Creating Usable Customer Intelligence from Social Media Data:

Clustering the Social Community

Rosaria Silipo Rosaria.Silipo@KNIME.com
Phil Winters Phil.Winters@KNIME.com
Killian Thiel Killian.Thiel@uni-konstanz.de
Tobias Kötter Tobias.koetter@uni-konstanz.de

Table of Contents

Social Intelligence: from Visualization to Clustering

This whitepaper is the continuation of a previous whitepaper titled "Creating Usable Customer Intelligence from Social Media data: Network Analytics meets Text Mining" and downloadable from the www.knime.org web site.

In that whitepaper, text mining and network analytics were combined together to provide a better description of each user of a forum in terms of leadership and sentiment. By using network analytics, an authority score was calculated for each forum user. Taking the authority scores into consideration, we could differentiate between users with a high level of leadership i.e. those with a high authority score and users with a high follower behavior i.e. those with a high follower score. Text mining was used to measure the attitude of each user in the forum. Using the attitude measure, we could identify positive, neutral and negative users across a wide spectrum of positive- and negativeness. Combining the authority/follower score with the attitude measure, we were finally able to visualize on a scatter plot a few users with a high authority and a positive or neutral attitude and a few less users with a high authority and a negative attitude.

While visualization of users on a scatter plot allows us to isolate the more interesting extreme users, it helps neither with an automatic user characterization nor with the description of the remaining more average users. Therefore, in this whitepaper we reach to traditional data analytics in order to define a few groups with more general user features. Indeed, we identified a number of different clusters, including a very large cluster of inactive neutral users, a smaller cluster with positive and very active users, and an even smaller cluster with negative very active users.

By identifying the cluster of very active and enthusiastic users, we have effectively identified "superusers" or "superfans". This group will be the most willing to support a topic (or product) and actively contribute to spreading the positive news. These are the users that you would want to keep informed about products and relevant community news. Particular care should be used towards the users in the small cluster of highly negative and active users, in order to listen to their complaints and act consequently if their complaints turn out to be real problems. And finally, neutral users or low activity users always need to be kept informed, especially if those are highly regarded users. Some of them might actually benefit from a little encouragement to move into the superfan circle.

Note: To enable our approach to be repeated by the reader, we have used the KNIME open source platform throughout this whitepaper. All data and workflows are available on the KNIME site at WWW.KNIME.COM.

Summarizing the User Categories

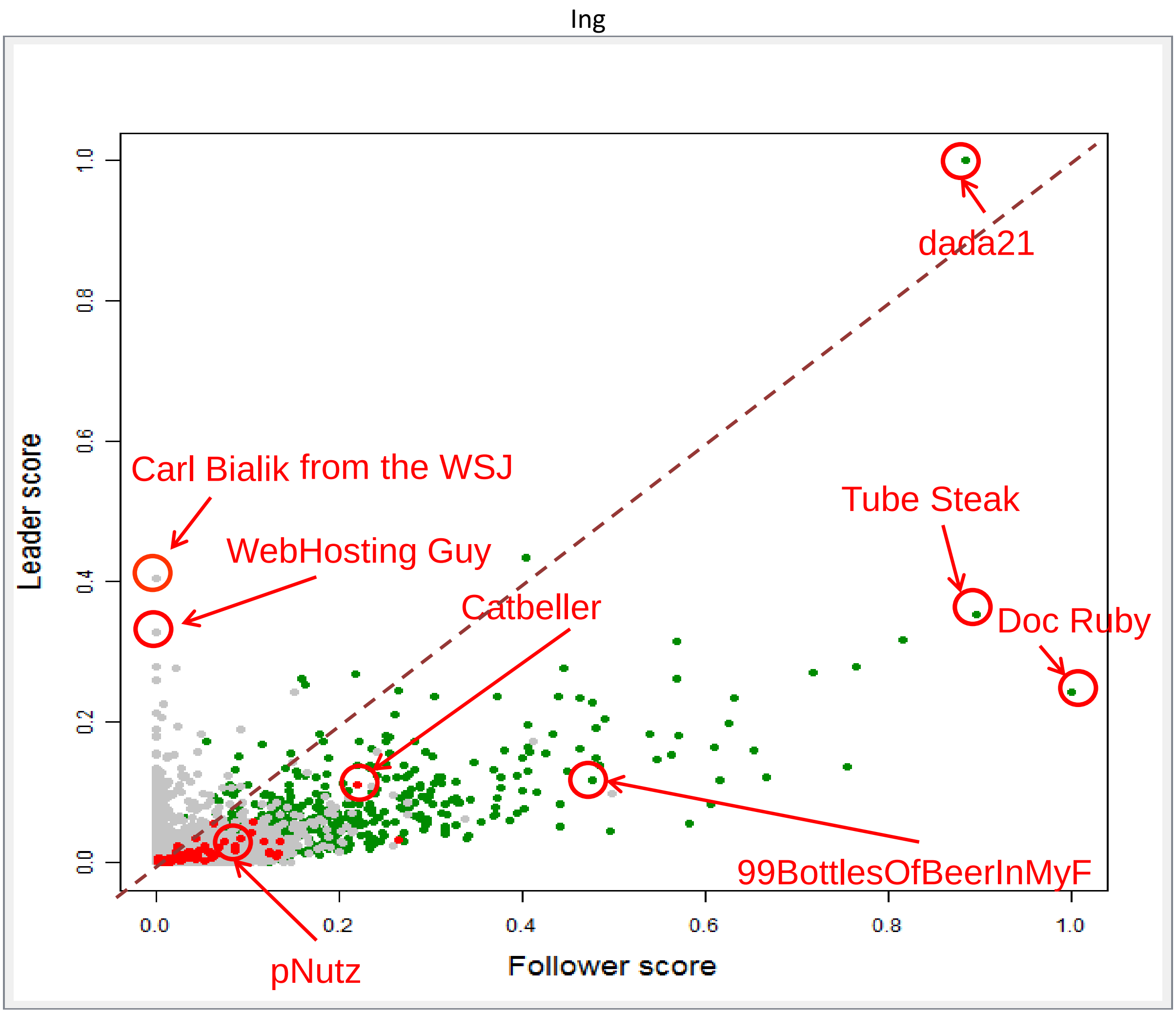
In the previous whitepaper "Creating Usable Customer Intelligence from Social Media data: Network Analytics meets Text Mining", we analyzed approximately 24,000 users participating in 496 discussions for a total of 140,000 comments around the topic of "Politics" from the Slashdot community. The data set was provided by the Fundación Barcelona Media4 (http://caw2.barcelonamedia.org/node/25).

