Identifying and Analyzing Researchers on Twitter

Asmelash Teka Hadgu
L3S Research Center
Appelstraße 4, 30167 Hannover, Germany
teka@l3s.de

Robert Jäschke
L3S Research Center
Appelstraße 4, 30167 Hannover, Germany
jaeschke@l3s.de

ABSTRACT

For millions of users Twitter is an important communication platform, a social network, and a system for resource sharing. Likewise, scientists use Twitter to connect with other researchers, announce calls for papers, or share their thoughts. Filtering tweets, discovering other researchers, or finding relevant information on a topic of interest, however, is difficult since no directory of researchers on Twitter exists.

In this paper we present an approach to identify Twitter accounts of researchers and demonstrate its utility for the discipline of computer science. Based on a seed set of computer science conferences we collect relevant Twitter users which we can partially map to ground-truth data. The mapping is leveraged to learn a model for classifying the remaining. To gain first insights into how researchers use Twitter, we empirically analyze the identified users and compare their age, popularity, influence, and social network.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Twitter; Computer Science; Classification; Social Network

1. INTRODUCTION

Twitter is a communication platform, a social network, and a system for resource sharing [8]. For scientists, it offers an opportunity to connect with other researchers, announce calls for papers and the like, communicate and discuss – basically: stay up-to-date. However, the exponential growth of information in society [7] does not exclude social media like Twitter: an abundant number of users court on one's attention which leads to the question of how (young) researchers can focus on the essential users and tweets?

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615676.

The classical approach in science to filter information is peer review: only information that is considered to be novel, sound, and significant by experts in the respective field is published. Currently, such a process is at most implemented manually: researchers can subscribe individually to other researcher's feeds by following them. However, there is no 'directory' of scientists on Twitter and finding feeds of experts in a specific discipline or area of interest is cumbersome.

Furthermore, the trend to consider visibility of scientific articles in the social web as a possible (and immediate) alternative or complement to citation counts [13] (with services like Altmetric[1] that provide counts for how often a scientific article has been mentioned on Twitter and other social networks) necessitates the need for peer-review-like mechanisms for the social web. Simple approaches purely based on the popularity of users, tweets, or URLs do not work as a tool for scientists to discover relevant research(ers), since popularity on the social web is fundamentally a matter of the crowd of non-scientists. Articles that are popularized by the media – often independent of their scientific significance – get superior attention compared to other, more important works. Consider the Ig Nobel Prizes,[2] whose winning (scientific) publications get quite some attention on the social web, e.g., the URL[3] of the winner of the 2012 physics prize [5] has been mentioned in more than 230 tweets.[4] Enabling users (and in particular researchers) to access the scientists' perspective in the social web and considering only tweets from physicists would provide a different and likely better picture.

Existing Twitter directories like Wefollow[5] rely on users' initiative to register and reveal their interests. This clearly limits the set of available profiles, since professionals have limited time and there is no immediate benefit for registration. Therefore, providing an automatically curated directory of scientists would simplify expert finding and the provision of topic-relevant feeds authored by peers. This approach requires to first identify scientists on Twitter and then classify their discipline, topics of interest, and expertise. Since only little is known about scientists on Twitter, such an endeavor should be accompanied by further steps to understand how Twitter is used by them.

In this work, we present an approach for the identification and classification of scientists on Twitter together with

[1] http://www.altmetric.com/
[2] http://www.improbable.com/ig/
[3] http://prl.aps.org/abstract/PRL/v108/i7/e078101
[4] http://topsy.com/trackback?url=http%3A%2F%2Fprl.aps.org%2Fabstract%2FPRL%2Fv108%2Fi7%2Fe078101
[5] http://wefollow.com/

an empirical analysis of researchers from computer science found on Twitter. We take a pragmatic approach on which users we regard as 'scientists': users being interested in the topics of the target discipline and having similar, Twitter-based features like users that have published scientific papers. We start with a list of seeds that are highly-relevant for the discipline of interest and use it to build and augment a set of candidate users that are likely scientists. For a subset of the candidates that we can match to ground-truth data from a digital library, we build a model for the classification of scientists. We can show that the model is very accurate and use it to classify all of our candidates. Both sets of users (matched and classified) allow us to perform an empirical analysis of scientists on Twitter.

The main contributions of this work are

- a complete framework for discipline-specific researcher classification on Twitter using a small set of seeds only,
- an automatic approach for the generation of ground-truth data by combining different data sources,
- an empirical analysis of computer scientists that are actively using Twitter, and
- the provision of the used datasets.

To the best of our knowledge, such an analysis has not been performed before. In addition, we publish the datasets of the different sets of users to foster research in the areas of expert finding and scientometrics.[6]

This paper is organized as follows: In Section 2 we review related work and in Section 3 we describe our classification approach and its concrete implementation. The results are presented in Section 4, accompanied by an empirical analysis of computer scientists on Twitter in Section 5. We draw conclusions about our approach in Section 6.

2. RELATED WORK

Several Twitter directories like Wefollow, Twellow, and JustTweetIt[7] list Twitter users by different areas of interest. There also exist more specific directories which, for example, list emergency physicians[8] or top Canadian politicians and keep track of what they and other citizens have to say on Twitter (and other social media) about politics.[9] In Wefollow, users provide their interests upon registration and are then ranked according to a prominence score that is computed similar to PageRank, restricted to the respective interest groups.[10] Even though the user interest is very accurate, because the users themselves provide the information, this approach is not scalable as it requires users to register at the web site and explicitly state their interests. Unlike Wefollow, our approach automatically builds profiles of Twitter users. In Twellow, user categorization is determined automatically using keyword/phrase matching on the users' Twitter profiles.[11] Our approach incorporates more features to get more accurate user classification. To the best of our knowledge, there is no directory that curates a list of scientists on Twitter. In this paper we present a general approach for generating a Twitter user directory and show its validity for computer scientists.

A good overview and one of the first comprehensive analyses of Twitter is [8] with findings on the distribution of followers and followees, tweets, trending topics and users, and retweet dynamics. The results suggest that Twitter, due to the speed of retweets, is a good medium for information diffusion from which scientists can benefit.

User classification in Twitter has been studied in [11] and [14]. Pennacchiotti and Popescu [11] propose a machine learning framework to perform large-scale user classification. They extract features from profile, content, and network connections of users and apply their framework to classify users by their political affiliation, ethnicity and affinity for a particular business. Similarly, Rao et al. [14] automatically infer users' latent attributes such as gender, age, regional origin, and political orientation on Twitter using features derived from tweet messages only. They use a focused crawling approach to build separate datasets for each attribute learning task. Starting from seed accounts of the target class, they gather more users by looking at their followers to manually build ground-truth data. They train an SVM model to perform the prediction. Another closely related work to our task of user classification deals with researcher home page classification. Gollapalli et al. [6] use URL-based features in addition to the content of a web page in a co-training scheme to classify web pages as academic or otherwise. In our approach, we use features that have been reported to work well in these three works.

