Mining Scholarly Communication and Interaction on the Social Web

Asmelash Teka Hadgu
supervised by Prof. Dr. Robert Jäschke
L3S Research Center
Appelstraße 4, 30167 Hannover, Germany
teka@l3s.de

ABSTRACT

The explosion of Web 2.0 platforms including social networking sites such as Twitter, blogs and wikis affects all web users: scholars included. As a result, there is a need for a comprehensive approach to gain a broader understanding and timely signals of scientific communication as well as how researchers interact on the social web. Most current work in this area deals with either a low number of researchers and heavily relies on manual annotation or large-scale analysis without deep understanding of the underlying researcher population. In this proposal, we present a holistic approach to solve these problems. This research proposes novel methods to collect, filter, analyze and make sense of scholars and scholarly communication by integrating heterogeneous data sources from fast social media streams as well as the academic web. Applying reproducible research, contributing applications and data sets, the thesis proposal strives to add value by mining the social web for social good.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*; H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Scholars; Scientific Content; Altmetrics; Social Web

1. PROBLEM

More and more researchers are using social media to connect with other researchers, to disseminate their research work and to keep up to date with the latest work in their research. The explosion of scientific publications coupled with new forms of interaction on the social web such as Twitter, blog posts, etc. pose several challenges to filter relevant scientific content, identify experts and measure impact.

Expert Finding.

Discovering experts across different areas of science, and even between fields of research within a discipline is not easy, e.g., young PhDs who are interested to follow the feeds of experts in a specific area do so manually. This requires the identification of researchers on the social web by combining heterogeneous data sources such as: Twitter, Wikipedia, digital libraries and blogs. The main challenge is disambiguating and linking entities across different sources.

Filtering Scientific Content.

Traditionally, peer-review has served to filter scientifically sound works from those that are not. With the increasing participation of researchers posting scientific content on social media streams, automated methods that serve peer-review like mechanisms are viable to filter scientific content. If we take a blog post or a tweet, the core challenges are: defining scientifically relevant content and building scalable methods to filter those scientific from non-scientific.

Measuring Public Attention and Impact.

With the increasing adoption of the social web by researchers as a means to disseminate their scientific discoveries, share important links, etc., there is a need for a broader, more comprehensive and timely approach to study the interaction among researchers and measure the attention and impact of their work beyond the traditional citation methods. The challenge is how to build models for drawing accurate conclusions about data gathered from heterogeneous sources.

Personalized Recommendation and Ranking.

Even after filtering experts and science related content from noise, the sheer amount makes it difficult on which essential users or content to focus. The problem is how to build recommender systems that consider the changing focus, level of expertise and interest to recommend diverse and relevant content and users.

The expected contributions of this proposal include:

- methods for profile linking and benchmark data sets about researchers on the social web
- automated methods for filtering, tracking and ranking scientific content on the social web
- a system that integrates the social web and the academic web to feature a directory of researchers, trending articles and personalized recommendations.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741749 .

2. STATE OF THE ART

In this section we survey the state-of-the-art in (i) entity disambiguation and linking (ii) making sense of social media and (iii) altmetrics that are the foundations for our work.

2.1 Named Entity Linking

Entities across different sources on the Web are difficult to track because an entity can be referred to by different strings, and the same string may be used to refer to multiple entities. Consider the name of a researcher e.g., Gregory Piatetsky on Twitter and Gregory Piatetsky-Shapiro on DBLP refer to the same person - an expert in data mining. This makes it difficult to study researchers across different sources without first linking them. Named Entity Linking (NEL) is the task of resolving named entity mentions in news documents, blog posts, tweets, queries etc. to entries in a knowledge base (KB), e.g., Wikipedia,[1] DBpedia,[2] YAGO,[3] Freebase[4] etc.

An exhaustive and detailed survey of entity linking systems is given in [19]. Here we mention works that are directly relevant for our research. Milne and Witten [12] used Wikipedia to identify significant terms in an unstructured text and link them to their corresponding Wikipedia articles. They start with unambiguous Wikipedia senses and compare each possible sense with its relatedness to the context sense candidates. In [13] Pilz et al. derive topics using Latent Dirichlet Allocation (LDA) to compare the entity mention's context with candidate entities in Wikipedia represented by their respective articles. Instead of linking name mentions in a document by assuming them to be independent, Han et al. [10] perform graph based collective entity linking where the name mentions in the same document are linked jointly by exploiting the interdependence between them. A comparison of three seminal named entity linking systems was performed by Hachey et al. [8]. They found that most NEL systems differ in their candidate entity search strategies instead of where most systems in the literature focus which is candidate ranking.

