Revealing User Influence in an Online Newspaper

Gilberto Flores Pochet, Carlos Kamienski
Universidade Federal do ABC (UFABC), Brazil,
{gilberto.pochet, cak}@ufabc.edu.br

Abstract – *Web-based online newspapers have become a very popular way to share information as well as to allow users to comment news and each other's comments. In such environments some users play a prominent role and eventually influence other users thus leading discussions as they please. So far, there is no way of identifying and quantifying such influence. This paper proposes and analyzes a methodology for identifying influential users that builds implicit social networks upon users who post comments in the same news, suggests three different ways of identifying influential users and measures influence based on similarity of comments. We applied it to data collected from a Brazilian online newspaper and results confirm its effectiveness by revealing a significant similarity between comments of identified influential users and the remaining ones.*

1. Introduction

User generated content and communication among users who produce content take place in different ways, such as in online news portals, blogs, messaging and social networks. An Online Social Network (OSN) generates a huge amount of data that can be used for different purposes, such as for understanding the reasons that make users like an idea, a product or a service. Users can share they thoughts freely and therefore they influence each other's opinions and decisions. Web-based newspapers have become an interesting way to share information in the past 15 years [13]. They provide Webpages and mobile applications where users can read a large variety of news, as well as post their comments and impressions about the news themselves and about other users' comments.

The effect of users who actively participate by posting comments in a social interaction driven by news is as significant as the effect of the content of the news itself [1]. The Internet may promote user active engagement because it allows users to access information on demand and to receive news in easier ways. Also, Web 2.0 features allow them to customize information to suit their interests and to provide deeper information about any topic as they see fit [3]. This flexibility can influence other users who are not only interested on news but also on other users' opinions.

In this paper we propose and analyze a methodology for identifying influential users, called Star Users, with six steps but with three key components. Firstly, we build implicit networks based on user participation on an online environment that is not officially a social network. Secondly, we propose three methods for identifying influential users: a) who posted comments in news together with the highest number of other users (Star-D); b) who posted the highest number of comments (Star-M) and; c) who posted comments in the highest number of news (Star-N). And thirdly, a method for measuring the level of similarity in the comments posted by the prospective influential users and the other remaining users, using text mining, data mining and similarity indexes. The key purpose of our work is to pinpoint more precisely the influential users. For the scope of this paper we define influence as the ability of certain users to generate discussion on a piece of news and particularly to lead other users to follow their line of

reasoning. We measured the former aspect by creating implicit social networks and identifying influential users and the latter by comparing comments using text mining.

We crawled public data from five different sections - Politics, Daily, World, Education and Technology - of the Brazilian Folha de São Paulo online newspaper. We collected news, comments and users for a period of three months, from November 2013 to January 2014 and only used public data, since privacy is a major concern of many users. Based on our results, we can see that the methodology works and confirms the influence inside the network, since the similarity metrics (Jacard's index and Dice's coefficient) yield high scores. For example, the Daily and Politics sections, the most popular ones, present scores as high as 0.76 for Dice and 0.41 for Jaccard. In other words, the ideas concealed by the Star user through their comments influence other users. Also, we observed that commenting the highest number of news (Star-N) and posting the highest number of comments (Star-M) generates more influence than having the highest degree (Star-D). The key contribution of this paper is the innovative methodology for identifying and measuring influence of users who comment news in an online newspaper.

The remainder of this paper is organized as follow. Section 2 discusses related work. Section 3 presents our methodology followed by the main results in section 4. Section 5 discusses lessons learned and section 6 draws some conclusions.

2. Related Work

There has been some research on identifying and measuring influence on online social networks, mainly analyzing Twitter due to the practicality of collecting data [18][2]. As stated by Castells [5], influence refers to the occurrence where the action of individuals can induce their friends to act in a related way. Centola [6] focused on the spread of behaviors among users and concluded that the adoption rate is higher when individuals receive multiple reinforcement signals from their neighbors. In our case, it may be mapped to a case where users start posting comments after a certain number of neighbors did it. Katona et. al [12] analyzed how highly connected users influence their neighbors, in the context of a traditional online social network, whereas our work analyzes an implicit network formed by users of an online newspaper. Chen et. al [7] studied influence maximization using greedy algorithms, but in a way not related to our problem.