For each community user we calculated an authority score and a hub score. The authority and hub scores were borrowed from a web analytics algorithm used to detect hub and source web pages. We also calculated the positive/negative attitude of each user by extracting his/her bag of words across all of the published posts and comparing it with the MPQA Subjectivity Corpus for sentiment analysis. Each community user was then described by means of an authority score (leadership), a hub score (follower), and an attitude level (positivity/negativity) and displayed as a point in a number of scatter plots.

In these scatter plots, we observed a few users who were easily distinguished from the crowd: like dada21 as the most positive leader and follower at the same time; or Catbeller as the most authoritative among the negative users; or pNutz as the most negative of all users; or Carl Bialik from WSJ as the most authoritative neutral user (Fig. 1).

However,in terms of marketing strategies it is not enough to simply look for the few users who stick out. We need a more automatic and general approach to identify the groups of users that might be interesting.

Figure 1. Leader vs. Follower Score for Negative (red), Neutral (gray), and Positive (green) Users.



In the previous whitepaper we identified a few possible categories of users by visual inspection:

- The followers with high hub scores. Followers do not have a high influence on other users and their attitude does not really propagate to other users. Therefore, focusing on followers with negative, positive or neutral attitude will not be of interest.
- The leaders with a high authority score. The leaders represent a category of interest, since their opinions travel quickly across a large number of users. Here we might differentiate among the three attitude-driven leader sub-categories:

- The positive leaders
- The negative leaders
- The neutral leaders

The focus was mainly on finding the leaders by visually identifiying them in the scatter plots. In this whitepaper, we focus now on grouping users based on leadership, follower, and attitude features.

Select the Data Mining Algorithm

In order to detect the negative or positive leaders, we might be tempted to proceed with a quick and dirty approach by defining a threshold system on authority and attitude. But as in all manual approaches, we must define exactly where leadership ends and following starts and precisely detect the boundary between negativity, positivity, and neutrality. The definition of these boundaries might end up being a very complex task.

In fact, while an authority score above 0.9 clearly indicates a strong leadership, most high authority scores lie below 0.5. What is the difference in leadership then between an authority score of 0.2 and an authority score of 0.3? What about an authority score of 0.2 and a hub score of 0 in comparison with an authority score of 0.3 and a hub score of 0.2? As we can see, the manual definition of leadership is a very complex task. Even more so is the definition of a negative or positive user.

As an alternative to manually defining a set of rules, we chose to use traditional predictive analytics tools for a more structured and controlled categorization of the users. All of these tools are available in KNIME.

We do not need an interpretable set of rules; we just need to identify all authoritative users. A multilayer perceptron or a decision tree technque were considered. However, both algorithms require a training set with known answers. Since we cannot define a priori exactly what "n authoritative negative/positive user" is, we cannot provide these algorithms with a meaningful training set.

We decided therefore to use an unsupervised training strategy that just aggregates data based on similarities and not based on known classes. From all available unsupervised training algorithms in the KNIME platform, we decided to use the k-Means algorithm because it produces highly interpretable results.

The k-Means algorithm calculates the Euclidean distance across all users and groups together with the most similar ones. Each cluster is then represented by a prototype, calculated as the average of all cluster data. At the end of the k-Means procedure we should obtain a number of clusters containing all similar users in terms of leadership and attitude.

Normalize the Data Features

At the end of the text processing and network analysis workflow, each user is represented by an authority score, a hub score, an attitude level, and an ID. Only numerical features can be used in the k-Means node; i.e. the authority score in [0,1], the hub score [0,1], and the attitude level in [-66, 1113] can be used to calculate the similarity between two different users.

The total number of users, after filtering out the anonymous user, reaches 22664. The authority score, the hub score, and the attitude level, this last feature being named Good.Bad.Rating, are distributed across the 22664 users as depicted in Figures 2, 3, and 4 respectively.