Another line of research related to our work deals with measuring user influence in Twitter [1, 2]. In [1] Bakshy et al. study the characteristics and influence of a large set of Twitter users by examining information cascades, more specifically the diffusion trees associated with tweets containing URLs. They found that predicting influential users or tweets with URLs in terms of generating large diffusion trees is unreliable. They conclude that to harness word-of-mouth in Twitter it is necessary to target a large number of potential influencers instead of just the top influencers. Cha et al. [2] study user influence by comparing directed links among users. They regard the three influence measures number of followers, retweets, and mentions. Among others, they found that the number of followers alone is not a good indicator of influence, i.e., popular users who have a large number of followers are not necessarily the most influential users when considering the number of retweets or mentions. We use these different influence metrics to identify prominent computer scientists on Twitter.

Closely related to our work is also research that studies the use of Twitter for academic activities and analyzes the spread of scientific tweets as an instrument for citation analysis [12, 4, 16, 9]. Priem and Hemminger [12] study whether and how scholars cite on Twitter by analyzing tweets from 28 scholars. They define a citation as a a tweet that contains a URL to a peer-reviewed scholarly article. They find that scholars use Twitter to cite articles and suggest to use this information to augment traditional scientometric methods. Weller et al. [16] propose a methodology to analyze citations in Twitter during scientific conferences. They manually inspect and classify URLs and retweets of users to conclude that citations on Twitter are different from classical citations. Another example is the work by Eysenbach [4], who explores which metrics could enable the prediction of cita-

[6] https://github.com/L3S/twitter-researcher
[7] cf. http://www.twellow.com/, http://justtweetit.com/
[8] http://emergencytwitter.ivor-kovic.com/
[9] http://politwitter.ca/
[10] http://wefollow.com/about/score
[11] http://www.twellow.com/faq

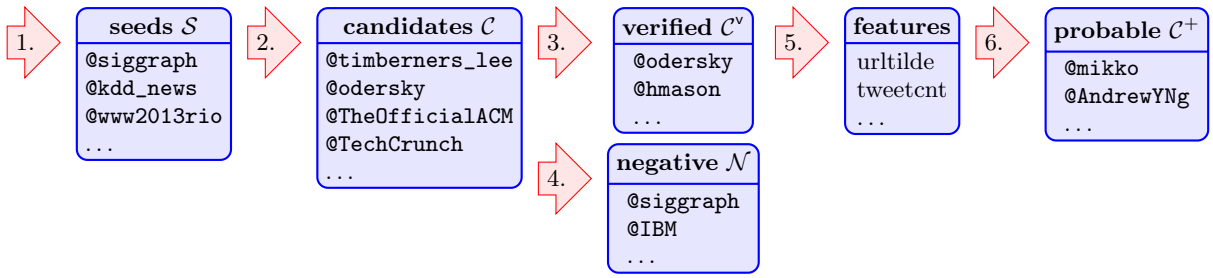


Figure 1: Overview of the processing pipeline.

tion counts based on tweets. Eysenbach computes metrics for 1 573 tweets that contain links to 55 articles of the Journal of Medical Internet Research and compares them with the citation counts. Letierce et al. [9] present the result of a survey to understand how Twitter is used for spreading scientific messages by semantic web researchers. These approaches can be seen as further motivation for our work, since we try to bridge the gap between Twitter and science by providing an automatic approach to identify scientists on Twitter. This task is not addressed by existing approaches, since the users and their (scientific) background is not considered. Furthermore, by focusing only on tweets that are linked to peer-reviewed articles, other forms of scientific discourse are ignored. Our approach to identify scientists, on the contrary, enables the analysis of a much larger share of academic content and its reception in Twitter.

3. APPROACH

In this section we describe our approach. We start with an overview of the processing pipeline and then explain the components in detail. The processing pipeline (cf. Figure 1) comprises the following six steps:

1. We start by building a *seed set* $\mathcal{S}$ of Twitter accounts for which we assume that they are typically followed by researchers. This set should be easy to obtain and should fulfill our assumption. By starting with a small set of seeds instead of (focused) crawling we favor precision over recall.
2. By collecting all users that follow the seeds or are otherwise related to them, we obtain an initial set $\mathcal{C}$ of *candidates*. This set can be further expanded by repeating the process several times with the candidates in place of the seeds.
3. We now match the candidates against ground-truth data $\mathcal{G}$. The subset $\mathcal{C}^{v} \subseteq \mathcal{C}$ of candidates for which we find a match with high confidence (*verified* candidates) allows us already to empirically analyze researchers on Twitter and will act as training data for Step 6.
4. The other part of the training data consists of *negative* examples $\mathcal{N}$, i.e., users that are not researchers. Basically, we are facing a one-class classification problem for which it is difficult to obtain ground truth data for the negative (outlier) class.
5. We identify and extract *features* for all users.
6. Based on the verified users $\mathcal{C}^{v}$ and negative examples $\mathcal{N}$ we train a model to classify the remaining candidates, which partitions them into users $\mathcal{C}^{+}$ that are *probably* researchers and users $\mathcal{C}^{-}$ that are *probably not* researchers.

Each step is a basic building block of our approach whose implementation could differ from the one we chose. In the following, we motivate and describe each step in more detail and present our implementation.

3.1 Generating a Seed Set of Twitter Accounts

The first step of our approach is the collection of Twitter users that serve as the seed set $\mathcal{S}$ for crawling further users in the next step by retrieving their friends, followers and retweeters. Our requirements to the seed set are similar to that of a focussed web crawl, where the crawling efficiency can be significantly influenced by the selection of seed URLs [17]: We want to ensure that the friends, followers and retweeters of the seeds have a high probability of being the target class, i.e., researchers, and that they are representative examples, e.g., have a good coverage over different sub-topics. Furthermore, the seed set should be small such that it could be curated manually.

In our implementation we chose Twitter accounts of *scientific conferences* as seeds, since we expect that these are typically followed or retweeted by computer scientists. Conferences are quite important and accepted in computer science, compared to other disciplines. To the best of our knowledge, such a list of accounts is not readily available. However, there exist some sources for lists of conferences, e.g., in Microsoft Academic Search,[12] or in Wikipedia.[13] Such lists can be taken as input for a subsequent manual or automatic collection of corresponding Twitter screen names. We decided to use the Wikipedia list, since it is maintained by the community and with 268 conferences it is quite comprehensive. Although the selection and completeness of this list could be questioned (as that of any list), it fulfills the initially stated requirements of being representative for computer science and of being small enough to acquire the corresponding seed set of Twitter accounts with high confidence. The choice of Wikipedia also makes our approach easily adaptable, since similar lists exist for other disciplines.

