IOP Conf. Series: Materials Science and Engineering **300** (2018) 012056 doi:10.1088/1757-899X/300/1/012056

Social network extraction based on Web: 1. Related superficial methods

Mahyuddin Khairuddin Matyuso Nasution*

Teknologi Informasi,Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan 20155 USU Medan Indonesia

E-mail: mahyuddin@usu.ac.id

Abstract. Often the nature of something affects methods to resolve the related issues about it. Likewise, methods to extract social networks from the Web, but involve the structured data types differently. This paper reveals several methods of social network extraction from the same sources that is Web: the basic superficial method, the underlying superficial method, the description superficial method, and the related superficial methods. In complexity we derive the inequalities between methods and so are their computations. In this case, we find that different results from the same tools make the difference from the more complex to the simpler: Extraction of social network by involving co-occurrence is more complex than using occurrences.

1. Introduction

The social network extraction is a model for constructing the data structure of the social community from the unstructured data, i.e. $\gamma_1 : A \to V$ and $\gamma_2 : A \times A \to E$, whereby A is a set of social actor names, V and E are the set of vertices and the set of edges, respectively, in $G = (V, E)$ as a graph [1]. So, as information source and social media, Web contain data about both the social actors and their relations [2]. The data in Web, over again, are growing very rapidly and dynamically changeable. In that order, there are some methods that have been developed by the researcher, and each has advantages and disadvantages [3]. In general, the methods for extracting social networks from the Web is by exploring the limited potential of search engine [4].

Extracting the social network is to establish a reliable method for revealing the trusted information about social network from dynamic media and large data like Web. Technology that may be able to optimally disclose the trusted information in certain constraint dimensions. This paper intends to analyze some methods of extracting social network from Web.

2. Related Work

The idea of the social network extraction from Web is based on the strength of relations by involving similarity distance. The strength relation comes from the results of the search of two occurrences of social actor names and co-occurrence of the social actor names. Therefore, if there are n social actor names differently, then we obtained $n(n-1)/2$ iterations computationally or exactly is as set $O(n^2)$ [5].

Some of systems have been implemented to generate social network from information sources are related Web. Therefore, the systems use the interrelated methods: The ReferralWeb [6]

Table 1. Iteration for queries and computation

No	Model	Iteration	Queries	Computation
1	Basic	$A = \{a_i \mid i = 1, \ldots, n\}$	n	
	Expanded	m keywords		$n(m(m-1)/2)$
2	Basic	$R = \{r_j \mid j = 1, \ldots, n(n-1)\}$	$n(n-1)/2$	$n(n-1)/2$
	Expanded	m keywords		$n(m(m-1)/2)$
	Threshold			$n(n-1)/2$

involved co-occurrence to find the experts on a sequence of social actors through the social network extraction. Although not disclosed that this system involves similarity distance, but the similarity by using search engine reveal clearly that results based on the implementation of similarity. On different occasion, one system has been developed (i.e. Flink) by a researcher to represent a social network of social actors through the concept similarities in the Web contents [7]. It powered by metadata such as FOAF. In other system, the researchers call it as POLYPHONET, systematically has been defined that to extract social network from Web needs a way of referring to the template, the theory of graph $G(V, E)$, which involves two steps: First define vertices (as social actors), and further define the relationship between them (the strength relations) [8]. To support the system, several measurements of co-occurrences take the decisive roles. One of them is Jaccard coefficient, i.e.

$$J_c = \frac{|X \cap Y|}{|X \cup Y|} \in [0, 1] \tag{1}$$

$|X \cap Y|$ is a cardinality of intersection of a set X and a set Y, and $|X \cup Y|$ is a cardinality of the joint of a set X and a set Y. The co-occurrence analysis is also used to extract the relations in heterogeneous community such as artists and the relations in complex community such as firms [9]. In this case, the analysis is done by adding keywords as appeals parameters in sentences that reveal the relations between firms [10, 11].