Other line of existing research uses qualitative approaches for analyzing motivation and analysis of discourse of users who comment about online news. For example, Weber [17] discusses factors influencing participation and interactivity in online news and Diakopoulos & Naaman analyze the quality of discourse in the comments posted by readers of online news. We infer user influence quantitatively analyzing a large amount of news and comments and modelling the problem as a graph.

To the best of our knowledge, this paper is the first source of information in the literature that particularly deals with online newspapers using a network-oriented approach. There are some other areas of related work as online social networks, data mining, but we did not find anything similar to our methodology and our evaluation based on real data taken from an online newspaper.

3. Methodology

Our methodology has been developed specifically for this research and therefore it is considered as a key contribution of this paper. Its main idea is the creation of implicit

social networks based on the behavior of users who post comments about news in an online newspaper. This methodology is based on a sequence of six steps: 1) data collecting; 2) building social networks; 3) exploratory data analysis; 4) selecting star users and groups; 5) text mining and discovery of word clusters; 6) finding similarities.

Together, all steps of our methodology contribute to better understanding the two related key challenges: a) to identify the most influential users among those who post comments on news in terms of driving discussions or calling the attention of other users; b) to learn how to measure and where influence can be fund more frequently inside a newspaper section.

3.1. Data Collecting

The newspaper chosen for gathering data is Folha de São Paulo (well-known as Folha), considered the number one in Brazil for both the traditional paper-based as well as the online web-based versions. Folha provides public and free access to most of its online news and comments posted by individual users. In order to obtain this data, we developed a crawler, using the Selenium tool and Java, that collected all news and comments from five Folha sections (Politics, World, Daily, Technology and Education) from November 1st 2013 to January 31th 2014. We collected 5,478 news and 37,672 comments coming from 6,950 different users. Please notice that the amount of data is low compared to what one can gather from online social networks. However, our goal was to analyze comments and influence in an online newspaper and not in Facebook or Twitter. We collected the following fields: user ID, date, news ID, total of comments by news and comments.

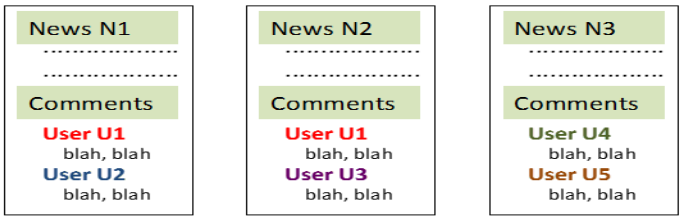
3.2. Building Social Networks

Here we propose a new method for building an implicit social network between users who post comments about news in a newspaper. The idea is that users who frequently interact by commenting the same news are in some way related to each other and form a social network. The more users comment the same piece of news, the more intense is their relationship. Therefore, we create a link between two users whenever they post comments in the same news. Please notice that we do not differentiate between commenting the same news and commenting each other comments, because users use both options interchangeably.

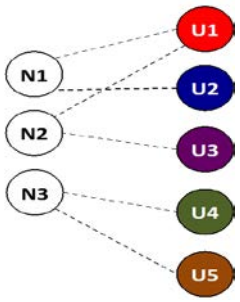
The process of creating implicit social networks from news, users and comments uses a bipartite graph G=(U,V,E) where U denotes the set of users, V the set of news and E the set of edges between disjoint sets V and U. As emphasized by Newman [15], some social networks actually are formed by affiliation when individuals join common group memberships. Therefore they are naturally modeled by bipartite graphs, as individuals and groups are represented by two different vertex types and edges between them represent group membership. In our case, whenever users post comments in a piece of news, they become members of that news group. This method creates a bipartite network that is used for building the implicit social network afterwards (also called one-mode projection [15]).

Figure 1 illustrates the process of generating implicit networks out of news, users and comments. Let us assume users U1 and U2 wrote comments about news N1, users U1 and U3 wrote comments about news N2 and users U4 and U5 wrote comments about news N3 (Figure 2a). In our method, first we generate a bipartite graph for the entire set of news and users (Figure 2b). Next, we generate links between users, i.e., user U1 is

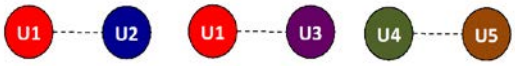
linked to user U2 via N1, user U1 is linked to user U3 via N2 and user U4 is linked to user U5 via N3 (Figure 2c). After analyzing all comments and creating all links, our method creates two separate networks, one connecting users U1, U2 and U3 and the other connecting users U4 and U5 (Figure 2d). This method aims at creating a single connected network, but depending on the user behavior some smaller disconnected networks may appear.



