A Text Analyser of Crowdsourced Online Sources for Knowledge Discovery

Ioannis Markou
Information and Communication Systems Engineering
University of the Aegean, Samos, Greece
e-mail: janis.markou@gmail.com

Efi Papatheocharous
Swedish Institute of Computer Science (SICS)
Kista, Stockholm, Sweden
e-mail: efi.papatheocharous@sics.se

Abstract—**In the last few years, Twitter has become the centre of crowdsourced-generated content. Numerous tools exist to analyse its content to lead to knowledge discovery. However, most of them focus solely on the content and ignore user features. Selecting and analysing user features such as user activity and relationships lead to the discovery of authorities and user communities. Such a discovery can provide an additional perspective to crowdsourced data and increase understanding of the evolution of the trends for a given topic. This work addresses the problem by introducing a dedicated software tool developed, the Text Analyser of Crowdsourced Online Sources (TACOS). TACOS is a social relationship search tool that given a search term, analyses user features and discovers authorities and user communities for that term. For knowledge representation, it visualises the output in a graph, for increased readability. In order to show the applicability of TACOS, we have chosen a real example and aimed through two case studies to discover and analyse a specific type of user communities.**

Keywords-User communities; Authorities; Social Network Analysis.

I. INTRODUCTION

Over the years, micro-blogging platforms have become a popular means of exchanging current crowdsourced information. Twitter, counting over 500 million tweets sent per day [1], holds a large share of that information traffic. One of Twitter's success factors is that the exchanged information is highly concentrated, as a tweet is limited to 140 characters in length. The unstructured nature of that information has led to the development of numerous content analysis methods. Such methods can be applied on tweet datasets to perform various tasks such as opinion mining and topic extraction. These tasks can be useful for various reasons, from discovering users' movie opinions to predicting voting outcomes [2].

Many of the available content analysis methods are quite accurate. However, they cannot evaluate comprehensively the credibility of the analysed information. Twitter, like all micro-blogging platforms, is challenged by the credibility of the information that is being exchanged [3]. In this work, we propose that the key to evaluating content lies in Twitter's second success factor, the open access to information. In particular, a Twitter user can access other users' feeds just by following them, requiring no approval from the user being followed. Additionally, a user can comment, like, reply and mention other users without the two-way friendship feature found in other social networks. This one-way relationship is the foundation of Twitter's rapid spread of information and it forms a unique way of evaluating a user's content by the Twitter community. As a result, users with higher evaluation have created more credible content. Consequently, it is imperative in order to discover credible information, to first focus on who creates content and then on what the content is about [4].

Most approaches analysing Twitter focus on the content rather than the users who create the content [5]. Even the few approaches that analyse user data [6]-[11], focus on identifying information about separate users and lack to provide information about the relationships between users. The importance of having relationship data lies in the need to understand people interactions and group effects over the Internet. Furthermore, by tracking relationships of authority users (i.e., users that influence the content and type of information spread), additional credible users can be discovered. Finally, relationship data can discover user communities, in which users share some features or communal goals.

Conclusively, analysing user data has become paramount to knowledge discovery, whether it targets building recommendation engines, marketing campaigns to specific audiences, predicting user trends or understanding buyers' behaviour. Based on the above mentioned reasoning, it is apparent that there is a need for approaches that are capable of complementing the current ways of analysing Twitter data in terms of content, by focusing on users and their relationships. In this work, we have targeted to address this gap and have developed an approach implemented in a software tool (named TACOS for Text Analyser of Crowdsourced Online Sources) that extracts user attributes from tweets and evaluates them to structure user and relationship data. We explain our approach and show the applicability of the tool developed through two explorative use cases.

The rest of this paper is structured as follows: in Section II, the related work is presented. In Section III, the steps taken in order to design and implement the TACOS tool are described. In Section IV, our approach is analysed in detail. In Section V, the tool is validated against two use cases and in Section VI, the results are discussed. Finally, in Section VII, some conclusions are drawn and future directions are suggested.