3. Kind of Approaches

Considering that the social network extraction from Web is so important [12]: Web as the information source, then be the knowledge, and based on trusted information the knowledge releases the needed wisdom [3]. Thus, the Web have some potentials that may serve as a basis for generating the social network. An extraction method which involves the search engine as the possible approach for crawling webpages [13, 14]. This approach we call as superficial method [15]. The method is to pick up information via the query q and generate hit count and snippet [16, 17, 18]. This is modeled as follows: For a pair of social actor names $a_i, a_j \in A$, query q assigned a_i or/and a_j, and any search engine will generate hit count $|a_i|$ and $|a_j|$ as occurrences and $|a_i \cap a_j|$ as a co-occurrence. So based on the queries, the search engine dedicate sa_i and sa_j as collections of snippets of a_i and a_j, respectively, and $sa_j \cap sa_j$ is a collection of snippets associated with a pair of actors [19].

Each snippet contains URL addresses, web title, and summary of webpage. The canonical form of URL address is

$$u = http(s) : //d_m \ldots d_2\dot{d}_1/p_1/p_2/\cdots/p_{n-2}/x \tag{2}$$

whereby $\{s, d, p, q\} = \{\text{scheme,authority,path,query}\}$, $x = p_{n-1}$ or $x = p_{n-1}q$. Each URL address has n layers where each layer is separated by a slash [15]. While web title and summary of webpage contain words, and each snippet contains $1, \ldots, max = \pm 50$ words [20].

Thus, the model expresses some methods for extracting social network from Web as follows.

Actors	Extraction Method	Occurrence		Co-occurrence		Scale of	
		Iteration	Content	Iteration	Content	Queries	Computation
n	BSM	n	a_i	$n(n\text{-}1)/2$	a_i,a_j	$(n^2+n)/2$	$n(n\text{-}1)/2$
	BSMv		a_i,kw		a_i,a_j,kw		
	PSM		"a_i"		"a_i","a_j"		
	PSMv		"a_i","kw"		"a_i","a_j","kw"		
	USM		ua_i	-	-	n	
	JUSM	-	-	$n(n\text{-}1)/2$	ua_i,ua_j	$(n^2+n)/2$	
	PUSM	n	$ua_{"i"}$	-	-	n	
	JPUSM	-	-	$n(n\text{-}1)/2$	$ua_{"i"},ua_{"j"}$	$(n^2+n)/2$	
	DSM	n	sa_i	-	-	n	
	JDSM	-	-	$n(n\text{-}1)/2$	sa_i,sa_j	$(n^2+n)/2$	
	DPSM	n	$sa_{"i"}$	-	-	n	
	JDPSM	-	-	$n(n\text{-}1)/2$	$sa_{"i"},sa_{"j"}$	$(n^2+n)/2$	

Figure 1. Type of methods for extracting social network from Web.

3.1. Basic superficial method

The basic superficial method (BSM) involves 3 (three) times to submit queries to the search engine for a pair of social actors and the computation by involving Eq. (1) [6, 9, 21], $j_c = |a_i \cap a_j|/(|a_i| + |a_j| - |a_i \cap a_j|)$, is done with BSM procedure: `BSM(`a_i, a_j`)`, (a) $|a_i| \leftarrow q = a_i$, (b) $|a_j| \leftarrow q = a_j$, (c) $|a_i \cap a_j| \leftarrow q = a_i, a_j$, and `return` j_c. Therefore, like Table 1, for n social actors, number of queries is $n + n(n-1) = n^2$ or if in symmetric condition the number of queries is $n^2/2 + n/2$, while the computation scale is $n(n-1)/2$ times, where $|a_i \cap a_j| \leq |a_i|$, $|a_i \cap a_j| \leq |a_j|$, $|a_i| + |a_j| - |a_i \cap a_j| \neq 0$.