Figure 2. Histogram of the Authority Score

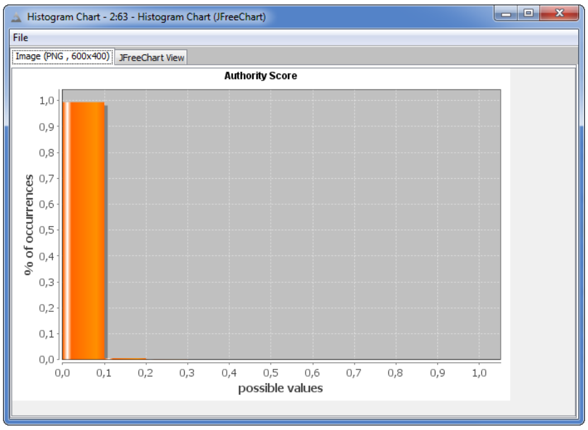


Figure 3. Histogram of the Hub Score

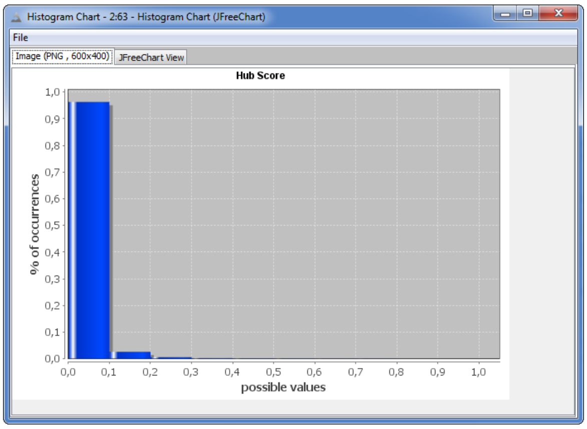
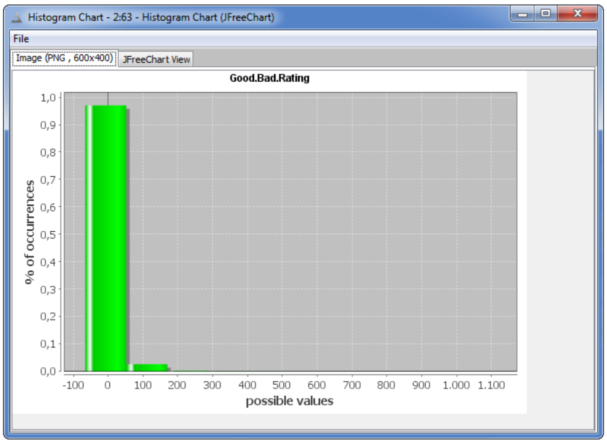


Figure 4. Histogram of the attitude level



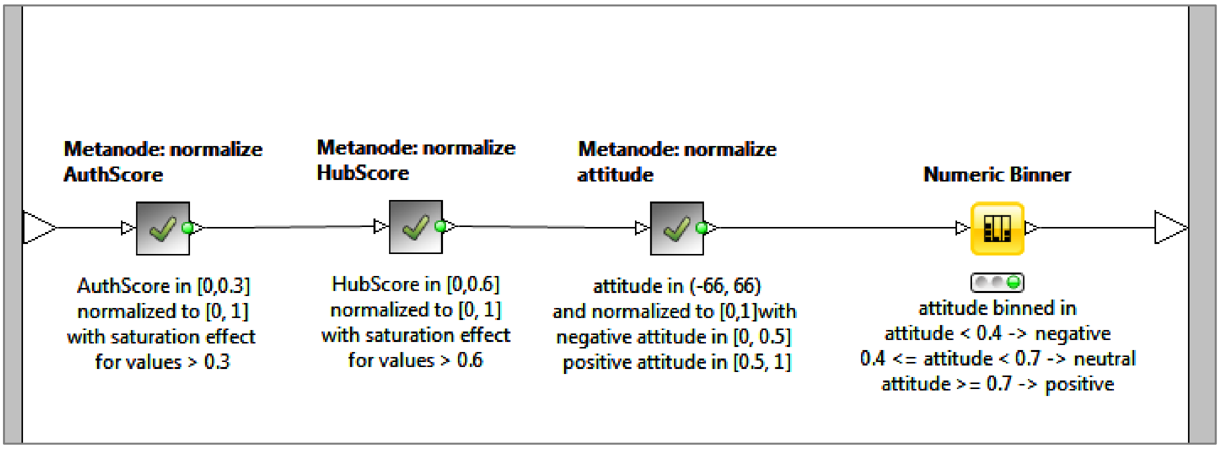
Even though nominally the values of the authority and hub score span the whole [0, 1] interval, almost all users have an authority and a hub score much below 0.5. The same thing happens for the attitude level. Even though nominally the attitude level spans an interval between -66 and 1113, almost all users show an attitude level between -66 and +100.

Since the k-Means algorithm is based on a Euclidean distance, a user with an attitude level of 1000 and a user with an attitude level of 500 would fall in two different clusters and would force most other users with lower attitude values into one cluster. This cluster configuration would then only describe the outlier users. In order to avoid outliers taking over the cluster structure and to enhance smaller differences among users, we restricted our analysis to a subset of each feature’s numerical range:

- [0, 0.3] for the Authority Score,
- [0, 0.6] for the Hub Score,
- [-66, 66] for the attitude level.

Feature values exceeding in both directions of the range of observation were saturated to the maximum or minimum value. All feature values were then normalized to fall in the [0, 1] interval.