II. RELATED WORK

Numerous Twitter analysis tools have been around since the popular micro-blogging platform was founded. Even though their implementations provide several advantages, they come with some limitations (described in Table I). The

 ISBN: 978-1-61208-486-2

lower part of the table contains three services that are recent approaches focusing on user attributes rather than the content of tweets. A significant benefit for targeting the analysis of Twitter users and their attributes is that it can offer insights on 'authorities' on a given topic as well as help discover user communities that are related to the topic.

In summary, all of these tools and services lack fundamental functionality related to the users, such as locating the most influential users, visualising their relations and community connections that could enable for example targeted advertising for businesses. These functionalities, can offer insights beyond who-follows-who and number of favourites. They can highlight users and user communities in a particular domain, by also pinpointing the closest users that authorities interact with. Such information can help identify reliable sources (i.e., authorities) that generate information related to a particular topic. The effect of acquiring this knowledge is particularly important, both for popular and not so popular topics. For not popular topics, the detection of even a single influential user is as valuable as finding numerous influential users for popular topics with thousands of daily generated tweets. The suggested approach, described in the rest of this paper, covers this limitation from the existing implementations and visualises the retrieved, analysed user data in user-relationship graphs, where authorities and user communities are easily distinguishable.

III. APPROACH

A. Requirements Collection

At the early stages of the project, we conducted a set of interviews with researchers and industrial practitioners. In the interviews, a total of 5 people were questioned about their perceived possible usage and usefulness of the developed tool. One of the interviewees was female and the rest were male. We performed two structured interviews with the two researchers and three semi-structured interviews with the three industrial practitioners. The structured interviews lasted for about an hour each and included open-ended questions, dichotomous questions as well as Likert questions. The semi structured interviews included an open discussion with practitioners in a small-to-medium start-up company working in social network analysis. The discussions took place during one of the authors' ex-job placement in the company and the interviewees were working in the field of linguistics for several years.

Both types of interviews gave a different flavour of opinions which served as valuable input to the requirements for the system developed. Researchers expressed an interest in detecting the users that post content related to a specific research domain. In Twitter, the homophily principle is observed [12], so discovering relationships between users for a specific domain could help researchers expand their contacts on that domain. Industrial practitioners stressed that customers were more interested in users that create trends rather than the actual trends. These observations directed our efforts in defining the requirements for the solution proposed, as well as understanding the arising challenges.

B. Challenges

Such challenges concerned mainly the process of structuring and analysing user data [13]. At first, 'users' in Twitter are abstract entities, since users might be individuals, groups or organisations. Additionally, according to a user's posted content and activity, the user can be considered, among others, an 'authority', a 'topic expert' or a 'spammer'. Another challenge is analysing, modelling, interpreting and quantifying abstract social phenomena such as 'authority', 'domain expertise' and 'influence' [14]. The challenge lies in defining the appropriate classifiers for labelling a user as an authority or a domain expert. A last challenge is that user analysis alone is not enough to provide actionable information to an end-user. In order to provide insights regarding users and their relationships, information needs to be represented in an intuitive manner. This can be achieved by creating visualisations of the analysed user data.

TABLE I. MOST POPULAR NON-COMMERCIAL TWITTER USER ANALYSIS TOOLS AND SERVICES

Name	Description	Limitation
Nokia Internet Pulse [6]	Detects the most popular words for a topic in Twitter and visualises them in word-clouds. Word-clouds can be used to find popular users.	Does not show relationships between users or user-communities. Optimised for Nokia-specific keywords, which can lead to bias.
CO GNOS [7]	Locates topic experts by analysing user generated lists. Improves upon Who To Follow ([9]) by focusing on all users related to a topic.	Ignores other relevant users on a subject. Does not analyse relationship attributes such as mentions, retweets and replies. Does not offer visualisation.
Twitter rank [8]	Measures users' influence for a given topic. First applies topic modelling. Then it analyses users' followers and friends lists to create relationship networks for each topic.	Uses only followers and friends attributes and ignores other relationship attributes such as favourites, mentions and replies. Does not offer visualisation.
Twitter Who to Follow (TWF) [9]	Detects suggested users to follow by analysing user-provided attributes such as e-mail, contacts and location.	Ignores topics and attributes such as followers, mentions and replies to suggest users.
Tweet Reach [10]	Analyses tweets relevant to a search term. It supports statistics for tweets' impressions as well as distribution of tweets through time and percentage of replies and retweets.	Does not provide a high level of user analysis, besides detecting top contributors for a term.
Tweet chup [11]	Analyses user, connections, keywords and hashtags. Offers a high level of detail to improve engagement between users.	User engagement is limited to retweets and mentions. Does not offer information on communities of users for a given search term.

