Information Evolution Modeling and Tracking: State-of-Art, Challenges and Opportunities

Ekaterina Shabunina and Gabriella Pasi

Università degli Studi di Milano-Bicocca, Dipartimento di Informatica Sistemistica e
Comunicazione, Viale Sarca 336, 20126 Milan, Italy
{ekaterina.shabunina, pasi}@disco.unimib.it

Abstract. In the Web 2.0, where everyone is the creator of content, information spreads and evolves rapidly through unpredictable paths of rebounds between news sources and Social Media. In this context, modeling, analyzing and tracking the information evolution through time offers unprecedented opportunities to diverse research fields, including Information Retrieval. In this paper we propose a synthetic analysis of the state-of-art on Information Evolution on the Web, and we summarize the interesting opportunities it offers to Information Retrieval.

1 Introduction

The emergence of the Web 2.0 has granted user the freedom to interact with other users and to contribute contents to the World Wide Web. Consequently, it has motivated novel research directions, and it has also provided new perspectives within existing ones. The most common way to propagate opinions and ideas on the Web is constituted by Social Media, which facilitates the creation and sharing of User Generated Content (UGC). Thus, Social Media provides the possibility to analyze the content generated by a vast number of users from different countries and social backgrounds to the aim of excerpting cultural trends and ideas that spread in real time. This enables unprecedented opportunities such as, for example, an early detection of social crisis, disasters and emergencies [2,10]. The identification of social phenomena in UGC can bring insight on the behavior of users in Social Networks, the patterns of their interactions, and the structure of the information spread depending on the phenomenon driving it. Thus, the study of the evolution of information on Social Media allows to track how information related to specific topics or events changes over time, and it makes possible to monitor the evolution of cultural, political and social ideas. Evidently, modeling, analyzing and tracking the evolution of information in time on the Web and, particularly, on Social Media promises a myriad of unprecedented opportunities and applications to several fields, among which Information Retrieval and Social Media Analytics.

In the following Sections we will present a synthetic analysis of the state-of-art in modeling, analyzing and tracking the evolution of information on the Web with respect to its main challenges (Section 2) and the opportunities it offers to Information Retrieval and related fields (Section 3).

2 State-of-Art and Challenges

Information Propagation (IP) aims to analyze the spread of information on the Web through time. This issue has been explored by a number of works in the literature. The majority of the proposed approaches has considered IP as a network-centered problem [6,13] The main focus of this line of research is to study how information spreads in a user network. Simultaneously, another line of works on IP aims to study how the content of a piece of information evolves in time [3,5,8]. The scope of the present paper is this latter approach, the objective of which is the quantitative and qualitative evaluation of the evolution of a stream of information.

In content-centered IP the common approach is to primarily identify the core units of information in a stream of Social Media posts. In the literature these units of information have been frequently referred to as "memes" [1,3,4,8,9,11,12], a notion coined by R. Dawkins in 1976 to refer to a unit of human cultural evolution, analogous to a gene in genetics [7]. Subsequently, the analysis of information evolution is performed on the identified core units of information of the studied information stream.

There are two main challenges in content-centered IP. The first core aspect is the formal representation of textual information. In the majority of works in the literature a unit of information is identified as a short, frequently quoted phrase and its slight variations [1,8,11,12]. In [8,12] it is formally defined as a phrase graph, with nodes representing the phrases and the edges representing the edit distance between the phrases. In [9] a unit of information is assumed to be represented by several objects such as "hashtags" and "mentions" on Twitter, URLs and the preprocessed text of the tweet itself. Similarly, in [4] different types of displays of memes from the Yahoo! Meme platform are considered: short snippets of text, photos, audio, or video, tokens in URLs, etc, which are represented as bag-of-words.

The work in [5] is among the ones that pioneered the issue of the identification and formal representation of an information granule in a Social Media stream; such information granule is defined by the authors as an *ememe* or "electronic meme", and is formally represented as a micro ontology generated by posts on the blogosphere by means of an OWL schema.

The study in [3] presents a semi-supervised attempt to represent memes in a set of documents as semantic networks, by extracting n-grams that co-occur in many documents and, subsequently, by constructing the semantic network where nouns and adjectives, from the extracted n-grams, are the nodes and verbs are the edges between them.