The BSM produces a strength relation between two social actors $a_i, a_j \in A$. This relation is influenced by social behavior [22], that is if the name of social actor is more than one word, then the other social actor name similar to one and more of the words (phrase) will be counted in, otherwise a name involves the matching pattern will generate exactly corresponding information [23, 24]. In last case, the name of the chosen social actor is a well-define name [25]. For example, the phrase Mahyuddin K. M. Nasution (without quotes) is a social name or in general a_i, "Mahyuddin K. M. Nasution" (in quotes) is a phrase in pattern form or "a_i". Generally there are not many of the same names in the form of patterns, whereas in the form of social names allows the similar names of social actors from anywhere to participate [26]. Therefore, to correctly identify the information source about a social actor, it requires the identity of each social actor, i.e. keywords kw [27]. Like Table 1, it is possible to have m keywords for each social actor, but the keyword serves to lift the related webpages from below of stack and reduces the ambiguity. Thus, BSM procedure can be expanded by adding keyword to each query: $|a_i| \leftarrow q = a_i, kw$; $|a_j| \leftarrow q = a_j, kw$; and $|a_i \cap a_j| \leftarrow q = a_i, a_j, kw$. Or we call it as the expanded basic superficial method or the basic superficial method with keyword (BSMv), see Fig. 1.

In co-occurrence concept, a name of social actor becomes a keyword (keyphrase) for the other name [28]. Therefore, if $|a_i \cap a_j| = 0$ or one actor has nothing to do with the other actor, then Eq. 1 be

$$j_m = \frac{|a_i \cap a_j|}{|a_i| + |a_j|} \in [0, 1] \tag{3}$$

where $j_m \leq j_c$. Moreover, computationally any strength relation generated from social names require a threshold [29]. It is to ensure the trusty of information about social network. In this case, j_m is the approximate value of threshold which is generally the lowest value of j_c. Thus, j_m is a threshold whereby $j_m \neq 0$ is the smallest for $n(n-1)/2$ potential strength relations. The addition of threshold computation on BSM will enhance BSM and is called as the enhanced basic superficial method (EnBSM) [30].

In case of using the name in pattern form and each query contains the couple among co-occurrence with keyword, then social network based on keyword be a community layer. In other words, there are m concentration of social actors based on m keywords, see Fig. 1.

3.2. Pattern superficial method

This method is as development of basic superficial method whereby each query contains entries in quotes. Therefore, complexity of the pattern superficial method (PSM) equal to BSM for n social actors. Similar to BSM, this method have the generated methods such as the pattern underlying superficial method (PUSM), the joint pattern underlying superficial method (JPUSM), the description pattern superficial method (DPSM), and the joint description pattern superficial method (JDPSM), see Fig. 1.

3.3. Underlying superficial method

The underlying superficial method (USM) involves twice submitting queries to the search engine for a pair of social actors [15]. Based on Eq. (1), each query q produces a list of snippets and then a list of URL addresses, or g_a URL addresses for a social actor, i.e. $L(u_a)$. Each URL has c duplication, m is a number of URL parts, thus u to be cm/n_i, n_i is layer of URL. Therefore, we get $|a_i| = \sum_{j=1}^{g_a}(c_j m_j/n_i)^2$. While each same URL of queries q_a and q_b produces the vector $qq_j = (c_{j_a} m_{j_a}/n_a)(c_{j_b} m_{j_b}/n_b)$, thus we get

$$|ab| = \sum_{j=1}^{g_{ab}}(c_{j_a} m_{j_a})(c_{j_b} m_{j_b})/(n_{i_a} n_{i_b}). \tag{4}$$

and then similarity based on USM as follows [31]

$$sim_{usm} = \frac{\sum_{j=1}^{g_{ab}}(c_{j_a} m_{j_a})(c_{j_b} m_{j_b})/(n_{i_a} n_{i_b})}{\sum_{j=1}^{g_a}(c_{j_a} m_{j_a}/n_{i_a})^2 + \sum_{j=1}^{g_b}(c_{j_b} m_{j_b}/n_{i_b})^2} \tag{5}$$

and the computation by involving Eq. (5) is done with USM procedure: `USM`(a_i, a_j), (a) $L(u_a) \leftarrow q = a$, (b) $L(u_b) \leftarrow q = b$, and `return` sim_{usm}.

The USM can also be expanded and/or enhanced by involving keywords and threshold usage. Keyword inclusion will reduce the number of URL addresses being processed while the involvement of threshold point measurements will reduce the similarity between the lower layers. In general, for n social actors it takes n times the submitting to a search engine, but computationally it takes $n(n-1)/2$ iterations potentially and is charged $g_{ab} = g_a g_b$ iterations for comparison between URL addresses based on a pair of social actors. In another case, co-occurrence also generates a set of URLs for two different social actors, and this will cause in the number of queries being $n^2/2 + n/2$ times submit to the search engine, even though the computation of the Eq. (4) will be simplified into $|ab_i| = \sum_{j=1}^{g_{ab}}(c_j m_j/n_i)^2$. See Fig. 1, for the generated and related methods.