C. Suggested Solution

In this work, we have considered and targeted to address all of these challenges mentioned, and we defined and quantified the abstract concepts (i.e., authority, domain

expertise and influence) by assigning a set of classifiers to them. In order to accurately quantify these concepts, weights have been used to assign the contribution of each values used to the estimation of the classifiers. These weights are based on the importance of each contributing factor in the equations. Classifiers were created by studying the different tweet and user attributes. By understanding what each attribute represents and how it is used by the users, we were able to create the formulae for each classifier. Weights were assigned to the classifiers by applying a classification algorithm to retrieved datasets. The confidence of the predictions was taken into account in order to assess weights to the selected classifiers. By doing so, we were able to determine which of the classifiers were most important for classifying a user as influential.

Prior to the analysis of the equations, it is important to define tweet types. Tweet types include *original*, *retweets* and *replies*. Original tweets are authored by the user that posted them. Retweets refer to tweets that are forwarded by a user that did not create the original content. Lastly, replies are tweets that their content refer to other tweet's content.

In detail, the influence score (*i*) shows the degree in which a user's tweets make other users interact with the content. It is calculated by:

$$i = 0.5 * a + 0.5 * de. \quad (1)$$

In (1), *i* is influence, *a* is *authority* and *de* is *domain expertise*. We define authority as the degree that a user posts original content that is shared by a large audience. It is calculated by:

$$a = 0.05*ps+0.35*rr+0.35*orr+0.05*pprr+0.2*vs. \quad (2)$$

To calculate (2), equations (3) – (6) were defined.

$$ps = (followers - friends)/\max(followers, friends), \quad (3)$$

$$rr=(authT - nonAuthT)/\max(authT, nonAuthT), \quad (4)$$

$$orr = (original - retweets) / \max(original, retweets), \quad (5)$$

$$pprr= (puR - prR) / \max(puR, prR). \quad (6)$$

In (3), *ps* is a user's popularity score and it is based on the observation that popular users have disproportionate number of followers and friends, with friends being a lot fewer than followers. *followers* is the number of the user's followers and *friends* is the number of the user's friends (i.e., the users that the user follows). In (4), *rr* is the retweet ratio of the user (i.e., the user's relevant-to-the-topic tweets that are retweeted). *authT* is the number of the user's authority tweets (i.e., tweets that are retweeted by other users) and *nonAuthT* is the number of non-authority tweets of the user. In (5), *orr* is the original to retweet ratio of the user and we defined it as a metric for tweets originality. *original* is the number of tweets that the user posted and *retweets* is the number of the tweets that user retweeted from other user profiles. Finally, in (6), *pprr* refers to the public replies (*puR*) to private replies (*prR*) ratio (i.e., the replies that a user made and are viewed by anyone following either one of these two users). Based on our observation, many authority users reply publicly by adding a '.' symbol before they mention the username that they are replying to. As a result, this metric takes that behaviour into consideration for the influence score. Last but not least, *vs* is the verification status of a user and shows if the user is verified by Twitter. If yes, *vs* = 1, otherwise *vs* = 0.

