Discovering scholarly communication on Twitter through keyword searches

Kim Holmberg

k.holmberg@wlv.ac.uk
School of Technology, University of Wolverhampton
Wulfruna Street, Wolverhampton WV1 1LY, UK

Abstract

Social media are changing the way we communicate with our friends and our colleagues. In academia various social media sites are being used for various purposes. One of these social media sites that researchers are increasingly using is Twitter. Twitter enables rapid sending of short messages to a group of followers, and hence Twitter could be an efficient tool in sharing research information and in conversations about research work. The present study investigates how well keyword searches could be used to identify and collect scientific content on Twitter. Both qualitative and quantitative methods were used. The results indicate that keyword searches can be used to identify scientific content but how accurate keyword searches are is influenced by the specific vocabularies of different research areas. The present research also demonstrates methods to map and analyse content of tweets.

Introduction

Social media are changing the way we communicate with our friends and our colleagues. We use social media to share our lives but also to seek for answers and help from our online social networks. This change is also happening in academia as more and more researchers are using an increasing number of different social media for various purposes (Rowlands et al., 2011). Researchers are using social media to schedule meetings, share information they have discovered, find new information they can use in their work and to discuss their work with colleagues, to name a few things for which social media is used. One of these new technologies that is being used both for sharing and finding information is Twitter. Twitter is a micro-blogging service where messages can be up to 140 characters long. Twitter's short messages seem to be well suited for rapid information sharing and scholarly communication, however, the extent to which researchers in different disciplines are using Twitter and how exactly they are using it is still somewhat unclear. The present research will investigate how well keyword searches on Twitter can be used to identify and collect tweets with scientific content and to map the content of the tweets discovered.

Background

In 2013 Twitter has over 500 million registered users that are all using the microblogging service in various ways. Some share events from their personal lives while some companies and organisations use it as a marketing tool. Twitter is somewhat different from other social media partly because of the limit of characters it has but also because of the specific features and conventions of use it has. One specific feature is RT, which in a tweet indicates that the tweet has been forwarded or retweeted. With the use of the @-sign followed by a username a Twitter user can include another user into the message and send the message so that the user will be notified of it. The use of the @-sign has been discovered to almost complete indicate conversational aspects of Twitter use (Honeycutt & Herring, 2009). Shortened URLs are also frequently shared on Twitter (Suh et al., 2010). The URLs are usually automatically shortened by Twitter so that the URLs do not take up more characters of the 140 character limit that is necessary. Hashtags are used to group related tweets together and to make it possible for other users to find them. At for instance conferences and other events hashtags are frequently used to group the tweets related to that event together. A data collection on Twitter can use these features to focus the collection on different aspects of Twitter use or to tweets from specific events or topics.

Twitter has been studied from a wide range of topics, for instance as a tool for rapid broadcasting of news (Harlow & Johnson, 2011) and marketing information (Jansen et al., 2009), and for personal conversations (Honeycutt & Herring, 2009). In academia Twitter has become a frequently used tool for backchannel conversations at conferences (Ross et al., 2010; Weller et al., 2011). During conferences the audience can in a way extend the onsite conversations online and even people that are not attending the conference can follow the tweets and participate in the conversations. Other frequently used ways to use Twitter during conferences are for instance sharing of notes and links, or use of Twitter as a personal notebook. Other studies about scholarly use of Twitter have found some evidence of a relationship between tweets and citations (Eysenbach, 2011) and that there may be disciplinary differences in the way researchers from different disciplines use Twitter for scholarly communication (Holmberg & Thelwall, 2013), but many questions about scholarly use of Twitter still remain unanswered.

Research goals

The aim of this research is to investigate scientific content on Twitter, using carefully selected keywords that are representative to respective research areas in the data collection. We are interested to see how well discipline specific keywords can be used to discover scientific content and whether there are disciplinary differences in the type of tweets and in the level of scientific content of the collected tweets. We chose three research areas that all have different but specific vocabularies to the discipline. The disciplines chosen for this research are biochemistry, cognitive science and sociology.