Figure 5. The meta-node performing the feature saturation and normalization


Metanode: normalize AuthScore
Metanode: normalize HubScore
Metanode: normalize attitude
Numeric Binner
AuthScore in [0,0.3] normalized to [0, 1] with saturation effect for values > 0.3
HubScore in [0,0.6] normalized to [0, 1] with saturation effect for values > 0.6
attitude in (-66, 66) and normalized to [0,1]with negative attitude in [0, 0.5] positive attitude in [0.5, 1]
attitude binned in
attitude < 0.4 -> negative
0.4 <= attitude < 0.7 -> neutral
attitude >= 0.7 -> positive

The histograms of the new values for the authority score, the hub score, and the attitude level after normalization are shown in Figures 6, 7, and 8.

Figure 6. Histogram of Authority Score after Normalization

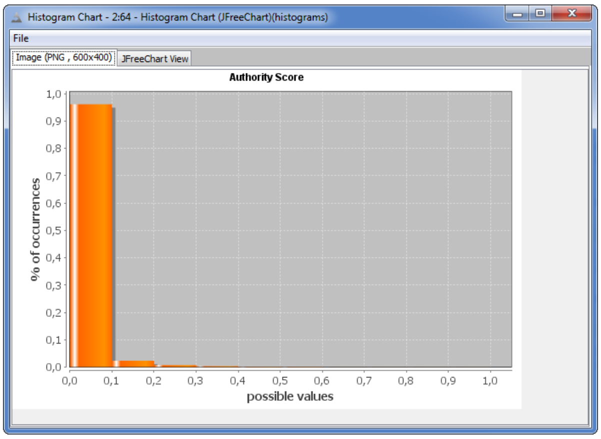


Figure 7. Histogram of Hub Score after Normalization

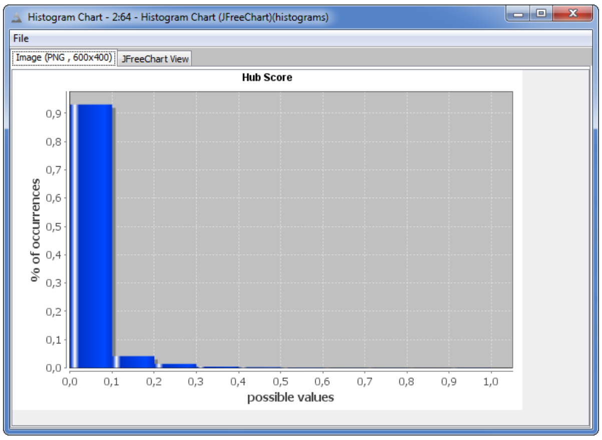
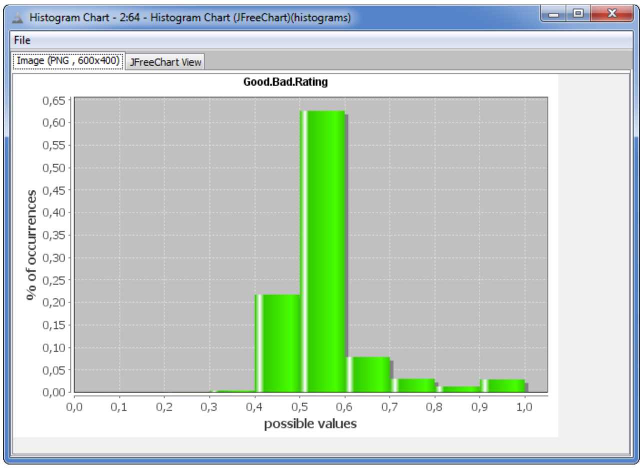


Figure 8. Histogram of attitude level after Normalization



In the previous whitepaper, the attitude level was binned for positive, neutral, and negative users following the histogram of the original data. However, since there are fewer negative users than positive users, the original binning produced only 58 negative users, which is not a very extensive group of users.

After the normalization of the attitude level, we re-binned the attitude level values using the following intervals based on the histogram in Figure 8.

IF attitude level in [0, 0.4[	THEN negative user
IF attitude level in [0.4, 0.7[	THEN neutral user
IF attitude level in [0.7, 1]	THEN positive user

These new binning intervals created 109 negative users, 20917 neutral users, and 1638 positive users.

Notice that these binning criteria, like the previous ones used for the scatter plots, are arbitrary and should not be used as correct answers for a supervised training algorithm. The k-Means algorithm does not need correct answers. These new bins are exclusively used to color and represent the final clusters in terms of attitude level on a chart.

Training Set and Number of Clusters

The first question e to be addressed is how many k-Means clusters need to be built and the second question is how many of the input data rows should be used to train the k-Means clusters.

It will not do to build too many clusters because then we lose the overview and therefore the interpretability of the cluster set. Building too few clusters is also a problelm since too many different users end up in a small number of clusters and these clusters no longer carry any clear meaning..

The goal with the input data is to obtain a good picture of how the users are distributed across the community. The only reason not to use all the data is that this could slow down the k-Means algorithm. In our case, the software used – KNIME – can easily analyze our 22664 data rows.

Thus, let’s establish a maximum number of k=30 clusters and use all the user data. We expect most users to fall in the category of the low follower, low leader, and neutral attitude. That is, we expect at least one big cluster to cover many low-activity users.

With k=30 clusters, we obtain circa 10 clusters with more than 1000 and less than 3000 users, and 20 smaller clusters. Of these 30 clusters, only 2 have prototypes with an attitude level below 0.4: one covers 37 more active and more negative users; the other one includes 207 more neutral and less active users. These two clusters, though small, clearly represent negative users.

With k=20 clusters, the number of clusters with more than 1000 users goes up to 5 while the remaining 15 clusters cover smaller groups of users. Only one cluster, with 50 users only, is represented by a prototype with a clear attitude level below 0.4. We gained in generality (bigger clusters), but we lost something in the representation of negative users.