Domain expertise (*de*) is defined as the degree that a user is involved in a topic as well as the quality of the content that the user shares. It is calculated by:

$$de=0.15*aes+0.1*rd+0.2*mr+0.25*tc+0.2*cq+0.1*ud. \quad (7)$$

To calculate (7), equations (8) – (16) were defined.

$$aes = 0.2 * rp + 0.8 * cp, \quad (8)$$

$$rp = (replies - repliesR) / \max(replies, repliesR), \quad (9)$$

$$cp= (convR - convNonR)/\max(convR, convNonR), \quad (10)$$

$$rd=times\_retweeted/retweeters*tweets, \quad (11)$$

$$mr = total\_mentions / num\_of\_relevant\_tweets, \quad (12)$$

$$tc = user\_associated\_tweets / relevant\_tweets, \quad (13)$$

$$cq = \sum_{i=1}^{k} favouritesNum_k * turtnr, \quad (14)$$

$$ud = total\_user\_relevant\_tweets / total\_user\_tweets, \quad (15)$$

$$total\_U\_relevant\_T = U\_DB\_relevant\_T\_before\_retrieval + relevant\_retrieved\_tweets \quad (16)$$

In (8), *aes* refers to audience engagement score and shows the degree that a user responds to conversations. In (9), *rp* refers to the replies participation of a user which is defined by the ratio of replies (*replies*) and replies received (*repliesR*). In (10), *cp* refers to the conversation participation of a user. It is defined as the ratio of conversations replied (*convR*) and conversations non replied (*convNonR*). Equation (11) calculates a user's retweets dedication (*rd*) and shows the degree in which users' retweets are retweeted by all users. In (12), *mr* is the mentions rate of a user and represents how often a user is mentioned. The *num_of_relevant_tweets* includes only original tweets. This means that it does not include retweets, as these are taken into account in other metrics. In (13), *tc* stands for topic contribution and represents the activity of a user for a particular topic for a single retrieval. Every day, a single retrieval is performed for each topic. By using this metric, a user's activity for a particular topic can be monitored through time. The variable *user_associated_tweets* refers to the total number of a user's relevant tweets, plus retweets and tweets retweeted by other users. In (14), *cq* refers to the content quality of a user's relevant to a topic tweets. *turtnr* refers to the number of a user's relevant tweets, without including retweets. In (15), *ud* reflects a user's dedication by

 ISBN: 978-1-61208-486-2

calculating how many of the user's total tweets are related to the particular term. *total_user_tweets* can be found in each user's attributes. Last but not least, in (16), *T* stands for tweet, *U* for user and *DB* for database.

The implemented tool (named TACOS for Text Analyser of Crowdsourced Online Sources) uses (1) – (16) to calculate these classifiers based on extracted user attributes from tweets. It then evaluates users in terms of authority, domain expertise and influence. The final influence score (*i*) is calculated using (1). Moreover, TACOS uses activity attributes to detect relationships between analysed users and evaluate them in terms of interactivity. By detecting relationships, user communities are discovered. The relationship score (*rs*) between two users is calculated as:

$$rs = max(A\_B\_Score, B\_A\_Score). \quad (17)$$

A_B_Score shows the degree in which user *A* interacts with user *B* and it is calculated by:

$$A\_B\_Score = 0.3 * A\_Mentions\_B + 0.3 * A\_Replies\_B + 0.05 * A\_Retweets\_B + 0.15 * A\_Favourites\_B + 0.2 * A\_Follows\_B. \quad (18)$$

A_Mentions_B shows the times that user *A* mentioned user *B* in relevant tweets and so on. Respectively, *B_A_Score* is similar with scores reflecting user B's activity. For example *B_Mentions_A* shows how many times user *B* mentioned user A in his/her tweets.

Interaction score *is* shows the degree that both users interact with each other. Let's assume that $A = A\_B\_Score$ and $B = B\_A\_Score$. Then, if A and B are 0, then $is = 0$. Otherwise,

$$is = max(A, B) - min(A, B) / max(A, B). \quad (19)$$

Finally, to offer intuitive analysis reports, TACOS presents results in the form of graph visualisations that show influential users and their communities. In that visualisation, nodes represent users and edges represent relationships between users. Influence score (*i*) is represented by the size of the node, with larger nodes belonging to more influential users. Authority and domain expertise are not apparent at first glance, but are available by clicking on a node, together with other information. The weight of an edge represents the relationship score between two users, with a thicker edge showing high activity, at least from one of the users towards the other. If there is activity from both users, then the relationship is considered as interactive, and the edge is shown in blue colour to represent that. Last but not least, relationship graphs in Twitter are not necessarily two-way, since a user might follow another user but not being followed by the second user. For our graphs, we wanted to emphasize on the flow of information, according to the HITS – hubs and authorities algorithm [15] [16]. For that reason, edges in our graphs don't show following status but influence, so the edges point towards the more influential between two users.

