Connecting the Dots to Infer Followers' Topical Interest on Twitter

Aastha Nigam, Salvador Aguinaga, Nitesh V. Chawla
Interdisciplinary Center for Network Science and Applications (iCeNSA)
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556
{anigam,saguinag,nchawla}@nd.edu

***Abstract*—Twitter provides a platform for information sharing and diffusion, and has quickly emerged as a mechanism for organizations to engage with their consumers. A driving factor for engagement is providing relevant and timely content to users. We posit that the engagement via tweets offers a good potential to discover user interests and leverage that information to target specific content of interest. To that end, we have developed a framework that analyzes tweets to identify the interests of current followers and leverages topic models to deliver a personalized topic profile for each user. We validated our framework by partnering up with a local media company and analyzing the content gap between them and their followers. We also developed a mobile application that incorporates the proposed framework.**

I. Introduction

Twitter, an online social network, enables individuals to write about their daily activities, express opinions, share information and connect with other users and businesses. According to a Twitter report[1], there are approximately 284 million monthly active users, and 500 million tweets shared every day. With such large scale exchange of information, it provides a unique opportunity for various businesses to promote their brands, reach out to their customers, increase awareness and interaction among users about a topic or product and keep them engaged in their content. Presumably, an individual's tweets are reflective of his or her interests, and it provides a compelling opportunity to cater content based on their preferences. This allows companies to not only achieve a broader reach of content but also more keenly engage the followers. Per Twitter[2], following Pareto principle, 80% of the content should be directed towards the current followers to keep them successfully engaged. Consequently, engaging and personalized content can improve the overall experience of a user on social media.

Twitter produces a list of popular trends[3] based on the location of the users and the accounts they are following. However, in order to understand what a company's followers are interested in, we also need to analyze their tweets and tweeting behavior. The relevant information cannot be captured by using only the hash-tags, mentions or keyword search; the objective is to capture user's interests based on their tweeting patterns and understand the latent topics embedded in the content of the tweets.

With the increasing volume of data being shared on social media and growing emphasis on improved user experience, there is a demand for an efficient and scalable framework to understand topical preference of each user. However, a user expresses an opinion about a topic using a limited number of words (140 characters) through a tweet. Identifying a topic or a group of topics from the limited text in a tweet becomes extremely challenging. Other than limited text, a tweet also suffers from ambiguity. An ambiguous tweet can be defined as one which can potentially have more than one meaning. Users have different ways to express their opinion which can lead to unclear topics.

To that end, we present a user centric topic discovery framework to identify content a user is, or might be, interested in. The proposed framework generates a topic profile for each user encompassing high level topics derived from their tweets. To combat limited text and ambiguity the model utilizes the top content delivered by the search engine in response to the keywords extracted from the tweet. This topic profile can then be leveraged to personalize content for the individual following the company or organization on Twitter. Our work has direct industry relevance and recommends a translation of academic research for business applications in collaboration with a company.

This personalized topic profile for each individual also allows companies to identify the gap between the content they publish and the information consumed by their users, giving businesses an opportunity to post more relevant content to keep their users engaged. We evaluate our work on the data acquired from a media company, which focuses on various business segments such as newspaper publishing, digital media, broadcasting radio and cable TV. The company uses social media platforms such as Twitter and Facebook to connect with their users by sharing articles on their social media pages. We focused on the twitter accounts for two of their digital properties: a local newspaper and a television station. These twitter accounts have a sizable follower base of twenty thousand and thirty five thousand followers, respectively. We have also developed a real-time mobile application prototype for analyzing the content consumed by followers based on time and location.

[1]Twitter internal data, 2014
[2]https://business.twitter.com/basics/how-to-create-a-twitter-content-strategy
[3]https://support.twitter.com/articles/101125

The main contributions of this work can be seen as follows:

1) A user centric topic based framework to identify, understand and potentially bridge the content gap between companies and their followers on Twitter.
2) Topic based personalized profiles for users to represent their interests leveraging open source information.
3) A real time mobile application for analyzing the content amongst the followers based on time and location.