Once we have a list of seed conferences, we want to link them to their Twitter accounts, if they have any. Rather than identifying these accounts manually, which requires a lot of effort and restrains reproducibility, we combine two search approaches:

1) Twitter Search. Motivated by the observation that many conferences have Twitter screen names that corre-

[12] http://academic.research.microsoft.com/?SearchDomain=2&entitytype=3

[13] http://en.wikipedia.org/wiki/List_of_computer_science_conferences

spond to their acronyms and are possibly followed by the year of the conference, we construct potential screen names for the years 2008 to 2014 by appending two and four digit numbers to each acronym, e.g., WWW is extended to @www09, ... and @www2009, ... Including the plain conference acronyms we issued in total 3 960 such queries from 264 acronyms[14] against the Twitter API[15] which returned detailed information for 1 652 valid Twitter screen names.

2) Web Search. We try to find links to Twitter from the web pages of the conferences. Therefore, for each seed conference, we perform a search with Microsoft Bing[16] to find its corresponding web site. We issue the conference's full name together with the years 2008 to 2014, e.g., for the WWW conference in 2009 we query for *World-Wide Web Conference 2009*. For each of the top three results of each query, we parse the web page for any link pointing to Twitter. From these links, we extract potential screen names for which we query the Twitter API to verify if they indeed represent Twitter accounts. Using this approach we could identify 260 accounts, some of which could not be found using the first approach, e.g., @www2012lyon or @www2013rio for the WWW conference.

The merged list of Twitter accounts from both approaches contained some noise which we removed by restricting to accounts that (a) contained one of the strings 'conference', 'symposium', 'workshop', or 'forum' in their profile description or (b) contained at least one of the phrases 'call for (papers|demos|tutorials|workshops)', '(paper|workshop) deadline', '(full|short|accepted) papers' and 'camera-ready' in their original (i.e., not retweeted) tweets. The resulting conference acounts were then validated by experts and only correct accounts were retained for the next step. A comparison of both approaches to generate seeds is given in Sections 4.1.

3.2 Generating Candidates

Having built a high-quality seed set, we gather candidate users $\mathcal{C}$ that follow the seeds or are otherwise related to them. In practice, there are other relationships we can leverage, e.g., users are followed by one of the seeds, or re-tweeted a seed's tweet. Other, more indirect relationships include the mention of a seed's name in a user's tweet or vice versa, the usage of mutual hash tags, or the following of common users. Our assumption is that the more direct and closer the relationship of the users to the seeds is, the more likely they are of our target class. Note that this expansion process can be repeated several times with the current set of candidates as input instead of the seeds. However, with every expansion round the distance to the seed set grows and thereby the likelihood that the users are researchers decreases.

In our implementation we follow an approach that is inspired by [15], where the retweet signal is used to collect politically interested Twitter users starting from a seed set of users. However, we also add users that follow or are followed by (i.e., are *friends* with) the users in the seed set – the latter because we observed that most conferences follow researchers. In our initial experiments we expanded the candidate set once. However, we considered the expanded set as both too large (more than 30 million users) and too broad to be useful and therefore omitted the expansion step in our implementation to favor precision instead of recall.

3.3 Matching Candidates With Ground-Truth

The preceding step generated a set of candidate users that are likely from our target class, i.e., researchers. To classify these users, training data for machine learning is required, i.e., a set $\mathcal{C}^{\mathrm{v}}$ of Twitter screen names known to be scientists. Since such a list does not exist and manually building it is tedious and error prone, we use an automatic method to generate the training data for classification. Therefore, we leverage the fact that a common goal of researchers is the publication of their results in journals or at conferences, and that meta data about publications and hence authors is often readily available from *digital libraries* (e.g., PubMed for medicine, or arXiv.org for physics). Thus, given a list $\mathcal{G}$ of authors in the discipline of interest, we can match their names against the real names from the Twitter profiles of the candidate users $\mathcal{C}$. Of course, using the real name to match persons has some drawbacks: people might use false names in their profile and many names are not unique. On the other hand, the real name is the best indicator which we have, since in our setup all of the candidates have specified their name.[17] Choosing both the seeds and the ground-truth data from the particular scientific discipline of interest we minimize the chance of mismatches. To further reduce errors, we omit candidates and authors whose names appear more than once in the respective set and match the names without any normalization (e.g., without abbreviating first names). Given these measures, we are confident that the matching provides high-accuracy ground-truth data.

In our implementation the candidates are matched against authors from DBLP [10], a computer science bibliography hosted at the University of Trier. We downloaded the XML dump of DBLP[18], parsed it and extracted 1 304 283 author names from all publications.

DBLP disambiguates authors with the same name by appending a number to their name, e.g., *Abbas Mohammadi*, *Abbas Mohammadi 0001*, and *Abbas Mohammadi 0002*. Furthermore, author names in DBLP are case sensitive. For instance, *BorMin Huang* and *Bormin Huang* represent distinct authors. We regard both cases as duplicates. After removal of duplicate names (13 688 in DBLP and 2 686 in our candidate set) we performed exact string matching (ignoring case) on the full names. We do not conduct more complicated matching operations as we want to collect matches with high confidence. This way, we could match 9 191 DBLP authors ($= \mathcal{C}^{\mathrm{v}}$) against Twitter users in our candidate set $\mathcal{C}$. The result of validating this mapping is given in Section 4.2.

3.4 Generating Negative Examples

Since we can identify and describe the positive examples and regard everything else as negative examples, we are basically dealing with a one-class classification problem. However, regardless of whether we apply a binary or unary clas-

[14] Four acronyms appeared twice, because they referred to different conferences (ISWC for *International Semantic Web Conference* and *International Symposium on Wearable Computers*) or the same conference appeared under different names but with the same acronym (USENIX, FSE, CHES).

[15] `http://dev.twitter.com/docs/api/1.1/get/users/lookup`

[16] `http://datamarket.azure.com/dataset/bing/search`

[17] In contrast, only 55% of the candidates have specified a web site which could be used for identification, though this would then require ground-truth data that includes researchers' web sites.

[18] `http://dblp.uni-trier.de/xml/dblp.xml`

sification approach, we still need negative examples to test and compare the performance of the learned model and the different algorithms. Since $\mathcal{C}$ is a biased sample of Twitter users mostly from our target class, it is not a good source for generating a representative sample of negative examples. Instead, we crawl users with Twitter's streaming API. From this set we remove all users contained in $\mathcal{C}$ and then create $\mathcal{N}$ as a randomly sampled subset.