IV. System Description

In this section, our approach's system design is described. TACOS consists of 7 modules. In Figure 1, our approach's system design is illustrated. The front end interacting with the user includes the Query Validator (QV) and Graph Visualisation (GV) modules. The QV consumes the user's input as a query. The query can consist of one or more words, it may contain hashtags, at-signs and special symbols to limit the search results, following the rules of the Twitter Search API [17]. The GV module is responsible for the graph visualisation, i.e., the final result produced. The user can interact with the result by zooming in and out, panning and clicking on nodes and edges to reveal information about users and their relationships. The GV module is using JavaScript frameworks and so the graph has the same functionalities in desktops, smartphones and tablets.

Moving on to the back end, the Gavagai Lexicon Connector (GLC) module handles the transactions between the Gavagai Living Lexicon API and our tool. The Gavagai Living Lexicon [18] is a tool that finds semantically similar and associatively related terms for a given topic. These terms are then presented to the users where they can choose to include some, all or none of them to the retrieval process. The Data Retrieval (DR) module includes all methods responsible for the communication with the Twitter Search API in order to retrieve tweets and users from Twitter. The Data Analysis (DA) module is the core module of the tool and is responsible for analysing the output of the DR module. The DA module handles operations such as extracting features from the retrieved data and linking tweets to users. Moreover, it gathers additional information about each user concerning their activity and finally calculates the influence and relationship scores, which are the input of the GG module. The Graph Generator (GG) module is responsible for preparing analysed data for visualisation. The output of the GG module is the the input of the GV module.

The last module is the MongoDB no-SQL Database (MDB) which is responsible for handling transactions with the QV, DR, DA and GG modules. In detail, the database holds collections of tweet, user and search documents. Each search document contains the ids of the tweets and users that are associated with it, as well as sub-collections such as influence and relationship scores. After each search, new users and tweets are added to the respective collections and existing documents are updated. These updates also serve a purpose of reducing the amount of requests that our tool must make and thus, minimising the times where the request limit is exceeded.