a) News, comments and users



b) Creating a bipartite graph



c) Creating links between users

d) Creating networks

Figure 1 – Building Social Networks

3.3. Exploratory Data Analysis

After the implicit social networks are created, we analyze our new expanded dataset by observing some key metrics:

- Number of comments posted by each user
- Number of news each user posted comments
- Number of implicit networks
- User degree in implicit networks (connectivity level)
- Number of comments posted for each news
- Number of users who posted comments in each news

The EDA step plays an important role here, helping us to better understand our dataset and eventually to remove outliers.

3.4. Selecting Star Users and Groups

Key to our methodology is to identify users with high potential to exert influence on other users, called Star Users. Although a variety of characteristics may be used to find Star Users, for the scope of our research we considered three classes:

- Star-M: users with the highest number of comments, measuring their absolute performance when it comes to commenting news or participation level;
- Star-N: users who posted comments to the highest number of news, measuring their interest spread;
- Start-D: users with the highest degree, i.e. who posted comments to news that attracted a high number of other users, measuring their socialization level.

Eventually, classes Star-M, Star-N and Start-D may select the same user or three different users, depending on the newspaper section. Whenever a set of keywords a star user posts overlaps the set of keywords posted by other groups of users in the same implicit social network, we assume the former has influence over the latter, i.e., we say that this user group is under the influence zone of the star user. The comments we collected are divided into three groups:

- Group-Star: comments posted by the star users;
- Group-First: comments posted by all users directly connected to star users, i.e. their first level neighbors;
- Group-Other: comments posted by all other users who do not belong to Group-Star and Group-First.

Our method of identifying influential users is based on comparing the words of Group-Star to the words of Group-First and Group-Other. The idea is that if there is a high overlap level between Group-Star and Group-First, it may mean that star users influence their first level neighbors. On the other hand, if there is a significant overlap between words of Group-Star and Group-Other it may mean that the spread of influence of star users goes beyond their first level. As this is our basic assumption, it may not be always true. However, this method yield results, as we advocate, confirming that star users have influence over other users.

Figure 2 illustrates a distribution of information into three star users and consequently into nine different groups for the World newspaper section.

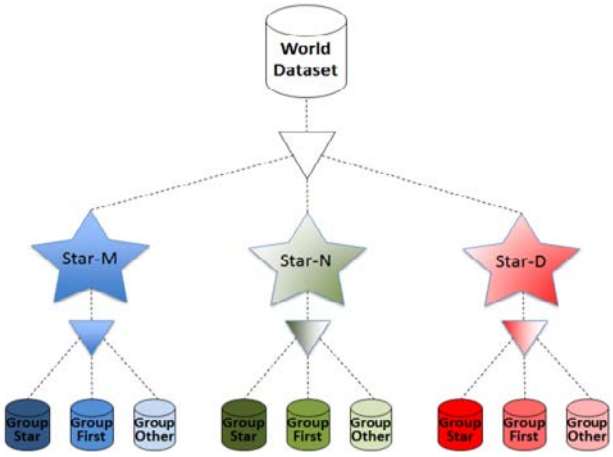


Figure 2 – Distribution of data into groups by section

Please notice that this arrangement means that we will end up with a range of information sets that may vary from 15 (5 sections x 1 star user x 3 groups) to 45 (5 sections x 3 start users x 3 groups). As mentioned before star users Star-M, Star-N and Star-D may be three different users at one extreme or they can be the same user in the other extreme.

A. Text Mining and Discovery of Clusters

Our methodology relies on well-known technique for text mining: tokenizing, stop list, stemming, creation of a bag of words and matrix of terms and TF-IDF. This method starts by reading the datasets and sequentially applying each of the text processing steps. We also remove terms with less than two characters, for reducing the size of the matrix of terms. The outcome of this step is a set of matrices of terms (between 15 to 45) ready for feeding the machine learning algorithms.

We use K-means [16] for machine learning, applying it over each matrix of terms coming from the previous text mining step. This process is comprised of three stages, where the first one is identifying the best number of clusters (K) [10]. We used the silhouette coefficient for finding the number of clusters. The second stage applies K-Means with the correct K [8] and finally the last stage obtains the results of the clustering method. Each cluster needs a depuration, which means that we need to extract their most frequent words for generating the Main Words or the ideas of the cluster (also keywords). At this point, one should recall that the number of cluster used here is the number of groups, which vary from 3 to 9 for each section of the newspaper, depending on the star users. We also considered other techniques in preliminary experiments, such as Latent Dirichlet Allocation [4], but since they yielded similar results, we stuck with K-Means, due to its simplicity.

