Visualizing the contributions of a user in the Social Web

AUTHORS:
Maria Schmidt
Maria-Stamatoula Karavolia

UNIVERSITY:
University of Fribourg
COURSE NAME:
Web Monitoring and Analysis
SUPERVISOR:
Aleksandar Drobnjak – Research Assistant

May 25, 2016

Contents

List of Figures

1 Introduction

In the recent decades, the notion of social networks and methods for social network analysis have attracted the interest of researchers and the curiosity of the social behavioral sciences, since there is a wide spread usage of the internet by people all over the world that interact with each other and exchange web content through numerous online communities such as Facebook[1], Twitter[2], Google+[3], LinkedIn[4], Pinterest[5] and many other. Such social networks have rapidly grown in popularity, because they are no longer constrained by the geographical limitations of a conventional social network in which interactions are defined in more conventional way such as face-to-face meetings, or personal friendships. The main interest is due to the attractiveness of analyzing relationships between entities, behavioral patterns that arise in such networks and the implications that are followed by such analysis. The social network perspective opens new directions in answering questions from social behavioral science by giving definitions to aspects of the political, economical, or social structural environment [23]. Consequently, the answers in such questions have great impact in the society, through understanding how individual behaviors are being expressed [3], how relationships between people are created or ended [2], which characteristics describe people [4] and how social communities are formed [8] and evolve over time [12].

In general, a social network is defined as a network of interactions or relationships, where the nodes typically may represent either users or web content (e.g. posts, news, videos, messages and other), and the edges consist of the relationships or interactions between nodes.

Social networks are typically rich in text, because of a wide variety of methods by which users can contribute text content to the network. For example, typical social networks such as Facebook allow the creation of various text content such as wall posts, comments, and links to blog and web pages. Studying the characteristics of content, for instance in the messages, becomes important for a number of tasks, such as topic detection, personalized message recommendation, friends recommendation, sentiment analysis and others. Topic models [5] are powerful tools to identify latent text patterns in the content. They are applied in a wide range of areas including recent work on Twitter [20].

With today‘s ubiquity and popularity of social network applications, the ability to analyze and understand large networks in an efficient manner becomes critically important. Visualization is becoming an important tool to gain insight on the structure and dynamics of complex social networks. Visualizing social networks is easy to understand and provides detailed information about the actual relations modeled in the data. Visualization of

[1] `http://www.facebook.com`
[2] `http://www.twitter.com`
[3] `http://plus.google.com`
[4] `http://www.linkedin.com`
[5] `https://www.pinterest.com`

social networks has a rich history, particularly within the social sciences, where node-link depictions of social relations have been employed as an analytical tool since at least the 1930s. Linton Freeman documents the history of social network visualization within sociological research, providing examples of the ways in which spatial position, color, size, and shape can all be used to encode information [9]. Freeman mentioned that visualizing social networks is more than simply creating intriguing pictures, it is about generating learning situations: "images of social networks have provided investigators with new insights about network structure and have helped them communicate those insights to others". Moreover, visualization can be very helpful in data analysis, for instance, for finding main topics that appear in larger sets of documents. Extraction of main concepts from documents using techniques such as Latent Dirichlet Allocation, can make the results of visualizations more useful.

The rest of the document is organized as follows. In Section 2 we provide the motivation and our research questions. In Section 3 we describe the visualization of the text corpus and we provide an overview of related work. In Section 4 we describe the Topic models that we used and their relative work. Next, in Section 5 we provide the similarity between short texts and the metric that we used. In Section 6, we introduce the steps that were followed for our prototype implementation, namely Framework, Data Collection and Data visualization. Finally, in Section 7 we conclude our document providing future work directions.

2 Motivation

Social networks have emerged as an important factor in information dissemination, search, marketing, expertise and influence discovery, and potentially an important tool for mobilizing people. More particularly, Twitter and Facebook have been a crucial source of information for a wide spectrum of users and are given access to massive quantities of data for further analysis. Thus, an interesting aspect in social networks is the visualization of the contributions of a user posts and their impact across different platforms. Over time the contributions of an user in social networks is growing and this creates the need to have a simple overview of the data by visualizing it. Data visualization offers a quick way to present the data in a way that can reveal valuable hidden insights. Thus, through the visualization, users can easily understand what are the hot posts, that is those posts are able to attract a greater attention or interest.