 ISBN: 978-1-61208-486-2

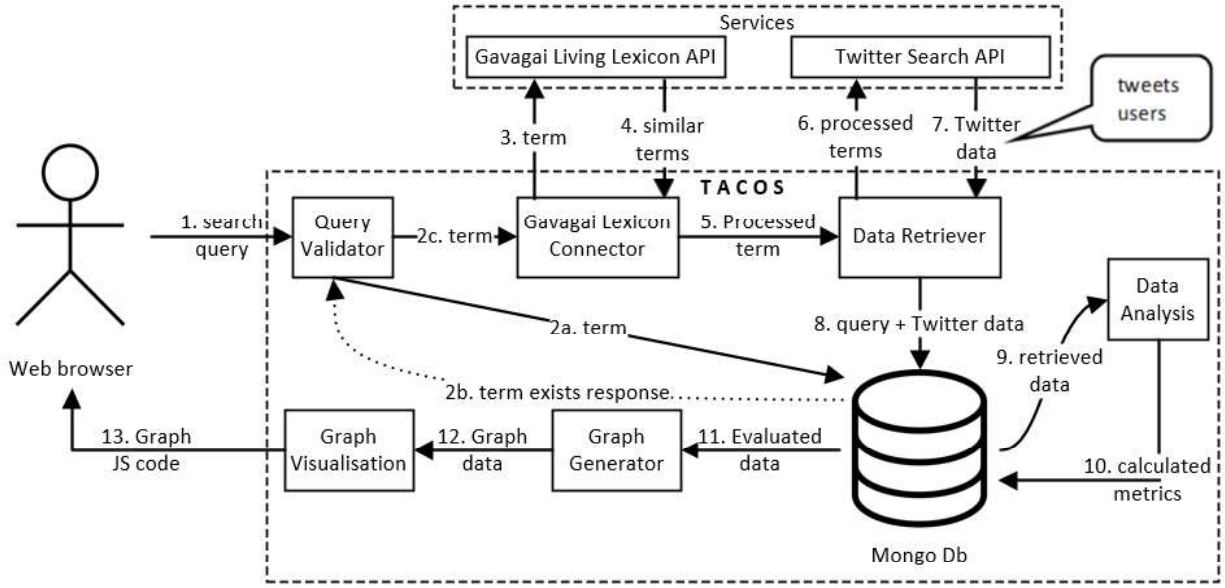

Figure 1. TACOS system design

V. VALIDATION

Due to the large amount of information posted on Twitter, it is challenging to aggregate tweet and user data for specific topics and events. In order to demonstrate our approach's ability to track the activity for specific events, we selected an event, the XP2015 conference on agile development practices, as an ideal scenario of use to validate our tool against, due to its manageable volume of data. With TACOS we retrieved about 500 users and 2100 tweets within a 25-day period, including the dates that the conference was held. We then analysed that data and detected that 'XPConf' was the most influential user of the event (something reasonable as it was the event organizer). Besides user activity, many trending terms linking to companies tweeting about using agile practices (Ericsson), blogs about agile development (42stc) and cities (Helsinki, which hosted the conference) were discovered. The term 'agile development' was one of the popular trends with a broader meaning. That motivated us to perform an additional analysis with the 'agile development' term. For that term, TACOS retrieved about 5000 users and 9500 tweets within a 25-day period. After analysing the retrieved data, the terms 'DevOps', 'cloud' and 'IoT' were discovered among the most popular trends. These results are valuable, since they describe the current state of agile development methodologies.

It is therefore evident from the above use case that our approach is performed well for following both specific and broad trends and understanding how trends are connected to through time.

Equally important to discovering popular trends is detecting influential users and relationships between users that post content related to those trends. Moreover, it is meaningful to present results in a readable way, such as graph visualisations, where user communities can be seen. In the generated graph, influential users and users with interactive relationships were easily distinguishable. By clicking on user nodes, additional information was available, including a link to the user's Twitter account. By visiting many accounts that our approach evaluated as influential, we realised that all of them had posted relevant content and that themselves were heavily involved in agile development practices and communities. As a result, our approach can successfully detect influential users and user communities for a given topic. At the same time, it can visualise that information in a readable and intuitive way.

VI. RESULTS AND DISCUSSION

Detailed results obtained from the use case example (described in the previous section) can be found in [19].

It is evident that our approach successfully identified top trending topics and users for both terms used for validation purposes. Moreover, it is important to highlight that our approach presents satisfactory results regardless of query type. As a scientific conference, '#XP2015' refers to a seasonal event – it has a narrow context, so this term is a navigational query – it seeks content of a single entity [20]. On the other hand, 'agile development' refers to a broader term and thus, it represents an informational query [20]. As indicated with the above use case, the type of the query affected both the dataset size and data type distribution in the results.

Specifically, regarding the dataset size, Figure 2 shows the amount of retrieved tweets for both terms for every day of retrieval. As expected, the first term ('#XP2015') refers to an event happening in a specific point in time, thus higher volume of tweets are retrieved around that time period. On the other hand, the second term ('agile development') is broader, and so the number of tweets retrieved is distributed more evenly across time. The total number of retrieved tweets

 ISBN: 978-1-61208-486-2

(9476) and users (5010) for the second term exceeds the respective numbers of the first (2106 and 496 respectively).

Data types can refer to tweets or users. Tweet types have been analysed in a previous chapter. User types include plain users, retweeters and domain experts. When a tweet is original but not retweeted, its user is classified as plain. Accordingly, when a tweet is original and retweeted, its user is classified as domain expert. Last but not least, if a user posts a non-original tweet, the user is classified as a retweeter. Figures 3 and 4 show tweet types (i.e., original, retweets and replies) and user types (i.e., plain users, retweeters and domain experts) distribution for both terms of the use case.

