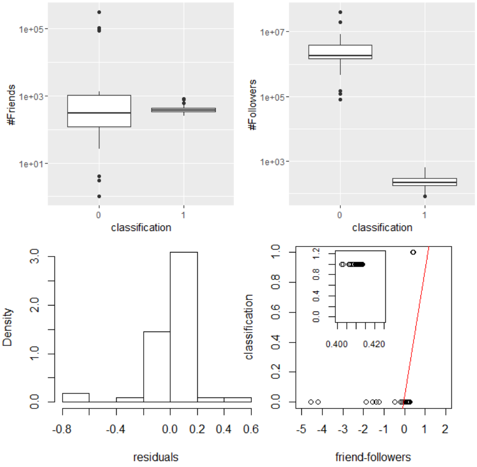

Figure 2. Top left: box plot of the number of friends for peer contacts (classification 0) and others (classification 1). Top right: box plot for the number of friends. Bottom right: fit residuals. Bottom left: classification versus friend-followers, points are data, red line is the regression line. The inset expands the range for some data points.

5. Model validation

To validate the model, 25 individuals of similar characteristics of those used in the training where asked to annotate their contacts as peers or information sources. The reduced sample is due to the high cost to obtain such sample because different individuals are personally requested to evaluate the results. Two kinds of validations can be performed. First, building the corresponding confusion matrix (Table 3).

Method	Precision	Recall	F-Factor
Clustering	0.94	0.75	0.83
Linear Regression	0.95	0.81	0.87
Reciprocity cut	0.96	0.80	0.90

The results show that precision and recall are quite high. Linear regression performance is quite similar to a simple cut based in the reciprocity of the contacts, that is used to asses the real discrimination power of the method. Both methods yield slightly better resuls than the clustering method. However, the linear regression method offers the possibility to rank the obtained results, so results can be ordered by “closeness” quite in the same ways as in other information retrieval systems (Lopresti 1998). When users were confronted with such ordering, usually expressed that the firsts results in the ranking were the most closest to their inner circle, while the results with the lowest ranks usually were newspapers or companies, with a perceived utility in such ranking.

6. Conclusions

For mass service oriented organizations interaction with their customers in the social networks services like Twitter, Facebook, etc… is becoming a core component of their customer relationship management. Big Data technologies allow for the massive analysis of the data generated during customer interaction. In this paper we have investigated an scalable automatic algorithm for one aspect of such interaction. Namely, to automatically disambiguate which part of the customer contacts in a social network relate to his peer circle and which part to other information sources used by the customer. Despite the reduce sample used in the evaluation, results show that a linear regression model based in robust observable variables like the number of contacts of each of the contacts of the customer itself can provide a ranking score automatically discriminating which contacts are peers and which ones other sources. The performance is similar to more straightforward rule based methods, but the linear regression offers a ranking that can be interpreted as a perceived “closenesses” similar to other information retrieval methods. Furthermore, the regression model can be further enriched with futher information. The obtained results show that this model could potentially be implemented at large scale yielding significant results.

Acknowledgements

The authors wish to thanks the colleagues from the GFT Data Practice Ignacio Sales, Sergio Gomez and Angel Rey for their enlighting discussions and support.

References

Agresti, A. & Kateri A.. (2011) *Categorical data analysis*. Springer Berlin Heidelberg.

Bielski, L. (2001). Giving your customer a face. *American Bankers Association. ABA Banking Journal*, 93.4, 49-53.

Boyd, D. & Ellison N. (2008). Social Network Sites: Definition, History and Scholarship. *Journal of Computer-Mediated Communication,* Vol 13, 210-230.

Heller C. & Parasnis G. (2011) From social media to social customer relationship management”, *Strategy & Leadership,* Vol 39, No 5, 30-77

Injazz J. C., & Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, Vol 9, No 5, 672-688.

Kanungo, T., Mount D. *et al*. (2002) "An efficient k-means clustering algorithm: Analysis and implementation." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7*, 881-892.

Lopresti, D., & Zhou. J. (1998) Document analysis and the world wide web. *Series in Machine Perception and Artificial Intelligence* 29 479-502.

Mitchell A. & Page D. (2015) The evolving role of news on Twitter and Facebook. *Pew Research Center.*

Myers S., Sharma A. *et al.* (2014) Information network or social network?: the structure of the twitter follow graph. *Proceedings of the 23rd International Conference on World Wide Web*. ACM New York, 493 – 498

Oberhofer M., Hechler E. *et al* (2015) *Beyond Big Data. Using Social MDM to Drive Deep Customer Insight*. Pearson plc publishing as IBM Press.

REST APIs, Twitter Developers (2016) https://dev.twitter.com/overview/api/users

Russell, M. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, Inc.

Zaharia, M. , Chowdhury M. *et al*. (2012) "Fast and interactive analytics over Hadoop data with Spark." USENIX; login 37.4: 45-51.

Zhang L., Zhao, J.*et al* (2015). Who creates trends in online Social Media: the crowd or opinion leaders. *Journal of Computer-Mediated Communication*, Vol 21, 1 1-16.