3.5. Finding Similarities

A final computational step is to take the main words and apply techniques for finding similarities between them, using Jaccard's index and the Dice's coefficient. Figure 3 depicts the process made in the last three steps, starting with the groups of information depicted by Figure 2, passing through the text mining and data mining steps and finally being processed by this step for finding similarities between different clusters of words. An additional non-computational step is the analysis and interpretation of the final results of the similarity metrics.

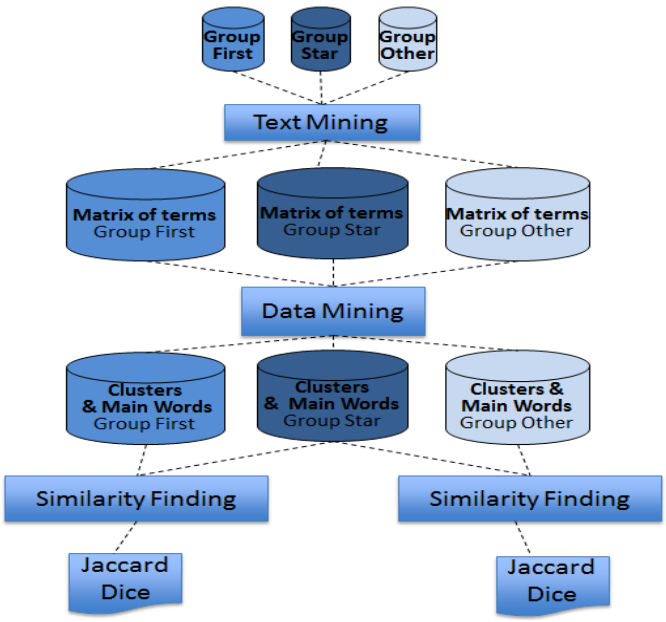


Figure 3 – Distribution of data into groups by section

Both Jaccard and Dice compare sets or strings and yield result values between 0 and 1 [11][14]. Jaccard compares words using exact matches, whereas Dice compares bigrams of each string of words. We use the results of Jaccard and Dice methods because they are two different techniques in order to obtain a clear decision for similarity. Jaccard is influenced by the number of words in a set, whereas Dice is flexible with the length of the string but is slightly less precise. Sometimes the results of the two techniques are similar, but other times they are completely different, which help us to understand better our results.

4. Results

4.1. Exploratory Data Analysis

The key results from the Exploratory Data Analysis come from Politics and Daily sections, the most popular ones. Figure 4 shows the histogram and ECDF for the Daily section. The Daily section was the most popular among users, maybe due to the bad news that always catch people's attention such as robbery, murder and poverty.

A total of 3,030 users posted 12,232 comments during the period measured, where 28 users commented more than 40 different pieces of news and 13 of them had more than one hundred comments.

As illustrated by Figure 4, more than 300 users have a degree higher than 181, which means that they commented news where this number of users also posted comments. This is the effect of 34 news that received comments from 51 different users. The whole Daily network has a clustering coefficient of 0.3, which means that it rendered a quite high triangulation of connections between users in the network.

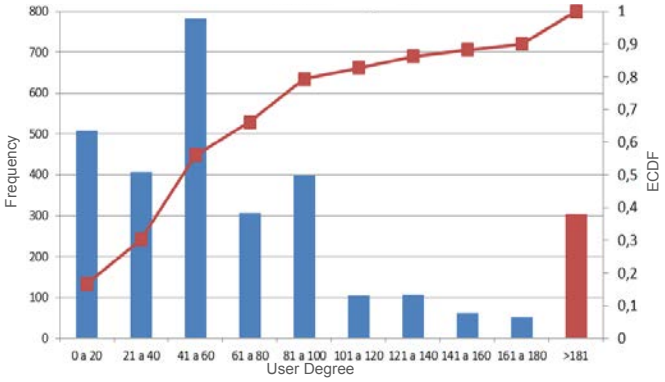


Figure 4 – Histogram/ECDF of user degree in the Daily section

On the other hand, the connectivity of the Politics section is smaller but much populated with a total of 2,560 participating users and 16,973 comments. Most news received between 0 and 10 comments, as shown by Figure 5.