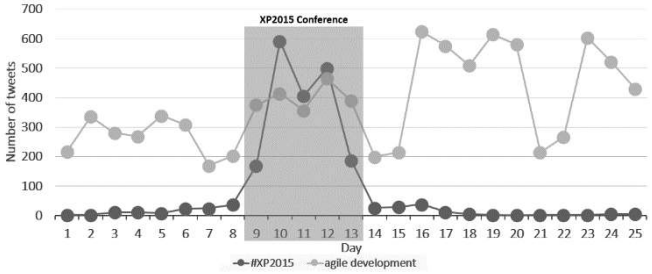

Figure 2. Retrieved tweets date distribution.

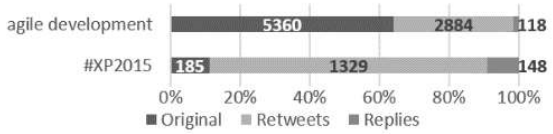

Figure 3. Retrieved tweets type distribution.

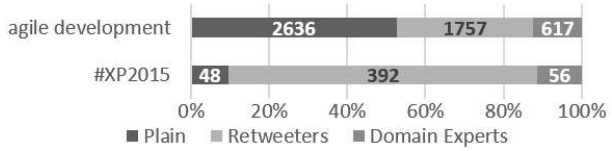

Figure 4. Retrieved users type distribution.

Seasonal events, such as the XP2015 conference, include fewer original tweets but a lot of retweets and replies. On the other hand, broader terms, such as 'agile development' include more original tweets and less retweets and replies. Consequently, from a user type perspective, seasonal terms include more retweeters and broad terms include more users posting original content. It can be therefore assumed that interaction between users is higher in seasonal terms than in broader ones. With our approach, the analysis and evaluation of tweets and users offered satisfactory results for both query types. Moreover, our approach visualises most popular users for the "#XP2015" term, in a clear way, as seen in Figure 5a. The produced graph is star-like, showing that all relationships have one user in common. However, this is rarely the case, especially for broad, not seasonal terms.

The second use case included the retrieval and analysis of data for the general term 'agile development' for a single day. We focused on a short timeframe in order to demonstrate our approach's ability to perform well, even when processing small datasets. In total, 100 tweets were retrieved and analysed, resulting in the evaluation of 125 users and the creation of 58 relationships between them.

For the produced graph, the most influential user is 'gclaps', as shown in Figure 5b. This user writes articles about agile development and start-ups. Several authority sources are also shown in the graph. Moreover, single-node and multi-node communities can be seen in the graph. Single-node communities are created because the analysis module checks for additional content for newly-retrieved users. Additionally, many of these single-node graphs represent new tweets that haven't been retweeted yet.

As can be observed, discovering influential users and their relationships with topic communities can be easily done with our approach. By using JavaScript, optional information is hidden for each node and edge, thus increasing graph readability considerably.

Regarding other data sources, our models can be modified to support other social networks as well. From the developer oriented StackOverflow to the topic-generic Reddit, it is possible to discover influential users by replacing Twitter attributes with the equivalent ones of each network and then assigning specific weights to them, based on the results of the classification algorithm. For LinkedIn the process should be easier since the attribute "Influencer" is already included in the social network's feature list.

VII. CONCLUSIONS

In this work, we described an approach for analysing Twitter data in order to model abstract social terms such as influence, authority and domain expertise, and apply that model to evaluate users and enhance knowledge discovery. We then validated our approach against two case studies that demonstrate the performance of our approach, regardless of the type of the search query, making it suitable for analysing Twitter data. Many approaches focus solely on tweets analysis or offer limited user data analysis features. Moreover, there is a need for proper visualisation of user communities in a way that can allow big datasets to be presented in a readable way. As it was demonstrated in our results section, our approach can successfully visualise users in terms of influence metrics so that user communities and relationships between users can be easily distinguished. Based on these results, we are confident that our approach can be used in many scenarios in industrial environments and academia. From gaining insights for a company's marketing campaign to complementing scientific material by supporting scoping studies [21], and other existing scientific research methods (such as mapping studies [22], literature reviews [23]). Scoping studies are concerned with contextualizing knowledge in terms of identifying the current state of understanding; identifying the sorts of things we know and do not know; and then setting this within policy and practice contexts.

