Mining and Profiling in Social Media

Xia Hu
Arizona State University
xiahu@asu.edu

Huan Liu
Arizona State University
huanliu@asu.edu

Word Count: 3,505

Keywords: Social Media, Data Mining, Profiling, Social Network Analysis, Sentiment Analysis, Cybersecurity, Community Detection

Abstract

The phenomenal rise of social media services in recent years presents new opportunities and challenges to both information consumers and service providers. With its growing popularity, social media has the potential to mine actionable patterns from a large amount of data to understand user behavior and to meet users' information needs. Profiling has the potential to better describe users and their relationships in social media. This entry introduces data mining and profiling in social media and discusses the characteristics of social media data in the context of research that is being undertaken in this area.

Social media services like Facebook and Twitter have emerged as important platforms for large-scale information sharing and communication in fields such as marketing, journalism, and public relations. Social media is transforming internet users from information consumers to producers by providing a way for users to collaboratively create content. Social media have become a major platform that enables people to communicate with each other.

The abundant data provided by social media is attracting the attention of researchers in many disciplines with the aim of understanding the behavior of social media users using data mining techniques. For example, O'Connor et al. (2010) analyzed online surveys of consumer confidence and political opinion and found correlations between survey results and sentiments contained in Twitter messages. Asur and Huberman (2010) exploited Twitter textual information to forecast movie box-office revenues, achieving better forecasts than traditional market forecasting methods.

Social Media

Social media can be defined as internet applications designed for social interactions, using highly accessible and scalable communication techniques that allow users to produce and share content of their interests. Both web-based and mobile

technologies are used to transform interactive dialogue in various social media services. Social media are immensely popular in news dissemination, information sharing, and event participation.

Data Mining and Profiling

Data mining techniques have been studied since the 1990s and have become effective tools for addressing real-world problems. Data mining is the computational process of eliciting useful patterns / knowledge from large-scale datasets (Han, Kamber, & Pei, 2006). The patterns / knowledge from these data can be used to gain advantage and for increased business intelligence. The knowledge discovery from data (KDD) process consists of three steps: pre-processing, data mining, and post-processing which may be combined in data analysis (Han et al., 2006).

Classification and clustering are typical data mining tasks. Classification, an example of supervised learning, involves training data using class labels and predicting class labels for new data. Typical applications in social media include sentiment analysis, spam detection, and graph classification. Clustering, an example of unsupervised learning, involves grouping data objects into clusters according to their similarity or dissimilarity with data objects. Classification and clustering are differentiated by whether or not the task uses pre-defined class label information for some of the data objects. Typical applications of clustering include community detection, outlier detection, and aggregated search. Semi-supervised learning, association rule mining, and feature / instance selection are also useful for data analysis.

Profiling aims to automatically generate a description of an object containing the most important or interesting information about it (Schiaffino & Amandi, 2009). It can also be defined as an automatic or semi-automatic process of collecting, cleaning and organizing information about an object for building its profile. The object of profiling, which varies from one application to another, includes user profiling, group profiling, and relation profiling, etc.

User profiling aims to build a user's profile by collecting user information. Accurate user profiling is essential in personalized recommendation systems and for effective advertising and marketing. For example, Twitter users can label their accounts with textual tags. The tags can provide a description of the users and be used to facilitate information retrieval and other applications as in the case of Yelp. Accurate profiling of a user’s preferences on food, for instance, will significantly improve performance on recommending restaurants to the user.

Characteristics of Social Media Data

Linked Data

A widely used assumption in traditional mining and profiling techniques is that data objects are independent and identically distributed (i.i.d.). This assumption does not hold for social media data where data objects are potentially networked through user-user connections. The connections between data objects may contain useful semantic information that are not available in i.i.d. data. For example, a user's preference for certain mobile phones might be similar to or influenced by their connected friends. The rationale underlying the assumption is explained by two theoretical assumptions: people

tend to befriend those who are similar to themselves (Homophily), and they tend to become more similar to their friends over time (Social Influence). The analysis of social connections in social media facilitates user behavior modeling for many applications.

