Leveraging Web Intelligence for Information Cascade Detection in Social Streams

Mohamed Cherif Nait-Hamoud[1,2(✉)], Fedoua Didi[1], and Abdelatif Ennaji[3]

[1] Department of Science Computing,
University of Abou Bekr Belkaid, 13000 Tlemcen, Algeria
fedouadidi@yahoo.fr
[2] Department of Mathematics and Science Computing,
University of Larbi Tebessi, 12000 Tebessa, Algeria
mc_naithamoud@hotmail.com
[3] LITIS Lab, University of Rouen, Rouen, France
ennaji@univ-rouen.fr

Abstract. In this paper, we present an approach for investigating information cascades in social and collaborative networks. The proposed approach seeks to improve methods limited to the detection of paths through which merely exact content-tokens are propagated. For this sake, we adopt to leverage web intelligence to the purpose of discovering paths that convey exact content-tokens cascades, as well as paths that convey concepts or topics related to these content-tokens. Indeed, we mine sequence of actors involved in cascades of keywords and topics extracted from their posts, using simple to use restful APIs available on the web. For the evaluation of the approach, we conduct experiments based on assimilating a scientific collaborative network to a social network. Our findings reveal the detection of missed information when using merely exact content propagation. Moreover, we noted that the vocabulary of actors is preserved mostly in short cascades, where topics become a better alternative in long cascades.

Keywords: Information cascade · Information diffusion analysis
Social and collaborative networks · Web intelligence
Information flow paths

1 Introduction

Information diffusion in online social networks (OSN) is a research field that aims at addressing concerns about what governs information spread in these networks. According to [1], information diffusion is categorized into four types: *herd behavior*, *information cascades*, *diffusion of innovation* and *epidemics*. Information cascades occur in OSN when actors adopt the behavior of their followees

A. Amine et al. (Eds.): CIIA 2018, IFIP AICT 522, pp. 56–65, 2018.
https://doi.org/10.1007/978-3-319-89743-1_6

or friends due to their actions influence. The taxonomy of information diffusion proposed in [2] reveals a set of challenges for the effective extraction and prediction of valuable information from the exchanged huge amount of data. The authors, in [2] have identified three main research axes: (1) interesting topics detection, (2) information diffusion modeling, and (3) influential spreaders identification. As an improvement for influential actors detection, the authors have suggested to take topics into account. Since that study, many works are dedicated to investigate the detection of influencers using contents of interactions taking into accounts the underlying network structure. Particularly, in [4] and later in [3] authors have studied the problem of information cascades, their work included mining paths that convey more frequent information and influence detection in the context of information flow in networks. In that work, both content of interactions and underlying network structure are considered. In addition, authors have designed an algorithm to detect information flow patterns in social networks called *InFlowMine*. Specifically, they first targeted the detection of paths (sequence of linked nodes of the social graph) through which exact content-tokens are propagated more frequently. They referred to these paths as *frequent paths*, considering the chronological order of actors posted information. Afterwards, they computed a score using the mentioned frequent paths to discover influencers in social and collaborative networks. The content generated by the different actors of the network was considered as a social stream of text content. Each element is a tuple that consists of a unit of information called *content-token* (hash tags, URLs, text in Twitter) and its originating actor. Formally, a social stream was defined as all couples (U_j, a_i) where U_j is the posted token and a_i is the originating actor.

In this paper, we propose to leverage web intelligence to extend the work of authors in [3,4] to the purpose of discovering paths that convey exact content-tokens cascades, as well as paths through which content-tokens referring concepts are propagated. Specifically, we mine sequence of actors involved in cascades of keywords and topics extracted from posts of social or collaborative networks actors, using simple to use restful APIs available on the web. Our proposal allows to reveal eventual missed paths that may not be detected when tracking merely cascades of exact contents.

The rest of this paper is organized as follows; in Sect. 2 we present necessary definitions for the clarity of the paper and the problem formulation; in Sect. 3 we introduce the details of our proposal; in Sect. 4 we present experiments and we discuss the obtained results. Finally, we conclude in Sect. 5 with some comments.

2 Problem Formulation

To make this paper self-contained and for the sake of clarity, we introduce below the basic concepts used in [3,4] to mine frequent paths and influencers in social networks. Some definitions were adapted to the purpose of the problem reformulation.