Methods

The keywords were selected partly by studying keywords used in scientific articles from each discipline and partly by searching Twitter for conversations about the selected research areas. Finally we chose 27 keywords in cognitive science, 69 in biochemistry, and 15 in economics. These were used to automatically collect tweets between 4 March 2012 and 16 October 2012 using Twitter's API. A total of 25,934 tweets in the economics, 22,999 tweets in biochemistry, and 9,018 tweets in cognitive science were collected. Of these a random 100 tweets from each research area were selected for classification and more qualitative content analysis.

With a multifaceted classification scheme tweets were classified according to 1) type and 2) scientific content. In facet 1 the tweets were classified in following classes: *Retweets* (identified by RT or some other clear sign indicating that the tweet was forwarded), *Conversational* (by the convention of the @-sign followed by a username), *Links* (tweets that were not retweets or conversational tweets, but that contained one or more URLs), and *Other*, for all remaining tweets. In facet 2 the tweets were classified into three classes according to scientific content or lack thereof; *Science* (for tweets that clearly contained scientific information), *Not clear* (for tweets for which the scientific purpose or value could not be clearly identified), and *Not science* (tweets that were clearly not about science). Examples for each tweet type in facet 2 are listed in table 1 below. The example tweets in table 1 also demonstrate the three different types of tweets; retweets, conversational tweets, and tweets containing URLs.

Table 1. Classification of tweets in facet 2

Facet 2	Example of tweet
Science	Oxidative Stress Modulates the Nitric Oxide Defense Promoted by Escherichia coli Flavorubredoxin http://t.co/tQiapQ1l [*Article in Journal of Bacteriology, vol. 194, no. 14.*]
Not clear	@[…] yeah. physiology and biochemistry too... T.T
Not science	RT @[…]: "Real stupidity beats artificial intelligence every time." ? Terry Pratchett

Another approach taken to analyze the most popular content of the tweets was a co-mention analysis of the most frequently used meaning bearing words in the tweets. The co-mention maps were created with VOSviewer (Van Eck & Waltman, 2009) and visualized using Gephi (Bastian et al., 2009). The co-mention maps were then qualitatively analyzed for topical clusters in the content of the tweets. A threshold of about 80-90 occurrences was chosen as a suitable level to focus on the very core content of the collected tweets and to filter out noise in the form of unrelated or insignificant tweets. Other thresholds were also tested but lower thresholds resulted in too few words for them to maintain meaningful clusters and higher thresholds resulted in maps with a very high density from which it was difficult to recognize any clear visible communities or statistically calculated communities.

Results

In facet 1 the tweets from each research area were classified according to tweet type (figure 1). The results show that sharing links is especially important in cognitive science where 54% of the tweets contained links. It is also important to recognize that the retweets and the conversational tweets could contain URLs, and hence the total number of URLs in tweets is probably significantly higher. In biochemistry and economics about one third of the tweets contained URLs. Conversations were not very important in any discipline, although slightly more important in economics (14%) and biochemistry (16%) than in cognitive science (6%). In economics about one third of the tweets were retweets, while only about a quarter of the tweets in cognitive science and biochemistry were retweets. Between 14% and 25% of the tweets were classified to the *Other* category.

