Towards an exhaustive framework for Online Social Networks user behaviour modelling

Towards an Exhaustive Framework for Online Social Networks User Behaviour Modelling

Alessia Antelmi
Università di Salerno, Dipartimento di Informatica, Fisciano, Italy
aantelmi@unisa.it

ABSTRACT

Since the advent of Web 2.0, Online Social Networks (OSNs) represent a rich opportunity for researchers to collect real user data and to explore OSNs user behaviour. Based on the current challenges and future directions proposed in literature, we aim to investigate how to comprehensively model OSNs user behaviours, by exploiting and combining user data of different nature. We propose to use *hypergraphs* as a model to easily analyse and combine structural, semantic, and activity-related user information, and to study their evolution over time. This novel user behaviour modelling technique will converge in open, efficient, and scalable libraries, which will be integrated into a modular framework able to handle the data crawling process from several OSNs.

KEYWORDS

User Modelling, Online User Behaviour, Online Social Networks

ACM Reference Format:
Alessia Antelmi. 2019. Towards an Exhaustive Framework for Online Social Networks User Behaviour Modelling. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3320435.3323466

1 INTRODUCTION

The exponential growth of Online Social Networks (OSNs) users - due to the ubiquitous online access and the increasing adoption of digital devices - is generating a huge amount of structured and unstructured data [24], that can be analysed to gather insights into many domains [2, 16]. Due to the enormous amount of accessible online services, the question of service personalization becomes an important requirement [11]. For this reason, the task of building user profiles able to accurately capture users' preferences and behaviours is currently a hot research topic. User profiling is the process of acquiring, extracting and representing the features of users [32]. The content and the purpose of a user profile strictly depend on the application domain and modelling user *behaviour* can be an important aspect to include [25]. In particular, OSNs user behaviour can be defined as the various social activities that users can perform online - friendship creation, content publishing, messaging, and commenting and its characterization can be useful for both OSN service providers and users [17]. Furthermore, accurate models of user behaviour in OSNs are crucial in social studies and viral marketing [7].

Problem Statement. In this doctoral project, we want to investigate how to comprehensively and effectively model OSNs user behaviour, by exploiting and combining user data of diverse nature and their evolution along the temporal axis. We plan to capture the OSNs user behaviour by modelling structural, semantic and activity-related user data through *hypergraphs*, able to handle general types of relations as a hyperedge can connect an arbitrary number of vertices. The proposed approaches will be integrated into a component-based framework to allow an easier combination and comparison of both data sources and behavioural profiling techniques. This research might be useful to a variety of tasks, such as getting insights into user engagement, link prediction, recommender systems, bot/spam detection, and community detection.

2 RELATED WORK AND RESEARCH AREAS

This Section provides a brief overview about how OSNs data have been explored in literature for the user modelling task and how hypergraphs can be used to model complex OSNs relationships. Some of the challenges that need to be faced when working with OSNs as data sources and future directions which should address open questions in this field will be further discussed. This Section also identifies the major research topics related to the user behaviour modelling problem. We plan to focus on selected aspects within these research areas.

OSNs user modelling. The copious information generated by users in OSNs and the variety of available user data create new opportunities for inferring user profiles in many applications. Halfaker et al. [15] identify the task of session identification as a common strategy to develop metrics for web analytics and behavioural analysis of user-facing systems. They employ user-initiated actions with timestamps for identifying clusters of user activity and evaluate their method on a variety of datasets, one of which is the popular question/answer website StackOverflow. Mac Kim et al. [18] exploit a collection of English Twitter profiles for detecting social roles - such as work, community, and familial roles - seen as a particular demographic characteristic, based on which social media content can be grouped. Giammarino et al. [14] propose a social recommender system (RecSys) able to handle both the temporal dynamics of user interests and the sentiment towards them; furthermore, integrating the underlying concept of *homophily* in a social network. Tagarelli et al. [29] exploit structural information to propose a topology-driven lurking definition. They also incorporate different temporal aspects concerning both the production and consumption of information. Several works also investigate

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3323466

the potentiality of combing OSN data of different nature. Skowron et al. [28], for instance, explore personality traits using together data derived from Instagram and Twitter, combining image and linguistic features from tweets and captions. They also include basic structural information, like the number of followers/friends. Xu et al. [30] propose a learning model to infer a user's topical expertise using four types of user data: tweets, friends, followers and lists. Other interesting works can be found in [10, 12, 23, 27].