Definition 1 *(Valid flow path)* [3,4]**.** *Let G(N,E) be a social or collaborative network where N is a set of nodes and E a set of edges, a valid flow path is an ordered sequence of distinct nodes* $n_1 n_2 \dots n_k, n_i \in \mathbb{N}$ *, such that for each t ranging from 1 to k−1 an edge exists between nodes* n_t *and* n_{t+1}*.*

Definition 2 *(Information flow frequency)* [3,4]**.** *The information flow frequency of actors* $n_1 n_2 \dots n_k$ *is the number f of content-tokens* $U_1 \dots U_f$ *given that for each* U_i *the following conditions hold:*

- *Each actor from the sequence* $n_1 n_2 \dots n_k$ *has posted* U_i
- *Each* U_i *was posted by the actors in the order* $n_1 n_2 \dots n_k$*.*

Definition 3 *(Frequent path)* [3,4]**.** *A sequence of actors* $n_1 n_2 \dots n_k$ *is defined as a frequent path of a frequency f, if the following two conditions hold:*

- *The sequence* $n_1 n_2 \dots n_k$ *is a valid flow path.*
- *The frequency of the actor sequence* $n_1 n_2 \dots n_k$ *is at least f.*

To the sake of reformulating mining frequent paths problem considered in this paper, we reformulate Definition 2 as follows:

Definition 4 *(Information flow frequency - reformulated)***.** *Information flow frequency of actors* $n_1 n_2 \dots n_k$ *is the number f of content-tokens* $U_1^{'}, U_2^{'}, \dots, U_f^{'}$ *representing the keywords and concepts extracted from actors posts using simple to use restful APIs available on the web, given that for each* $U_i^{'}$ *the following conditions hold:*

- *Each actor from the sequence* $n_1 n_2 \dots n_k$ *has posted a token* $U_i^{'}$
- *Each* $U_i^{'}$ *was posted by the actors in the order of the sequence* $n_1 n_2 \dots n_k$

Problem 1 (*Information flow paths mining - Extended*). Given a graph G, a stream of content propagation and a frequency f, the problem of information flow path mining is to find the frequent paths in the underlying graph G using keywords and concepts extracted from actor posts instead of using content-tokens. Keywords and concepts are obtained using available restful APIs.

3 Proposed Approach

The content-tokens considered in [3,4] may represent the results of the tokenization of each actor post (i.e.; text message) after the removal of stop words. To focus only on important words or composed words of the post, we propose to consider the keywords extracted from the post as content-tokens when mining frequent paths of exact content cascades. Indeed, the gain is substantial in the extent that only the important part of posts is considered; this leads to the extraction of more accurate frequent paths.

As an attempt to capture semantic content of posts, authors in [3,4] proposed to use a vocabulary for all content-tokens of a given topic or content-specific flow mining; this way, the occurred content-tokens in posts are treated as regular

tokens (i.e.; U_r). In order to seek for all topics and not only specific ones while keeping track of the propagation of important exact content, we propose to leverage web intelligence to extract concepts and keywords from actors posts that will replace the original content-tokens of each actor's post.

Afterwards, we propose to use the *InFlowMine* algorithm [3,4] to mine frequent information flow paths. For this sake, as in [3,4], we use a hash table to track the sequence of actors for each extracted concept C_i and keyword $U_i^{'}$ from the post containing the original set of content-tokens $\{U_i\}$. Each slot of the hash table $h(C_i)$ corresponds to an ordered list of actors ordered chronologically on the basis of posting the concept C_i. Similarly, each slot of the hash table $h(U_i^{'})$ corresponds to an ordered list of actors in chronological order of posting the keyword $U_i^{'}$. Figure 1 depicts the proposed approach.