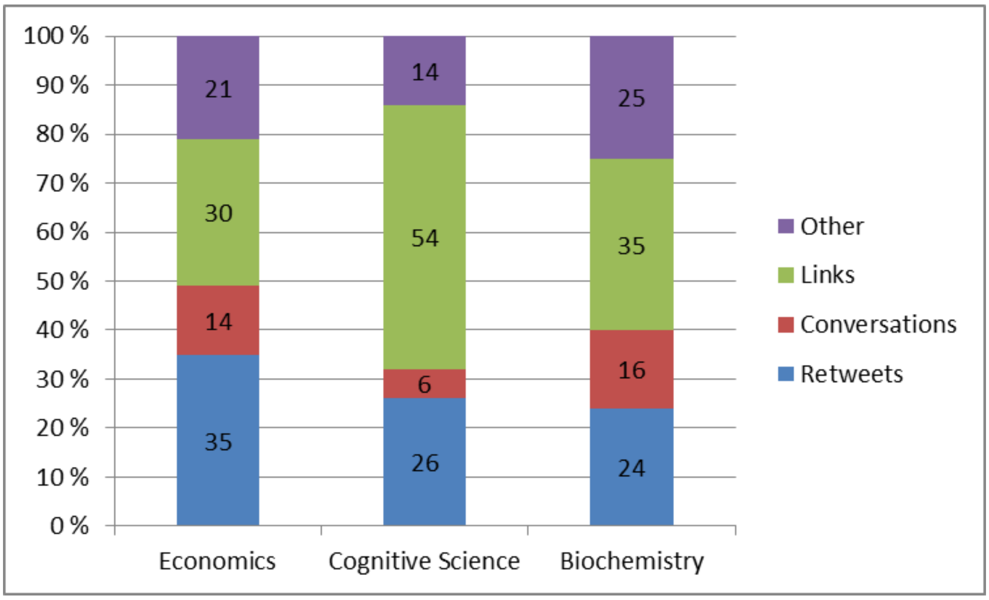

Figure 1. Type of tweets by discipline

In facet 2 the tweets were classified according to content (figure 2). Some evidence of scientific content or scientific information sharing was discovered in cognitive science tweets (23%) and somewhat less in biochemistry tweets (16%), but no evidence of scientific content could be discovered in economics tweets. It is however important to note that the number of unclear tweets is significant in both economics (45%) and in cognitive science (51%). A conservative approach (meaning that when in doubt, a less scientific class was chosen) was taken in the classification of the tweets to prevent exaggeration of the results, and hence some of the unclear tweets may in fact contain content of scientific value. At the same time the amount of tweets that were clearly not scientific was significant in both economics (55%) and in biochemistry (61%).

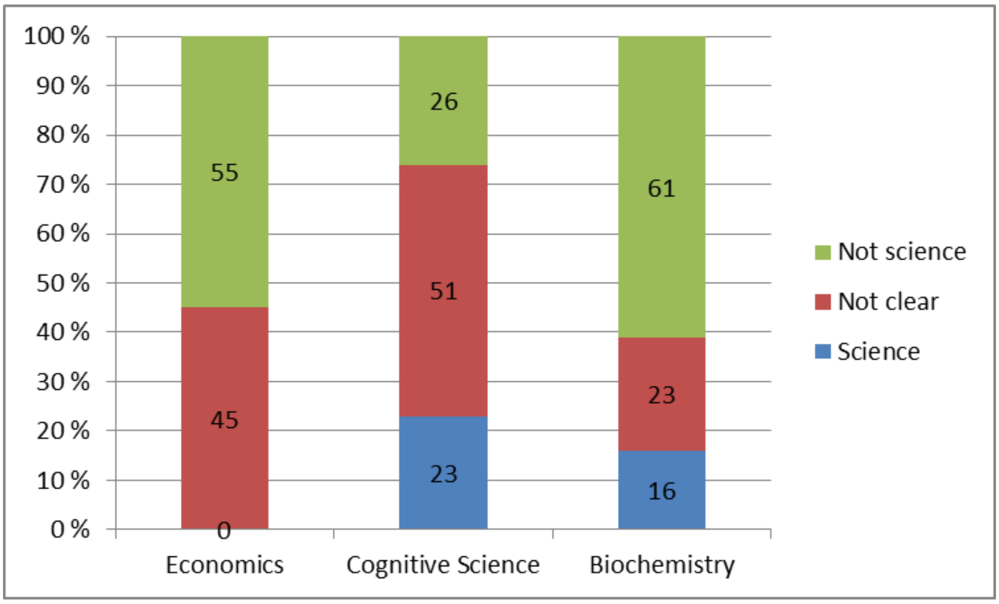

Figure 2. Scholarly communication by discipline and tweet type

A content analysis of the tweets showed that although the number of tweets containing scientific content was relatively low, a number of tweets in the *Not clear* category could have been about science or have had some scientific value. Some tweets clearly belonged to the respective research area but they contained links to articles in popular science magazines, newspapers or blogs, with now reference to scientific articles or scientific content. The scientific value of the content on these is hence unclear. In some cases the tweet contained too little information to make the judgment whether the tweet was part of scientific conversations and contained some scientific value or not. This is a problem when only fractions of the conversations can be collected. For instance, without the whole conversation it is not possible to tell whether the tweet: “*@[...] how so?*“ is scholarly communication or not.