The second challenge in the content-centered study of information diffusion concerns the methods for measuring, evaluating and analyzing the information evolution in time. In [5] a set of operators is proposed in the context of semantic web ontologies aimed at measuring some useful properties of ememes such as fidelity (the degree to which a meme is accurately reproduced, computed as the fuzzy matching between the given blog post and the original ememe description), mutation (the difference between the maximum and the minimum fidelity among the instances of the ememe), spread (reproductive activity of a meme, calculated as the number of instances of the ememe in the searched source), and longevity (the time duration of the ememe's life span, which is the difference between the dates of the most recent and the oldest posts that contain the ememe instance). Similarly, in [3] three meme metrics are proposed in the context of semantic networks: longevity (alike *longevity* in [5]), fecundity (alike *spread* in [5]) and copy-fidelity (alike *fidelity* in [5]). One of the most recent and large-scale studies on memes, presented in [1], analyzes the mutation and replication rates in memes evolution with the Yule process. The work in [11] presents a study on the changes introduced in quoted texts as they diffuse through time; the authors examine properties of the quoted texts variants and uncover patterns in the rate of appearance of new variants, their length, the types of changes introduced, their popularity and the type of sites that are replicating them. The temporal patterns of variations in quoted phrases are studied in [8], by extracting the temporal threads of all blogs and news media sites that mention the meme phrase, identifying the patterns and time lags of quoting between them, as well as analyzing their change in time in the whole thread.

3 Opportunities

As previously outlined, the possibility to track through time the evolution of information on the Web can bring large and unprecedented benefits and opportunities to many research areas such as Information Retrieval, Social Network Analysis and others. Here we emphasize some promising directions for the exploitation of content-centered IP in the context of IR.

User profiling for personalized search is commonly performed by tracking the user's activities on the Web to infer a representation of the user's interests. More recently, users' profiles have been defined based on the content generated by users in Social Media [14]. Commonly, user interests are dynamic and they evolve in time. Thus, user profiling presents a natural scenario for an application of the automatic analysis of the evolution of an information stream. Independently from the means by which the user's topical interests are gathered, either by query logs or as the content generated by the user on Social Media, the methods for tracking the evolution of information in time can be successfully applied to update the formal representation of the user model.

Additionally, the exploitation of the evolution in time identified in the users topical interests, can help in dealing with the "filter bubble" problem in personalized search and personalized recommendation by introducing a diversification in the retrieved results through the natural and non-evident change in the information.

Another interesting application of tracking the evolution of textual information is the analysis of queries formulated by the user over a given time interval. This could bring insights on how users' interests change in time.

References

1. L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng. Information evolution in social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 473–482, New York, NY, USA, 2016. ACM.
2. M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1749–1758, New York, NY, USA, 2014. ACM.
3. H. Beck-Fernandez and D. F. Nettleton. Identification and extraction of memes represented as semantic networks from free text online forums. In *MDAI 2013 - Modeling Decisions for Artificial Intelligence*, Barcelona, 20/11/2013 2013.
4. F. Bonchi, C. Castillo, and D. Ienco. Meme ranking to maximize posts virality in microblogging platforms. *Journal of Intelligent Information Systems*, 40(2):211–239, 2013.
5. G. Bordogna and G. Pasi. A fuzzy approach to the conceptual identification of ememes on the blogosphere. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8, July 2013.
6. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.
7. R. Dawkins. *The Selfish Gene.* Oxford University Press, Oxford, UK, 1976.
8. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.
9. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM.
10. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.
11. M. P. Simmons, L. A. Adamic, and E. Adar. Memes online: Extracted, subtracted, injected, and recollected. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.
12. C. Suen, S. Huang, C. Eksombatchai, R. Sosic, and J. Leskovec. Nifty: A system for large scale information flow tracking and clustering. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1237–1248, New York, NY, USA, 2013. ACM.
13. J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 599–608, Washington, DC, USA, 2010. IEEE Computer Society.
14. A. Younus, C. O'Riordan, and G. Pasi. A language modeling approach to personalized search based on users' microblog behavior. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 727–732, 2014.