With 10 clusters, we have 2 clusters with more than 5000 users, 2 clusters with less than 100 users, and 6 clusters covering between 100 and 5000 users each. We have no prototype anymore with an attitude level below 0.4. Negative users are mixed up with positive and neutral users in all clusters.

k=10 clusters seems to offer the most compact overview of the user space, with all very populated clusters. However, k=10 does not produce a dedicated cluster for negative users. 10 clusters seem to be just too few when considering all users, especially since most users are either neutral or positive.

The negative users are such a small proportion of the total number of users that they do not represent a consistent part of any of the 10 final clusters.

On the other hand, 10 clusters only are much easier to analyze than 30. We can then keep the specification k=10 and feed the k-Means algorithm with a more balanced training set. That is, the neutral and the positive users are under-represented to the level of the negative users. To do that, we use an “Equal Size Sampling” node. Now, we re-train the k-Means node with k=10. This time we get 2 clusters, covering 30 and 80 data rows each, with a prototypical attitude level below 0.4, i.e. mostly negative users.

The Final k-Means Clusters

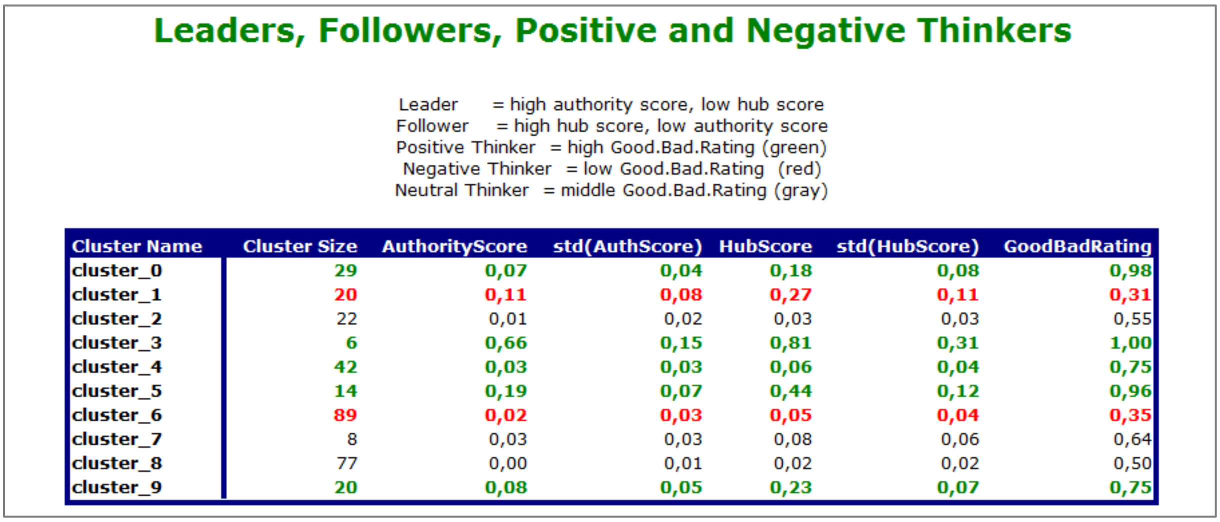
Figure 9 shows the first page of the report associated with the clustering workflow and contains a table with the prototype values for the 10 clusters obtained with the k-Means node. The same workflow also calculated the standard deviation of the Authority Score and of the Hub Score in each cluster.

Each row of the table then describes a cluster in terms of:

- Cluster size, i.e. how many users are covered by the cluster;
- The Authority Score of the cluster prototype;
- The standard deviation of the Authority Score in the cluster;
- The Hub Score of the cluster prototype;
- The standard deviation of the Hub Score in the cluster;
- The attitude level of the cluster prototype, named “Good.Bad.Rating”.

First of all we notice that throughout the table a high authority score goes hand in hand with a high follower score. This seems to indicate that it is hard to isolate a group of pure leaders. The leadership and the follower score together seem more likely to describe an active sort of user rather than a specific social role in the community.

Figure 9. This is the first page of the report associated with the clustering workflow. The table reports the prototypes of the 10 clusters built by the k-Means algorithm.



Leaders, Followers, Positive and Negative Thinkers

Leader = high authority score, low hub score
Follower = high hub score, low authority score
Positive Thinker = high Good.Bad.Rating (green)
Negative Thinker = low Good.Bad.Rating (red)
Neutral Thinker = middle Good.Bad.Rating (gray)

Cluster Name	Cluster Size	AuthorityScore	std(AuthScore)	HubScore	std(HubScore)	GoodBadRating
cluster_0	29	0,07	0,04	0,18	0,08	0,98
cluster_1	20	0,11	0,08	0,27	0,11	0,31
cluster_2	22	0,01	0,02	0,03	0,03	0,55
cluster_3	6	0,66	0,15	0,81	0,31	1,00
cluster_4	42	0,03	0,03	0,06	0,04	0,75
cluster_5	14	0,19	0,07	0,44	0,12	0,96
cluster_6	89	0,02	0,03	0,05	0,04	0,35
cluster_7	8	0,03	0,03	0,08	0,06	0,64
cluster_8	77	0,00	0,01	0,02	0,02	0,50
cluster_9	20	0,08	0,05	0,23	0,07	0,75

The red rows in Figure 9 represent clusters with a low attitude level in their prototype (< 0.4) i.e. clusters composed mainly of negative users. There are two negative clusters. The cluster named cluster_6 has mostly inactive users (low values for the Authority Score and for the Hub Score) and mildly negative users. The other cluster, cluster_1, contains more active users (with higher values for the Authority and the Hub Score) and more negative users.