Existing mining and profiling methods focus on building a sophisticated feature space or effective learning model for data analysis. It is challenging to apply the methods directly in social media applications to handle linked data. Some efforts have been made to explore social connections for different tasks in social media. Jiliang Tang and Liu (2012) investigated whether linked data can be used in a new feature selection framework. They studied four types of relations based on correlations among indicators extracted from social media data, i.e., CoPost, CoFollowing, CoFollowed, and Following. Exploration of linked data also includes location-based social network analysis, recommendation, spammer detection, active learning, and text mining.

Unstructured Data

An important distinction between social media data and data derived from other platforms is the unstructured form. This feature can be interpreted in two ways. First, social media data can be represented as features of heterogeneous information sources. For example, a Twitter user can be represented by word features from the tweets he or she posted, by Scale Invariant Feature Transform (SIFT) algorithmic features derived from pictures uploaded, and by friendship features derived from his social network. Integrating features from heterogeneous space in a unified learning framework is an open problem for many applications. Second, user-generated data in social media are informal. For example, tweets like “Happy Bday” and “so coooool” are intuitive. The unstructured but intuitive format presents great challenges for a learning algorithm to accurately measure the semantic meaning of the posts.

Typically, there are two ways of processing data that consists of features derived from heterogeneous information sources: the concatenating strategy and the separation strategy (Jiliang Tang, Hu, Gao, & Liu, 2013). The first involves projecting all the features into a unified feature space and the second involves treating every source independently in preprocessing and then combining them for learning. While the concatenating strategy ignores the differences between features from different sources, the separation strategy does not consider the correlations among different sources. Jiliang Tang et al. (2013) investigated ways to exploit relations among different views to help select features in social media and the proposed method performed well on Flickr and BlogCatalog datasets.

Tackling the noisy form of text data in social media is attracting much attention from Natural Language Processing specialists using lexical normalization and semantic analysis. Lexical normalization converts the unstructured forms to standard forms. For example, lexical normalization converts coooool to its standard form cool. Semantic analysis builds connections between features that are semantically correlated. Topic modeling, e.g., Latent Dirichlet Allocation (LDA), has been extensively used for semantic analysis. The use of semantic knowledge to build semantic connections between textual features in data derived from sources such as Wikipedia, WordNet, DBPedia and Web Search Engine.

Dynamic Data

In many social media services user-generated content evolves very quickly as do social networks. For example, when composing a microblogging message, users may coin new abbreviations or acronyms that are rarely used in other formal documents. For example, "spoopy" is an intentional misspelling of the word "spooky" to describe a situation that induces fear and laughter at the same time. Slang terms are very popular and keep changing in social media.

The groups in a network, memberships in a group, and the influence of individuals in a network may change over time. For example, in a social network of machine learning researchers, the community of deep learners may grow significantly, along with the performance of deep learning methods.

Preliminary work tackled the content and network evolution in social media. Amiri and Chua (2012) proposed to mine slang and urban words and phrases from cQA services using a semi-supervised learning framework. Their studies show that a learned lexicon can effectively capture newly emerged slang words and help with the text mining task. For community detection, some evolutionary clustering algorithms have also been proposed to study community evolution in multi-mode networks. Given a sequence of network snapshots, the methods aim to extract more accurate community information.

Social Media Mining

Data mining techniques have been used in many social media applications.

Social Network Analysis

Social network analysis aims to understand structural or topological characteristics of social networks consisting of nodes (representing individual users) and ties (representing relationships between the users). Computer and social scientists are interested in studying the topological structure of social networks and verifying whether social theories hold for online social media interactions. Bond et al. (2012) used a randomized controlled trial of political mobilization messages sent to 61 million Facebook users during the 2010 US congressional elections to examine the effects of social influence in online social media. The results suggested that these messages influenced political self-expression and real-world voting behavior (Bond et al., 2012).

Some efforts have been devoted to improving social media applications via social network analysis. For example, finding influential people in a social network can help develop innovative business opportunities, forge political agendas, and lead to discussion of social and societal issues. The identification of influential nodes, e.g. bloggers and tweeters, in social networks has attracted many attentions recently. A variety of metrics such as degree centrality, betweenness centrality and page rank are widely used for link prediction, election prediction, and social network sampling.

