UDK 004.773.2

Z. Holub

THE ALGORITHM FOR DETECTING ONLINE DISCUSSION FRAGMENTS CONTAINING INFORMATION AND PSYCHOLOGICAL MANIPULATION

Annotation. *The paper scrutinizes the algorithm for detecting suspicious fragments of online community discussions that potentially contain information and psychological manipulation precedents. The classes of criteria for detecting suspicious discussion fragments are presented; the system of filters for detecting the fragments is described. The stages of the algorithms for detecting suspicious fragments are detailed.*

Keywords: *information and psychological manipulation, suspicious discussion fragment, online community*

Introduction

The internet means of communication are the most common and widely spread way of communication nowadays. Information transmission speed, vast audience, source diversity, the possibility of overcoming territorial limits make the internet means of communication attractive for informing and information search.

Due to the little number of requirements for publishing information (e.g. registration) and a great number of informers and recipients, the significant deal of internet-communication takes place on online community platforms. However, online community members could be exposed to the harmful effect of information and psychological manipulation. Information and psychological manipulation (IPM) is a deliberate influence on the subconsciousness of online community members by means of information and utilizing psychological mechanisms with the aim to affect thought process and reach one party vested goal.

Formulation of problem

The high speed of information sharing is typical for online communities. Therefore the prerequisite of IPM prevention is the timely detection of IPM precedents. The last requires the large unit of people with specific qualifications or application of automated methods and means.

The aim of the paper is to develop the algorithm for detecting discussion fragments that potentially contain IPM. The algorithm foresees application of the system of filters for detecting suspicious discussion fragments [1]. This process takes place at the second stage of the algorithm for monitoring online community with the aim of IPM detection [2]. The output of the considered in the paper algorithm is a set of suspicious discussion fragments, which are sent to the next stage of the algorithm for monitoring online communities with the aim of detecting IPM in order to identify IPM precedents.

Analysis of recent research and publication

Research in different fields and areas is aimed at increasing effectiveness of internet communication [3], in particular, increasing and using advantages and preventing negative phenomena of internet communication. Numerous studies are devoted to linguistic [4], intentional [7], behavioral [5, 6] and psychological mechanisms and means of communicative acts realization in internet media.

The basic material

The detection of suspicious discussion fragments is one of the two interim tasks fulfilled at the second stage of the algorithm of monitoring online communities with the aim of detecting IPM. The two interim tasks of the stage are detecting suspicious discussion fragments and detecting of precedents of IPM in online discussions.

Suspicious discussion fragments are the sets of logically connected messages, which quantitative and qualitative features as well as quantitative and qualitative characteristics of the profiles of their authors are characteristic to messages containing IPM.

Criteria for detection of a suspicious discussion fragment

The system of filters and means for detecting IPM tactics are based on different criteria. The system of filters base on static criteria that are utilized for suspicious profiles detection and dynamic criteria of the surface level that are used for detecting clusters of suspicious messages [9]. Meanwhile, means for detecting IPM tactics are based only on dynamic criteria.

Criteria that signal the potential existence of IPM are classified into two temporal classes, namely dynamic and static. The feature of the classification is a time period required for gathering and processing information necessary for computing the criterion.

- Static criteria are computed on the basis of members activities during the set period of time.
- Dynamic criteria do not require gathering data over the set time period, the presence of IPM can be stated right after determining them.

By means of static criteria information activity of a member is considered in relation to three structural and organizational levels of the content of an online community. According to the formal model of an online community, these three levels are community level, discussion levels, and message level. [1, 9]. Criteria of these three organizational and structural levels differ from one another in the computation mechanism and importance.

The values of criteria of the discussion and community levels are considered in relation to a community member. By means of these criteria, suspicious community members are detected. The activity of the latter is to be analyzed for IPM precedents.

Static criteria of the messages level point at elements of the content of the discussion that are to be analyzed for IPM presence.

Dynamic criteria are used for analysis of the particular act of information activity. They contain no generalized information about the role and behavior of the member of an online community.

The algorithm for detecting a suspicious discussion fragment

The analysis of an online community with the aim of detecting suspicious fragments of a discussion starts with processing discussion by filters of the highest level, namely community level. The criteria of community level do not demand complicated calculations, therefore these criteria make it possible to analyse the bulk of information without significant time and resource spending. If the fragment of a discussion is identified as manipulative by the present number of filters of the level N^{Filter}, it is not send to the filters of next levels, but saved in database as a suspicious discussion fragment.