Modelling OSNs with hypergraphs. In the past decades, the theory of hypergraphs found application in many real-world problems where complex relationships between the objects in the system play a dominant role [8]. With the increasing complexity of the underlying nature of OSNs, in terms of both content and possible activities, researchers started using hypergraphs also in the Social Media context. Amato et al. [3] propose a model to represent different kinds of relationships typical of a Multimedia Social Network - such as among multimedia contents, among users and multimedia content, and among users. On the top of the hypergraph structure defined, they analyse the topic of the textual content and define similarity values between users, multimedia objects and topics. In another work [4], the same authors investigate the model designed for the lurker detection task. Yang et al. [31] apply the hypergraph model in the context of viral marketing to maximize the profit generated by a group of activities. Fang et al. [13] use this structure to capture multi-type relations - including links among images, and social links between users and images - to mine topic-sensitive influencers. Other interesting work can be found in [9, 19, 26].

Challenges. The massive quantity and dynamicity of User Generated Content (UGC) available from a variety of social media platforms provide new possibilities for a more accurate construction of a user profile but, at the same time, challenge the current personalization techniques [1, 32]. The dynamicity is an important element to consider as users' preferences and behaviour can change with time and this transformation needs to be reflected in their profiles [1, 17]. Dealing with the *time dimension* means not only to incorporate, for instance, a decay factor to reduce interests' weights by time but also handling the data collection and the analysis tasks in reasonable time. Furthermore, collecting dynamic data raises challenges for information storage [17].

OSNs themselves present multiple issues related to the *crawling process*. APIs fragmentation, rate limits, distinct access levels (based on licenses or collaboration agreement), and frequent change in API protocols and terms of service make difficult the creation of a common crawling mechanism and the access to the available data [21]. The decision to no longer expose platform data through public APIs (as in the case of Facebook and Instagram) can further impact social media crawlers. Another issue is the *ephemeral nature* of the shared content: it can be removed either by the users or by the OSN for violating its guidelines and terms of service, making referencing or documenting content difficult [21]. *Ethical* and *legal requirements* (e.g. GDPR) have considerable implications for the design and operation of social multimedia mining [21]. They also impact the possibility of sharing the crawled datasets for validation purposes. Finally, working with UGC is a complex task due to the *nature* of its content: short messages and rich in emojis/slang.

Open Questions. In their survey, Abdel-Hafez and Xu [1] proposed three future directions with respect to the user modelling task in OSNs: (1) more dynamicity, (2) more enrichment, and (3) more comprehensiveness. According to Piao and Breslin [22], many efforts have been made towards the second direction in the last years, but the first and the third proposed directions have not made much progress. On the top of the work of Abdel-Hafez and Xu [1] and based on the recent literature, Piao and Breslin [22] recommend several new directions, related to (1) mining user interests, (2) multi-faceted user interests, (3) comprehensive user modelling, and (4) evaluation of user modelling strategies. They detail the need for more sophisticated approaches for understanding the semantics of UGC. For instance, to infer implicit user interests, they suggest to consider the context referring to some previous microblogs posted by the user and to leverage collective knowledge via frequent pattern mining approaches. As Abdel-Hafez and Xu [1], they indicate that more comprehensive user modelling strategies should be investigated by considering different dimensions of user modelling - obtained from several data sources - together and whether there is a synergistic effect on application performance by their combination. They further propose to combine different data sources not only from multiple OSNs but also inside the target platform itself. Finally, they observe how the lack of common benchmarks and datasets hinders comparison among the approaches proposed in previous studies. To this end, they suggest to provide all approaches as user modelling libraries - publicly available - so that other researchers can easily reimplement them for comparison.

3 A FRAMEWORK FOR OSNS USER BEHAVIOUR MODELLING

Based on the future directions outlined by Piao and Breslin [22] and considering the current state of the art, research questions to fill these research gaps can be summarized as follow. i) Are hypergraphs a valid tool to comprehensively model a user behaviour profile? ii) How and to what extent can this model be enhanced with activity-related and semantic user data? iii) Is there a positive effect on the performance of a given application by combining several data sources and modelling approaches?