2.1 Research Questions

Several research questions arise for visualizing the contributions of user posts in social networks that require additional work. Three interesting research questions are provided below:

1. How can we visualize the contribution of a user in social networks?
2. How can we interrelate the posts published by the same user?
3. How can we find similar posts in different social networks?

Answering these questions is important to understand how much a user contributes to a social network but, also to understand the impacts of his posts. By visualizing his contribution over time we can see for instance, in which month of 2015 had the most posts in Facebook or Twitter. Moreover, By interrelating his posts we can gain an interesting view about the topics that a user is discussing. Last, by finding the similarity of his posts we can see how many likes has a similar or a same post in Facebook and Twitter.

contributons visualization

3 Visualize text corpus

3.1 Literature review

3.2 Graph based visualization

4 Topic models

As our collective knowledge continues to be digitized and stored—in the form of news, blogs, web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search and understand these vast amounts of information. To this end, machine learning provides topic models which are a suite of algorithms for discovering themes (or topics) that spread through a collection of documents.

In this section, we provide an overview of existing literature on topics models and then we describe two types of existing topic models: (i) the probabilistic model: Latent Dirichlet allocation and (ii) the non-probabilistic model: Non-negative Matrix Factorization.

4.1 Literature review

Topic modeling is gaining increasingly attention and is applied in a wide range of areas including on social networks such as Twitter and Facebook. From the view of methodology, topic models are separated into two groups: the non-probabilistic and probabilistic approaches.

Most of probabilistic approaches are based on Latent Dirichlet Allocation (LDA) [6], which is the most popular standard tool in topic modeling. As a result, LDA has been used and extended in a variety of ways, and in particular for social networks and social media, so a great number of research papers that deal with LDA have been proposed.

Ramage et al. [20, 21] extended LDA to a supervised form and studied its application in micro-blogging environment. More particularly, in [21] the Labeled LDA, a novel model of multi-labeled corpora that directly addresses the credit assignment problem is introduced. In [20], a scalable implementation of a partially supervised learning model (Labeled LDA) is proposed for discovering topics in microblogs like Twitter. This model maps the content of the Twitter feed into dimensions such as substance, style, status, and social characteristics of posts. Thus, this approach helps to efficiently characterize selected Twitter users along these learned dimensions and indicates that topic models can provide interpretable summaries of users' tweet posts. Phan et al. [19] studied the problem of modeling short text through LDA. In particular, a general framework, based on LDA, for building classifiers with hidden topics discovered from large-scale data collections that can deal successfully with short and sparse text & Web segments.

Moreover, Chang et al. [7] proposed a novel probabilistic topic model to analyze text corpora and infer descriptions of the entities and of relationships between those entities on Wikipedia. McCallum et al. [15] proposed a probabilistic generative model to simultaneously discover groups among the entities and topics among the corresponding text. Zhang et al. [26] introduced a model to incorporate LDA into a community detection process.

More specifically, in this paper they designed a hierarchical Bayesian network based approach, namely GWNLDA(Generic-Weighted Network-LDA) which is inspired by LDA for discover probabilistic communities from complex networks. Similar work can be found in [14] and [17].

Standard LDA is often less coherent when applied to microblog content like Twitter because tweets are short. To overcome this difficulty, some previous studies proposed to aggregate all the tweets of a user as a single document. In [1], a topical classification of Twitter users and messages is provided. This paper deals with the problem of using topic models in microblogs by proposing schemes based on LDA and one extended model based on the Author-Topic model. It also presents that topic models aims some classification problems by indicating that topic models learned from aggregated messages by the same user obtaining higher accuracy. Also, a different approach on topic modeling of Tweets is provided in [16]. This paper focus on how to improve clustering metrics and topic coherence with existing algorithms. More specifically, it provides two novel schemes that lead to significantly improved LDA topic models on Twitter content without requiring any modification of the underlying LDA machinery. The first one is about pooling tweets by hashtags that yields a great improvement in all metrics for topic coherence across three diverse Twitter datasets, and the second is an automatic hashtag assignment scheme further improves the hashtag pooling results on a subset of metrics.

