Twitter for Scientific Communication: How Can Citations/References be Identified and Measured?

Katrin Weller
Dept. of Information Science
Heinrich-Heine-University Düsseldorf
Phone: 0049 (0) 211 8110803
weller@uni-duesseldorf.de

Cornelius Puschmann
Dept. for English Language and Linguistics
Heinrich-Heine-University Düsseldorf
Phone: 0049 (0) 211 8115927
cornelius.puschmann@uni-duesseldorf.de

ABSTRACT
This paper discusses 'citations' and 'references' within the microblogging service Twitter with the aim to provide measures for scientific communication on this platform. It provides definitions for different types of citations on Twitter and discusses general difficulties in accessing scientific tweets. Furthermore, two different datasets that represent scientific usage of Twitter have been analyzed with respect to citation counts.

General Terms
Measurement, Human Factors

Keywords
Microblogging, Twitter, scientific communication, citations, references, scientometics..

1. INTRODUCTION
Scientific communication is typically perceived as a process of publishing scientific publications and of citing other scientists' publications. The disciplines of bibliometrics and scientometrics have established procedures for measuring scientific output based on publications and scientific reputation based on citations. Informetric citation analysis distinguishes citations from references [11]: A citation is a formal mention of another work in a scientific publication – viewed from the cited work's perspective. A reference is the same mention of a work but viewed from the citing work's perspective (typically in form of a reference section in a publication). Thus, citations and references are two sides of the same coin. Slightly inconsistently, the term 'citation' is also used as a broader term that subsumes both the dimension of citations as well as the dimension of references. This paper investigates whether comparable structures of citations and references can also be identified in microblogging environments, particularly in the microblogging service Twitter

WebSci '11, June 14-17, 2011, Koblenz, Germany.

The Web as a medium for information exchange and communication has lead to the investigation of new metrics (webometrics) in addition to classical bibliometric and scientometric indicators [12]. While classical webometrics mainly consideres hyperlink structures between Websites, recent Web 2.0 tools that enable novel forms of social interaction have brought about a range of new aspects that can be measured and evaluated (e.g. relating to access and usage, Web publication behavior, user interrelations). [12] explains that measuring Web 2.0 services offers new ways for data mining; it can help to gain insights to "patterns such as consumer reactions to products or world events" [12]. [7] provide an overview on Web 2.0 services that may be of interest for new scientometric indicators by measuring publication impact based on social mentions. One of these social software scenarios is microblogging.

Within this paper, we investigate Twitter usage in scientific contexts and consider Twitter as a means for scientific communication. The scientific use of Twitter has received some attention in previous work: [4] and [5] have performed several automatic analyses of tweets collected for different conference hashtags, including for example time series and lists of most active twitterers. [3] and [9] have furthermore carried out manual analyses of tweet contents for conference tweet datasets to determine, what conference participants are tweeting about. [10] are developing automatic methods for extracting semantic information from conference tweets. [6] have focused on tweets published by a set of manually identified scientists and have investigated their citation behavior.

[6] define Twitter citations as "direct or indirect links from a tweet to a peer-reviewed scholarly article online" and distinguish first- and second-order citations based on whether there is an "intermediate webpage between the tweet and target resource". Within this paper, a broader approach is applied. Two fundamental types of citations are distinguished: external citations are all links included in tweets; internal citations are retweets within the Twitter platform.

The paper will discuss these two types of citations and will focus on their implications and challenges for informetrics (section 3). But first of all it will have to start with the general problem in analyzing scientific impact of Twitter: how can scientific contents be actually identified on Twitter (section 2)? We will furthermore present our current approaches to citation analyses on Twitter for

two different types of datasets. Section 4 describes how these datasets were gathered, section 5 presents very preliminary results. Our overall aim is to better understand how scientists use Twitter and whether traditional patterns of scientific communication are being mapped to microblog communications or whether entirely new practices emerge. This paper should primarily be viewed as exploratory research in the field of informetrics for microblogging. It may provide a basis for future work on developing novel informetric indicators or for the development of applications that make use of these indicators, e.g. for identifying and ranking popular tweets, popular twitterers or external resources, as well as for displaying user networks based on co-citation or bibliographic coupling.