The main drawback of these approaches is that they are targeted at linking entities to a general KB and need to be adapted to work well for linking (i) researcher profiles and (ii) scientific articles across different sources.

2.2 Making Sense of Social Media

Social media streams such as Twitter are ideal for capturing real-time reactions to every day communication. Twitter has been used for studying a wide range of applications including: politics [2], economy [1], health and well being [16] and romantic relationships [4]. In this proposal, we argue that we can leverage social media for science.

At their core, these applications depend on making sense of these short 140 character posts. However, standard NLP tools that perform very well on edited content such as news corpora degrade in performance when applied to short posts such as tweets. Ritter et al. [15] used LabeledLDA to exploit Freebase dictionaries to build a tool for POS tagging, chunking and named entity recognition on Twitter. Similarly in [5] Gimpel et al. developed a POS tagger for Twitter. They developed tagsets that are tailored for Twitter, annotated 1,827 tweets, engineered features including regular expression style rules to detect mentions, hashtags and URLs and built a POS tagger in a supervised machine learning scheme using conditional random fields.

2.3 Altmetrics

The explosion of scientific publications implies we can no longer rely on traditional filters such as peer-review and citation counts alone. However, the emergence of researchers embracing the social web allows us to look into new ways to filter and track attention of scientific content. These altmetrics[5] reflect broader and timely impact of scientific communication. There have been previous works comparing citation and article mentions on Wikipedia, blogs and Twitter.

Samoilenko et al. [17] studied whether having a presence in Wikipedia correlates with higher academic ranking as measured by citation counts. They examined 400 biographical Wikipedia articles on academics from four scientific fields and found no statistically significant correlation between Wikipedia articles metrics and academic notability. Similarly, Shuai et al. [20] investigated if scholarly references and mentions on Wikipedia correlate to scholarly citation. Contrary to [17], they found that academic and Wikipedia impact are positively correlated. In [18] Shema et al. investigated ResearchBlogging.org[6] to investigate whether articles receiving blog citations close to their publication time receive more journal citations later than the articles in the same journal published in the same year that did not receive such blog citations. They found that blog citations can be used as alternative metric source. Eysenbach [3] explored how a citation in a tweet, mentioning a journal article URL, compares to citations in peer-reviewed articles. He found that these metrics are correlated and tweets can predict highly cited articles. Similarly in [14] Priem et al. studied the use of social media to explore scholarly impact. Among others, they found that altmetrics and citations track different but related forms of impact.

The main limitations of the current studies are: (i) lack of shared data sets and hence conflicting results and (ii) comparing citation against Wikipedia, blogs or Twitter and not investigating how these signals complement each other.

3. PROPOSED APPROACH

Here we outline a holistic approach to study scholarly communication on the social web. Our general framework shown in Figure 1 consists of four components: (i) linking researcher profiles, (ii) filtering scientific content, (iii) measuring public attention and impact and (iv) ranking and recommendation of scholars and scientific content.

3.1 Linking Researcher Profiles

Problem Statement: Given two facets of researchers: the publishing record from digital libraries such as DBLP on the one hand and their social media use, for instance, on Twitter, on the other, the task is to link profiles representing the same person.

Proposed Solution: Reza Zafarani and Huan Liu [25] provide a methodology for mapping users across several social media sites. They formalize the user linking problem as fol-

[1] http://www.wikipedia.org

[2] http://dbpedia.org

[3] www.mpi-inf.mpg.de/yago-naga/yago

[4] https://www.freebase.com

[5] http://altmetrics.org/manifesto/

[6] http://researchblogging.org/

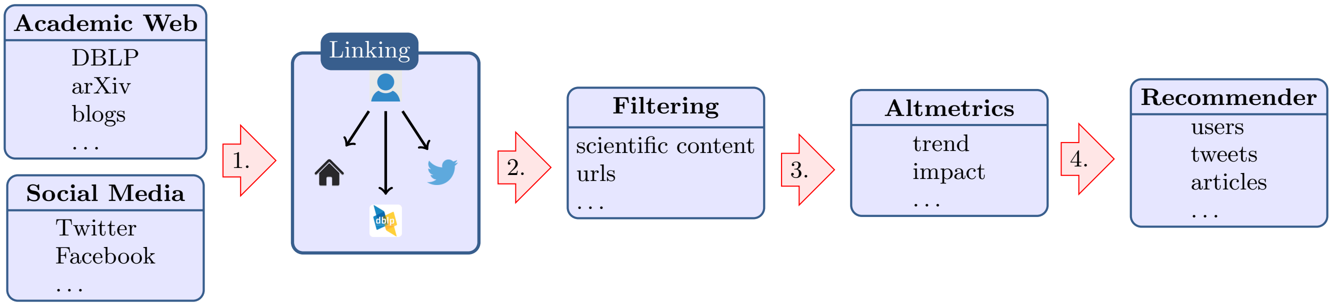

Figure 1: Pipeline of proposed approach.