Non-probabilistic topic models are also very popular. One of the most known representative model is the Non-Negative Matrix Factorization (NMF) [18, 13]. Yan et.al [25], proposed a novel term weight called Ncut-weighted, which measures term's discriminability according to the words cooccurrences, for short text clustering. More particularly, the experiments show that the clustering performance of NMF is greatly improved with terms weighted by the Ncut-weight. Due to the severe sparsity of short texts, in[24], a different approach on the non-negative matrix factorization is introduced. This approach first learns topics from term correlation data using symmetric non-negative matrix factorization, and then infers the topics of documents. The experimental results on three short text data sets show that this method provides substantially better performance than other baseline methods like LDA.

4.2 Latent dirichlet allocation

The idea behind Latent Dirichlet allocation (LDA) [6, 10, 22], which is an unsupervised machine learning technique, is to model documents as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms. Specifically, we assume that K topics are associated with a collection, and that each document exhibits these topics with different proportions. The interaction between the observed documents and hidden topic structure is manifest in the probabilistic generative process associated

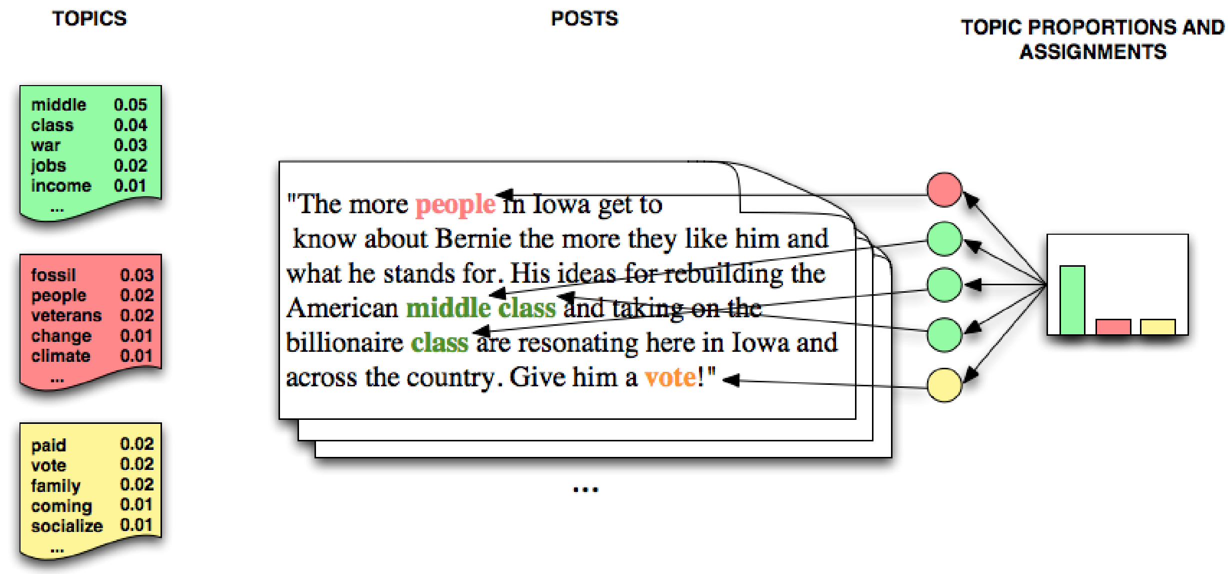

Figure 1: An LDA example.

with LDA. This generative process is as follows:

To generate a document:

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary

So, this statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).

In Figure 1, an LDA example is illustrated. We assume that we have three topics, which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right) and then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. As we can see, in this example the particular document is assigned to the first topic with higher probability compare to the others.