3.4. Description superficial method

Similar to USM, the description superficial method (DSM) also involves 2 (two) queries. Each query produces a list of snippets, or it produces a group of the words sets. Thus, each set of words has probability of sets and probability of word, with which each word has a weight. In other word, based on list of snippets, we have a set of vocabularies with their weights, we call it as the weighted words. The words in snippets are the description of a social actor. By using a similarity measurement towards two lists of snippets, we get the description relation between two social actors [14]. Let W_a and W_b are the sets of weighted vocabularies for social actor $a, b \in A$, respectively. A set of weighted vocabularies for both social actors a and b in A is W_{ab},

then we have a relation based on Eq (1), i.e. $sim_w = W_{ab}/(W_a + W_b - W_{ab})$. To get a weight from two weights of same words of the different social actors we propose using average of two weights, i.e. $w_{ab} = (w_a + w_b)/2$, $w_{ab} \in W_{ab}$, $w_a \in W_a$ and $w_b \in W_b$ [32]. See Fig. 1, for the generated and related methods.

The computational implementation procedures for this method involve the same concepts like USM either involving just occurrence or by involving co-occurrence as the treatment, and likewise about behaviors of complexity of this method is dependent on the use of same treatment.

3.5. Seed based superficial method

This method is developed with the concept of association rule. By the intent that search engine and query are used to obtain a URL address corresponding to the pattern of the social actor name and a name of the referenced online database. The database contains a list of scientific papers: author names, title, year, and events. Social network of authors are formed following a record of online database, and the descriptions of social network refer to titles of scientific paper. Therefore, through the titles that consist of words, we can use the concept of DSM for generating the clue of concentration of the scientists community [25]. We call this method as the seed based superficial method (SSM).

For n social actors this method employs n queries. Each social actor will generate $0, \ldots, m$ others social actors, and SSM will produce $0, \ldots, m$ relations with $0, \ldots, p$ edges, $p \leq m$, because each edge may be consists of more than one relations between two social actors. So, SSM first produces a social network based on the social actors or a social network with one social actor as central, and for n social actors this method generate the social networks. Therefore, the complexity computationally of SSM is $n(m_i)$, $i = 1, \ldots, n$, m_i is number of record rows in online database.

4. Conclusion

Using the same source, Web, there are several methods that can be employed to extract social networks. Each method has a different complexity than others, and they computationally involve multiple iterations. In general, methods involving sources related to snippets will have lighter complexity than others, as well as iterations used. Based on their complexity, the integration of methods is likely to be mutually supportive for generating credible social networks, but this needs the comparison study.

References

[1] P Mika (2007) *Social Networks and the Semantic Web*. Springer-Verlag: Berlin.
[2] M K M Nasution and S A Noah 2012 Information retrieval model: A social network extraction perspective *Proceedings of 2012 International Conference on Information Retrieval and Knowledge Management* (CAMP'12).
[3] M K M Nasution 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)**.
[4] Y Matsuo, J Mori, M Hamasaki, T Nishimura, T Takeda, K Hasida, and M Ishizuka 2007 POLYPHONET: An advanced social networks extraction system from the Web *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* **5**.
[5] M K M Nasution, Marischa Elveny, Rahmad Syah, and S A Noah 2015 Behavior of the resources in the growth of social network *5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community* (ICEEI).
[6] Henry Kautz, Bart Selman, and Mehul Shah 1997 ReferralWeb: Combining social networks and collaborative filtering *Communication of the ACM* **40(3)**.
[7] Peter Mika 2005 Flink: Semantic Web technology for the extraction and analysis of social networks *Web Semantics: Science, Services and Agent on the World Wide Web* **3 (2-3)**.
[8] Yutaka Matsuo, Junichiro Mori, Masahiro Hamasaki, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka 2006 POLYPHONET: An advanced social network extraction system *Proceedings of the 15th International Conference on World Wide Web 2006* (WWW 2006).