At the next level, namely discussion level, at first are analysed the fragments which were trapped by the number of filters closest to N^{Filter}. If the sum total of all filters they were trapped by at all levels equals, N^{Filter}, then fragments are marked in database as the one’s that contain IPM. For instance the number of filters that trapped a discussion fragment at the community level and at the discussion level equals the predefined by expects threshold value, then the fragment is considered to be suspicious and is sent to the next level for detecting IPM precedents by applying deep analysis.

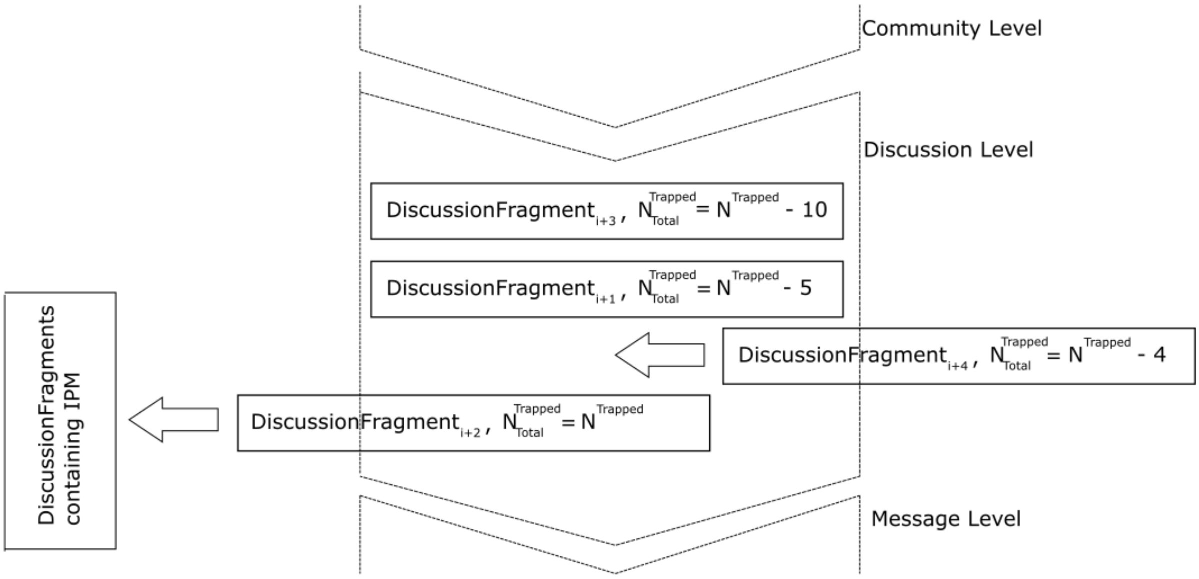
Community Level
Discussion Level
$\text{DiscussionFragment}_{i+3}, N^{Trapped}_{Total} = N^{Trapped} - 10$
$\text{DiscussionFragment}_{i+1}, N^{Trapped}_{Total} = N^{Trapped} - 5$
$\text{DiscussionFragment}_{i+4}, N^{Trapped}_{Total} = N^{Trapped} - 4$
DiscussionFragments containing IPM
$\text{DiscussionFragment}_{i+2}, N^{Trapped}_{Total} = N^{Trapped}$
Message Level

Figure 1. Sorting discussion fragments

Every criterion used by the filter for detecting specious precedents has different value in light of IPM detection process. Therefore a filter is assigned a value, predefined by experts.

The filters of each succeeding level require more data for calculations than of the preceding level, therefore it is reasonable to apply them to discussion fragments that were checked, but not trapped by the filters of other previous levels.

The subsequent application of criteria of the next levels only to the discussion fragments that were not suspected in presence of IPM, decreases the volume of information queuing to be checked for containing IPM. Furthermore, within each level simple filters that do not require complicated calculations are exploit first. In case of not detecting enough indications of IPM, they are subjected to further check by more sophisticated compound filters. This is done in order to increase the efficiency of the performance of the system of filters.

The number of community discussions a member posts in is an example of a compound filter. If a member takes an active part in numerous discussions, but his posts are relevant only to a small number of the discussions, then the member is considered to be a manipulator.