2. IDENTIFICATION OF SCIENTIFIC TWEETS

As Twitter is not dedicated to one particular application scenario and thus includes users with various backgrounds and different motivations, it is difficult to identify scientific tweets or twitterers. It is not yet defined in the research community what actually classifies as scientific Twitter usage or as a scientific tweet. There are also no reliable statistics about how many scientists use Twitter (and consequently no insights to how many of them do so for science-related communication). Empirical studies (quantitative and qualitative designs) that investigate scientists' motivations for using Twitter are missing – analyses are so far mainly based on the data delivered by Twitter. So far, there are basically two different ways to compose scientific tweet datasets [13]: a) based on hashtags and b) based on persons. Theoretically, a third way would be to collect all tweets with scientific content or that link to scientific content. This, however, is almost impossible to achieve, as it would require either manual identification of tweet contents or elaborated computer-linguistic automated methods as well as an elaborated definition for 'scientific contents'.

2.1 People-Based Approach

Analyses of scientific Twitter behavior may be based on a collection of tweets published by a scientist. Similar approaches are frequently applied in analyses of scientific blogging. Yet, the definition of 'scientist' in this context is not always consistent. It may for example be a narrow definition only including members of universities or a broad one including also, e.g., teachers and science journalists. Analyzing (micro)blogs based on users is depending on the availability of biographical information provided by the blog authors or twitterers. Furthermore, a selection of users will have to be made manually. [6] have applied this approach and have manually identified 28 twittering scientists (using a snowball system) to analyze their citation behavior. [14] has identified twitterers with academic background by examining the list of followers of the Chronicle of Higher Education's Twitter account. The most notable effort in collecting scientific twitteres has been made by David Bradley, who identified more than 500 scientific twitter accounts [2].

One problem in people-based approaches is that a twitter account may also be shared by a group of people. For example, a research group may have a twitter account and several members of that group may access this account to report their latest efforts. Other official institutional accounts (e.g. for a university) may be completely taken care of by a single person. In many cases it is not possible to distinguish whether a twitter account is used by a single person or a group. To our awareness, there are so far no studies that exclusively analyze Twitter accounts belonging to scientific groups or institutions.

2.2 Hashtag-Based Approach

The more common way to compose datasets for scientific Twitter analyses is to collect tweets for specific (science-related) hashtags. Only in rather rare cases, scientists announce official hash-tags for their projects or topics of interest. One recent prominent example is the hashtag "#altmetrics" which is introduced by [8] for work on measuring scholarly impact on the Web. But much more frequently, specific hashtags are announced for scientific conferences (some of them officially proposed by the organizers, e.g. "#websci10", some are arranged by the participants of a conference during the event). Most studies on scientific microblogging have used datasets collected via conference hashtags [3, 4, 5, 9, 10, 13]. This approach always has to accept, that tweets might be "lost". If twitterers engage in the discussion without using the respective hashtag, their tweets cannot be included. The same holds for tweets in which the hashtag is misspelled (e.g. "#websci2010" instead of "#websci10"). Still, it enables us to compose datasets for a relatively consistent subset of Twitter users, namely people interested in the contents of a particular scientific conference.

3. CITATION ANALYSIS ON TWITTER

Sets of scientific tweets may be analyzed with different objectives. Our main question within this paper is whether scientific tweets include citation structures similar to traditional information flows in scientific literature. [6] define Twitter citations as "direct or indirect links from a tweet to a peer-reviewed scholarly article online". They distinguish first- and second-order citations based on whether there is an "intermediate webpage between the tweet and target resource". In their sample of tweets collected from 28 academics they discovered that of all tweets including an URL, 6% fit into their definition of twitter citations, i.e. they linked directly or via an intermediate page (like a blog post) to a peer-reviewed article. Within our previous work [13] we suggested alternative definitions and different dimensions of citations in Twitter.

3.1 External Citations

We consider all URLs included as a form of citation: the tweet includes a reference in form of a URL and a certain website obtains a citation through this tweet. URLs in tweets act as external citations as they link Twitter content with external websites. Analyses may focus on the types of resources that are referenced in URLs [13]. For purely scientometric analyses, references to scientific publications are of highest interest, but references to

scientific blog posts or presentations slides may also be valuable information. For more general informetric analyses, references to all other websites may provide additional value.

3.2 Internal Citations

3.2.1 Retweets

Retweets (RTs) can be interpreted as a form of inter-Twitter citation (*internal citations*). A user who retweets another one's tweet publishes a reference, the retweeted user gets a citation. As analyzed by [1], users retweet for different reasons like information diffusion or use retweets as a "means of participating in a diffuse conversation". This should be investigated in more detail for scientific tweets. Yet, retweet analyses are not easy to perform, due to the lack of format standardization. Not all twitterers retweet with the standard "RT @user" format.