In our implementation we crawled 1 000 000 users with the Twitter streaming API and removed all users from $\mathcal{C}$ and all their followers and friends. Therefrom, we sampled 1 500 users which serve as the set $\mathcal{N}$ of negative examples. In addition to these 1 500 users, we added the seeds $\mathcal{S}$ to $\mathcal{N}$, since they are on the one hand closely related to our target class (by the topics of their tweets and their relationships with other users) and on the other hand not our target. We observed that our candidate set contained quite some technology companies. Thus, we identified companies from the Forbes Global 2000 list[19] whose names matched with our candidates after filtering duplicates. We found 24 such companies and added them to $\mathcal{N}$.

3.5 Feature Generation

Building upon ideas from [11] and [14], we generate features from *profile* and *content* information. These groups mimic semantically the steps a human surfer would normally perform, if asked to determine whether a given Twitter account corresponds to a researcher or not: The *profile* features are derived from the top information that is displayed on the web page of a user's Twitter account. These include name, location, URL, description and global counts like the number of tweets, followers, and friends. Ideally, these are the fields that represent the identity of a user and should be sufficient to determine who is who. In reality, the profile information is not enough since some fields are missing or they are not specific enough. With the *content* features, we consider the user's tweets. They provide information about the topics the user is interested in which allows us to decide whether an account belongs to a researcher or not.

In our implementation we use different features for each group which we explain in the following.

Profile Features. The *number of tweets, number of followers, number of friends*, and the *ratio of followers and friends* can be regarded as global indicators that capture how active the user is in the social media platform. Researchers are professionals and hence we capture how well organized the profile is with the boolean features *location, profile picture, description* which indicate whether the corresponding fields have been set. Inspired by [6] we constructed features that capture if a website is given in the profile and if it likely points to a researcher's web page (*website exists, website contains tilde character, website contains academic top-level domain .edu or country code second-level domain .ac (e.g., .ac.uk)*). Keywords such as *phd*, *researcher*, or *scientist* are a strong signal that the user is a researcher. We build a set of keywords that can be found in the bio of our verified users $\mathcal{C}^{\text{v}}$ but not in the negative examples $\mathcal{N}$ by employing the following steps: 1) we generate a list of top k% terms from the bio fields of users in $\mathcal{C}^{\text{v}}$, 2) we generate a list of top l% terms from the bio fields of users in $\mathcal{N}$ that serve as our 'stop words', and 3) we remove these stop words from the terms generated in the first step. The thresholds $k = 5\%$ and $l = 5\%$ were determined experimentally during the learning phase of the classification using cross-validation. The final terms are *architect, assistant, associate, author, candidate, co-founder, cs, designer, developer, director, engineer, fellow, founder, geek, graduate, lecturer, manager, phd, ph.d, prof, professor, programmer, researcher, scientist, senior*. These result in a boolean feature which indicates if the *bio contains keywords* from the above list.

Content Features. We include features that quantify the activity and topical interests of the users, such as the number of *original tweets, retweets, retweets to tweets ratio, tweets containing URL(s), fraction of tweets with URL(s), tweets containing hashtags, distinct hashtags* and *fraction of distinct hashtags used*. A good signal to distinguish researchers from other users is the fraction of tweets that are related to science. Since hashtags are often used to define the topic of a message, we can use them as an approximation of the topics a user is interested in. Therefore, we bootstrap the top hashtags used by the seeds $\mathcal{S}$ to gather similar scientific hashtags. These are typically conference acronyms or scientific terms preceded by the hash symbol, e.g., #siggraph2013, #machinelearning, ... In a political context, Conover et al. [3] showed that it is possible to extend seed political hashtags using co-occurence patterns to gather more political hashtags. Unfortunately, this method does not work for scientific hashtags. Whereas it is usual to use similar and conflicting political hashtags such as #obama #romney in a single tweet like "How could Watson and Big data help pick a better US president http://bit.ly/PFo3me #obama #debate #romney", it is unlikely that researchers use different conference acronyms as hashtags (e.g., #www2013, #siggraph2013) in the same tweet. Instead, we consider terms that occur most often with the seed hashtags and collect other hashtags that occur in a similar context. More precisely, we implemented the following approach to gather more scientific hashtags from a small seed of hashtags:

1. We build a set of seed hashtags by collecting the most frequent hashtag of each seed conference.
2. We identify the unigrams in tweets that contain one of the seed hashtags and remove the most frequent unigrams we can find in random tweets – these act as stop words. This way, we generate a set of terms that frequently co-occur with the hashtags from Step 1, e.g., *papers, workshop, keynote, poster, etc.*
3. We gather all hashtags that co-occur with these terms. (e.g., #wsdm2011, #websci13, #machinelearning, ...)
4. We remove the most common hashtags from random tweets which again act as stop words. This removes very general hashtags such as #ff, #followfriday.

With the final set of 1 872 hashtags, we can leverage the *number of tweets containing scientific hashtags* as feature for classification. Finally, we count *how often a user mentions other users that have used one of these hashtags*.

In preliminary experiments we also considered *network features* like the *number of seeds* that have been *mentioned*, are *followed*, or whose tweets have been *retweeted*, and the *number of candidates* that have been *retweeted* or *mentioned*, or are *followers* or *followees*. Although these would allow us to capture the notion that our target users more likely connect to the conferences in $\mathcal{S}$ and to each other, the features are biased towards our approach to gather the candidates

[19] `http://www.forbes.com/global2000/list/`

Table 1: Overview on the used datasets.

dataset	date	#users	#tweets
seeds $\mathcal{S}$	Nov 2013	170	23 843
candidates $\mathcal{C}$	Nov/Dec 2013	52 678	54 146 027
ground-truth $\mathcal{G}$	Dec 2013	1 304 283	–
verified $\mathcal{C}^{\mathrm{v}}$	Jan 2014	9 191	7 726 905
negative $\mathcal{N}$	Jan 2014	1 694	3 639 650

Table 2: Results of the automatic seed generation.