Organization. The remainder of this paper is organized as follows. In Section II, we review related work conducted in this domain. We then describe the data used for this study and the pre-processing steps in Section III. In Section IV, we present our framework to understand topic preferences for users. Then in Section V we present our results and illustrate the visualization platform. Lastly, we provide our conclusions in Section VI.

II. Related Work

One of the initial approaches to understand topics from text leveraged the term frequency and the inverse document frequency model [1]. The most discriminative sets of words obtained from the method were deemed as the topic. However, usage of single words as topics could not represent complicated topics and were not able to reduce the text. Blei *et al.* [2] developed a probabilistic topic model which discovered underlying thematic (or semantic) information in large collection of documents. One of the most commonly used topic models is called the Latent Dirichlet Allocation (LDA). Using LDA, the structure of the document is characterized by three phenomenas: distribution of topics across the corpus inferred from a fixed vocabulary, the distribution of topics in one document and a weight assignment for each word in each document to depict which topic it belongs to.

Tweets pose an interesting challenge as they are much shorter than traditional text and also capture diverse topics [3]. Research has been done to analyze the content of tweet using various approaches. Probabilistic topic models are not able to perform well in situations where there is not enough data. Consequently, direct application of such models to tweets does not yield good results [4]. Common approach to combat this problem is to combine all tweets by a user into one document and then perform LDA on the user aggregated profiles. This is sometimes also referred to the author-topic model [5]. Twitter-LDA [6] is another approach which assumes that a single tweet is usually about a single topic and compares the data on Twitter with other traditional forms of media such as newspaper articles. On the other hand, Bernstein *et al.* [7] proposed a technique called TweeTopic that uses parts-of-speech tagger on the tweet to create a query which is fed into the Yahoo! Build Your Own Search Service. The pages returned from this search are used to understand the topic using inverse document frequency. Similarly, Micheson *et al.* [8] extracted the named entities in a tweet to understand the higher level concepts using Wikipedia categories.

In our work, we firstly assume that a tweet can capture an array of topics based on a user's opinion. Secondly, we augment the user's opinion with open source information to understand the context better. In addition, we perform topic modeling to discover latent topics embedded in the text and do not depend on any predefined categories. This helps in building a comprehensive and personalized topic profile for each user. Lastly, we provide a convenient way to monitor the trends and changing interests of users using a mobile application.

III. Data

In this section, we explain our data sources and various stages of cleaning and pre-processing.

A. Collection

The data used for this research originated from a local U.S.-based news and information company operating in radio, TV and news media segments. To communicate and connect with their users, each of their businesses (newspaper or TV or radio) have a Twitter account through which content (news articles and videos) published is shared. The company annotates each Twitter follower based on their activity as prolific or average user. A prolific tweeter can be defined as a follower who (re)tweets the company's content and has a network of followers significantly larger when compared to an average user. A list of prolific tweeters and their tweets (obtained using Twitter Streaming API) were provided to us. However, to study the general content gap between companies and users, we do not focus on prolific users as it subjects the study to a biased perspective. To analyze an average user we collected random follower tweets using the Twitter API. Twython[4], a Python wrapper for Twitter API, was used for the aforementioned task. In addition, to contrast the content between the producer and consumer, we collected tweets for overlapping time periods from the business's Twitter accounts.

Tweets from users and companies are sampled on a periodic basis. We use a *cron* job to sample the last ten tweets from users every hour. Sampling tweets from companies was done in a similar fashion. The interval is a parameter setting that can be easily modified depending on the size of both the company and the user set. Table I shows a snapshot of our data set, depicting the number of tweets for prolific users and average users sampled for a 24 hour period.

TABLE I
Number of tweets collected for average and prolific users in a 24 hour period.