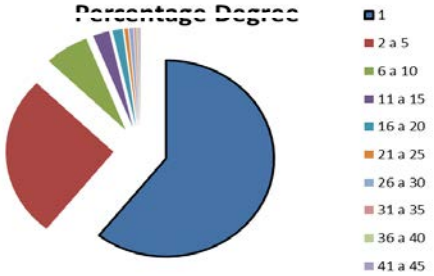


Figure 5 - Degree distribution of the Politics section

4.2. Clustering

Figure 6 presents the K-Means plot of a single star user (the same user was identified as Star-M, Start-N and Star-D) for the World section with clusters of the most common words. In this plot, we can observe a clear separation between two groups of words, where the first one in the left is larger with more ideas and documents, and the second group in the right contains less ideas and documents.

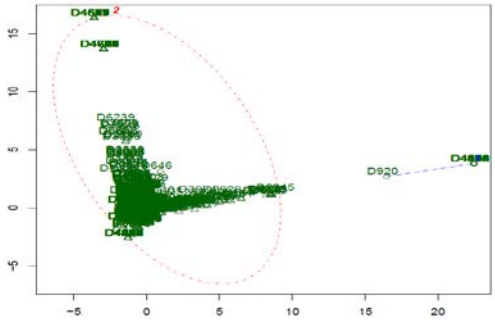


Figure 6 - K-Means plot for Group-Star – World Section

In addition, Table 1 presents a list of main words generated in the analysis of each cluster, for groups Star, First and Other. This is just an example to help understanding which words were mostly used, since the most significant results are generated by the similarity indexes. In this case, it is clear that most comments are related to war, which is quite reasonable.

For all five sections of the newspaper the results were similar to Figure 6 and Table 1. The purpose of Figure 7 is to highlight how results were obtained for Daily and Politics sections. These sections generated three Star Users (M, N and D) each, whereas for Technology, Education and World sections the three methods coincided in the same star user.

Table 1 - World section Main Words

Group	Cluster 1 Main words	Cluster 2 Main words
Other	Espionage,white,cryptology,security, service,software,stealth,companies, President	Brazil,government,Americans,military, president,Snowden,invade,trust,conflicts, nation,war,policy
First	Terror,American,military,technology, espionage,ronald,merkel,angel	Americans,fear,government,spying,speech, motives,Brazil,president,fidel,venezuela, canada,israel,Australia
Star	Religion,Planet,Terror,Americans,funding, tech,spy,software,flag	Church,Fidel,government,Islam,religion,abor tion,Catholic,Vatican,Venezuela,Brazil,Cuba n,war,Israel

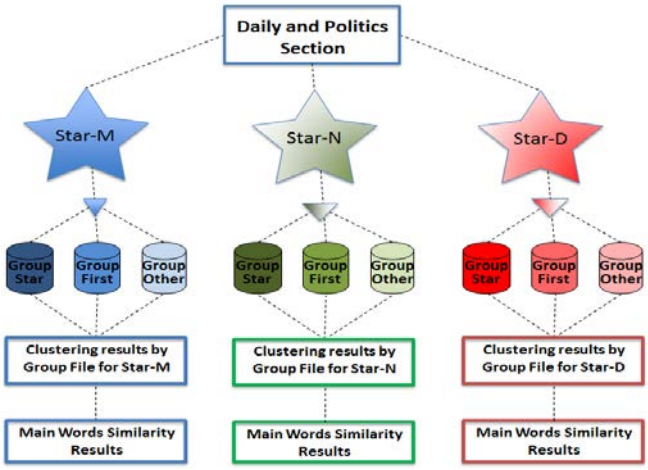


Figure 7 - Structure Results for Politics and Daily sections

4.3. Similarity results: Jaccard and Dice

Using the main words for the five sections we proceed to the next step and obtain the similarity indexes using Jaccard and Dice. In order to do this, we compare the main words (as a string) of all clusters of Group-Star with the main words of Group-First and Group-Other. For example, assuming Group-Star and Group-First have 2 clusters each, we make four cross comparisons. Next, we select the highest value of the similarity metrics among all clusters as our final result. Since clusters of words come from the same news, this procedure reveals whether star users effectively exerted influence over the remaining group of users.