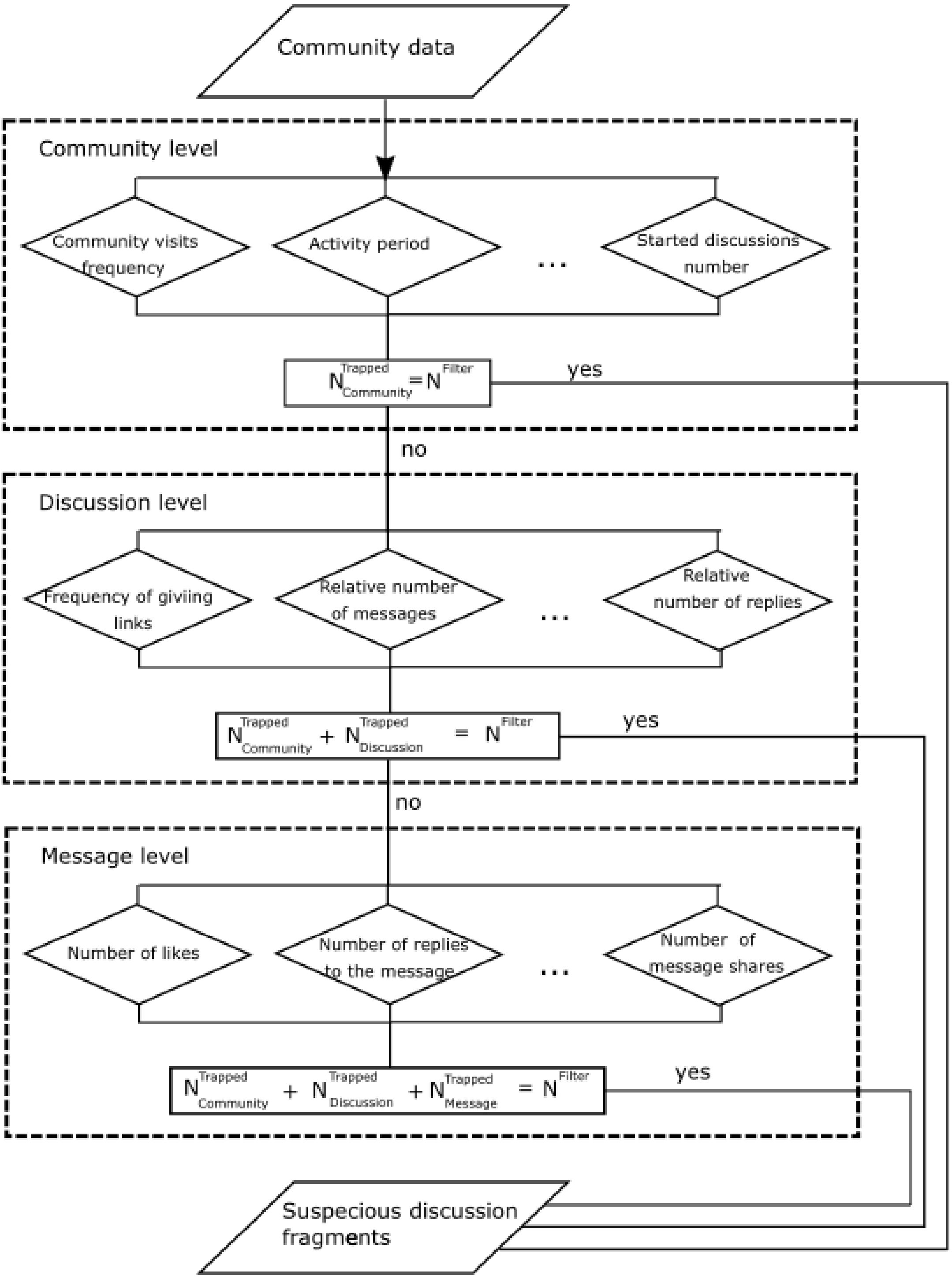
Community data
Community level
Community visits frequency
Activity period
...
Started discussions number
$N^{Trapped}_{Community} = N^{Filter}$
yes
no
Discussion level
Frequency of giviing links
Relative number of messages
...
Relative number of replies
$N^{Trapped}_{Community} + N^{Trapped}_{Discussion} = N^{Filter}$
yes
no
Message level
Number of likes
Number of replies to the message
...
Number of message shares
$N^{Trapped}_{Community} + N^{Trapped}_{Discussion} + N^{Trapped}_{Message} = N^{Filter}$
yes
Suspecious discussion fragments

