Social network extraction based on web: 2. Strategies in superficial methods

To cite this article: M K M Nasution 2018 J. Phys.: Conf. Ser. 1116 022029

View the article online for updates and enhancements.
Social network extraction based on web: 2.
Strategies in superficial methods

M K M Nasution

Technologi Informasi, Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan 20155 USU, Medan, Indonesia
E-mail: mahyuddin@usu.ac.id

Abstract. Different superficial methods for extracting social network from Web have common approach, but also developed based on different strategies. This difference is due to the characteristics of the source of information that is the source of the social network. There is a hit count, URL address and snippet as a result of any search engine restore, it can efficiently be used to extract social networks with the right strategy. This paper describes some of the results based on the strategies that have been developed.

1. Introduction
Historically, the social networks based on the Web are continuously growing and bringing different meanings from the ordinary social networks [1]. The extraction of social networks from dynamic sources of information shows the meaning of trends, and this depends on the observed community, with which the trend can be predicted, so it is possible to determine alternatives to the problems faced by any community [2].

In an approach involving superficial methods as a primary alternative to extracting social network from the Web, in addition to understanding the behavior of each information source, it is also necessary to characterize it so that it is possible to set the social network extraction strategy of the Web [3]. However, the limited access of methods to information resources requires strategies that need to optimize the use of such access tools. This paper outlining some mutually supportive strategies in superficial methods to extract social network from Web.

2. Review and Motivation
Historically, the approach that describes the superficial method of extracting social network from the Web first aims to seek out the relation of social actors [4]. This method involves the concept of similarity to reveal information about the relationship of hidden information behind the Web [5]. Some extraction concepts are developed based on the social network requirements that need to be disclosed from information sources [6]. A social network of academics and researchers involves a simple concept involving hit counts in singleton [7] or doubleton [8], so the relationship between two academic actors is based on the strength of the relationship generated by the Jaccard coefficient

\[ s_{jc} = \frac{|\Omega_a \cap \Omega_b|}{|\Omega_a| + |\Omega_b| - |\Omega_a \cap \Omega_b|} \in [0, 1] \] (1)
where $|\Omega_a|$ and $|\Omega_b|$ are the hit counts in singleton for the social actors $a$ and $b$, respectively, and $|\Omega_a \cap \Omega_b|$ is the hit count in doubleton for both social actors [9].

Some similarity measurements are used as the strategies to ensure the correctness of the resulting information [10], such as cosine

$$s_{\text{cos}} = \frac{|\Omega_a \cap \Omega_b|}{\sqrt{|\Omega_a||\Omega_b|}} \in [0, 1],$$

mutual information

$$s_{\text{mi}} = \log \frac{n|\Omega_a \cap \Omega_b|}{|\Omega_a||\Omega_b|} \in [0, 1],$$

overlap coefficient

$$s_{\text{oc}} = \frac{|\Omega_a \cap \Omega_b|}{\min(|\Omega_a|, |\Omega_b|)} \in [0, 1],$$

dice coefficient

$$s_{\text{dc}} = \frac{2|\Omega_a \cap \Omega_b|}{|\Omega_a| + |\Omega_b|} \in [0, 1],$$

where $|\Omega_a| > 0$ and $|\Omega_b| > 0$. While matching coefficient is to ensure the existence of the relationship between two actor socials, i.e.

$$n_{ab} = |\Omega_a \cap \Omega_b|,$$

Eqs. (1), (2), (3), (4), and (5) can be used if Eq. (6) is greater than zero [11], we call it as the positive relations. Thus, different strategies are expressed by different researchers to gain social networks [12].

By involving different forms of query content, although resulting in a hit count, analytics and deployment produce different social networks [13]. In this case, usage of the name pattern and the social name in the query produce the different social networks [14]. The analysis of information generated by search engine has also been revealed by the researchers, this spawned some characteristics of the information source, thus composing different strategies for each information source [15]. Computation of URL addresses has resulted in different social networks in the quantity of relationships between social actors [16]. Of course, computing the power of relationships only can be generated by involving special similarity measurements $s_{\text{our}}$ [17].

Based on the information source generated by search engines, there is a set of words that are at the core of the events in which social actors exist [18, 19]. Therefore, the strength relation between social actors is based on the probability of each word to either the number of snippets or tfidf [20]. The use of snippet concepts to extract social networks is more sensitive than others when viewed from the point of truth, since the tendency of people involving almost identical words in their daily lives is likely to be higher [21, 22]. This sensitivity, especially needs to be considered for communities outside of academia or researchers [23].