Since social networks may be comprised of multiple types of objects, heterogeneous information network analysis is attracting attention. Jie Tang, Lou, and Kleinberg (2012) developed a framework to classify social relationships in a social network by learning across heterogeneous networks. The basic idea is to learn from a source network to infer the type of social relationship in a target network in order to demonstrate the effectiveness of information resources drawn from a variety of sources.

Sentiment Analysis

Sentiment analysis aims to automatically examine the opinion expressed in a given document. Sentiment analysis requires three stages. The first is to retrieve topic-based messages. For example, to analyze opinions about the iPhone 5s on Twitter, a specific crawler must be designed to retrieve iPhone 5s related tweets. The second is to determine whether a given document is objective or subjective using a trained classifier and only the latter is analyzed further in sentiment analysis applications. The third step is polarity sentiment classification to classify documents as positive or negative or along other dimensions.

Social media has been extensively used for users to share their opinions. Sentiment analysis is being used as a tool to understand opinions of individuals or to gauge aggregated mass sentiment. It attracts many attentions to conduct sentiment analysis on the Tweets associated with certain events such as stock market, poll rating, death of celebrities, movie box-office revenue and presidential debate. Besides the applications on Twitter, Siersdorfer, Chelaru, Nejdl, and San Pedro (2010) analyzed the influence of sentiment as expressed in comments on ratings on YouTube demonstrating that community feedback on rated comments can help to filter unrated comments or to recommend useful comments.

To make use of linked data in social media, Hu, Tang, Tang, and Liu (2013) examined how social network data can support sentiment analysis by developing a sociological approach that proposes that the sentiments of two messages posted by the same user are more likely to be consistent than those in two random messages (Sentiment Consistency), and the sentiments of two messages posted by friends are more likely to be similar than those of two random messages. In addition, to explore unstructured contextual information from heterogeneous information sources, Hu, Tang, Gao, and Liu (2013) proposed a unified unsupervised model to capture emotional signals assessing emotion indication correlations with the sentiment expressed in the social media posts.

Security Issues

With the growing popularity of social media, security issues have become important for platform owners and users. Many (fake) accounts, known as spammers, are employed to send other users unwanted information and launch attacks such as disseminating pornography, viruses, or phishing. Significant effort has been devoted to detecting spammers. One method is to analyze network structures based on the assumption that spammers cannot establish an arbitrarily large number of social trust relations with legitimate users. This traditional assumption, does not hold for many social media services. It has been found that, on Facebook and Renren, spammers can have their friend requests accepted by many other users and can successfully blend into the social matrix, quickly obtaining a large number of friends, many of which are legitimate users. Thus, network-based methods become less effective in social media.

Different from Facebook-like online social networks that need prior consent from the followee, microblogging systems feature unidirectional user bindings since anyone can follow anyone freely. Hu, Tang, Zhang, and Liu (2013) introduced this phenomena as "reflexive reciprocity". Spammers can acquire a large number of legitimate followers, especially those referred to as 'social capitalists', who increase their social capital by following anyone following them. In addition to using algorithms to study social

networks, others study characteristics related to tweet content and user social behavior. Effective spammer detection methods need to use a large amount of labeled data which is time consuming and labor intensive to obtain from social media but some work has been done using Amazon and Blogosphere data.

Profiling in Social Media

Profiling techniques are playing an increasingly important role in many social media applications.

User Profiling

More accurate user profiling may help in user behavior modeling which is essential for various social media applications. A user profile can consist of various kinds of information that can be divided into implicit and explicit information. Topics and social dimensions are two types of implicit information in user profiling. Topics, also known as interests, behavior, or activities, can be obtained through topic modeling or dimensionality reduction algorithms. A typical approach is to apply Latent Dirichlet Allocation (LDA) or its variants on a user-word matrix of user posts to obtain the latent topics of the user. In addition to content information, social information is also very important for user profiling. L. Tang and Liu (2009) proposed to extract latent social dimensions based on network information and then to use these as features for discriminative learning based on the rationale that a user might belong to different affiliations and latent affiliations might be a better indication of a user's profile than simply using friends as features. The constructed social dimensions have been shown to describe diverse social communities well and experimental results demonstrate their effectiveness on Blogcatalog and Flickr data. The techniques of mining latent topics and social dimensions use similar ideas as those used to project the original social media feature space onto a latent low-dimension space for user profiling.