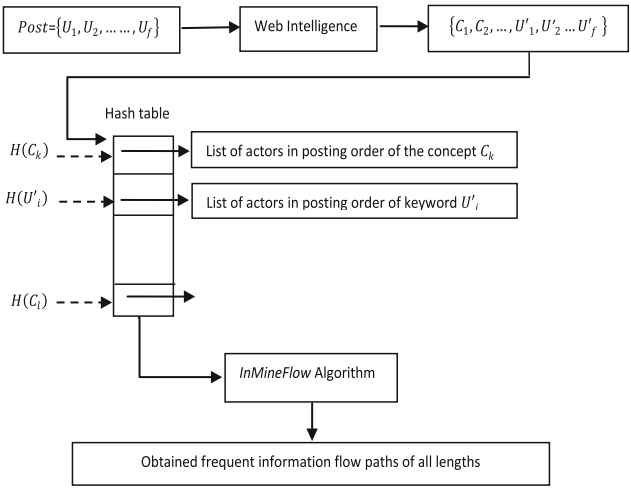

Fig. 1. Diagram chart of the proposed approach

The web intelligence phase mentioned in Fig. 1 and depicted in Fig. 2 is crucial; it consists of using IBM Watson Natural Language Understanding service (NLU) [5] that offers text analytics through a simple to use restful API framework. The IBM Watson NLU allows developers to leverage natural language processing techniques such as: analyzing plain text, URL or HTML and extracting meta-data from unstructured data contents for instance concepts, entities and keywords. Note that, it is not obvious extracting concepts and keywords from one single content-token U_i or a low content string, unless the content-token is itself a keyword. In this case, a solution is inspired by the works in [6,7], a context may be created using Google search restful API [8]. This way, the titles and the snippets of the k-top results of search are used to build an enhanced text.

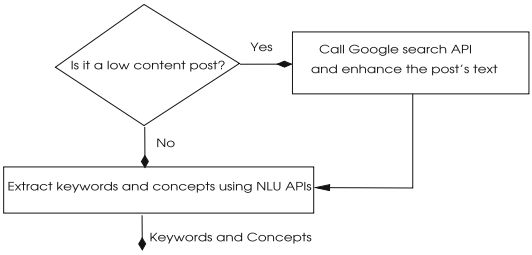

Fig. 2. Web intelligence phase

4 Experiments and Results

Due to Twitter restrictions, datasets used in previous works are no longer available nor sharable. The only free way to access Twitter data is through their restful APIs that deliver amongst others information; tweets and followers or friends of a given user. However, Twitter APIs or methods have restrictions, every method allows only 15 requests per rate limit window of about 15 min. Hence, it will take long time to get a small dataset with some data lost between calls. For more details about the problem one could face when conducting research on Twitter we refer the reader to [9]. To tackle this problem and to test our proposed approach, we have conducted our experiments on a collaborative dataset namely Core Dataset [10,11] assimilating it to a social network. Effectively, as an underlying graph, we have considered the co-authorship graph that we have built using the meta-data field "author list" of the repository records. Moreover, we have considered as a message the concatenation of the title and the abstract of each paper; messages are posted by the first author to all its co-authors that are assimilated to followers. Note that co-authors are not considered as followers of each other, they are only direct followers of the first author of a given paper.

It should be noted that, an important motivation of our choice to conduct experiments on a collaborative network, is the fact that in this kind of networks the co-authors share more likely the same topics.

Core dataset is a collection of open access research outputs from repositories of journals worldwide. It allows free and unrestricted access to research papers. The Core Dataset offers two datasets, (1) a meta-data file of 23.9 million items and a content dataset of 4 million items. The former contains meta-data on scientific papers in JSON format structured in 645 repositories, and the latter contains full articles. We carried out our experiments using *repository 2* of the last dump of the Core dataset (i.e., dump of October 2016). Figure 3 depicts an example of a meta-data dataset item. As a first step, we imported the repository 2 file as a collection to the NoSQL database *MongoDB* [12]. Secondly, we sorted the information about the papers in ascending order by publication date ("dc:date" as shown in Fig. 3) to conserve the chronological order of posting. Afterwards, we used the IBM Watson NLU APIs to generate keywords and concepts; then, we built the hash table as a python dictionary. Finally, we used

{"identifier":30474512,
"ep:Repository":2,
"dc:type":{},
"bibo:shortTitle":"Learning from triads: training undergraduates in counselling skills",
"bibo:abstract":"Background:\\ud\nResearch has shown that counselling skills training....
.....tutors must act proactively to ensure a safe learning environment",
"bibo:AuthorList":("Smith, Kate"),
"dc:date":"2015-05-06",
"doi":"10.1002/capr.12056",
"bibo:cites":{},
"bibo:citedBy":{},"
similarities":{}}

Fig. 3. Example of a Core dataset meta-data item

the *InFlowMine* algorithm to extract frequent paths from the hash table built earlier.

4.1 Results and Discussion

As first experimental tests of our approach, we have set frequency f to 10 (ten content-tokens were diffused through all the detected paths) and obtained all paths of length 1, 2, 3 and 4 as maximal possible length extracted from the repository 2 of the Core dataset. Figure 4 depicts some of the detected length-3 paths. A line of the results below represents a path of length-3, with each author separated by a semi-colon.