	Twitter Search	Web Search	Inter-section	Union
queries	3 960	1 869	–	–
screen names	1 652 (90%)	260 (14%)	69 (4%)	1 843 (100%)
filtered screen n.	135 (63%)	139 (65%)	60 (28%)	214 (100%)
valid screen n.	122 (72%)	107 (63%)	59 (35%)	170 (100%)
conferences	74 (76%)	75 (77%)	51 (52%)	98 (100%)

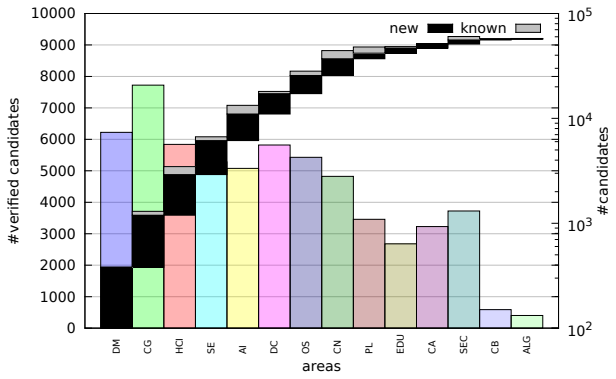


Figure 2: Distribution of the number of (verified) candidates over the different computer science areas.

and the negative examples. Since the examples were randomly sampled, their chance to be well-connected with the seeds or the candidates is very low, though for the candidates necessarily the opposite is true.

3.6 Classification

The target for the classification are the candidates that could not be matched against ground-truth data, i.e., $\mathcal{C} \setminus \mathcal{C}^{\mathrm{v}}$, where we expect that many of them are also researchers. Having identified verified candidates $\mathcal{C}^{\mathrm{v}}$ and some negative cases $\mathcal{N}$, we use these users to train a machine learning algorithm to classifiy the remaining unknown users in our candidate set. Any binary classification algorithm can be used, alternatively, one-class classification could be performed.

In our implementation we chose the classification algorithms Support Vector Machines (SVM), Random Forest (RF), Classification and Regression Trees (CART), and Logistic Regression (LR) using the implementations *e1071, rpart, randomForest and glm* available in R, the free software for statistical computing. We performed a stratified 10-fold cross-validation to train the models on 2 000 random users from $\mathcal{C}^{\mathrm{v}}$ and all users from $\mathcal{N}$, results are given in Sec. 4.5. Finally, the best performing algorithm is selected and trained on the complete training set to classify the remaining users.

4. RESULTS

In this section we analyze the implementation of our approach and present the results of the classification step. An overview of the datasets used is given in Table 1.

4.1 Seeds

In total, we found 170 Twitter screen names for 98 conferences (37% of the 268 conferences) using either Twitter Search or Web Search, cf. Table 2. The number of accounts found by both methods is 60. This shows that the methods are complementary in that we find Twitter accounts for conferences with one approach that we can not find with the other. On the other hand, in the intersection of both approaches (after filtering) almost all seeds are valid screen names of conferences, since only one of the 60 screen names was judged to be a false positive. Thus, if the effort of manual validation is too high and one can accept a smaller seed set, the process can be automated by using the screen names that are returned by both approaches only. We also note that Twitter Search has much more false positives before filtering which is caused by users that have screen names that resemble conference acronyms, e.g., @cikm, @www, etc. For Web Search the experts judged more of the filtered screen names to be false positives. One reason is that it yielded some accounts of research databases and conference organizers, e.g., @msftacademic, @ieeeorg, or @globaleventlist that can not be easily filtered by keyword matching on profiles and tweets. Overall, both approaches have a similar performance, although Web Search covers slightly more conferences with less screen names.

Since on the list from Wikipedia the conferences are partitioned into sub-areas of computer science[20] we can plot the contribution of each area to the sets $\mathcal{C}$ and $\mathcal{C}^{\mathrm{v}}$ in Figure 2. Sorted by the number of verified candidates that each area contributes, the black bars indicate the contribution relative to the previous area in the list (e.g., *DM* contributes most, followed by *CG*). The grey bars extend the black bars and thereby show the overal number of verified candidates from each area. We shifted the bars up such that they reach the sum of 9 191 in $\mathcal{C}^{\mathrm{v}}$ to the very right. The colored bars in the background show the (logarithmic) number of users each area contributes to $\mathcal{C}$.

Figure 3 shows a similar plot for each of the 98 seed conferences where the color of each conference matches the areas in Figure 2. The conference with by far the most followers is SIGGRAPH with 19 394 followers, followed by SC (5 229) and CHI (4 016). While following a conference clearly shows interest in it, retweeting one of its tweets is an even stronger signal.[21] Ranked by retweeters to followers ratio the order of conferences is: ICAC (38%), I3D (36%), ECOOP (33%), and AOSD (32%). Though ICAC has only 8 followers, the other conferences have more than 75 followers. On average, a user follows or retweets 1.1 conferences. Broken down by conference, ECCOP, ICDE and UIST have the most diverse users that follow or retweet on average 3.9, 3.5, and 3.0 con-

[20] namely, Data Management (DM), Computer Graphics (CG), Human-Computer Interaction (HCI), Software Engineering (SE), Artificial Intelligence (AI), Concurrent, Distributed and Parallel Computing (DC), Operating Systems (OS), Computer Networking and Networked Systems (CN), Programming Languages (PL), Education (EDU), Computer Architecture (CA), Security and Privacy (SEC), Computational Biology (CB), and Algorithms and Theory (ALG)

[21] Due to scarcity of space, we present only selected results.

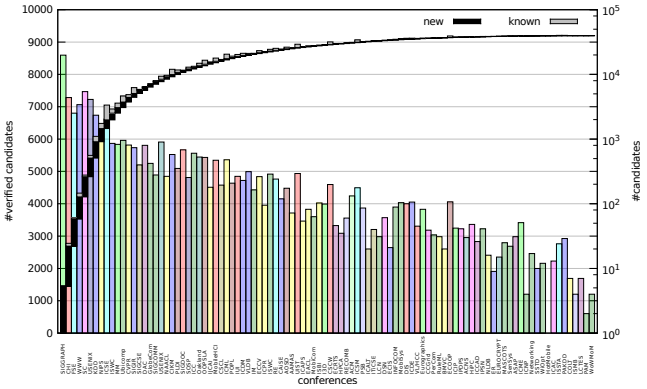


Figure 3: Distribution of the number of (verified) candidates over the conferences.