We plan to continue carrying out work in this domain to further study the results and improve our models' accuracy. Moreover, we plan to conduct a user study in order to determine the best way to release our tool as an open-source web application. Lastly, we plan to target other audiences, like software developers [24], and thus, enhancing our retrieval and analysis modules to support a wider range of social networks, blogs and forums.

 ISBN: 978-1-61208-486-2

Figure 5. (a) Most popular users for the '#XP2015' term. (b) The most influential user for the second use case for the 'agile development' term.

REFERENCES

[1] S. Haustein, et al., "Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter." Journal of the Association for Information Science and Technology 67.1, 2016, pp. 232-238.

[2] A. Tumasjan, T. O. Sprenger, P. G. Sandner, I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment.", ICWSM, 10, May 2010, pp. 178-85.

[3] S. Ravikumar, R. Balakrishnan and S. Kambhampati, "Ranking tweets considering trust and relevance." In Proceedings of the Ninth International Workshop on Information Integration on the Web, ACM. Vancouver, 2012, May, p. 4.

[4] J. Brown, A. J. Broderick, N. Lee, "Word of mouth communication within online communities: Conceptualizing the online social network." Journal of interactive marketing, 21(3), 2007, pp. 2-20.

[5] My Top Tweet – https://mytoptweet.com, retrieved: May, 2016.

[6] J. J. Kaye, et al., "Nokia internet pulse: a long term deployment and iteration of a twitter visualization." In CHI'12 Extended Abstracts on Human Factors in Computing Systems, ACM, 2012, May, pp. 829-844.

[7] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, K. Gummadi, "Cognos: crowdsourcing search for topic experts in microblogs." In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM. Vancouver, 2012, August, pp. 575-590.

[8] J. Weng, E. P. Lim, J. Jiang, Q. He, "Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining. ACM, Vancouver, 2010, February, pp. 261-270.

[9] About Twitter's suggestions for who to follow | Twitter Help Center – https://support.twitter.com/articles/227220?lang=en, retrieved: May, 2016.

[10] How Far Did Your Tweets Travel? | TweetReach – https://tweetreach.com/, retrieved: May, 2016.

[11] Twitter analytic tool | Tweetchup – http://tweetchup.com/, retrieved: May, 2016.

[12] D. Stojanova, M. Ceci, A. Appice, S. Džeroski, "Network regression with predictive clustering trees." Data Mining and Knowledge Discovery, 25(2), 2012, pp. 378-413.

[13] R. Zafarani, M. A. Abbasi, H. Liu, "Social media mining: an introduction." Cambridge University Press, 2014.

[14] F. H. Khan, S. Bashir, U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision Support Systems, 57, pp. 245-257, 2014.

[15] F. Fouss, M. Saerens, J.M. Renders, "Links between Kleinberg's hubs and authorities, correspondence analysis, and Markov chains." In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, Vancouver, 2003, November, pp. 521-524.

[16] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM), 46(5), 1999, pp. 604-632.

[17] The Search API | Twitter Developers – https://dev.twitter.com/rest/public/search, retrieved: May, 2016.

[18] Gavagai Living Lexicon online - Gavagai – Next generation text analytics. – https://gavagai.se/blog/2015/05/01/gavagai-living-lexicon-online/, retrieved: May, 2016.

[19] TACOS – http://johnmarkou.com/tacos/agile_development/, retrieved: May, 2016.

[20] B. J. Jansen, D. L. Booth, A. Spink, "Determining the user intent of web search engine queries." In Proceedings of the 16th international conference on World Wide Web, ACM, 2007, May, pp. 1149-1150.

[21] H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework." International journal of social research methodology, 8(1), 2005, pp. 19-32.

[22] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, "Systematic mapping studies in software engineering." In 12th international conference on evaluation and assessment in software engineering, 2008, June, Vol. 17, No. 1, pp. 1-10.

[23] S. Keele, et al., "Guidelines for performing systematic literature reviews in software engineering." In Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.

[24] A. Zagalsky, O. Barzilay, A. Yehudai, "Example overflow: Using social media for code recommendation." In Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering, 2012, June, pp. 38-42, IEEE Press.

 ISBN: 978-1-61208-486-2