To answer these questions, we intend to design and implement a framework to comprehensively model a user profile. This framework will thus allow i) an easier *combination* and *comparison* of several data sources (within the same platform and/or across various OSNs), and ii) the *construction* and *comparison* of different user behaviour models according to the selected application task. As a novel behaviour profiling technique, we propose *hypergraphs* to combine: i) *activity-related* data - in terms of activity typology and interaction patterns, ii) *structural* data - related to a user activity and friendship networks - and iii) *semantic* data - related to the UCG (we will be focusing on textual content, in terms of both topics discussed and sentiment expressed). Each dimension will be evaluated on the temporal axis to consider the dynamic nature of the user behaviour. We further plan to propose *hybrid* modelling approaches based on hypergraphs and machine learning (ML) techniques.

The development of this framework will reflect the phases summarized as follows: i) *design* and *tuning* of methods and techniques to implement, ii) *implementation* of the approaches to profile OSNs

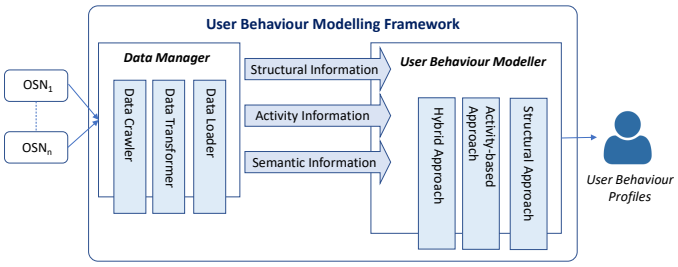

Figure 1: The process of generating user behaviour profiles through our component-based framework.

users and iii) *experimentation* using platform data that allow or will allow in future to access them. The architectural *modularity* is a fundamental requirement for the proposed system in order to: (i) be *independent* from both the OSNs and the domain of application, and (ii) guarantee the easy extension to other data sources and approaches. The extensive amount and the dynamicity of the required data highlight the need to implement such a framework in a distributed manner in order to guarantee *efficiency*. Fig. 1 presents the general process of generating user behaviour profiles, showing the component-based architecture of our framework.

3.1 Data Manager Component

Working with OSNs as data sources implies to take into account the mechanisms made available by the queried platforms to retrieve the necessary information and manage their rate limits. In previous studies [5, 6], we crawled the microblogging platform Twitter due to the openness of its APIs and its popularity. During this doctoral project, we will keep monitoring the evolution of the OSNs and the mechanisms that they make available to crawl their data in order to include them in our framework. As storage solution, we chose the document-based model MongoDB as it fits with the requirements of our framework and it is widely used in literature [20]. We will evaluate and introduce new storage technologies as the size of our dataset will start increasing and we will have to handle non-textual data. This data manager component will act as a common interface to query a subset of OSNs to (i) make the design and the execution of the same experiment easier on several OSN platforms and (ii) aggregate data from multiple OSNs in a simpler way.

3.2 User Behaviour Modeller Component

This Section describes in more detail our strategy to analyse OSNs user behaviour and the progress made to date towards this direction. This component will consist in scalable and efficient libraries to improve the reproducibility and the evaluation of the designed experiments. We plan to support several independent approaches (e.g. structural, hybrid) to construct user behaviour profiles, whose results can be easily combined and compared within the framework.

Activity-based approach. In a first approach to characterize OSNs users' behaviour, we considered the information related to the *volume* and *typology* of the activities made by the OSNs users, examining both *static* [5] and *dynamic* [6] data, respectively computed from a user's profile and posts history. Our main purpose was to understand how well this kind of information reflects different levels of user activity in online platforms, in terms of *active* and *lurking behaviours*. In the online platforms context, a *lurker* is a user who does not contribute to any content, but he/she only reads it.

Experimental setup. As described in Section 3.1, we conducted our analyses on the microblogging platform Twitter. To build a dataset as heterogeneous as possible, we collected four sub-datasets of users: a random sample of Twitter users (122,894) and three groups representing as many Twitter communities (around 300,000 users each). In our first work [5], we analysed static snapshots of a user's profile based on his/her whole history of interaction, considering the total number of likes, statuses, friends, followers and, subscription lists. We analysed all datasets independently. In our second work [6], we investigated two typologies of features: (i) the number and frequency of the activities, and (ii) their typology, evaluated during a given period of time. In this second case, we analysed only the random dataset. The analysis of the temporal dimension will be presented in a journal article, currently in preparation.

Experimental evaluation. Since this kind of datasets lacks the presence of ground truth, we used the K-means algorithm - a common data mining approach to extract grouping patterns without any prior knowledge of their characteristics. We clustered our data according to the sets of features previously described. We repeated this process varying the features values over the observation period and analysed the differences among the partitions produced to evaluate migration patterns.