The data was also analyzed using a co-mention analysis of the most frequently used meaning bearing words in the tweets and the results were visualized as co-mention maps using Gephi. Some clear clusters were visible in each discipline and these communities were confirmed statistically with Gephi’s built-in function for modularity classes (Blondel et al., 2008; Lambiotte et al., 2009) which finds the highly connected communities in a network. These networks and the detected communities in them are visualized in figures 3-5 below.

In biochemistry four communities were detected (figure 3). The largest of these communities contained words related or connected to metabolism (top of graph in figure 3). The tweets creating this cluster were not scientific to their content, but rather people talking about the impact of metabolism on weight loss. An example of such a tweet would be: *“@[...] you are soo lucky you have a fast metabolism.*” Lower part of the graph is occupied with more scientific tweets. There are three communities here; one community surrounding topics such as *diabetes*, *cancer*, *atherosclerosis*, and *cell studies*, another community surrounding topics such as *biochemistry, molecular biology*, *research* and *biology*, and a final smaller but tightly connected community caused by a frequently retweeted tweet: "*RT @[…]: Dark, organic, unprocessed chocolate has been found to benefit your glucose metabolism, blood pressure, and cardiovascular health.*"

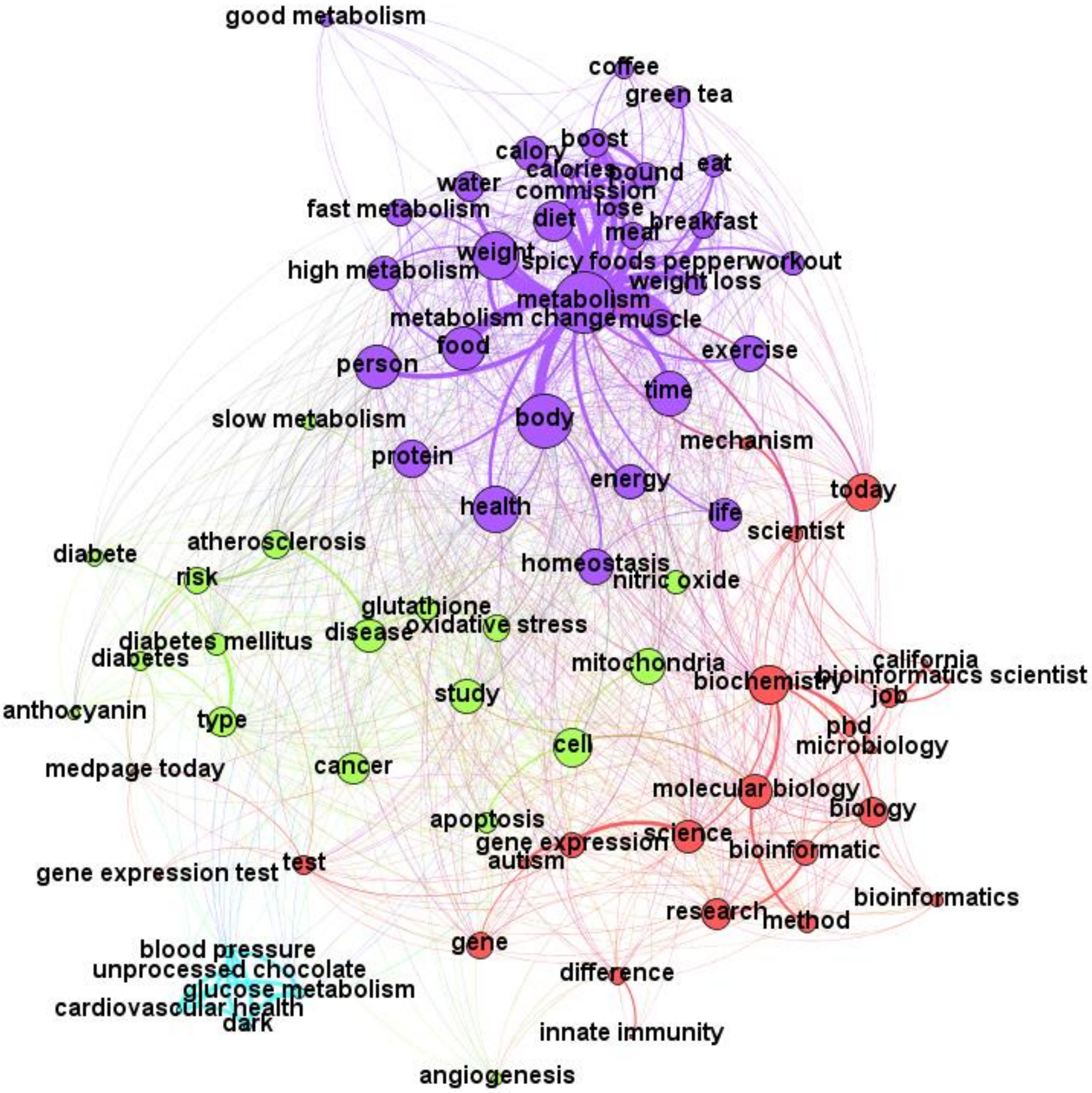

Figure 3. Co-mention map in biochemistry

In cognitive science (figure 4) the graph is influenced by the 100th anniversary of Alan Turing's birthday, which coincided with the data collection period. Turing is widely recognized as the father of artificial intelligence, a topic that is strongly connected with cognitive science. Another topic that gained plenty of visibility in the data was an unbeatable robot in the rock-paper-scissors game. Tweets like: "*Robot Hand Beats You at Rock, Paper, Scissors 100% Of The Time* http://t.co/Pz1xxyZK" were frequently sent during the time of data collection. Another topic that is visible in the graph is caused by tweets about Google creating artificial intelligence that can identify a cat. This topic was frequently shared in tweets like: "*Google develops Artificial Intelligence to identify a cat* http://t.co/FL7Mif6K".

The scientific tweets can be found in the top of the graph and these contain research related to *cognition*, *technology*, *memory*, and *information processing*. There are three frequently used words that although in the graph belong to different communities, are still highly connected across the whole graph. These words are also in the very core of cognitive science: *artificial intelligence*, *brain*, and *cognition*.