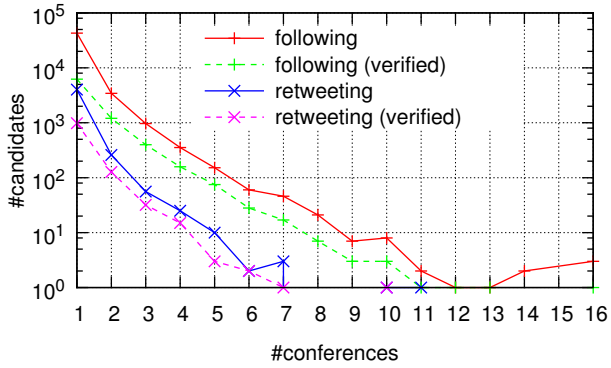


Figure 4: Distribution of the number of (verified) candidates that follow/retweeted one or more conferences.

ferences. SIGGRAPH and DAC have the most homogeneous users that, on average, follow 1.08 and 1.13 conferences, respectively. If we restrict the followers and retweeters to the verified candidates $\mathcal{C}^{\text{v}}$, SIGGRAPH still leads with 1 408 followers but is then followed by CHI (1 182) and SC (649). The retweeter/follower ratio is lead by I3D (77%), followed by RULEML (44%) and AOSD (40%).

4.2 Candidates

In the candidate generation step, we found 52 678 candidates, of which 47 870 follow at least one of the seed conferences. An additional 1 180 users retweeted at least one of the seeds' tweets but did not follow them at the time of our crawl, and 3 741 users were followed by one of the seeds (but the users neither followed nor retweeted the seeds). Most candidates are interested in one conference only (around 83% for both following and retweeting), though some are interested in five or more conferences (cf. Figure 4). Since we are interested in the researchers among the candidates, we present more insights in Sections 4.5 and 5.

4.3 Training Data

The training data consists of verified candidates $\mathcal{C}^{\text{v}}$ and negative examples $\mathcal{N}$. Overall, we could find 9 191 users (17.46%) in $\mathcal{C}$ whose names match with those in our ground-truth data from DBLP. They form the set $\mathcal{C}^{\text{v}}$. The fraction

Table 3: Performance of the feature groups using Random Forest.

feature group	P	R	F1	Acc	TNR
profile	0.91	0.89	0.90	0.91	0.92
content	0.94	0.88	0.91	0.92	0.95
profile + content	**0.96**	**0.92**	**0.94**	**0.95**	**0.97**

of them that follow more than one conference is 10% higher than for the whole candidate set (cf. also Figure 4).

The fraction of candidates of a conference that we could verify using DBLP varies strongly, as can be seen in Figure 3. SIGGRAPH is very popular among non-computer scientists, since only 7% of its candidates are in $\mathcal{C}^{\text{v}}$, but still 1 449 are verified – the largest absolute value of all conferences. On the other extreme are HOTMOBILE with 1 of 1 verified candidate, ICNP which has 75% verified candidates (3 of 4) and COLT, where 7 of the 13 users could be verified. Of the conference accounts with more than 100 candidates, ECOOP (53/107 ≙ 50%), ICSM (85/173 ≙ 49%), CSCW, (91/196 ≙ 46%), and SOSP (110/240 ≙ 46%) have a ratio of at least 45% verified candidates.

We evaluated the quality of our mapping by randomly selecting 150 users from $\mathcal{C}^{\text{v}}$ and asking three experts to verify if each candidate Twitter account (e.g., `https://twitter.com/odersky`) belongs to the identified DBLP author (e.g., `http://dblp.uni-trier.de/pers/hd/o/Odersky:Martin.html`). For 87 of the 150 users the experts came to the same decision (76 correct match, 4 wrong, 7 unknown) which underlines the difficulty of the matching task. For the 63 remaining cases, the experts jointly re-performed the task and reached an agreement (33 correct match, 17 wrong, 13 unknown). In summary, 109 (73%) of the matches were identified to be correct, 21 (14%) wrong, and 20 (13%) unknown. Many of the wrong and unknown matches either had different DBLP pages, or there was not enough evidence to confidently link their Twitter and DBLP accounts. However, most of these users were still researchers in computer science.

4.4 Features

We study the importance of features and feature groups for the classification. Using cross-validation as described in Section 3.6 to train and test the models, we restrict the features to single feature groups and the combination of both groups. For each set of features we learn models using 10-fold cross-validation (with stratified sampling). The classification accuracy of the feature groups for the best algorithm Random Forest is shown in Table 3. We can observe that *profile* and *content* alone yield a comparable good performance while their combination yields even better results. To gain more insights into the importance of individual features, we investigate the feature ranking as provided by Random Forest. The top ten important features are given in Table 4. The mean decrease accuracy (MDA) shows how much using the feature in the classifier reduces the classification error. Most important is the *number of tweets*, followed by more specific features targeted towards researchers.

4.5 Classification

The performance of the different classification algorithms during cross-validation is shown in Table 5. Random Forest is the best algorithm in all performance measures (precision (P), recall (R), F1-measure (F1), accuracy (Acc) and

Table 4: Individual features in order of importance by their mean decrease accuracy (MDA).

rank	feature	MDA	group
1	#tweets	54.57	profile
2	#tweets with scientific hashtags	49.35	content
3	friend/follower ratio	40.86	profile
4	bio contains keywords	40.33	profile
5	#conference mentions	39.53	content
6	#original tweets	34.14	content
7	#friends	34.04	profile
8	fraction of distinct hashtags	30.90	content
9	fraction of tweets with a URL	30.89	content
10	#tweets with a URL	27.57	content

Table 5: Performance comparison of the algorithms.

algorithm	P	R	F1	Acc	TNR
SVM	0.90	0.89	0.90	0.91	0.92
Random Forest	**0.96**	**0.92**	**0.94**	**0.95**	**0.97**
CART	0.88	0.90	0.89	0.90	0.90
Logistic Regression	0.88	0.87	0.88	0.89	0.90

true negative rate (TNR)). As a baseline, we trained an SVM classifier with a simple bag-of-words model and TF-IDF weighting. It yielded an F1-measure of 0.93. Given that 94% of all candidates have tweets, this approach is a viable alternative due to its simplicity and good performance. Finally, we retrained Random Forest on all the training data and used the model to score the remaining candidates. From a total of 43 383 unverified users in our candidate set, it classified 38 368 as researchers ($\mathcal{C}^+$) and the remaining 5 015 as non-researchers ($\mathcal{C}^-$).

5. RESEARCHERS ON TWITTER

In this section we empirically analyze computer scientists on Twitter. We consider this as a first important step towards a better understanding of how Twitter is used by researchers and how science can benefit from it. More specifically, we answer the following research questions:

- What kind of computer scientists use Twitter? We explore this from two perspectives namely, age and productivity.
- Which of Twitter's activities are used most frequently between researchers?
- Who are the most influential researchers on Twitter? Can we characterize these users?
- What are the most important scientific topics treated by computer scientists on Twitter?