lows. Given a set of n user names $U = \{u_1, u_2, \ldots, u_n\}$, which is possibly a set of names, representing an individual I in one system and a candidate user name c on another platform. A user identification procedure attempts to learn an identification function $f(\ldots)$ such that

$$f(U,c) = \begin{cases} 1: & \text{If } c \text{ and the set } U \text{ belong to } I; \\ 0: & \text{Otherwise;} \end{cases} \quad (1)$$

E.g., in equation 1, U may represent the publication records of a researcher under possibly multiple names. They construct behavioral features from user names only in a supervised scheme to connect identities. Goga et al. [6] exploit geo-location, time stamp and writing style of posts to link user accounts across social media sites Yelp, Flickr and Twitter. Tang et al. [21] tackled the the problem of researcher profiling using a unified approach to extract profile information such as homepages, affiliation, address, emails, publications etc. from the Web. They used a supervised approach using Conditional Random Fields (CRF) to identify homepages. Finally, to integrate publications from the existing bibliography data sets, they used a constraints-based probabilistic model to name disambiguation. We build on these previous approaches and extend them in the following two important ways: (i) We recognize names are the strongest indicators of a match; however, unlike in [25], we leverage additional features derived by analyzing content such as language and expertise, e.g., conferences they mention. (ii) We relate structural similarities across networks. We hypothesize that researchers that co-author or attend conferences together are more likely to follow, retweet and mention each other. We can combine these features in a machine learning framework to build the identification function $f(\ldots)$.

3.2 Filtering Scientific Content

Problem Statement: How can we identify scientifically relevant content on the social web?

Proposed Solution: Weller et al. [23] list several aspects that describe a scientific tweet: (i) having scientific content or linking to a peer-reviewed scientific article on the web, (ii) any tweet published by a scientist or (iii) a tweet containing science-related hashtags. We build on this intuition and propose a model that aggregates these different aspects. Many unsupervised approaches use pattern matching approaches that do not require labeled examples. They rely on matching certain keywords, hashtags and tuning parameters to classify a tweet as scientific or not. The actual challenge is how to gather science related keywords and hashtags. Using domain knowledge to curate such lists is very expensive and not scalable. One way to build an academic lexicon is to gather terminologies from abstracts of papers in the domain of interest. Another approach is to look into co-occurrence patterns. In [2] the authors start from a small seed of hashtags and gather more of them by looking into other hashtags that co-occur with them. In our case, we can start from, e.g., conference hashtags, and use a similar approach to gather more science related keywords and hashtags. Automatically telling whether a link points to a scientific resource (e.g., scientific blog, peer-reviewed article) on the web can be treated as a classification problem. In [7] Gollapolli et al. used some URL patterns such as the URL domain, existence of tilde character and so on as features to build a classifier for researcher homepage classification. Besides using such features, we propose to build a predefined list of web pages (e.g., digital libraries, slide sharing websites and so on), from which we can perform pattern matching. Finally we can aggregate the signals derived from the previous approaches to build a robust model to identify scientific tweets. This avoids a selection bias which is commonly the case when we take, for instance, only tweets containing certain hashtags, keywords or URL patterns.

3.3 Measuring Public Attention and Impact

Problem Statement: How can we track and measure the public attention and impact of scientific content at scale?

Proposed Solution: Ultimately, scientific discoveries should reach the public. Social media opens up this opportunity. In the same way, public attention can serve as a feedback to complement traditional impact measures. Previous studies have looked into how citation correlates with Wikipedia [17, 20], blog [18] and Twitter [3] mentions. In particular Priem et al. [14] used multiple social media as sources to assess impact of scholarly impact. We build on these works and propose to extend them in the following key aspects. First, by dealing with more researchers and additional conferences and journals. This requires building robust automatic methods of identifying and linking scientific content across social media streams as we pointed out in the previous sub section 3.2. Second, integrating these signals to have a more holistic view of the impact of scientific content taking into account (i) the user reputation, e.g., a computer scientist sharing a link to an article should weigh more than a conference bot sharing the same link and (ii) the weight given to the sources for article mentions, e.g., a link from Wikipedia should weigh more than a link from a personal blog post or a tweet.