Havard, Catriona;Memon, Amina;Gabbert, Fiona;
Williams, P. K. G.;Stark, Craig R.;Helling, Ch.;
George, Keith P.;Grant, Marie Clare;Baker, Julien S.;
Ivanova, Iva;Pickering, Martin J.;McLean, Janet F.;
Tummala, Hemanth;Khalil, Hilal S.;Mitev, Vanio;
Clifford, Brian R.;Havard, Catriona;Memon, Amina;
Clifford, Brian R.;Memon, Amina;Gabbert, Fiona;

Fig. 4. Example of length-3 paths extracted from the repository 2 of the Core dataset

The path "Havard, Catriona; Memon, Amina; Gabbert, Fiona;" shown in Fig. 4 means that the author "Havard, Catriona" was the first author of the paper that she co-authored with "Memon, Amina". In her turn, "Memon, Amina" was the first author of another paper that she co-authored with "Gabbert, Fiona". The order of the positions of the authors in the path indicates the chronological order of posting (i.e., who first has emitted this content-token) of the propagated content-token. This result means that ten different content-tokens were cascaded trough this path with respect to the order of appearance of authors in the path.

Unlike the work proposed in [3,4] that uses a vocabulary to capture semantic aspects of posts, our proposal permits, in addition, the extraction of vocabularies. The concepts and the keywords collected in the hash table could be used to

build a vocabulary of the whole analyzed texts. The resulted vocabularies allow, among others, the detection of topical similarities between different posts of social networks or research papers of collaborative networks.

To assess the improvement gained by our approach, we started by considering, as a first experiment, the extracted keywords from the text with frequency set to 10. As a second experiment, we considered concepts keeping the same frequency as in the first experiment. Table 1 below shows early results in terms of number of mined frequent paths.

Table 1. Early results of our proposed approach

Path length	# mined frequent paths - frequency = 10	
	Using keywords	Using concepts
2	438	634
3	20	23
4	1	2

Basically, it fall in common sense that actors in a social or a collaborative network may use their own vocabulary while keeping cascading the same topic. We expected that the length and the number of mined frequent paths may increase if concepts or topics are used for mining instead of considering keywords. However, we noted that with low frequencies the numbers of mined frequent paths of short length are better if keywords are used as shown in Table 2. We explain these findings by the fact that the vocabulary (keywords) of actors tend to be conserved in their neighborhood. Moreover, with high frequencies, concepts give better results in terms of path numbers than keywords. This is due to the fact that the extracted keywords represent only a subset of representative terms of a given concept.

Table 2. Results with low frequencies

Path length	# mined frequent paths - frequency = 2	
	Using keywords	Using concepts
2	2756	2113
3	112	93
4	7	10

Table 3 shows a comparison of the results obtained in case of maximal length frequent paths. These results reveal that concepts performs well than keywords in case of long cascades. Hence, in the case of longer cascades the vocabulary vanishes and only concepts (topics) persist. In spite of the aforementioned cases,

considering topics in frequent paths detection reveals valuable information, that could be missed when using merely the exact content.

Table 3. Mined paths of maximal detected length with different frequencies

Frequency	# mined frequent paths of maximal length	
	Using keywords	Using concepts
1	9	10
2	7	10
3	6	8
4	4	7
5	4	6
6	4	6
7	3	5
8	2	2
9	1	2
10	1	2

Figure 5 depicts all the extracted frequent paths of maximal length and their respective propagated concepts with a frequency set to 3. We have noted when analyzing the abstracts and the short titles that in the case of the 8th path, the propagated concepts are somewhat generic which is due to the employed technique of concepts extraction.

1. Tinlin, Rowan M.;Watkins, Christopher D.;DeBruine, Lisa M.;Jones, Benedict C.;
2. Quist, Michelle C.;Watkins, Christopher D.;DeBruine, Lisa M.;Jones, Benedict C.;
3. Clifford, Brian R.;Havard, Catriona;Memon, Amina;Gabbert, Fiona;
4. Davies, J. W.;Butler, D.;Jefferies, Christopher;Duffy, A.;
5. Simpson, Edward;Gilmour, Daniel J.;Blackwood, David J.;Isaacs, John P.;
6. Phillips, P. J.;Jamison, S. P.;Berden, G.;van der Meer, A. F. G.;
7. Phillips, P. J.;Jamison, S. P.;Berden, G.;MacLeod, Allan M.;
8. Scott-Brown, Kenneth C.;Gilmour, Daniel J.;Blackwood, David J.;Isaacs, John P.;