Experimental results. Our analyses showed that profile-based information is useful in characterizing users *influence* level and highlighting *famous* accounts in a Twitter online community. Nevertheless, they fail to capture whether a user is active/passive. On the other hand, timeline-based information captures a strong separation of users in four levels of activity (*high*, *medium*, *low*, *no-activity*). Analysing the *time* dimension, we further observed that users tend not to change their active or passive behaviour. Our data-driven findings suggest that only 3 users out of 4 are lurkers - interestingly, less than the proportion suggested by the 90-9-1 rule about participation inequality [1] in online communities.

Structural approach. For the structural approach, we aim to explore which insights a user activity and interaction networks can offer about his/her online behaviour. Our idea is to describe these relations through *hypergraphs*, able to model more general types of relations than graphs do [8]. A hypergraph is a graph in which an edge can connect more than two vertices (in other words, an edge is a subset of vertices). This mathematical object can thus capture if a vertex shares a common property/relation with others. In practice, we can model - for instance - if two or more OSNs users comment the same post or have the same interests.

We are currently implementing an open library[2] to model hypergraphs, written in Julia. The Julia programming language is getting more attention by the scientific community thanks to its flexibility as a dynamic language, appropriate for scientific and numerical computing, and its performance comparable to traditional statically-typed languages. The designed model aim to be as

[1] https://www.nngroup.com/articles/participation-inequality
[2] https://github.com/pszufe/SimpleHypergraphs.jl

general as possible in order to be easily instantiated according to the application requirements. It further allows the user to attach metadata to both vertices and hyperedges to appropriately represent additional real-world problem data. A line of inquiry we are pursuing is related to understand to what extent a hypergraph can give qualitative better information respect to its graph representations - such as line graphs, incidence graphs, and two-section graphs. We are using the 2019 Yelp dataset challenge to validate our hypothesis, which contains more than 9GB of information related to users, businesses, and reviews. As near future development, we plan to further investigate the Yelp dataset in contexts of RecSys and community detection. On top of our research about user engagement and recent literature, we will continue working on the lurker detection task.

Hybrid approach. As a future direction, our intent is to exploit hybrid approaches that combine either user data of different nature and modelling techniques. Our purpose is to evaluate how and if the combination of different kinds of data and models can enhance the construction of a user behavioural profile in several application domains (e.g. lurker detection). For instance, we plan to integrate activity-related and semantic user information in the construction of the hypergraph. We further intend to combine the proposed modelling techniques based on hypergraphs with ML approaches.

4 CONCLUSIONS AND WORKING TIMELINE

In this work, we have outlined a collection of open research problems and challenges that need to be faced when dealing with the user profiling task. The main contributions of our research will consist in proposing novel techniques to comprehensively model OSNs user behaviours by exploiting hypergraphs. These proposed approaches will be integrated into a component-based framework that will allow to easily combine and compare both heterogeneous data sources and approaches for OSNs user behaviour modelling.

During the first year of the PhD programme, we have planned to (i) conduct a comprehensive literature review, (ii) have a stable version of the library to model hypergraphs, (iii) deliver the Data Manager component, and (iv) implement the structural approach. In the second year we plan to (i) merge the activity-based and structural approaches, and (ii) study how to model and integrate the semantic information. During the third and last year of the PhD programme we plan to extend our research and experiments to other OSNs and hybrid approaches.

REFERENCES

[1] A. Abdel-Hafez and Y. Xu. 2013. A survey of user modelling in social media websites. *Computer and Information Science* 6, 4 (2013), 59–71.

[2] I. Adaji, K. Oyibo, and J. Vassileva. 2018. The Effect of Gender and Age on the Factors That Influence Healthy Shopping Habits in E-Commerce. In *26th Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 251–255.

[3] F. Amato, L. Barolli, V. Moscato, A. Picariello, and G. Sperlì. 2018. Strategies for Social Networks Modeling. In *International Conference on Advanced Information Networking and Applications Workshops*. 681–686.

[4] F. Amato, V. Moscato, A. Picariello, F. Piccialli, and G. Sperlí. 2018. Centrality in heterogeneous social networks for lurkers detection: An approach based on hypergraphs. *Concurrency and Computation: Practice and Experience* 30, 3 (2018), 41–88.

[5] A. Antelmi, D. Malandrino, and V. Scarano. 2018. Characterizing Twitter Users: What do Samantha Cristoforetti, Barack Obama and Britney Spears Have in Common?. In *2018 IEEE International Conference on Big Data*. 3622–3627.