The green rows represent clusters with a high attitude level (> 0.7) i.e. clusters with mainly positive users. There are five such clusters.

Three clusters, cluster_0, cluster_3, and cluster_5, have an attitude level close to 1, which indicates that they contain the most enthusiastic users: the superfans. Cluster_3 and cluster_5 especially contain very active users (very high values for the Hub and the Authority Score) and very enthusiastic users. These are the teams you would like to call for help to spread news and explain new products to other users. Cluster_0 contains still very positive users, but they are not as active as the ones in cluster_3 and cluster_5.

The other two green clusters, cluster_4 and cluster_9, refer to a still quite enthusiastic user basis (attitude level barely above 0.7), but not as enthusiastic as in the previous three clusters. In addition, their level of activity is greately reduced with respect to the users for example in cluster_3. They are still fans of the topic, but somehow more moderate and less active than the previous ones.

There are three clusters with a prototypical neutral attitude (>0.4 and <0.7). The corresponding rows in the table are shown in gray. However, most neutral clusters show very low activity scores (Authority and Hub Score), indicating that neutral users are also rarely involved in discussions.

The clusters are also represented by means of a bubble chart, where the Authority Score is on the Y-axis, the Hub Score on the X-axis, the standard deviation of the Authority Score produces the bubble size, and the attitude level the bubble color (Fig.10).

From Figure 10 and from the table in Figure 9, we observe the following facts.

- **The Superfans.** Cluster_3 contains the users with the highest activity and most positive attitude. These users are your superfans. You should leverage their enthusiasm to spread the news and to talk about new products. For this reason, they must always be kept informed and encouraged in their positive attitude.

- **The Active Users.** In Figure 10, cluster_5, cluster_9, cluster_1, and cluster_0 are the clusters positioned in the center of the chart: i.e. with the highest values of the Authority and of the Hub Score. They are also the clusters with the biggest bubbles around their prototypes, meaning that they a wider range of different users.

    - **The Positive Active Users.** Cluster_0, cluster_5, and cluster_9 collect mainly positive, especially cluster_5, but equally active users. While these users do not compare with the users of cluster_3 in terms of positiveness or of activity, they are still quite some fan of the topic they are discussing and should be kept in the loop about news and new product features.

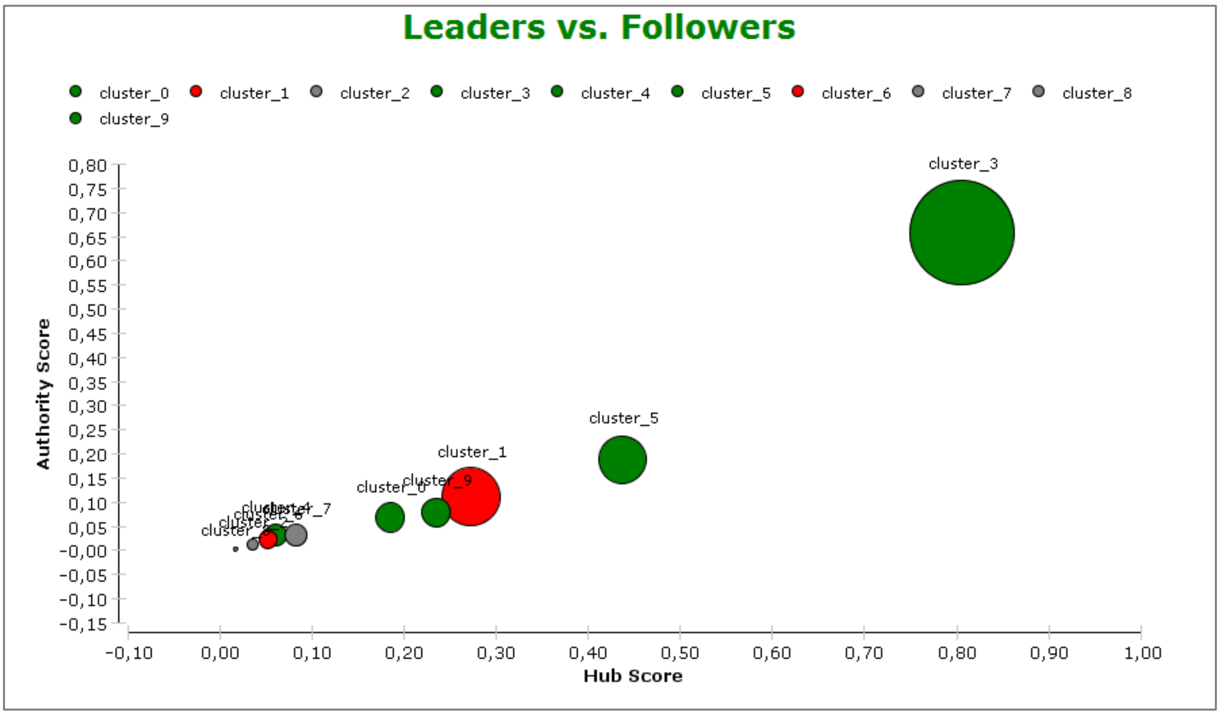
    - **The Negative Active Users.** Cluster_1, on the other hand, collects the very active and negative users. This cluster is the most important one for the negative users, since its users are quite negative and quite active at the same time. These users need some listening to their complaints and then either a solution to their problems or some kind of help and encouragement to change their negative sentiment about the discussion topic.