Average User Tweets	
Property	Count
Local newspaper Twitter page	1425
Local Television Twitter page	1335
Prolific User Tweets	
Local newspaper Twitter page	1878
Local Television Twitter page	1885

[4]https://github.com/ryanmcgrath/twython

Tweet:
'I have completed the quest 'Another Try 'in the #Android game The Tribez. http://t.co/uHpflTNi57 #androidgames, #gameinsight'

Preprocessed and cleaned:
i have completed the quest another try in the android game the tribez URL androidgames gameinsight

Broken down to its components:

```
[(u'i', 'PRP'),(u'have', 'VBP'),
(u'completed', 'VBN'), (u'the', 'DT'),
(u'quest', 'JJS'), (u"'another", 'JJ'),
(u'try', 'NN'), (u'in', 'IN'), (u'the',
'DT'), (u'android', 'JJ'), (u'game', 'NN')
, (u'the', 'DT'), (u'tribez', 'NN'),
(u'URL', 'NNP'), (u'androidgames', 'VBZ'),
(u'gameinsight','NN')]
```

Fig. 1. Breakdown of a tweet from it raw text form to its tagged components.

B. Pre-processing

Each retrieved tweet is associated with varied attributes such as: tweet ID, time of the creation, text of the tweet, re-tweet count (the number of times it has been re-tweeted), URLs shared in the tweet, hash-tags, users mentioned in the tweet, favorite count and if the tweet was in reply to some other user. We extracted tweet ID, text and the URLs present in the tweet. Punctuations and emoticons can be indicative of user's sentiment towards a given topic, however do not communicate any information about the content of the tweet. Therefore, the tweet was cleaned out of any unnecessary characters using regular expressions.

Figure 1 shows a brief example of different stages of pre-processing. We begin with a collected tweet which has been cleaned out of characters such as '#'and ',. We also convert all words to the lower case. We will describe the last step shown in Figure 1 in Section IV.

IV. Framework

Tweets have a noisy and complex structure [8]. To understand latent topical interests of a user, the most representative words from the text of a tweet were extracted. Previous research has demonstrated noun phrases to capture main concepts of text better than other parts of speech [9]. Therefore, we use a Parts-Of-Speech (POS) [10] tagger, that classifies each word in a sentence based on its syntactic function into categories such as noun, verb and adjective, to extract the nouns. Short text (140 characters) and insufficient information (context) restricts the performance of most commonly used implementations [11]. However, CMU ARK Twitter Part-of-Speech Tagger [12] performed well with our data since it has been trained over a collection of tweets.

We leverage the search engine's ability to augment our limited knowledge with the database of millions of documents. We built a query using the nouns extracted from each tweet [8] and arranged them in the search query maintaining the same order as they appeared in the tweet. We refer to the query built using the nouns in a tweet as the *tweet query*. If a tweet was found to have no nouns, the entire tweet text was utilized as the search query. The tweet query was fed into the Google search engine[5] [13] to retrieve the top 20 documents, ranked based on their relevance. Although the utilization of only nouns may cause a loss of context for the tweet, but this is compensated for as we believe the topic is centralized around nouns of the tweet. Since Twitter is a real time information network, we argue that the true context can be captured from the results returned by the search engine. Additionally, we do not bias the search engine results by specifying any time or location variables since we want to retrieve the most related content and keep the upper limit of documents to top 20. This not only helps in enriching the text of tweet with publicly available information on Google, but we are also able to improve any ambiguity in the context of the tweet since we leverage the most current and trending information on the web. After retrieving at most top 20 documents for each tweet, we crawl each url to get the content on that page. This article extraction was done in Python using the newspaper library[6]. The text of each tweet is appended with data from the top 20 searches to create a *tweet document*. We define the *tweet document* as a document containing an enriched tweet, mainly a tweet appended with related information obtained from the search engine. By performing this, we are no longer restricted by the limited content in a tweet and can utilize existing topic models. We then performed some pre-processing tasks on the data such as removing the stop words. Before applying topic modeling on the *tweet document*, we further performed data cleaning using the Term Frequency-Inverse Document Frequency (TF-IDF) score [14]. To decide the tf-idf score threshold we experimented with different percentile values and found that 95 percentile was the most favorable. All words below this threshold value were removed.