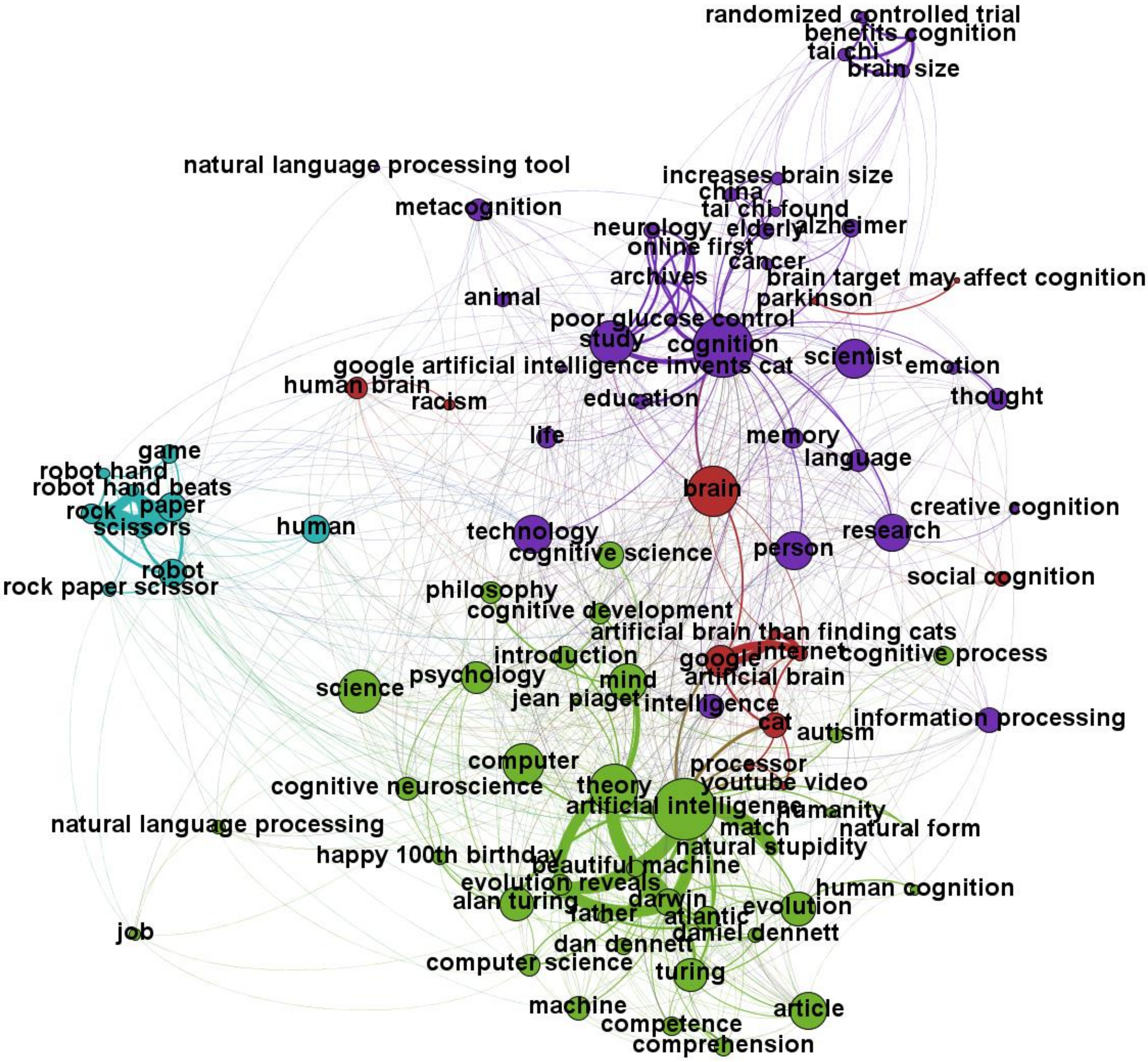

Figure 4. Co-mention map in cognitive science

Economics was the topic where it was most difficult to find or identify scientific content from the tweets. The most popular topics and words are visualized in figure 5 below. In the very core of the graph we can find a small but frequently mentioned community containing words such as *economic*, *time*, *politic*, *Obama*, *Europ*e, and *analysis*. However, the graph is dominated by a large and not as tightly connected community containing words such as *economics*, *year*, *person*, *government*, *today*, *money*, *book*, and *game theory*. Thematically these two communities are closely related and could probably be merged into a single community, but statistically they create separate communities as shown in the graph. Three smaller but tightly connected communities are also present in the graph. The first of them in the bottom left of the graph is caused by a frequently retweeted tweet originally tweeted by The Wall Street Journal: "*RT @WSJ: A single person spends 76 minutes more on personal care, sleeping and leisure than a married person. http://t.co/FN8rDR99*". The second tightly connected community is caused by another popular tweet originally published by The Wall Street Journal: "*RT @WSJ: Most patents that come out of major American universities have at least one foreign-born creator, says report. http://t.co/Af4LURb8*". The final tightly connected but small community is created by tweets sharing information about Reuters' reports on various economic indicators from different countries.

