

# Mining of the social network extraction

M K M Nasution<sup>1\*</sup>, M Hardi<sup>2</sup>, R Syah<sup>3</sup>

<sup>1,2,3</sup>*Informatin Technology, Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan USU Medan 20155, Indonesia.*

\*Email: mahyuddin@usu.ac.id

**Abstract.** The use of Web as social media is steadily gaining ground in the study of social actor behaviour. However, information in Web can be interpreted in accordance with the ability of the method such as superficial methods for extracting social networks. Each method however has features and drawbacks: it cannot reveal the behaviour of social actors, but it has the hidden information about them. Therefore, this paper aims to reveal such information in the social networks mining. Social behaviour could be expressed through a set of words extracted from the list of snippets.

## 1. Introduction

Based on the superficial method for extracting social network from information sources we have the social network as a resultant [1]. This method is pre-processing for transforming the raw data into the social network whereby Web as one of information sources [2]. Web be a picture of social change dynamically, and Web therefore represent the activities of social actors [3]. In other words, exploring Web as the social media is to allow us acquire the behaviour of either social actor personally or the community of social actors [4]. However, there is no extraction of social networks connected with the exploration of the behaviour of social actors. Thus, this paper aims to disclose formally the mining of the social network extraction: That is to develop the concept formally for extracting social network and propose some problems as basic concept and motivation; to propose an explanation of the problems and prove its importance in an approach systematically; and to conduct the extraction of social network and the related behaviour in an example.

## 2. Basic Concept and Motivation

Each network naturally we can model it as a graph  $G<V,E>$  whereby  $V$  is a set of vertices  $\{v_i|i=1,...,I\}$  and a set of the name labels  $\{a_i|i=1,...,I\}$ , and  $E$  is a set of edges  $\{e_j|j=0,...,J\}$ , a set of weights  $\{b_j|j=0,...,J\}$  and a set of labels  $\{l_k|k=0,...,K\}$ , or we can define the template of network as follows [5].

*Definition 1.*  $TN = G<V,E,L_1,L_2,B>$  is a template of network with conditions as follows

- T1  $V = v_i|i=1,...,I\}$  as vertices in  $TN$ .
- T2  $E = \{e_j|j=0,...\}$  as edges in  $TN$ .
- T3  $L_1 = \{a_i|i=1,...,I\}$  as labels of vertices.
- T4  $L_2 = \{l_k|k=0,...\}$  as labels of edges.
- T5  $B = \{b_j|j=0,...\}$  as weights of edges.

Each network not only has a set of vertices and a set of edges, but also it has a set of labels for vertices/edges and a set of weights for edges. Therefore, to develop a social network from information

sources, we choose the tools that are used to access information. Suppose for a social actor 'Mahyuddin K. M. Nasution' we use search engines, if query  $q_1$  = 'Mahyuddin K. M. Nasution' or  $q_2$  = "Mahyuddin K. M. Nasution" submitted to any search engine, then we get the hit counts like Table 1. So there is a difference between  $q_1$  and  $q_2$ . The results of search engine have been contaminated by ambiguity in documents / web pages [6] or the semantic problems: meronymy, holonymy, hyponymy, synonymy, or polysemy [7]. The result of  $q_1$  explained that each social actor socially has many names, aliases or attributes, while the result of  $q_2$  still contains the possibility of same name for different people, although the exact pre-defined name for somebody. Therefore, to gauge out the appropriate information from the stack of documents, we always add an eligible keyword [8].

**Table 1.** Hit counts

Search engine	$q_1$	$q_2$
Google	16,100	1,110
Yahoo	1,980	244
Bing	213,000	2,090
Yandex	57,000	634

In the extraction of social networks, first we declare the social actors with names that are properly defined, to ensure every vertex  $v$  in  $V$  in template  $TN$  represents a social actor: A set of actors  $A = \{a_i | i=1, \dots, I\}$ . So there is a function  $\gamma_1$  that maps one by one actor to vertex, or  $\gamma_1(1:1): A \rightarrow V$  [9]. In this case, if a query  $q$  contain a term for  $a_i$  in  $A$ , then the query classify web pages into one class that is recognizable as a singleton event  $\mathcal{Q}_{ai}$  whereby number of occurrences (hit counts) is  $|\mathcal{Q}_{ai}|$  [2].

*Lemma 1.* If every web page representing the activity of a social actor, then the class of web pages represent the behavior of a social actor.

In addition, for  $|\mathcal{Q}_{ai}| > 0$  the search engine also bring back (return) a list of snippets, every snippet contains  $\pm 50$  words, and the label of social actors can be generated from this list [5].

In the classical social networks, the relations between social actors was transformed directly from the data, but in the extraction of a social network, a collection of relationship that may be contained in the document and translated through a search engine in the form of co-occurrence [5]: If the query  $q$  contains two terms  $a_i, a_j$  in  $A$ , then the query classifying web pages into one class that can be recognized as a doubleton event  $\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}$  whereby number of co-occurrences (hit counts) is  $|\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}|$ , with conditions  $|\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}| \leq |\mathcal{Q}_{ai}|$  and  $|\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}| \leq |\mathcal{Q}_{aj}|$  [2].

*Lemma 2.* If every web page representing the communication of social actor, then the class of web pages represents the behavior of the relation between social actors.

Let  $|\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}| > 0$ , the search engine returns a list of snippets that also contains information about 50 words in each snippet, and label of a relation can be generated from it through a procedure. All relations that may exist between social actors are accumulated into one and recognizable as weight of the strength relation by using the concept of similarity, for example Jaccard coefficient

$$sim_j = |\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}| / (|\mathcal{Q}_{ai}| + |\mathcal{Q}_{aj}| - |\mathcal{Q}_{ai} \cap \mathcal{Q}_{aj}|) \quad (1)$$

*Definition 2.*  $SN_e = G \langle V, E, A, R, \Gamma \rangle$  is an extracted social network with an operation  $\Gamma$  which satisfies the conditions

C1  $\gamma_1(1:1): A \rightarrow V, \gamma_1$  in  $\Gamma$ .