Using the set of *tweet documents* we built a dictionary of words and represented each tweet document in a vector format. We then performed Latent Dirichlet Allocation (LDA) [2], a popular topic modeling approach, on the document corpus. LDA is a generative probabilistic model where we assume the data is created from a generative process over observed and hidden random variables. In this case, the observed variables are the words of the documents and the hidden variable is the topic structure. In a generative process, a model calculates the joint probability distribution to understand the conditional probability of the hidden variable given the observed variable. We used the gensim[7] implementation of LDA in python. Using LDA, the topic structure of the document is characterized by three phenomenas: distribution of topics across the corpus inferred from a fixed vocabulary, the distribution of topics in one document and a weight assignment for each word

[5]https://www.google.com
[6]https://pypi.python.org/pypi/newspaper
[7]https://radimrehurek.com/gensim

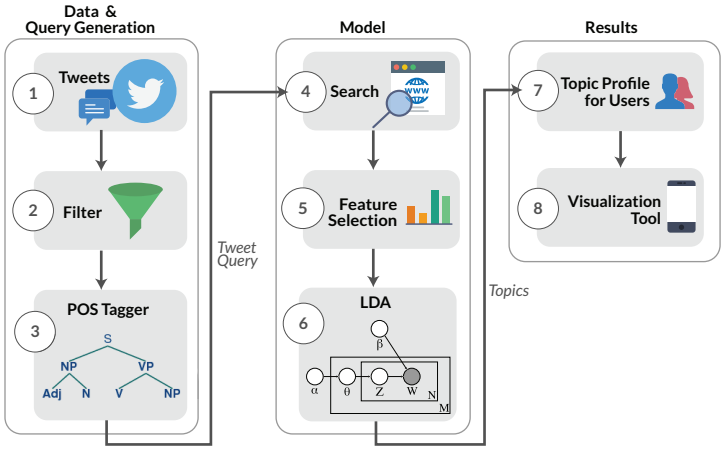

Fig. 2. Framework: An overview of the user centric topic based framework captured in steps [1-8] from data collection to visualization platform. For better comprehension the steps have been broadly separated into: 1) Data & Query Generation 2) Model and 3) Results.

in each document to depict which topic it belongs to. Since each document is represented as a mixture of topics with a probability distribution of words constituting the topic, we obtain a list of topics over our entire tweet data set and also get the topic contribution in each tweet.

Our framework as shown in Figure 2 gives a distribution of various topics constituting a tweet and in turn enables us to define a topic profile for each user. A topic profile for a user characterizes the main topics inferred from an individual's tweeting behavior. This permits us to have a more personalized user centric topic profile to better understand the content preferences of the followers. With large number of documents or in case of a stream of documents, Online LDA [15] has proved to scale well.

Figure 2 summarizes our entire framework, which can be captured by the following steps:

1) Collect tweets for the business and their followers.
2) Clean out tweets of emoticons and unwanted characters using regular expression. Extract the text of the tweet.
3) Identify the nouns of the tweet using the CMU ARK Twitter Part-of-Speech Tagger.
4) Generate a *tweet query* using the nouns in the tweet (keeping the same order as seen in the tweet). Feed the tweet query into any search engine such as Google and obtain at most top 20 results (urls).
5) Crawl each of the urls returned for each tweet to create a *tweet document*. Perform basic pre-processing on the document such as removing stop words. Perform feature selection using tf-idf score.
6) Use the processed *tweet documents* as the corpus to train a topic model - Latent Dirichlet Allocation (LDA).
7) Identify the topics in each tweet based on the results of LDA. Based on the topics identified in the tweets, generate a topic profile for each user.
8) Visualize the results on the mobile application.

V. Results

Having explained the framework, we present a case study with the local media company and introduce a mobile application that we developed for inferring follower's interests.

A. Personalized User Profile

As described earlier in Section III-A, we collected a set of tweets from the followers of the company. In this section, we test our framework on a collection of 880 tweets from a sampled user set for a given time period. From each tweet, a *tweet query* was generated for the Google Search Engine. The cleaned and filtered data from top retrieved pages are condensed into a *tweet document* per tweet. The LDA model is trained on this corpus of *tweet documents*. We retrieve 100 topics from the corpus which represent an overall topical interest of the users. On experimentation, we found 100 topics were most ideal for our tweet data set. Given the list of tweets for an individual and a distribution of topics for the tweet as obtained from LDA, we can infer the prominent topics a user usually tweets about or is most interested in. This results in a user centric topic profile for each individual to capture their preferences.