3.2.2 @mentions

@mentions of usernames within tweets also sometimes resemble references, e.g. in tweets like "Just read an interesting paper by @sampleuser". Yet, they can currently not be automatically distinguished from other @messages and will thus have to be excluded from current analyses.

4. DATA COLLECTION

Within our previous work [3, 13] we have exclusively worked with scientific tweets collected via conference hashtags. We now want to extend this and include additional data collected via a list of scientific twitterers.

4.1 Hashtag-Based Collection

During our previous work [3] we have collected tweets from four scientific conferences. Conferences were selected based on two features: size and discipline. We have chosen two smaller conferences (<500 participants) and two major conferences (>1.000 participants). One small and one larger conference was on topics from (digital) humanities and one small and one larger conference was located in the field of computer sciences. In [3] we performed intellectual analyses of tweets in these conference datasets. In [13] we continued this work and performed additional intellectual analysis of URLs included in tweets and first citation analyses. Within this paper and the respective poster we now want to consider the results of citation analyses from the hashtag-based dataset in comparison to additional data collected with a people-based approach.

Currently, we have restricted our citation analyses to data from two conferences out of the initial set of four conferences, as the methodology is still subject to refinements and should be improved after discussion in the scientific community. We have chosen the two larger conferences: one from computer science (the World Wide Web Conference 2010, WWW2010, hashtag #www2010), and one from humanities (the Modern Language Association Conference 2009, MLA 2009, hashtag #mla09). Table 1 presents an overview of the key information about the selected conferences and their respective hashtags. We deliberately concentrated on the main hashtag for each conference in order to achieve uniform preconditions for each set (we did not include spelling variants or hashtags for associated or co-located events).

Table 1. The test dataset for tweets with conference hashtags #mla09 and #www2010

Hashtag	**#www2010**	**#mla09**
Conference	World Wide Web Conference (WWW2010), Raleigh, NC, USA.	Modern Language Association Conference (MLA 2009), Philadelphia, PA, USA.
Conference dates	26.-30. April '10	27.-30. Dec. '09
Discipline	Computer science	Linguistics, literature, (digital humanities)
No. of tweets from two weeks before until two weeks after the conference	3,358 [during period: 13. April 2010-14. May 2010]	1,929 [during period: 15. Dec. 2009-14. Jan. 2010]
Total no. of unique twitterers (average no. of tweets per twitterer)	903 (∅ 3.72)	369 (∅ 5.23)
Total no. of tweets during actual conference days only	2,425 [26.-30. April 2010]	1,206 [27.-30. December 2009]

4.2 People-Based Collection

We assume that scientists tweet differently during conferences than in every-day situations. To fully support this, broad additional studies with data collected from scientific twitterers are needed. In order to start first analyses in this regard we have started to set up a sample collection of tweets by scientists.

We used the list of scientific twitterers by Bradley [2] and modified it; we added some more twitter accounts which we had manually identified as belonging to scientists. Scientists in this context are not purely university staff but may also be (graduate) students or researchers in companies. Some twitter accounts may not belong to individual persons but to scientific groups. Altogether, we obtained a set of 589 unique users. We then collected all the tweets from these 589 Twitter accounts during the period January 7, 2010 until August 31, 2010. The total number of tweets for this dataset is 410,609 tweets.

5. FIRST RESULTS

Within this poster paper we will only be able to give a very first insight to our overall results. More detailed data will be presented in the poster. A first result of high interest can be found in the pure counting of URLs as external citations. We counted the numbers of URLs (identified as strings starting with "http(s)://" or "www." followed by text) in different ways. Table 2 shows how many tweets in the #www2010, the #mla09 and the people-based

dataset contain at least one URL. As some Tweets contain more than one URL, the total number of URLs is also listed. For the two conference datasets we have also resolved the shortened URLs to count the number unique URLs: the #www2010 dataset includes 574 unique URLs, the #mla09 dataset includes 199 unique URLs. Table 2 shows that the people-based dataset includes a much higher percentage of Tweets with URLs than the conference-based datasets. That also shows that in general, scientists post a URL in more than 55% of their published tweets. During conferences, the number of non-URL tweets increases – we assume that this is due to a higher number of "conversational" tweets during social events like conferences and will investigate this in more detail.