3.4 Ranking and Recommendation

Problem Statement: How can we leverage signals from the social web to recommend scientific content (blogs, tweets, articles) and scholars?

Proposed Solution: Having a better understanding of the underlying researcher population and a mechanism to identify scientific content from the previous tasks, we are equipped to tackle interesting challenges for recommending scholars and scholarly content. Kywe et al. [11] provide a comprehensive survey of recommendation systems in Twitter. They provide a taxonomy of several recommendations such as (i) followee recommendation: who to follow? (ii) tweet recommendation: what to tweet about? What URL, hashtag to include in a tweet and (iii) mention and retweet recommendations. Wang and Blei [22] use a collaborative topic regression model, a combination of collaborative filtering and probabilistic topic modeling, to recommend scientific articles. In [24] Younus et al. use tweets within a topic modeling framework to recommend scientific articles. We propose to build user models for researchers in terms of their stage and hierarchy, e.g., academia (PhD student, postdoc, professor) or industry (researcher, senior scientist) and their area of expertise through their publications. We can model the user-to-user relationship among researchers, leveraging their academic relationships such as the co-authorship, citation network and their affiliation besides the Twitter generated interactions. Finally, we plan to implement automatic methods for peer-review-like mechanisms to rank scientific content and scholars by taking into account what articles and researchers get retweeted, favorited and so on by the identified researchers and their research communities in contrast to counts based on the crowd.

4. METHODOLOGY

In this section we describe the methodology and experimental design for the proposed research.

4.1 Data sets

The proposed research requires the integration of heterogeneous data sources mainly from: social media data particularly Twitter and the academic web including digital bibliographic resources such as DBLP,[7] arxiv.org,[8] PubMed[9] etc. We have been collecting Twitter data from the one percent public stream[10] which returns a small random sample of all public statuses since January 2013. Besides the public stream, we have a focused crawler of computer scientists that were identified in [9]. Finally, we have snapshots of university websites from Germany - the German academic web and a subset of the Internet Archive data on the Web about Germany.

4.2 Evaluation

Linking Researcher Profile Ideally, we would like to build a validation data set by asking researchers themselves. For this purpose, we developed a survey application[11] that asks researchers in computer science to validate their DBLP profile by logging in with their Twitter accounts. Another alternative is to use crowd sourcing platforms such as Mechanical Turk.[12] Entity linking systems use standard evaluation measures: *precision*, *recall*, *F_1 measure* and *accuracy*.

Filtering Scientific Content: We plan to perform large-scale labeling of tweets and blog posts as scientific or not using crowd sourcing platforms to validate our models.

Measuring Public Attention and Impact: Some websites such as altmetric.com[13] track article level metrics on the web. We can use such systems to compare and evaluate tracking mentions as well as different experimental setups of computing impact.

Ranking and Recommendation: Evaluating the results of ranking and recommending scientific content on the social web requires a gold standard data set. A first step towards this will be considering the 'like', 'retweet', or 'favorite' counts of articles by researchers.

To measure the effect of the integrated solution, we plan to perform a large-scale user study. We hypothesize that solving the problems in the proposed pipeline brings about a holistic solution to the broader problem than the individual components the way it is done now.

5. PRELIMINARY RESULTS

The proposal is at its early stage. Here we describe a general framework we developed to identify researchers on Twitter [9]. The framework was applied to identify researchers in Computer Science.

The processing pipeline begins with a *seed set* which is used to generate possible candidates. We used Twitter accounts corresponding to computer science conferences as seed users. We collected this from the list of computer science conferences in Wikipedia.[14] We used automatic methods to link them to their Twitter accounts using (i) web search, i.e., searching for the official page of the conference and then extracting the Twitter account from the page, and (ii) Twitter search, i.e., searching on Twitter using potential screen names built from acronyms of conferences with years appended. The next step is *generating candidates.* we gathered candidate researchers that follow the seed accounts or retweeted any of their tweets. Our approach treats the identification task as a classification problem. For this, we generated *labeled examples.* To generate positive examples, we mapped our candidates to DBLP by looking exact name matches after removing duplicates from each set before matching. For negative examples, we collected one million users through the Twitter streaming API and then removed candidates and their followers from the set. Finally, we generated *features* and learned a model to classify the remaining users. The resulting data sets and features used are available on Github.[15] To gain first insights into how researchers use Twitter, we empirically analyzed the identified users and compared their age, popularity, influence, and social network. Our results show that, increasingly researchers are using social media that could be used for expert finding, tracking attention and personalized recommendation.