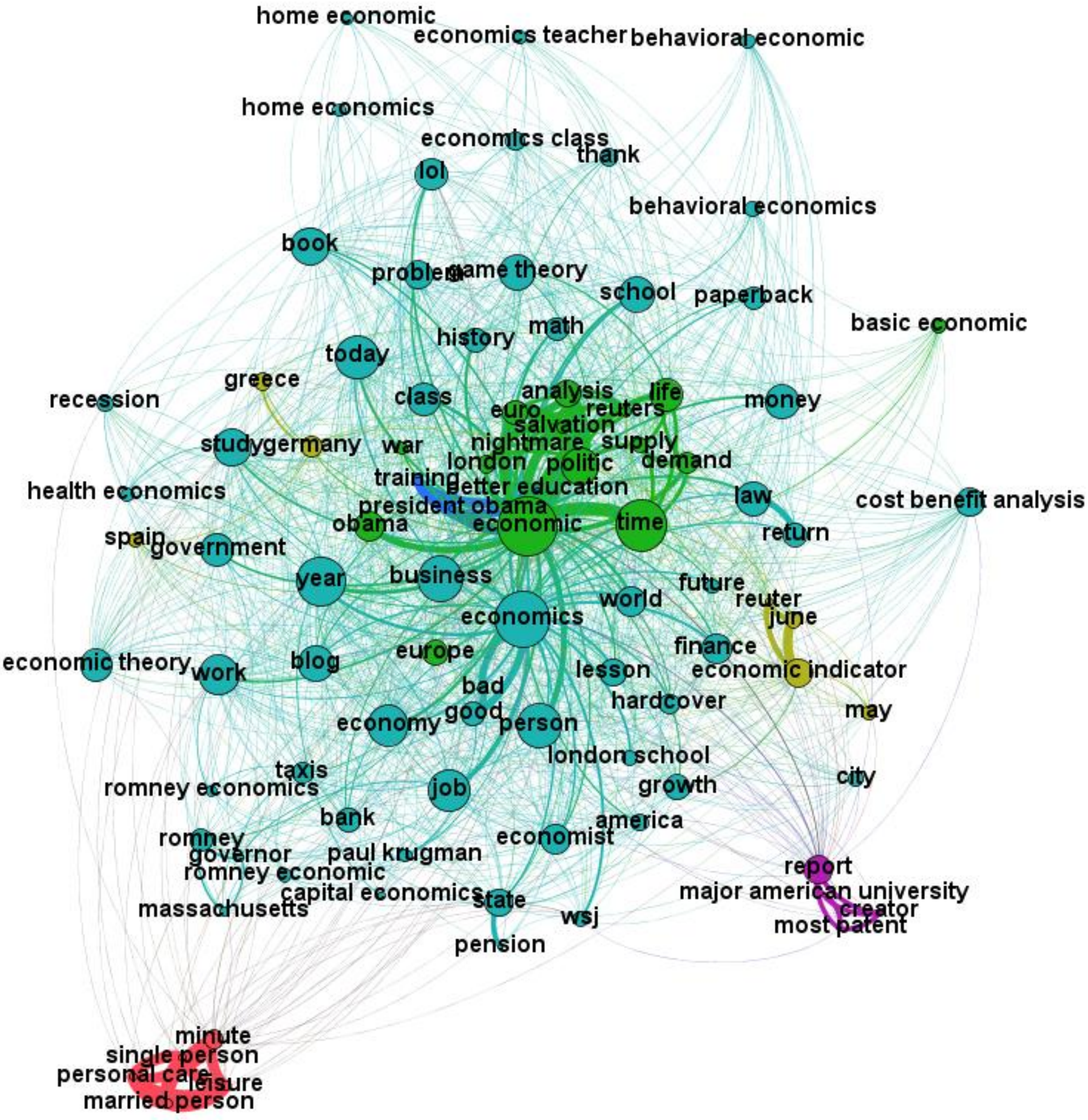

Figure 5. Co-mention map in economics

Discussion and conclusions

Using keywords specific to certain research areas proved to be a possible method to discover scientific content and to investigate scientific information sharing on Twitter, as the present study found evidence of both. However, the success of using keyword searches to discover scientific content on Twitter is highly connected to the keywords chosen, and for some disciplines it may be difficult to find unique and specific keywords that would not appear in everyday talk not related to scientific content. The co-mention analysis and the visualized co-mention graphs give an overview of the most popular topics in the tweets. In some cases it is also possible to recognize scientific content in tightly connected communities in the graphs, but in order to get an overview of the scientific content the data should be cleaned from unrelated tweets prior to analysis. Automatic filtering of tweets by what is relevant to a specific analysis and what is not is still a great challenge. Hence more research is required before anything conclusive can be said about the use of research specific keywords in mining Twitter for scientific content.