In order to clarify the results, it is important to understand that Jaccard usually renders lower values than Dice, which does not mean that the former is worst than the latter. Jaccard is deeply affected by the length of the strings, so that we consider that a value higher than 0 means some similarity. On the other hand, Dice is not affected by the length of the string and therefore renders higher scores. In addition, sections Technology and Education did not generate convincing results due to the lower participation level, so that results are quite inconclusive, unfortunately.

Figure 8 presents encouraging results for the World section, where Jaccard and Dice produced scores of 0.31 and 0.65 respectively when compared the words from Group-Star to Group-First. It means that there is relevant similarity in the comments posted by the prospective influential users and their neighbors. Also, the results show that for Dice influence was spread from the Star user also beyond their first level neighbors, i.e., to Group Other.

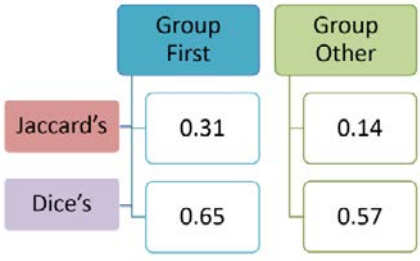
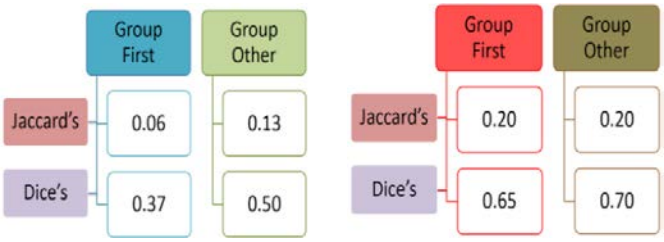
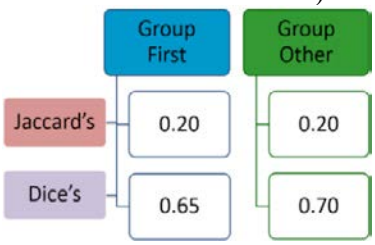


Figure 8 – Similarity results for the Star user – World section

Figure 9 presents the Jaccard and Dice results for the Daily section. Starting from Figure 10a, for Star-D, one may observe that its results are the lowest from this section, which is significant because it means that the highest degree user is not as influential a 4.4.s the users who posted the highest number of messages (Star-M) or commented the highest number of news (Star-N). For Star-M we have excellent results of 0.29 for Jaccard and 0.65 for Dice. For Star-N we have the best results from the Daily section, with 0.20 for Jaccard 0.71 for Dice. We observed something peculiar in this network, because the overlap of words coming from the Star users is more similar to Group-Other than to Group-First.



a) Star-D b) Start-N

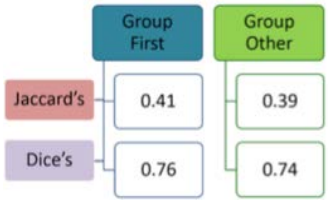


c) Star-M

Figure 9 - Similarity results for Star users – Daily section

Figure 10 presents results for the Politics section, which are similar to those of the Daily section presented in Figure 9, where Star-D users have the lowest similarity scores and Star-M and Star-N the highest ones. It should be noticed that Star-D has a higher Jaccard score for Group-Other than for Group-First, which means that their influence is spreading into the network. The highest scores for the Politics section are for the Star-M user, although close to the similarity for Star-N so that it is not possible to state which one is more significant.

a) Star-D b) Start-N



c) Star-M

Figure 10 - Similarity results for Star users – Politics section

5. Discussion

The key contribution of this paper is an innovative methodology for identifying and measuring influence of users who comment news in an online newspaper. Based on our results, we can observe that the methodology works because the similarity metrics present high scores. For example, the Daily and Politics sections, the most popular ones, present high scores for the similarity metrics, such as 0.76 for Dice and 0.41 for Jaccard. In other words, the ideas concealed by the Star user through their comments influence other users such as Group-First and Group-Other. Also, we observed that commenting the highest number of news (Star-N) and posting the highest number of comments (Star-M) generates more influence than having the highest degree (Star-D).

As an initial and innovative methodology, there is plenty of room for improving and adapting it for being more precise in identifying influential users. For example, in this paper we assume that whenever two users commented the same piece of news, there is a connection between them. An interesting and simple variation is to consider that an edge in the implicit network will only be created when two users commented together a number of n pieces of news, where $n \geq 1$. For $n > 1$ the resulting network will be less connected, i.e., the clustering coefficient will be lower, but the relationship between those users will be stronger.