[7] http://dblp.org/

[8] http://arxiv.org/

[9] http://www.ncbi.nlm.nih.gov/pubmed

[10] https://stream.twitter.com/1.1/statuses/sample.json

[11] http://researchersontwitter.appspot.com/

[12] https://www.mturk.com

[13] http://www.altmetric.com

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

[15] https://github.com/L3S/twitter-researcher

6. CONCLUSION AND FUTURE WORK

In this thesis proposal, we have identified an area of research that bridges the gap between, on the one hand, the long standing tradition of scientific publishing using sources from digital libraries and the academic web which contains mainly researchers' university web pages and scientific blogging; and on the other hand, the emerging practice of social media by researchers to connect and disseminate scientific content. We proposed a holistic approach to study this wide spectrum. This systematic approach contains four pillars that are grounded on the following research challenges: linking researchers' profiles; identifying and filtering scientific content; measuring public attention and impact; and building scalable ranking and recommender systems.

We envision to build a system that harnesses the power of the social web and combines it with scholarly publishing in the academic web using data-driven approaches to make it easy to filter and tap into the best in humanity - science and scientific discoveries.

7. ACKNOWLEDGEMENTS

This work was performed in the context of the Leibniz Research Alliance 'Science 2.0'.[16]

8. REFERENCES

[1] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *Proc. 5th Intl. Conference on Weblogs and Social Media*, 2011.

[3] G. Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *J Med Internet Res*, 13(4), 2011.

[4] V. R. K. Garimella, I. Weber, and S. Dal Cin. From "i love you babe" to "leave me alone"-romantic relationship breakups on twitter. In *Social Informatics*, pages 199–215. Springer, 2014.

[5] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL (Short Papers)*, pages 42–47. The Association for Computer Linguistics, 2011.

[6] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *WWW*, pages 447–458. International World Wide Web Conferences Steering Committee / ACM, 2013.

[7] S. D. Gollapalli, C. Caragea, P. Mitra, and C. L. Giles. Researcher homepage classification using unlabeled data. In *WWW*, pages 471–482, 2013.

[8] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating entity linking with wikipedia. *Artif. Intell.*, 194:130–150, 2013.

[9] A. T. Hadgu and R. Jäschke. Identifying and analyzing researchers on twitter. WebSci '14, pages 23–32, New York, NY, USA, 2014. ACM.

[10] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *SIGIR*, pages 765–774. ACM, 2011.

[11] S. M. Kywe, E.-P. Lim, and F. Zhu. A survey of recommender systems in twitter. In *SocInfo*, volume 7710 of *Lecture Notes in Computer Science*, pages 420–433. Springer, 2012.

[12] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, pages 509–518. ACM, 2008.

[13] A. Pilz and G. Paaß. From names to entities using thematic context distance. In *CIKM*, pages 857–866. ACM, 2011.

[14] J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*, 2012.

[15] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, pages 1524–1534. ACL, 2011.

[16] A. Sadilek and H. Kautz. Modeling the impact of lifestyle on health at scale. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 637–646, New York, NY, USA, 2013. ACM.

[17] A. Samoilenko and T. Yasseri. The distorted mirror of wikipedia: a quantitative analysis of wikipedia coverage of academics. *EPJ Data Science*, 3(1):1–11, 2014.

[18] H. Shema, J. Bar-Ilan, and M. Thelwall. Do blog citations correlate with a higher number of future citations? research blogs as a potential source for alternative metrics. *JASIST*, 65(5):1018–1027, 2014.

[19] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, Feb. 2015.

[20] X. Shuai, Z. Jiang, X. Liu, and J. Bollen. A comparative study of academic and wikipedia ranking. In *JCDL*, pages 25–28. ACM, 2013.

[21] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 292–301, Oct. 2007.

[22] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD*, pages 448–456. ACM, 2011.

[23] 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.

[24] A. Younus, M. A. Qureshi, P. Manchanda, C. O'Riordan, and G. Pasi. Utilizing microblog data in a topic modelling framework for scientific articles' recommendation. In *Social Informatics*, pages 384–395. Springer, 2014.

[25] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. In *KDD*, pages 41–49. ACM, 2013.

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