Social network extraction strategies for non-academic communities have been developed in several approaches [24]. Homogeneously, the strategy for extraction the social network of artists is expressed by involving the overlap principle and taking into account the applicable thresholds [12]. This is to ensure that the closeness between the artists is well measured [25]. In the community, artists are in the same community and almost have no different behavior, but have noting in common in implementing their lives. Thus, the form of the relationship is considered weak, and this requires learning to specifically characterize the similarities revealed based on homogeneity. Heterogeneously, the strategy of extracting social network between companies from the Web is seen as more difficult, although the same concept based strategy is also used and developed to approach the characteristics of the form of relationship to be revealed. The
form of relationship between companies in this case that has been disclosed is an alliance between the companies and the formal relationship between the companies [26].

Generally, to develop the approaches for extracting social network from Web is by taking into account the strategy of dealing with the characteristics of the information source.

3. An Approach
Each strategy is involved in dealing with the characteristics of the information source so that the desired results are obtained optimally and efficiently. In the extraction of social networks from the Web, the superficial methods have been developed based on approaches to obtaining reliable information [10]. By involving any search engine, a strategy of using the seeds is developed to extract social networks [20], and based on the seeds pick up the other social actors from the Web, to build an extended social network [23].

By involving the information generated search engine, we can develop strategies to reinforce the resulting social network [14]. By involving the URL addresses built social networks, generally it provide more complex number of relationships. The extraction of meaning from the environment of each social actor, i.e. by involving the snippets is to do vectorization about each word so that it has a good value based on the probability or frequency of the words [27]. Furthermore, based on word vectors, the social network extraction strategy of the Web is validated to affirm the meaning of any relationship between social actors [28]. Based on Fig. 1, some modules perform appropriately the tasks to get the information: hit counts, lists of URLs,
and lists of snippets for each social actors \((a_i, a_j \in A)\) or a pair of social actors, where \(A\) is a list of social actors. The relations based on their information we aggregate to be the trust relation or an edge \(e(a_i, a_j) = v_iv_j\) for vertices \(v_i\) and \(v_j\). While the meaning of relations based on snippets semantically expressed by words \(w_k, k = 1, \ldots, K\).

4. Discussion: Role of strategies

Each strategy has its own role to support the approach involved to extract social networks. Besides the power relationships in social networks take on different meanings, different strategies give different social networks. Of the 469 social actors tested, we have 31,623 edges were extracted based on the strategy of using URLs from the Web. Fig. 2 shows a role based on relationship between Eq. (1) and Eq. (4) in the relative probability, where the relative probability is a ratio of any doubleton and maximum doubleton. Fig. 3 shows a role of \(s_{our}\) that group the strength relations into classes, while Fig. 4 group strength relations based on their strength meaning.

At each measurement, we determine the limits of the lower and upper bounds. The lower bound \((lb)\) is 0.05% of the maximum singleton while the upper bound \((ub)\) is 10% of the maximum singleton. Suppose categorized every seed into an academic role: professor (Prof), associate
Figure 4. Edges in relation between the meaning, $S_{soc}$ and relative probability

Table 1. Number of edges based on strategies and categories

| Parameters | $< lb$ | $lb < \ldots < ub$ | $> ub$ |
|------------|--------|---------------------|--------|
| Strategy   | Seeds  | Vertices | Edges | Vertices | Edges | Vertices | Edges |
| 1 Prof     | 9      | 797      | 1     | 33       | 0     | 0        |
| Ass Prof   | 19     | 1,183    | 1     | 17       | 0     | 0        |
| Lect       | 35     | 2,228    | 4     | 46       | 0     | 0        |
| 2 Prof     | 0      | 0        | 10    | 546      | 0     | 0        |
| Ass Prof   | 10     | 460      | 11    | 331      | 0     | 0        |
| Lect       | 33     | 1,533    | 5     | 211      | 0     | 0        |
| 3 Prof     | 6      | 608      | 4     | 125      | 0     | 0        |
| Ass Prof   | 17     | 1,013    | 2     | 39       | 9     | 0        |
| Lect       | 36     | 1,676    | 2     | 46       | 0     | 0        |
| 4 Prof     | 5      | 436      | 5     | 183      | 0     | 0        |
| Ass Prof   | 20     | 1,252    | 1     | 38       | 0     | 0        |
| Lect       | 34     | 1,662    | 3     | 56       | 0     | 0        |

professor (Ass Prof) and lecturer (Lect), the query content is based on the social name, pattern name, and has meaningful direction (the word as a keyword), we have social networks based on strategies like Table 1.