5.1 Demographics

We start with the question *whether there is a bias of Twitter usage towards young researchers*. Although we do not have the birth dates of the researchers, we can leverage the fact that we mapped a portion of them to DBLP and use the year of their first publication as a proxy for their age. Let us first have a look at the corresponding distribution for *all* authors from DBLP, i.e., the set $\mathcal{G}$ (+): Figure 5 shows for each year between 1960 and 2013 the number of authors whose first publication was published in that year. We can see that the number of authors increases over the years with

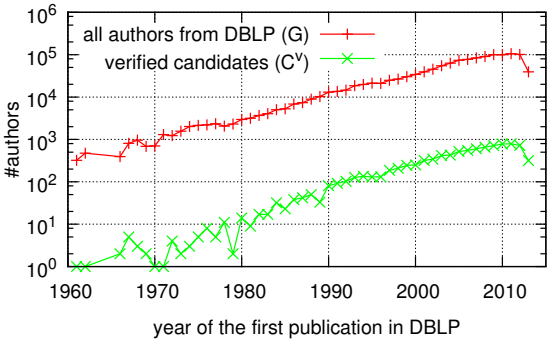


Figure 5: Distribution of the no. of authors per year.

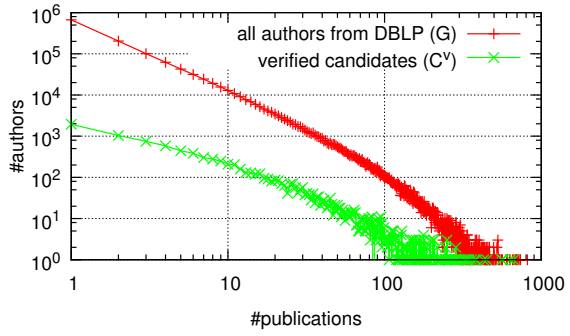


Figure 6: Distribution of the no. of publications per author.

a peak in 2011 and then a drop in 2013 which is caused by the yet incomplete data for this year. On the one hand, the increase reflects the real growth of the number of computer scientists (and publications) over the last decades, and on the other hand, it is possibly also influenced by the coverage of DBLP. When we compare this to the distribution of users from $\mathcal{C}^v$ (×) we can see that both distributions are very similar. This means that the subset of authors from DBLP that can be found in our sample of Twitter users has a similar distribution of first publication years as that of the average computer scientist from DBLP. We conclude that – at least for computer science – we can not find a difference in Twitter use between younger and older researchers.

As a measure for productivity we consider the number of publications an author has written. The plot in Figure 6 shows the long-tail distribution of the number of publications for the ground-truth authors from DBLP (+). It indicates that most authors (665 949 or 51%) have written only one paper and that only very few authors have written 100 or more. This can largely be explained by the large portion of young authors who just started to publish and probably by other cases like keynote speakers from industry, co-authors from other disciplines, etc. We get a different picture for the verified candidates (×): the curve is less steep and only 21% (1 949) of the candidates have published just one article. Although at first sight this might raise the idea that scientists on Twitter are more productive, another explanation could be that many of the 'one-paper-authors' in DBLP indeed are researchers from other disciplines. They are less likely related to our seeds and therefore under-represented in our

candidate set – which would support our assumption that conferences are good seeds for computer science.

5.2 Relationships

How are computer scientists connected with each other on Twitter? Analyzing the different ways people connect through the follow, mention, or retweet relation lays the foundation for understanding how the different connections on social media could correspond to traditional citations on peer-reviewed publications. Therefore, we quantify which fraction of positive and verified candidates follows, mentions, or retweets which fraction of the same or the other set (cf. Figure 7). From the relative values, follow and mention are very similar. The values for follow are much higher than for mention (at least 10 percentage points), except for the followed verified candidates that are followed by verified candidates (69.4%). The amount of users following one of the positive candidates is very high (91.8% and 93.3%) although the conferences are not contained in $\mathcal{C}^+$. In general, the order of activity is follow, mention, retweet.

5.3 Popularity

Who are the most influential computer scientists on Twitter? We apply some commonly used and intuitive influence measures used in [2]. These measures are the number of followers, the number of retweets and the number of times the user is mentioned in other users' tweets. However, unlike in [2] our use of the measures is contextualized to the community of interest (i.e., computer scientists). Thus, the numbers are restricted to $\mathcal{C}^v \cup \mathcal{C}^+$, e.g., 'retweet' is the number of retweets by users in $\mathcal{C}^v \cup \mathcal{C}^+$. Table 6 shows the ranking positions for the top 20 users from $\mathcal{C}^v$ by the number of retweets, followers, and mentions. As was previously found in [2], we observe that the number of followers alone is not a good indicator of the popularity of users. We can see that users with a top ranking by followers are not necessarily among those with a top ranking by retweets or mentions. The fact that many users with a rather low number of publications on DBLP appear among the top indicates on the one hand that people from industry are relevant for researchers and on the other hand that the number of publications in DBLP could be used to further adjust the focus to scientists, though without considering the age of publications it would mean that most younger researchers would be omitted.

Next, we inspect the 'bio' field in the profiles of the top influential users in $\mathcal{C}^v \cup \mathcal{C}^+$ to understand how they differ from the rest. Figure 8(a) shows the words for the top 200 influential researchers by all three measures, and Figure 8(b) the top words in the profiles of the remaining users. Influential researchers tend to be *professors*, *scientists*, and *researchers*. The remaining users are *PhD* students in *computer science*, *software developers*, *engineers*, *designers*, and *artists*.

(a) top 200 (b) rest

Figure 8: Keywords in the bio of positive candidates.

Table 6: Popularity of researchers in $\mathcal{C}^v$ ranked by the number of researchers ($\mathcal{C}^v \cup \mathcal{C}^+$) who retweet (r), follow (f) and mention (m) them, and the number of publications (p) in DBLP.

		ranking by			
screen name	name	r	f	m	p
@timoreilly	Tim O'Reilly	1	2	1	16
@billgates	Bill Gates	2	1	2	1
@hmason	Hilary Mason	3	9	3	2
@zephoria	Danah Boyd	4	7	6	24
@csoghoian	Christopher Soghoian	5	51	12	5
@doctorow	Cory Doctorow	6	16	4	2
@ioerror	Jacob Appelbaum	7	30	7	5
@mattmight	Matthew Might	8	47	16	34
@kentbeck	Kent Beck	9	18	17	35
@mattcutts	Matt Cutts	10	15	9	2
@timberners_lee	Tim Berners-Lee	11	3	5	35
@codepo8	Christian Heilmann	12	87	14	1
@mattblaze	Matt Blaze	13	60	25	72
@digiphile	Alex Howard	14	42	13	1
@paulg	Paul Graham	16	11	10	23
@migueldeicaza	Miguel de Icaza	15	32	21	1
@mralancooper	Alan Cooper	17	80	27	5
@werner	Werner Vogels	18	14	19	36
@timbray	Tim Bray	19	37	18	7
@pogue	David Pogue	20	8	8	8

Table 7: Most retweeted tweets – scientific above, non-scientific below.