- **The Inactive Users.** The inactive users represent the majority of the users who can have different opinions about the discussion topic. The bubble size also decreases for the clusters with inactive users, showing a bigger homogeneity among such users.

- **The Neutral Users.** The only important cluster among the clusters with inactive users is cluster_7. Cluster_7, in fact, contains relatively active users (in the middle of the chart in Figure 10) and relatively different users (medium bubble size). In addition, cluster_7 contains mainly neutral users. These users need to be kept up to date in a non-intrusive way. This means that they should not be acknowledged directly, like your superfans, but they might be invited for example to special events for free, which, as in our case, would be political events. Neutral users are also not always neutral; they just balance out over time. So if we can get them to be slightly more positive towards our message, brand, company, that can only help the business.

- **The Very Inactive Users.** The remaining clusters cover very inactive users of all opinions: negative, positive, and neutral.

Figure 10. Bubble chart with Hub Score on the X-axis, Authority Score on the Y-axis, attitude level as color, and standard deviation of authority score as bubble size.



The same clusters can also be inspected in a different bubble chart, as in Figure 11.

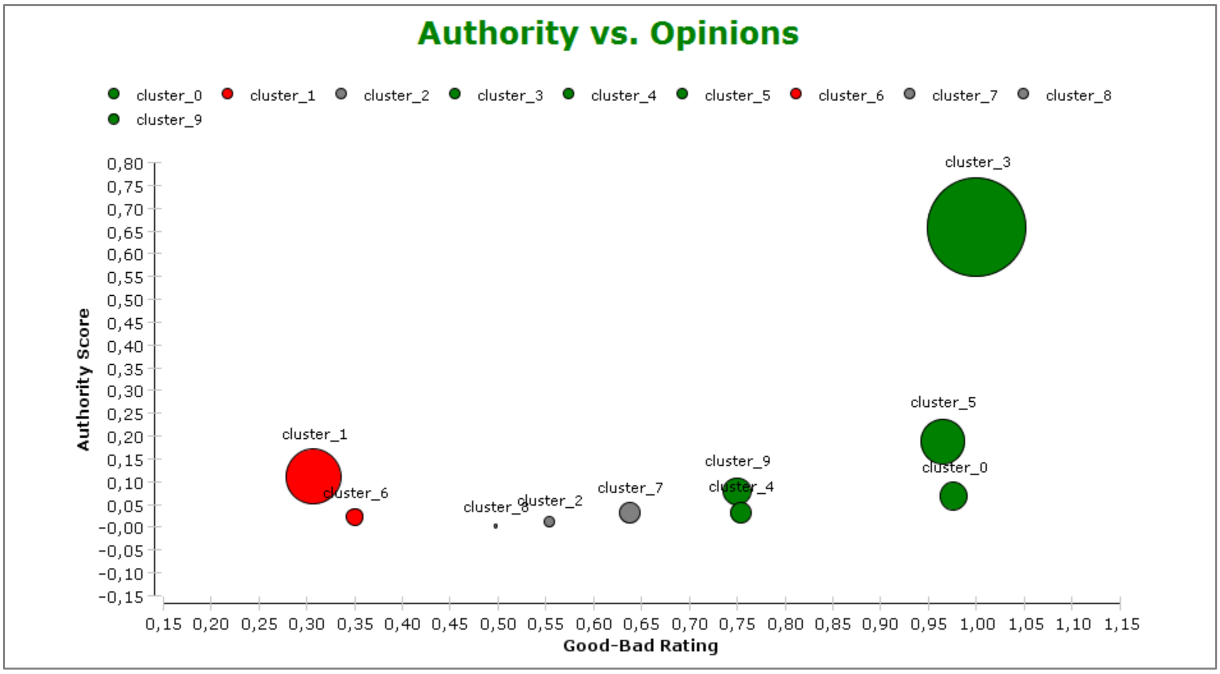
Here, the X-axis indicates the attitude level and the Y-axis the Authority Score. Since the Authority Score and the Hub Score generally move together, representing only one of them as an activity index should already give an idea of how important the users of a given cluster are.

Cluster_3 is again in the top right corner of the chart, given its very high activity score and its very high positiveness. If we color the bubble depending on the attitude level (red if attitude level < 0.4, green if attitude level > 0.7, gray otherwise), we see each cluster clearly located according to the attitude level and the activity score of its prototype.

Besides cluster_3 all other clusters contain users with similar levels of activity. All other clusters are now located in the bottom part of the chart and from left to right depending on their attitude level.

Of course, increasing the number of clusters k, would produce a more detailed picture of the community.

Figure 11. Bubble chart with attitude level on the X-axis, Authority Score on the Y-axis, attitude level again as color, and standard deviation of authority score as bubble size.