5. Conclusion

Based on information source already has been derived some strategies in superficial method by using some similarity measurements and involving characteristics of information source. So, they claim that some methods have produced the extracted social networks in difference. By categorizing social actors and the also different measurements also to type of information source we obtained different social networks.

References

[1] Nasution M K M, Elveny M, Syah R and Noah S A 2015 Behavior of the resources in the growth of social network Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community (ICEEI).
[2] Nasution M K M and Noah S A 2012 Information retrieval model: A social network extraction perspective
[3] Nasution M K M, Hardi M and Sitepu R 2016 Using social networks to assess forensic of negative issues Proceedings of 2016 4th International Conference on Cyber and IT Service Management (CITS)M.

[4] Kautz H, Selman B and Shah M 1997 ReferralWeb: Combining social networks and collaborative filtering Communications of the ACM 40(3).

[5] Kautz H, Selman B and Shah M 1997 The hidden web AI Magazine 18(2).

[6] Mika 2007 P Social networks and the Semantic Web Springer.

[7] Nasution M K M 2012 Simple search engine model: Adaptive properties Cornell University Library arXiv:1212.3906 [cs.IR].

[8] Nasution M K M 2012 Simple search engine model: Adaptive properties for doubleton Cornell University Library arXiv:1212.4702 [cs.IR].

[9] Nasution M K M 2018 Semantic interpretation of search engine resultant IOP Conference Series: Materials Science and Engineering 300(1)

[10] Matsuo Y, Mori J, Hamasaki M, Nishimura T, Takeda H, Hasida K and Ishizuka M 2007 POLYPHONET: An advanced social network extraction system from the Web Web Semantics: Science, Services and Agents on the World Wide Web 5(4).

[11] Deza M-M and Deza E 2006 Dictionary of Distances Elsevier.

[12] Jin Y Z, Matsuo Y and Ishizuka M 2007 Extracting social networks among various entities on the Web ESWC 2007 4519 LNCS.

[13] Nasution M K M 2018 Social network extraction based on Web: 1. Related superficial methods IOP Conference Series: Materials Science and Engineering 300(1).

[14] Nasution M K M and Noah S A 2017 Social network extraction based on Web. A comparison of superficial methods Procedia Computer Science 124.

[15] Nasution M K M, Sitompul O S and Noah S A 2018 Social network extraction based on Web: 3. The integrated superficial method. IOP Conf. Series: Journal of Physics: Conf. Series 978.

[16] Nasution M K M and Noah S A 2010 Superficial method for extracting social network for academics using web snippets Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6401 LNAI.

[17] Mahyuddin K M N, Sitompul O S, Nasution S and Ambarita H 2017 New similarity IOP Conference Series: Materials Science and Engineering 180(1).

[18] Nasution M K M 2014 New method for extracting keyword for the social actor Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8397 LNAI(PART 1).

[19] Finin T, Ding I, Zhou I and Joshi A 2005 Social networking on the semantic Web. The Learning Organization 5(12).

[20] Nasution M K M and Noah S A 2011 Extraction of academic social network from online database 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011.

[21] Nasution M K M, Hardi M and Syah R 2017 Mining of the social network extraction Journal of Physics: Conference Series 801(1).

[22] Nasution M K M, Syah R and Elveny M 2017 Studies on behaviour of information to extract the meaning behind the behaviour Journal of Physics: Conference Series 801(1)

[23] Tang J, Zhang J, Yao L, Li J, Zhang L and Su Z 2008 ArnetMiner: Extraction and mining of academic social networks Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[24] Kwak H, Lee C, Park H and Moon S 2010 What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web.

[25] Jin Y Z, Matsuo Y and Ishizuka M 2008 Ranking entities on the Web using social network mining and ranking learning WWW 2008.

[26] Jin Y Z, Matsuo Y and Ishizuka M 2007 Extracting inter-firm networks from World Wide Web The 9th IEEE International Conference on E-Commerce Technology and the 4th International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007).

[27] Nasution M K M and Sitompul O S 2017 Enhancing extraction method for aggregating strength relation between social actors Advances in Intelligent Systems and Computing 573.

[28] Nasution M K M 2016 Social network mining (SNM): A definition of relation between the resources and SNA International Journal on Advanced Science, Engineering and Information Technology 6(6).