[6] A. Antelmi, D. Malandrino, and V. Scarano. 2019. Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. In *Companion Proceedings of the The Web Conference 2019*. Accepted.

[7] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2009. Characterizing User Behavior in Online Social Networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*.

[8] A. Bretto. 2013. *Hypergraph Theory: An Introduction*. Springer.

[9] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He. 2010. Music Recommendation by Unified Hypergraph: Combining Social Media Information and Music Content. In *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, New York, NY, USA, 391–400.

[10] A. Caliò, R. Interdonato, C. Pulice, and A. Tagarelli. 2018. Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks. *IEEE Trans. on Knowledge and Data Engineering* 30, 12 (2018), 2421–2434.

[11] A. Cufoglu. 2014. User Profiling - A Short Review. *International Journal of Computer Applications* 108, 3 (2014), 1–9.

[12] A. Delic, J. Masthoff, J. Neidhardt, and H. Werthner. 2018. How to Use Social Relationships in Group Recommenders: Empirical Evidence. In *26th Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 121–129.

[13] Q. Fang, J. Sang, C. Xu, and Y. Rui. 2014. Topic-Sensitive Influencer Mining in Interest-Based Social Media Networks via Hypergraph Learning. *IEEE Transactions on Multimedia* 16, 3 (2014), 796–812.

[14] D. Giammarino, D. Feltoni Gurini, A. Micarelli, and G. Sansonetti. 2017. Social Recommendation with Time and Sentiment Analysis. In *25th Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 376–380.

[15] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. Nguyen, K. Shores, A. Uduwage, and M. Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *Proceedings of the 24th International Conference on World Wide Web*. 410–418.

[16] T. Hecking, V. Dimitrova, A. Mitrovic, and U. Ulrich Hoppe. 2017. Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching. , 326–335 pages.

[17] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. 2013. Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine* 51, 9 (2013), 144–150.

[18] S. Mac Kim, S. Wan, and C. Paris. [n. d.]. Detecting social roles in twitter. In *Conference on Empirical Methods in Natural Language Processing* (2016). 34–40.

[19] L. Li and T. Li. 2013. News Recommendation via Hypergraph Learning: Encapsulation of User Behavior and News Content. In *ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 305–314.

[20] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. 2015. CrowdPulse: A framework for real-time semantic analysis of social streams. *Information Systems* 54 (2015), 127 – 146.

[21] S. Papadopoulos and Y. Kompatsiaris. 2014. Social Multimedia Crawling for Mining and Search. *Computer* 47, 5 (2014), 84–87.

[22] G. Piao and J. G. Breslin. 2018. Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction* 28, 3 (2018), 277–329.

[23] V. Pourheidari, E. S. Mollashahi, J. Vassileva, and R. Deters. 2018. Recommender System based on Extracted Data from Different Social Media. A Study of Twitter and LinkedIn. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. 215–222.

[24] M. Salampasis, G. Paltoglou, and A. Giahanou. 2011. Using Social Media for Continuous Monitoring and Mining of Consumer Behaviour. *IJEB* 11 (2011), 85–96.

[25] S. Schiaffino and A. Amandi. 2009. *Intelligent User Profiling*. Springer Berlin Heidelberg, Berlin, Heidelberg, 193–216.

[26] J. Silva and R. Willett. 2008. Detection of anomalous meetings in a social network. In *2008 42nd Annual Conference on Information Sciences and Systems*. 636–641.

[27] G. Silvestri, J. Yang, A. Bozzon, and A. Tagarelli. 2015. Linking Accounts across Social Networks: the Case of StackOverflow, Github and Twitter. In *KDWeb*. 41–52.

[28] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 107–108.

[29] A. Tagarelli and R. Interdonato. 2015. Time-aware analysis and ranking of lurkers in social networks. *Social Network Analysis and Mining* 5, 1 (11 Aug 2015), 46.

[30] Y. Xu, D. Zhou, and S. Lawless. 2017. User Expertise Inference on Twitter: Learning from Multiple Types of User Data. In *25th Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 395–396.

[31] W. Yang, J. Yuan, W. Wu, J. Ma, and D. Du. 2019. Maximizing Activity Profit in Social Networks. *IEEE Trans. on Comp. Social Systems* 6, 1 (2019), 117–126.

[32] X. Zhou, Y. Xu, Y. Li, A. Josang, and C. Cox. 2012. The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review* 37, 2 (2012), 119–132.