Table 2. Different URL Counts

	#www2010	#mla09	Scientists
Number (and %) of tweets including at least one URL	1,338 (39.85%)	525 (27.22%)	227,550 (55.42%)
Number of total URLs	1,460	551	234,731

6. CONCLUSION AND OUTLOOK

The poster presentation will include additional results: the investigation of the types of Websites that the URLs link to, the highly cited URLs from the conference datasets, the number of retweets for the different datasets, highly retweeting and retweeted users. Altogether citation behavior in Twitter is different from traditional scientific publication and citation behavior and need specific standards for analysis and metrics.

7. ACKNOWLEDGMENTS

Many thanks to Evelyn Dröge who worked with us during earlier phases of this project. Thanks to Julia Verbina and Parinaz Maghferat for their contributions to data collection. Thanks to Bernd Klingsporn for advice and support and to Wolfgang G. Stock and Isabella Peters for critical remarks. Financial support from the Heinrich-Heine-University Düsseldorf for the Research Group "Science and the Internet" is greatly acknowledged.

8. REFERENCES

[1] Boyd, D., Golder, S. and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In R. H. Sprague (Ed.), Proceedings of the 43rd Conference on System Sciences (HICSS 10), Honolulu, Hawaii, USA . Piscataway, NJ: IEEE.

[2] Bradley, David (no year). Hundreds of scientific Twitter friends. Retrieved May 6, 2011, from http://www.sciencebase.com/science-blog/100-scientific-twitter-friends.

[3] Dröge, E., Maghferat, P., Puschmann, C., Verbina, J. and Weller, K. 2011. Konferenz-Tweets: Ein Ansatz zur Analyse der Twitter-Kommunikation bei wissenschaftlichen Konferenzen. In Joachim Griesbaum, Thomas Mandl, Christa Womser-Hacker (Eds.), Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (pp. 98-110). Boizenburg: VWH.

[4] Ebner, M. and Reinhardt, W. 2009. Social networking in scientific conferences: Twitter as tool for strengthen a scientific community. In U. Cress; V. Dimitrova, & M. Specht (Eds.), Learning in the Synergy of Multiple Disciplines.4th European Conference on Technology Enhanced Learning, EC-TEL 2009 Nice, France. Berlin: Springer.

[5] Letierce, J., Passant, A., Decker, S. and Breslin, J. G. 2010. Understanding how Twitter is used to spread scientific messages. In Proceedings of the Web Science Conference (WebSci10): Extending the Frontiers of Society On-Line, Raleigh, NC, USA.

[6] Priem, J. and Costello, K. L. 2010. How and why scholars cite on Twitter. In C. Marshall; E. Toms, & A. Grove (Eds.), Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, Pittsburgh, PA, USA (pp. Article No. 75). New York, NY: ACM.

[7] Priem, J. and Hemminger, B. M. 2010. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. First Monday, 15(7). Retrieved January 06, 2011, from http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570.

[8] Priem, J., Taraborelli, D., Groth, P. and Neylon, C. 2010. Alt-metrics: A Manifesto. Retrieved January 13, 2011, from http://altmetrics.org/manifesto/.

[9] Ross, C., Terras, M., Warwick, C. and Welsh, A. 2011. Enabled backchannel: Conference Twitter use by digital humanists. Journal of Documentation, 67(2), 214–237.

[10] Stankovic, M., Rowe, M., and Laublet, P. 2010. Mapping tweets to conference talks: A goldmine for semantics. In Proceedings of the Third Social Data on the Web Workshop SDoW2010, collocated with ISWC2010, Shanghai, China.

[11] Stock, W.G. (2007): Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

[12] Thelwall, M. 2008. Bibliometrics to webometrics. Journal of Information Science, 34(4), 605–621.

[13] Weller, K., Dröge, E., and Puschmann, C. 2011 (in press): Citation Analysis in Twitter. Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. In Proceedings of Making Sense of Microposts Workshop (#MSM2011). Co-located with Extended Semantic Web Conference, Crete, Greece.

[14] Young, J. R. 2009. 10 High Fliers on Twitter: On the microblogging service, professors and administrators find work tips and new ways to monitor the world. The Chronicle of Higher Education, 31, A10, April 10, 2009. Retrieved January 11, 2011, from http://chronicle.com/article/10-High-Fliers-on-Twitter/16488/.