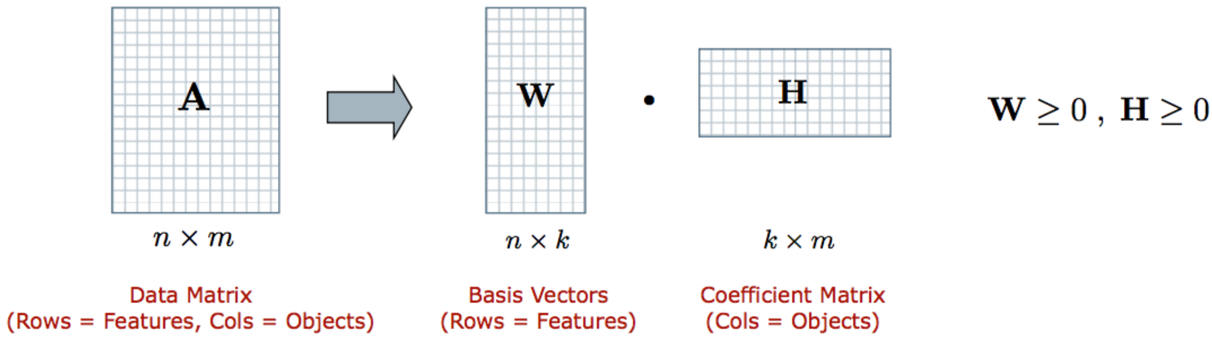

Figure 2: Illustration of approximate non-negative matrix factorization: the matrix A is represented by the two smaller matrices W and H, which, when multiplied, approximately reconstruct A

4.3 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is an unsupervised family of algorithms from linear algebra that simultaneously perform dimension reduction and clustering. NMF was first introduced by Paatero and Tapper [18] as positive matrix factorization and subsequently popularized by Lee and Seung [13].

In Figure 2, NMF takes a non-negative matrix **A** as an input, and factorizes it into two smaller non-negative matrices **W** and **H**, each having k dimensions. When multiplied together, these factors approximate the original matrix **A**. The specified parameter k controls the number of topics that will be produced. The rows of the matrix **W** provides weights that indicate the strength of association between documents and topics. The columns of the **H** that indicate the strength of association between terms and topics. By ordering the values in a given column and selecting the top-ranked terms, we can produce a description of the corresponding topic.

5 Similarity between short texts

5.1 Literature review

5.2 Cosine Similarity

For creating the user timeline, we need to find similar posts that a user share both in Twitter and Facebook. For this purpose, we calculate the cosine similarity metric.The cosine similarity [11] between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents.

Given two posts t_a and t_b, their cosine similarity is

$$\cos(\mathbf{t_a}, \mathbf{t_b}) = \frac{\mathbf{t_a}\mathbf{t_b}}{\|\mathbf{t_a}\|\|\mathbf{b}\|} \tag{1}$$

where t_a and t_b are m-dimensional vectors over the term set $T = t_1, ..., t_m$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between $[0, 1]$.

We predefine a threshold to accept two similar posts to have similarity at least 60%. This portion lets us a great amount of similar posts and also recognize twitter posts that have urls, hashtags and mentions.

6 Prototype implementation

In this section we provide a detailed description of our framework and the technologies that used in the front and back ends, the data collection and last the different aspects of data visualization.

6.1 Framework

Our web application is based on Django[6], which is a free and open-source web framework, written in Python, and follows the model–view–controller (MVC) architectural pattern. We used HTML, CSS and JavaScript in the front-end and Python, JavaScript in the back-end. First, the application is using the retrieved data for politicians and athletes from JSON files for all the tasks that will be performed. Then, the application analyzes the data further for visualizing different aspects of user contributions. For producing the dynamic, interactive data visualizations we use the JavaScript library D3[7]. The application provides an overview of that data with different visualizations for presenting the user contribution over the time, user's summary of posts, impact of posts and the topics that a user is discussing.

6.2 Data Collection

Our data collection consists of real-world data from Facebook and Twitter and focus on 28 public persons such as, politicians and athletes because they tend to post the same content on Twitter and Facebook more than a normal user. The attributes of the data along with their definitions are displayed in Table 1. The below data was saved in different files for each user in JSON format.