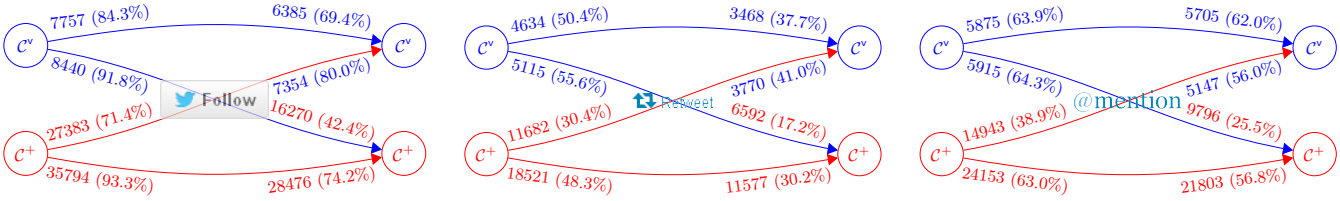
@jimmy_wales	Student warning! Do your homework early. Wikipedia protesting bad law on Wednesday! #sopa
@timberners_lee	The Web Index launched just now http://t.co/CyJH7gRb @webfoundation well done! #opendata #oneweb
@kapravel	probably the saddest graph ever shown in a CS conference. #Linux #tshirtsize pic.twitter.com/RWBnTSJ2ip
@josh_wills	Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
@timberners_lee	Aaron dead. World wanderers, we have lost a wise elder. Hackers for right, we are one down. Parents all, we have lost a child. Let us weep.
@jaykreps	Trick for productionizing research: read current 3-5 pubs and note the stupid simple thing they all claim to beat, implement that.

5.4 Topics

One interesting question to address is what the popular tweets are among computer scientists. Broadly, we can classify the content as being scientific and non-scientific. We define a scientific tweet as one that contains at least one scientific hashtag or mention (as described in Section 3.5). Table 7 shows popular tweets ranked by the number of retweets from $\mathcal{C}^v \cup \mathcal{C}^+$. The scientific talk mostly concerns top events, e.g., #sopa and the dissemination of scientific articles from conferences, whereas the non-scientific talk is mostly humor and private life (e.g., *death*) of researchers.

6. CONCLUSION & FUTURE WORK

In this work we presented a framework for the generation of discipline-specific directories of researchers using Twitter. We proposed automatic methods for assisting the tedious task of finding seeds by searching for conference accounts using Twitter and web search. We also showed ways to au-



(a) Who follows whom? (b) Who retweets whom? (c) Who mentions whom?

Figure 7: Relationships between computer scientists on Twitter.

tomatically leverage ground-truth data by matching against author names in digital libraries. Using machine learning we were then able to build a set of probable computer scientists' Twitter accounts. Our approach lays a foundation for more detailed analyses and understanding of researchers and science in social media at a large scale.

As next steps, we plan to investigate the impact of the seed selection process and of different numbers of expansion rounds in Step 2 on our findings. Likewise, we want to verify the negative examples and improve the matching accuracy in Step 3. A grouping of researchers by their areas of interest (e.g., artificial intelligence, databases, etc.) would help us to answer questions such as *Which differences in the activity of the different research areas are there on Twitter?*, *How diverse or homogeneous are users in a given area?*, or *Which relations exist between the communities of interest?* For the verified users, the publications they have written are a good source to identify their expertise and interests. Another resource to identify expertise are Twitter's *lists*. A first analysis revealed that the 9 191 verified candidates maintain 12 826 lists with 45 270 unique users. Since only 1 150 (3 369) of those users are contained in $\mathcal{C}^{\mathrm{v}}$ ($\mathcal{C}$), the coverage will be lower than our approach. We further want to investigate if our approach can be transferred to other disciplines such as the humanities. This would help us to investigate the connections between different disciplines. Finally, we want to build a web application – a directory of researchers – which features different disciplines and the recommendation of tweets, users, and posted URLs.

7. ACKNOWLEDGEMENTS

This work was performed in the context of the Leibniz Research Alliance 'Science 2.0'.[22]

8. REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proc. 4th Int. Conf. on Web Search and Data Mining*, WSDM '11, pages 65–74. ACM, 2011.

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proc. 4th Int. Conf. on Weblogs and Social Media*, pages 10–17. AAAI, 2010.

[3] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on Twitter. In *Proc. 5th Int. Conf. on Weblogs and Social Media*, pages 89–96. AAAI, 2011.

[4] G. Eysenbach. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4), 2011.

[5] R. E. Goldstein, P. B. Warren, and R. C. Ball. Shape of a ponytail and the statistical physics of hair fiber bundles. *Phys. Rev. Lett.*, 108(7):078101, Feb. 2012.

[6] S. D. Gollapalli, C. Caragea, P. Mitra, and C. L. Giles. Researcher homepage classification using unlabeled data. In *Proc. 22nd Int. Conf. on World Wide Web*, WWW '13, pages 471–482. International World Wide Web Conferences Steering Committee, 2013.

[7] M. Hilbert and P. López. The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. Int. Conf. on World Wide Web*, pages 591–600, 2010.

[9] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to widely spread scientific messages. In *Proc. Web Science Conf.*, 2010.

[10] M. Ley. DBLP: some lessons learned. *Proc. VLDB Endow.*, 2(2):1493–1500, Aug. 2009.

[11] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in Twitter. In *Proc. Int. Conf. on Knowl. Discovery and Data Mining*, pages 430–438, 2011.

[12] J. Priem and K. L. Costello. How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.

[13] J. Priem and B. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7), 2010.

[14] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in Twitter. In *Proc. 2nd Int. Workshop on Search and Mining User-Generated Contents*, pages 37–44. ACM, 2010.

[15] I. Weber, V. R. K. Garimella, and A. Teka. Political hashtag trends. In *Proc. European Conf. on Information Retrieval Research*, pages 857–860, 2013.

[16] K. Weller, E. Dröge, and C. Puschmann. Citation analysis in Twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In *Proc. ESWC 2011 Workshop on 'Making Sense of Microposts'*, pages 1–12, 2011.

[17] J. Wu, P. Teregowda, J. P. F. Ramírez, P. Mitra, S. Zheng, and C. L. Giles. The evolution of a crawling strategy for an academic document search engine: whitelists and blacklists. In *Proc. 3rd Annual Web Science Conf.*, pages 340–343. ACM, 2012.

[22] `http://www.leibniz-science20.de/`