User profiling with explicit information focuses on inferring social media user demographic variables, including gender, race, location, political orientation, etc. Burger, Henderson, Kim, and Zarrella (2011) proposed a statistical model for determining the gender of uncharacterized Twitter users based on the construction of a large, multilingual dataset labeled with gender information. In addition, home location estimation has become a hot topic on location-based social networks recently. Many content-based, social-based and temporal-based methods have been proposed to better profile users in the context of geographical locations.

Group Profiling

The object in group profiling is a group of users, also known as a community in social network analysis. To perform group profiling, one solution is to conduct community detection and then use methods similar to user profiling to collect the community information. A task in social network analysis and community detection is to identify groups of nodes that have high intra-group similarity and different inter-group properties.

Community detection has been extensively studied. Airoldi, Blei, Fienberg, and Xing (2008) introduced a class of variance allocation models for pairwise measurements: mixed membership stochastic block models and demonstrated the effectiveness of the

proposed method in social network analysis. In addition, latent space models, spectral clustering, and modularity-based methods are also used for community detection in various social media sites. In group profiling, community detection serves as the first step in defining groups in a network with the next being similar to that discussed for user profiling and the extraction of explicit and implicit information of a group.

Future Research

Interesting directions for further exploration include: methods for analyzing physical world events and social media influence together for social media mining; methods for conducting multi-source and multi-view analysis for profiling in social media; methods for handling large-scale data produced by social media services; and the evaluation of the performance of social media applications in the absence of a deterministic or bounded dataset.

More algorithms and applications are likely to emerge to help understand and make use of social media data as these issues are addressed.

See Also: Social network analysis; social media; semantic web; privacy

References and suggested Readings

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *the Journal of machine Learning research, 9*, 1981-2014.

Amiri, H., & Chua, T.-S. (2012). *Mining slang and urban opinion words and phrases from cQA services: an optimization approach.* Paper presented at the Proceedings of the fifth ACM international conference on Web search and data mining.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295-298.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*: Morgan kaufmann.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). *Unsupervised sentiment analysis with emotional signals.* Paper presented at the Proceedings of the 22nd international conference on World Wide Web.

Hu, X., Tang, J., Zhang, Y., & Liu, H. (2013). *Social spammer detection in microblogging.* Paper presented at the Proceedings of the Twenty-Third international joint conference on Artificial Intelligence.

Hu, X., Tang, L., Tang, J., & Liu, H. (2013). *Exploiting social relations for sentiment analysis in microblogging.* Paper presented at the Proceedings of the sixth ACM international conference on Web search and data mining.

Schiaffino, S., & Amandi, A. (2009). Intelligent user profiling *Artificial Intelligence An International Perspective* (pp. 193-216): Springer.

Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. (2010). *How useful are your comments?: analyzing and predicting youtube comments and comment ratings.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Tang, J., Hu, X., Gao, H., & Liu, H. (2013). *Unsupervised Feature Selection for Multi-View Data in Social Media.* Paper presented at the SDM.

Tang, J., & Liu, H. (2012). *Feature Selection with Linked Data in Social Media.* Paper presented at the SDM.

Tang, J., Lou, T., & Kleinberg, J. (2012). *Inferring social ties across heterogenous networks.* Paper presented at the Proceedings of the fifth ACM international conference on Web search and data mining.

Tang, L., & Liu, H. (2009). *Relational learning via latent social dimensions.* Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.

Author Biography:

Xia Hu is a research assistant of Computer Science and Engineering at Arizona State University. His research interests are in cybersecurity, text analytics in social media, social network analysis, machine learning, sentiment analysis, etc. As a result of his research work, he has published more than 30 research papers in several major academic venues. Before joining ASU, he obtained his MSc and BSc from Beihang University. He also worked as a research intern at Microsoft Research and National University of Singapore.

Huan Liu is a professor of Computer Science and Engineering at ArizonaState University. He obtained his Ph.D. in Computer Science at the University of Southern California and B.Eng. in Computer Science and Electrical Engineering at Shanghai JiaoTong University. His research interests are in data mining, machine learning, social computing, artificial intelligence, and investigating problems that arise in real-world, data intensive applications with high-dimensional data of disparate forms, such as social media. He is an IEEE Fellow and an ACM Distinguished Scientist.