Assigning and Predicting Users to the Closest Cluster

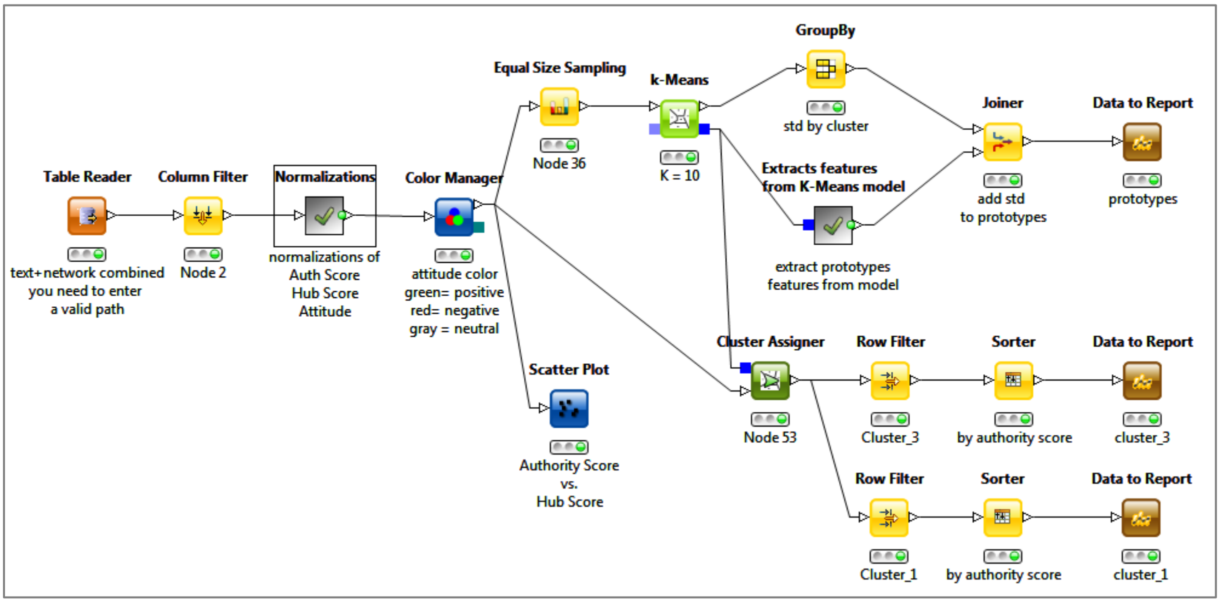
Our initial goal was to cluster all influential users of the community with the goal of contacting them with different modalities according to their attitude. This means, for example, that all users in cluster_3, cluster_5, cluster_1, cluster_9, and cluster_7 must be monitored, kept up to date, and in some cases contacted with special promotions or special initiatives.

Now, not all possible users were in the training set; that is the cluster organization that we have built does not represent all possible users: for example it does not cover future users. In order to cluster all users in our user basis as well as in future data, we need to classify every unknown user against this cluster set on the basis of a distance measure. By means of the “Cluster Assigner” node in KNIME it is possible to assign any new user data to the closest cluster and therefore detect immediately which incentive program should be associated with each new user.

The last page of the report associated with the clustering workflow shows the list of all users classified as belonging to cluster_3 or to cluster_1. These are the users identified as active negative or as superfans. Two different promotional programs are associated with these kinds of users. In particular, we find Catbeller as one of the users with the highest activity in cluster_1 and dada21 as the most active user in cluster_3.

Please note that while cluster_3 contains almost exclusively positive users, cluster_1 contains also a number of neutral users. The final workflow is shown in Figure 12.

Figure 2. The clustering workflow



This technique can be used not only to assign new users to the appropriate cluster, but to identify changes in existing user authority and opinion. By capturing the social media data over time, a window could be set up to process the data and rescore each user, say weekly or monthly, and compare the previous categorization and the new categorization.

In this way, old “superfans” who were no longer active or people who had previously been only slightly active and positive, who suddenly started to contribute much more to become new “superfans” would all be identified at a very early stage and appropriate actions could be taken.

The KNIME Advantage and Conclusion

In the previous whitepaper, combining sentiment analysis from online forum posts together with reference structures from the quotation network allowed us to position negative and positive users in context with their relative weight as influencers or followers in the underlying discussion forum.

In this whitepaper we used the attitude level and the follower and leadership scores built previously to cluster the community. After normalizing the data and producing a training set with equal distribution of all sentiments, we ran a k-Means algorithm and built a few clusters across the community users. The cluster set with 10 clusters turned out to be the easiest to interpret.

One of the first clusters to distinguish itself from the rest was populated by superfans who are very positive and very active users both in terms of follower and leadership scores. These are the users you want to take care of, in order to help you in your marketing strategy by spreading news and supporting your products. The cluster of the “just fans” appeared next in whichstill very positive users yet not as active as the superfans could be found. A big cluster mainly of very active and negative users was also built. These users should at least be closely monitored. They should be listened to and, if their issue is with your product or service, contacted and if possible helpedin solving the problem. A number of other smaller clusters were also found. Those smaller clusters are mainly populated by not so active users and therefore less interesting.

To demonstrate the techniques, we built this set of clusters based on the data of a public forum focused exclusively on the topic of politics.. Howeverthe ability to group individuals into clearly defined social media segments and constantly monitor them for change has a strong relevance for all companies that already use data mining and customer intelligence techniques within their organization. In fact, a good understanding of the social media segments can provide an invaluable contribution to the decision as to how to invest and shape the company's social media and marketing strategies.

By using the KNIME platform, which allows not only traditional data mining, predictive analytics and machine learning, but also text mining and network analysis, we have been able to quickly and effectively create this new understanding of the social media segmentation with a minimum of delay.

Final Note: We used KNIME open source software and publicly available data sources, therefore the complete workflows, also including the data, can be downloaded free of charge from www.knime.com .