The key idea is to be able to pinpoint more precisely the influential users. Our methodology has six steps but we can identify three main components that make it appropriate for what it is aimed at. Firstly, the building of implicit networks based on participation on an environment that is not officially an online social network. Secondly, the three methods for identifying influential users, which are those who posted comments in news together with the highest number of other users (Star-D), those who posted the highest number of comments (Star-M) and those who posted comments in the highest number of news (Star-N). Of course different methods may be applied, such as the well-known centrality metrics. And thirdly, a method for measuring whether there is an overlap in the comments posted by the prospective influential users and the other remaining users, using text mining, data mining and similarity indexes.

Finally, we consider that influence is evidenced by the number of similar words used by users in their comments. However, we did not explore the cases when all words forming a cluster coincide with those of the news itself. In that case, the evidence might be less strong, since users might be just repeating the same words. In other words, we do not distinguish correlation between words from causality, i.e., one user influencing others, which we considered a derived open research problem.

6. Conclusion

This paper proposed and evaluated the effectiveness of a methodology for identifying influential users, called Star Users, who comment news in an online newspaper. This methodology builds implicit social networks upon users who post comments in the same news, propose three different ways of identifying influential users and measure influence based on an overlap between comments of these users.

Our methodology was validated by collecting news, users and comments for three months of five sections of the Brazilian Folha de São Paulo online newspaper. Results show that there is effectively a high overlap between comments of influential users and

the remaining users, up to scores of 0.76 for Dice's coefficient and 0.41 for Jaccard's index.

As future work we intend to confirm the influence of Star Users by observing their behavior over time and to confirm the effectiveness of our methodology applying it to other datasets.

References

[1] Adamic, L. A., Adar, E., "How to Search a Social Network", Social Networks, 27(3), July 2005, p 187–203, 2003.

[2] Bakshy, E., Hofman, J., Mason, W., Watts, D., "Everyone's an Influencer: Quantifying Influence on Twitter", ACM WSDM'11, pp. 65-74, February 2011.

[3] Bampo, M, Ewing. M. T., Mather, D. R., Stewart, D. e Wallace, M., "The Effects of the Social Structure of Digital Networks on Viral Marketing Performance", Information Systems Research, Vol. 19, No. 3, p. 273-290, 2008.

[4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

[5] Castells, M. "The Information Age: Economy, Society and Culture", Oxford, 1998.

[6] Centola, D., "The Spread of Behavior in an Online Social Network Experiment", Science 329, 1194, 2010.

[7] Chen, W., Wang, Y., Yang, S., "Efficient influence maximization in social networks", KDD '09, pp. 199-208

[8] Chiang, M. M. T., & Mirkin, B., "Intelligent choice of the number of clusters in K-Means clustering: an experimental study with different cluster spreads", Journal of classification, 27(1), 3-40, 2010.

[9] Diakopoulos, N., Naaman, M., "Towards Quality Discourse in Online News Comments", ACM 2011 Conference on Computer Supported Cooperative Work (CSCW 2011), pp. 133-142, March 2011.

[10] Frahling, G., & Sohler, C., "A fast k-Means implementation using coresets", International Journal of Computational Geometry & Applications,18(06), 605-625, 2008.

[11] Huang, A., "Similarity measures for text document clustering", 6th New Zealand computer science conference, pp. 49-56, 2008.

[12] Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. Journal of Marketing Research, 48(3), 425-443.

[13] Katz J. "Here Come the weblogs", Slashdot.org, 1999.

[14] Majumder, P., Mitra, M., & Chaudhuri, B. B., "N-gram: a language independent approach to IR and NLP", ICUKL 2002, November 2002.

[15] Newman, M. E., "The Structure and Function of Complex Networks", SIAM Rev., 45(2), 167–256, 2003.

[16] Teknomo, K., "K-Means clustering tutorial", Medicine, 100(4), 3, 2006.

[17] Weber, P, "Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments", New Media & Society 16(6), pp. 941-957, August 2013.

[18] Ye, Shaozhi, Wu,S. F., "Measuring Message Propagation and Social Influence on Twitter.com", 2nd Intl Conference on Social Informatics (SocInfo'10), pp. 216-231, October 2010.