Table 1: Description of the attributes of data

Attribute	Description
ID	The id of the post
date	The date of the post
text	The text of the post
likes	The number of likes of a post

[6] `https://www.djangoproject.com/`
[7] `https://d3js.org/`

6.2.1 Retrieve data from Facebook

The data was retrieved by requesting the Facebook Graph API for each user using JavaScript.

6.2.2 Retrieve data from Twitter

The data was obtained by quering the timeline API of Twitter with the username of each person related to politicians and athletes. For this procedure we used the Tweepy[8], which is a Python library for accessing the Twitter API. We were able to collect a fixed number of tweets because Twitter only allows access to a users most recent 3240 tweets.

6.3 Data visualization

6.3.1 Contribution over time

6.3.2 Summary of posts

6.3.3 Topic models

6.3.4 Impact of posts

[8] `http://www.tweepy.org/`

7 Conclusion

In this document, we analyzed posts from Facebook and Twitter for politicians and athletes by visualizing them for different aspects, namely (i) contribution over time (ii) summary of posts (iii) identify topics from posts and (iv) impact of posts. So, visualization helps to better analyze the data and reveal some hidden insights of them.

Moreover, we provided detailed description about the topics models, such as Latent Dirichlet Allocation and Non-negative Matrix Factorization that used for extracting topics from user's posts. This approach aims to have an interesting view about the themes (or topics) a user is discussing in social network. According to find the similarity of a user's posts, the cosine similarity metric is used. In this way, we had a view about how popular is a similar or same post in a social network.

As a future work, it would be interesting to find similar posts among users on a socials network. This will aim to observe if the users are interested in talking about similar topics and then it is quite possible that they are interested in similar things. Another interesting direction is to take into account more values for the impact of posts such as comments, shares, retweets etc. By looking, for instance the comments of a user, we can gain an interesting view about the sentiment of the post i.e if it positive or negative. Lastly, it would also be interesting to investigate other approaches to visualize posts of a user as well as use the different visualizations together.

References

[1]

[2] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):9:1–9:33, 2012.

[3] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, pages 49–62, Chicago, Illinois, USA, 2009.

[4] S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. In C. C. Aggarwal, editor, *Social Network Data Analytics*, pages 115–148. Springer US, 2011.

[5] D. Blei and J. Lafferty. Topic models. *Text Mining: Theory and Applications*, 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[7] J. Chang, J. L. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In J. F. E. IV, F. Fogelman-Soulié, P. A. Flach, and M. Zaki, editors, *KDD*, pages 169–178. ACM, 2009.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3,Äì5):75 – 174, 2010.

[9] L. C. Freeman. Visualizing social networks. *Journal of Social Structure*, (1), 2000.

[10] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[11] A. Huang. Similarity measures for text document clustering. pages 49–56, 2008.

[12] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 611–617, New York, NY, USA, 2006. ACM.

[13] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[14] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 665–672, New York, NY, USA, 2009. ACM.

[15] A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. In *Proceedings of the 2006 Conference on Statistical Network Analysis*, ICML'06, pages 28–44, Berlin, Heidelberg, 2007. Springer-Verlag.

[16] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM, 2013.

[17] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[18] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[19] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 91–100, New York, NY, USA, 2008. ACM.

[20] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proc. ICWSM 2010*. American Association for Artificial Intelligence, May 2010.

[21] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[22] M. Steyvers and T. Griffiths. *Probabilistic topic models*, volume 427. Handbook of Latent Semantic Analysis, 2007.

[23] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.

[24] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the SIAM International Conference on Data Mining*, 2013.

[25] X. Yan, J. Guo, S. Liu, X.-q. Cheng, and Y. Wang. Clustering short text using ncut-weighted non-negative matrix factorization. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2259–2262, New York, NY, USA, 2012. ACM.

[26] H. Zhang, C. L. Giles, H. C. Foley, and J. Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, pages 663–668. AAAI Press, 2007.