Figure 1. Filtering discussion fragments according to the organization and structural levels

The number of IPM markers identified at the preceding interim stage defines the order of processing the discussion fragments on the next interim stage. For instance, if a suspicious discussion fragment was detected by filters of the community level, this fragment is analysed before the fragment that was detected by filters of community and discussion levels. The sequence of fragments according to which they are going to be analysed for the presence of IPM precedents, depends on the relation of the quantity of filters, which trapped the discussion fragment to the number of filters the fragment was processed by. The smaller the value of the relation is the sooner the fragment is check for presence of IPM precedents.

The output of the system of filters is the list of discussions that potentially contain IPM. The discussion fragments in the list are arranged according to the urgency of their check for IPM precedents presence.

Conclusion

The key prerequisite of an effective functioning of an online community is detection and neutralization of harmful information activities, that are beneficial only to their stakeholders. Existence of IPM precedents in an online community causes the decrease in content quality, member's credit and as a result leads to the member's quitting the community and the end of active functioning of the community.

Owing to the criteria devised on the basis of different user's profile characteristics and applying the system of filters the system for detecting suspicious discussion fragments is developed and the algorithm for accomplishment of the interim stage is suggested. The system of filters and the algorithm for detection suspicious discussion fragments enables the monitoring of online communities with the aim of IPM detection. The drawbacks like time spending and human resources required to check a huge bulk of information are excluded. The results of the system of filters are sent to the succeeding stage of the algorithm for monitoring online community, which foresees the deep analysis of content with the aim to detect IPM precedents and further identification of the IPM tactic.

REFERENCES

1. Peleschyshyn, A. Methods of real-time detecting manipulation in online communities / A. Peleschyshyn, Z. Holub, and I. Holub // Scientific and Technical Conference "Computer Sciences and Information Technologies (CSIT), 2016 XIth International. – IEEE, 2016 – pp.15-17.
2. Голуб З. Розроблення алгоритму виявлення шкідливих інформаційно-психологічних маніпуляцій в онлайн-спільнотах ВНЗ / З. Голуб // Вісник Національного університету «Львівська

політехніка». Серія: Інформатизація вищого навчального закладу : збірник наукових праць. – Видавництво Львівської політехніки 2017 – № 879 – с. 33-41

3. Krombholz K. Fake Identities in Social Media: A Case Study on the Sustainability of the Facebook Business Model / K. Krombholz, D. Merkl, E. Weippl [electronic source]. https://www.sba-research.org/wp-content/uploads/publications/krombholzetal2012.pdf
4. Ritter, A., (2010) Unsupervised Modeling of Twitter Conversations / A. Ritter, C. Cherry, B. Dolan // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. - 172–180.
5. Angeletou S. Modelling and Analysis of User Behaviour in Online Communities / S. Angeletou, M. Rowe, H. Alani [electronic source]. https://pdfs.semanticscholar.org/c8e4/1ccbfd8b953464b8eaa58c9af6c3cf851661.pdf
6. O'Donovan F. T.Characterizing user behavior and information propagation on a social multimedia network / F. T. O'Donovan, C. Fournelle, S. Gaffigan, O. Brdiczka, J. Shen, J. Liu, K. E. Moore [electronic source]. https://arxiv.org/ftp/arxiv/papers/1305/1305.2091.pdf
7. Dey L. Opinion Mining from Noisy Text Data / L. Dey, S. K. Haque, A. Mirajul // Proceedings of the second workshop on Analytics for noisy unstructured text data – 2008. – p.83-90.
8. Iqbal, S. The survey of sentiment and opinion mining for behavior analysis of social media / S. Iqbal, A. Zulqurnain, Y. Wani et al.// International Journal of Computer Science & Engineering Survey (IJCSES). – 2015. – Vol. 6 p.21-27. – doi: 10.5121/ijcses.2015.6502
9. Peleschyshyn, A. Development of the System for Detecting Manipulation in Online Discussions / A. Peleschyshyn, Z. Holub // International Conference on Systems, Control and Information Technologies 2016. – Springer International Publishing, 2016. – vol. 543 – pp. 111-117.