Respective propagated Concepts:

1. Anorexia nervosa; Nutrition; Obesity;
2. Characteristic; Histrionic personality disorder; Evidence;
3. Immune system; Pharmacist; Mycelium;
4. Evidence; Critical thinking; Major;
5. Human resources; Management; Project;
6. Holography; Fundamental physics concepts; Optics;
7. Optics; Systems of measurement; Measurement;
8. Management; Higher education; Learning;

Fig. 5. Extracted paths and their respective propagated concepts- frequency set to 3

In this specific case there are no propagated keywords; what explains that the path was not detected when using keywords as shown in Fig. 6.

```
1.  Isaacs, John P.;Blackwood, David J.;Gilmour, Daniel J.;Falconer, Ruth E.;
2.  Tinlin, Rowan M.;Watkins, Christopher D.;DeBruine, Lisa M.;Jones, Benedict C.;
3.  Quist, Michelle C.;Watkins, Christopher D.;DeBruine, Lisa M.;Jones, Benedict C.;
4.  Simpson, Edward;Gilmour, Daniel J.;Blackwood, David J.;Isaacs, John P.;
5.  Stojanovic, V.;Blackwood, David J.;Gilmour, Daniel J.;Falconer, Ruth E.;
6.  Clifford, Brian R.;Havard, Catriona;Memon, Amina;Gabbert, Fiona;

Respective propagated keywords

1.  rural development project; rural planning projects; inclusive decision making;
2.  Positive correlations; facial characteristics; facial attractiveness propose;
3.  dominance questionnaire; average facial characteristics; facial masculinity;
4.  post project review; formal learning context; large redevelopment project;
5.  GPU shader programs; rendering methods; Short paper version;
6.  TP line-up; video identification parade; TP line-ups;
```

Fig. 6. Extracted paths and their respective propagated keywords- frequency set to 3

5 Conclusion

In this paper we have presented an approach to extend previous works on information flow frequent paths detection. The main objective is to capture in addition semantics of propagated information in a given social stream, considering in that both underlying graph structure and the content of interactions. Our experiments were conducted on a collaborative dataset that we have assimilated to a social network; the results showed an improvement that might reveal useful information missed when considering only cascades of exact content. As findings, we have noted that the vocabulary (keywords) of actors tend to be reused in their neighborhood, but vanishes in the case of long cascades and considering concepts remain the best alternative.

References

1. Zafarani, R., Abbasi, M., Liu, H.: Social Media Mining. Cambridge University Press, Cambridge (2014)
2. Guilles, A., Hacid, H., Fabre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. ACM SIGMOD **42**(2), 17–28 (2013)
3. Subbian, K., Aggarwal, C., Srivastava, J.: Content-centric flow mining for influence analysis in social streams. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 841–846 (2013)
4. Subbian, K., Aggarwal, C., Srivastava, J.: Mining influencers using information flow in social streams. ACM Trans. Knowl. Discov. Data **10**(3), Article 26 (2016)
5. IBM NLU API. https://www.ibm.com/watson/developercloud/natural-language-understanding/api/v1
6. Singhal, A.: Leveraging Open Source Web Resources to Improve Retrieval of Low Text Content Items, Ph.D. thesis in Department of Computer science, university of Minnesota, Minneapolis, MN, 144 (2014)
7. Singhal, A., Kasturi, R., Sivakumar, V., et al.: Leveraging web intelligence for finding interesting research datasets. In: IEEE/WIC/ACM Proceedings of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 321–328. IEEE (2013)
8. Google search API. https://developers.google.com/custom-search/json-api/v1/overview

9. Kelly, P.G., Sleeper, M., Cranshaw, J.: Conducting research on twitter: a call for guidelines and metrics. In: CSCW Measuring Networked Social Privacy Workshop (2013)
10. Knoth, P., Zdrahal, Z.: CORE: three access levels to underpin open access. D-Lib Mag. **18**(11/12) (2012). Corporation for National Research Initiatives
11. Knoth, P.: From Open Access Metadata to Open Access Content: Two Principles for Increased Visibility of Open Access Content. Open Repositories 2013, Charlottetown, Prince Edward Island, Canada (2013)
12. Chodorow, K.: MongoDB The Definitive Guide. O'Reilly Media Inc., Sebastopol (2013)