On an average, each tweet was assigned 1.116 topics. The maximum number of topics assigned to any tweet was 6. The tweets were identified as covering mostly one topic. We also studied the distribution of the 100 topics across the corpus. Figure 3 illustrates the presence of each topic across the data set. From amongst the 100 topics, Table II lists three of the topics. The first deals with national politics involving words such as Obama, immigration and tax. The other two topics cover two national incidents. The first one covers the Ferguson unrest[8] as can be seen from the word usage such as police, Wilson and Ferguson. The second topic in the national news captures the murder of Skylar Neese[9] including words such as Skylar Neese and the two suspects, Rachel and Shelia.

[8]http://en.wikipedia.org/wiki/2014_Ferguson_unrest
[9]https://en.wikipedia.org/wiki/Murder_of_Skylar_Neese

TABLE II
AN ILLUSTRATION OF 3 TOPICS FROM A 100-TOPIC MODEL FOR THE TWITTER DATA SET. EACH TOPIC IS SHOWN WITH TOP WORDS THAT CONSTITUTE IT AND THEIR ASSOCIATED PROBABILITIES. EACH TOPIC HAS ALSO BEEN GIVEN A HIGHER-LEVEL LABEL FOR EASIER COMPREHENSION.

National Politics		National news		National news	
word	**probability**	**word**	**probability**	**word**	**probability**
obama	0.023	police	0.014	skylar	0.016
immigrants	0.015	brown	0.012	neese	0.013
immigration	0.015	wilson	0.007	shelia	0.013
president	0.013	switch	0.007	learn	0.011
tax	0.007	ferguson	0.007	rachel	0.010
learn	0.007	stores	0.007	stores	0.009
ocean	0.005	learn	0.006	switch	0.009
border	0.005	sorry	0.006	world	0.008
country	0.05			hell	0.007
nation	0.05			needs	0.007

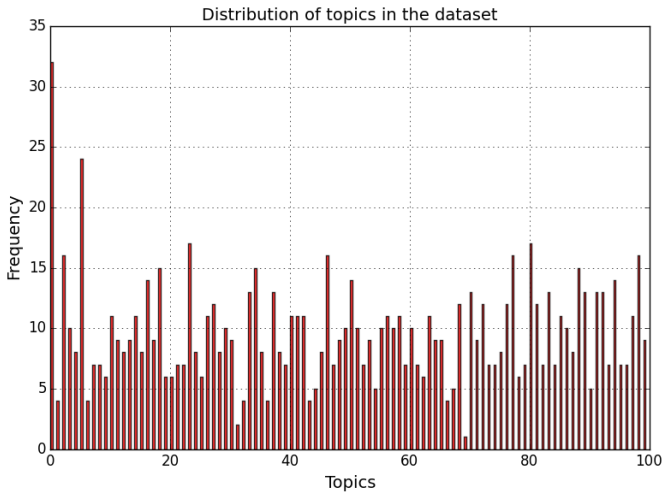

Fig. 3. The plot depicts the prevalence of 100 topics obtained from LDA across our data set.

On comparing the content with the information being shared on Twitter by the company, it was seen that the followers were talking more about news at a national level (as shown in Table II) but the company was mostly sharing tweets about local news and local sports. This clearly demonstrates a gap in the content being distributed by the local news media versus the topics of potential interest to the users. We posit that armed with such information the company can be more targeted in engaging with its Twitter followers not only with the right content but also timely content.