The present study is not without weaknesses, of which the size of the sample is perhaps the most significant. However, the present sample already indicated differences in how scientific content in different research areas can be discovered on Twitter. A possible future research direction would be to expand on the current study to include more research areas and to use larger samples of tweets. For future data collection the number of keywords used should be addressed. In the present study the number of keywords between the disciplines varied from 15 to 69, which may have had some impact on the results. A final weakness of the present study is that the classification was done by only a single researcher. However, classification by type of tweet is fairly straightforward and it does not leave much room for discussions, and in the classification of the tweets by content a conservative approach was taken in order to prevent exaggeration of the results. Hence the amount of scientific content may be higher than presented in the results of this study.

Future research could investigate possible relationships between different ways of data collection from Twitter. Data can be collected using keywords, hashtags, or by focusing on retweets or conversational tweets. It is also possible to collect tweets sent by a select group of Twitter users. This type of research could help focus the data collection to scientific content.

Acknowledgments

This manuscript is based upon work supported by the international funding initiative Digging into Data. Specifically, funding comes from the National Science Foundation in the United States (Grant No. 1208804), JISC in the United Kingdom, and the Social Sciences and Humanities Research Council of Canada.

References

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. In Proceedings of the International AAAI Conference on Weblogs and Social Media. Retrieved on June 7, 2013, from http://gephi.org/publications/gephi-bastian-feb09.pdf.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, vol. 13, no. 4. Retrieved on February 9, 2013, from http://www.jmir.org/2011/4/e123/.

Harlow, S. & Johnson, T.J. (2011). Overthrowing the protest paradigm? How the New York Times, Global Voices and Twitter covered the Egyptian revolution. *International Journal of Communication*, vol. 5, pp. 1359-1374.

Holmberg, K. & Thelwall, M. (2013). Disciplinary differences in Twitter scholarly communication. To appear in the *Proceedings of International Society for Scientometrics and Informetrics 2013*, Vienna , Austria.

Honeycutt, C. & Herring, S.C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*. Retrieved March 29, 2011, from http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf.

Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169-2188.

Lambiotte, R., Delvenne, J.-C., Barahona, M. (2009). Laplacian Dynamics and Multiscale Modular Structure in Networks. In *arXiv*:0812.1770 [physics.soc-ph].

Ross, C., Terras, M., Warwick, C. & Welsh, A. (2010). Enabled backchannel: conference Twitter use by digital humanists. *Journal of Documentation*, vol. 67, no. 2, pp. 214-237.

Rowlands, I., Nicholas, D., Russell, B., Canty, N. & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, vol. 24, no. 3, pp. 183-195.

Suh, B., Hong, L., Pirolli, P. & Chi, E.H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of IEEE International Conference on Social Computing*, 2010. Retrieved January 17, 2013, from http://web.mac.com/peter.pirolli/Professional/About_Me_files/2010-04-15-retweetability-v18-final.pdf.

Van Eck, N.J., & Waltman, L. (2009). VOSviewer: A computer program for bibliometric mapping. In B. Larsen, & J. Leta (Eds.), Proceedings of the 12th International Conference on Scientometrics and Informetrics (pp. 886-897).

Weller, K., Dröge, E., & Puschmann, C. (2011). Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. In M. Rowe, M. Stankovic, A.-S. Dadzie, & M. Hardey (Eds.), *Making Sense of Microposts* (#MSM2011), Workshop at Extended Semantic Web Conference (ESWC 2011), Crete, Greece (pp. 1–12). CEUR Workshop Proceedings Vol. 718. Retrieved January 17, 2013, from http://ceur-ws.org/Vol-718/paper_04.pdf.