[9] YingZi Jin, Y Matsuo, and M Ishizuka 2006 Extracting a social network among entities by web mining *Workshop on Web Content Mining with Human Language* (ISWC 2006).
[10] YingZi Jin, Y Matsuo, and M Ishizuka 2007 Extracting social networks among various entities on the Web *ESWC 2007* LNCS **4519**.
[11] F Benhawi, N Mohamad Ali, and H M Judi 2012 User engagement attributes and levels in facebook *Journal of Theoretical and Applied Information Technology* **41(1)**.
[12] Jason Potts, Stuart Cunningha, John Hartley, and Paul Ormerod 2008 Social network markets: A new definition of the creative industries *Journal of Cultural Economics* **32(3)**.
[13] M K M Nasution 2017 Modelling and Simulation of Search Engine *Journal of Physics: Conference Series* **801(1)**.
[14] M K M Nasution, M Hardi, and R Syah 2017 Mining of the social network extraction *Journal of Physics: Conference Series* **801(1)**.
[15] M K M Nasution and S A Noah 2010 Superficial method for extracting social network for academics using web snipppets *Lecture Notes in Computer Science* **LNAI 6401**.
[16] M K M Nasution 2012 Simple search engine model: Adaptive properties *Cornell University Library*, arXiv:1212.3906v1.
[17] M K M Nasution 2012 Simple search engine model: Selective properties *Cornell University Library*, arXiv:13033964v1.
[18] M K M Nasution 2017 Semantic interpretation of search engine resultant *InteriOR*.
[19] M K M Nasution 2012 Simple search engine model: Adaptive properties for doubleton *Cornell University Library*, arXiv:1212.4702v1.
[20] M K M Nasution 2014 New method for extracting keyword for the social actor *Lecture Notes in Computer Science* **LNAI 8397 (PART 1)**.
[21] M K M Nasution, S A M Noah, and S Saad 2011 Social network extraction: Superficial method and information retrieval *Proceedings of International Conference on Informatics for Development* (ICID11).
[22] M K M Nasution, R Syah, and M Elveny 2017 Studies on behaviour of information to extract the meaning behind the behaviour *Journal of Physics: Conference Series* **801**.
[23] YingZi Jin, Y Matsuo, and M Ishizuka 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).
[24] YingZi Jin, Y Matsuo, and M Ishizuka 2008 Extracting inter-Firm networks from the World Wide Web using a general-puspose search engine *Online Information Review* **32(2)**.
[25] M K M Nasution and S A Noah 2011 Extraction of academic social network from online database *2011 International Conference on Semantic Technology and Information Retrieval*(STAIR 2011).
[26] Y Matsuo, H Tomobe, and T Nishimura 2007 Robust estimation of Google counts for social networks extraction *AAAI'07 Proceedings of the 22nd National Conference on Artificial Intelligence* **2**.
[27] D Bollegala, Y Matsuo, and M Ishizuka 2007 Measuring semantic similarity between words using web search engine *Proceedings of the 16th International Conference on World Wide Web* (**WWW'07**).
[28] R M Alguliev, R M Aliguliyev, and F S Ganjeliyev 2011 Extracting a heterogeneous social network of academic researchers on the Web based on Information Retrieved from multiple sources *American Journal of Operations Research* **1**.
[29] M K M Nasution and S A Noah 2012 A methodology to extract social network form the Web Snippet *Cornell University Library*, arXiv:1211.5877.
[30] M K M Nasution, R Syah, and M Elfida 2018 Information retrieval based on the extracted social network *Applied Computational Intelligence and Mathematical Method, Advances in Intelligent Systems and Computing* **662**.
[31] M K M Nasution, O S Sitompul, Sawaluddin Nasution, and H Ambarita 2017 New Similarity *IOP Conference Series: Materials Science and Engineering* **180(1)**
[32] M K M Nasution and O S Sitompul 2017 Enhancing extraction method for aggregating strength relation between social actors." *Advances in Intelligent Systems and Computing* (AISC) **573**.