B. Visualization

In addition to providing a solution for finding trending topics amongst the followers, we also provide a prototype for visualizing and analyzing the content in real time. We explore an application-specific data visualization interface for a mobile platform. Section IV described the framework that yields various outputs organized for consumption on mobile devices. The organization of the output, centers around the presentation of *what followers are talking about*. The mobile application enables the companies to see the trending topics across their followers in real time. The visualization displays the set of words associated with each topic and also the topics attributed to a given tweet. The words and topics are visualized as a table-view and a word (or tag) cloud. The size of the word in the tag cloud indicates the importance of the word in the topic. The application also entails a user centric view where we list the most trending topics of the user. Additionally, on selection of a topic, the associated actual tweets are listed for reference. The visualization is prototyped on the iOS platform. Figure 4 shows screenshots of the actual implementation running on the iOS Simulator. Another feature of the app, currently under exploration, is geo-location visualization showing tweet provenance. The mapview offers a picture of where the tweets originate from. This type of information can be used to explore topics emerging around local news or events.

Fig. 4. Application prototype to visualize the trending topic amongst the followers of a company.

VI. CONCLUSION

In the previous section we discussed the experiment setup and the visualization prototype. In this section, we discuss the usability of our framework, conclude and discuss future work in this direction.

A. Potential users

As an application of our framework, we see an opportunity to identify potential new followers for a company's Twitter

page. We leverage the network of the company's current followers by building a topic profile for each potential user and studying the overlap in the interests between the user and the company.

As described in Section IV, the framework outputs a user centric topic profile for each follower. To find the potential followers for a company, we look at the prolific users (as described in Section III-A). Since the aim of the company is to increase their follower base, we assume that it will be most effective to consider the users who are most active. Therefore, to generate a candidate set of followers, we can crawl the follower network of each prolific user. In order to understand, which user would be most likely interested in consuming the content of the company, we need to understand about the topics the candidate follower is interested in. We can apply the discussed framework to build a user based topic profile and contrast the user's interest with the company's content. The candidate users with the maximum overlap would be the potential followers.

B. Summary

Twitter is a great medium for corporations to connect with users and to share content with them. In order to sustain the engagement of the users, it is imperative to keep pushing content which interests the users. Many-a-times companies are unaware what exactly excites the user, therefore it is crucial to understand what the followers want to read about. Our work focuses on bridging this content gap between the two groups: users and company. We present a framework that constructs a personalized topic profile for each user based on their tweeting habits, which enables the company in understanding what interests their followers on social media. In order to validate our framework, we present a real life application of our work on the Twitter pages of a local media company. We also provide a real time visualization application to monitor the change of topical interests in the users over time.

C. Future Work

We would like to incorporate a method to compare two content models and score them based on their (dis)similarity to quantify the interest gap over time. In addition, analyzing how topical preferences change over time amongst the users would be helpful. We would also like to include additional search engines such as Bing to minimize any bias in data collection and utilize other social network sites such as Facebook and Google+, to obtain a more holistic view of the consumer base.

ACKNOWLEDGMENT

The research is supported in part by NSF grant ACI-1029584 and IIS-1447795.

REFERENCES

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(88)90021-0

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[3] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 379–388. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002521

[4] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proc. ICWSM 2010*. American Association for Artificial Intelligence, May 2010. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=131777

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964870

[6] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ser. ECIR'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 338–349. [Online]. Available: http://dl.acm.org/citation.cfm?id=1996889.1996934

[7] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '10. New York, NY, USA: ACM, 2010, pp. 303–312. [Online]. Available: http://doi.acm.org/10.1145/1866029.1866077

[8] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: A first look," in *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, ser. AND '10. New York, NY, USA: ACM, 2010, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/1871840.1871852

[9] M. Bendersky and W. B. Croft, "Discovering key concepts in verbose queries," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 491–498. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390419

[10] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, ser. EMNLP '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 63–70. [Online]. Available: http://dx.doi.org/10.3115/1117794.1117802

[11] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, ser. CSLDAMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 80–88. [Online]. Available: http://dl.acm.org/citation.cfm?id=1866696.1866709

[12] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1524–1534. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145595

[13] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the Seventh International Conference on World Wide Web 7*, ser. WWW7. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117. [Online]. Available: http://dl.acm.org/citation.cfm?id=297805.297827

[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[15] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," in *In NIPS*, 2010.