C2  $\gamma_2: R \rightarrow E, \gamma_2$  in  $\Gamma$ .

Gamma function contains a set of processes to gain information from the sources:  $\gamma_1$  for example is to declare the social actors based on MM method [8], and  $\gamma_2$  is the steps that must be done for exploring the relations between the social actors like the superficial method [10].

*Definition 3.*  $SN_e = G\langle V, E, A, R, I \rangle$  is an labeled social network based on extraction method  $\Gamma$  which satisfies the conditions C1, C2 and/or

C3  $\gamma_3(M:1): L_1 \rightarrow V, \gamma_3$  in  $\Gamma$  and  $L_1 = \{l_j | j=0, \dots, J\}$  as labels of vertices, or

C4  $\gamma_4: B \rightarrow E, \gamma_4$  in  $\Gamma$  and  $B = \{b_k | k=1, \dots, K\}$  as weights of edges, or

C5  $\gamma_5(M:1): L_2 \rightarrow E, \gamma_5$  in  $\Gamma$ , and  $L_2 = \{l_t | t=0, \dots, T\}$  as labels of edges.

*Theorem 1.* If the behavior of a cluster describes the behavior of a social actor, then the behavior of other actors expressed by the relationship between the clusters.

### 3. An Approach

In Web as social media, each web page can represent an actor in a way that is the name of actor is a part of text in web page and other words in the same web page as actions, ethics and attitudes of the actor [15]. In literal, the name of actor is a term  $t_a$  consists of at least one or a set of words or  $t_a = (w_1, \dots, w_l)$ ,  $l \leq k$ ,  $k$  as a number of words  $w$ ,  $l$  as a number of vocabularies in  $t_a$ ,  $|t_a| = k$  is the size of  $t_a$ , and a cluster based on a term we can generate it as follows [12].

*Definition 4.* Let  $\Omega$  as a space of the search engine for indexing the web pages. For a search term  $t_a$ , where  $t_a$  in  $\Sigma$ , that is a set of singleton search term of search engine. There are a dynamic space  $\Omega_a$  containing the ordered pair of the term  $t_{ai} \ i=1, \dots, I$  with web pages  $\omega_{aj} \ j=1, \dots, J$ , i.e.  $(t_{ai}, \omega_{aj}) = (t_a, \omega_a)_{ij}$  such that  $\Omega_a = (t_a, \omega_a)_{ij}$  is a subset of  $\Omega$  is a singleton search engine event of web pages (briefly we call it as *singleton*) that contains the occurrences of  $t_a$  in  $\omega$ .

A set of web pages as the cluster is the representation of the actor formally (Definition 4), whereby  $|\Omega_a|$  is the cardinality of  $\Omega_a$  (*hit count*), while the expression of behavior we can mine from a list of snippets.

*Proposition 1.* If  $|\Omega_a| > 0$  for a search term  $t_a$ , then there is a set of the weighted words  $W = \{(w_j, b_j) | j=1, \dots, J\}$ ,  $w_j$  is a word and  $b_j$  is a weight of word.

Proposition 1 as the direct consequence of  $\gamma_3$  in Definition 3 whereby  $b_j$  of words are generated by using term frequency ( $tf$ ) or probability of word  $p(w)$  in collection of words from the list of snippets. Therefore, we can rewrite Lemma 1 formally as  $|\Omega_a| > 0 \rightarrow \Omega_a \wedge \{(w_j, b_j)\}_a$  and Proposition 1 as the proof of Lemma 1.

*Definition 5.* [13] Let  $\Omega$  as a space of the search engine for indexing the web pages. For two search terms  $t_{ai}$  and  $t_{aj}$ , that is a set of doubleton search term of search engine. There are a dynamic space  $\Omega_{ai} \cap \Omega_{aj}$  containing the ordered pair of the terms  $t_{ai} \ i=1, \dots, I$  and  $t_{aj} \ j=1, \dots, J$  with web pages  $\omega_{aij}$ , i.e.  $(\{t_{ai}, t_{aj}\}, \omega_{aij}) = (t_a, \omega_a)_{ij}$  such that  $\Omega_{ai} \cap \Omega_{aj} = (t_a, \omega_a)_{ij}$  is a subset  $\Omega$  is a doubleton search engine event of web pages (briefly we call it as *doubleton*) that contains the co-occurrences of  $t_{ai}, t_{aj}$  in  $\omega$ .

*Proposition 2.* If  $|\Omega_{ai} \cap \Omega_{aj}| > 0$  for two search terms  $t_{ai}$  and  $t_{aj}$ , then there is a set of the weighted words  $W = \{(w_k, b_k) | k=1, \dots, K\}$ ,  $w_k$  is a word and  $b_k$  is a weight of word.

Similar to  $\gamma_3$ , Proposition 2 as a consequence of  $\gamma_5$  in Definition 3 and it prove Lemma 2 if  $|\Omega_{ai} \cap \Omega_{aj}| \leq |\Omega_{ai}|$  and  $|\Omega_{ai} \cap \Omega_{aj}| \leq |\Omega_{aj}|$  whereby  $|\Omega_{ai} \cap \Omega_{aj}|$  is the cardinality of  $\Omega_{ai} \cap \Omega_{aj}$ , while  $\{(w_k, b_k) | k=1, \dots, K\}$  as the expression of the relation behavior between actors [14].

$|\Omega_{ai} \cap \Omega_{aj}| > 0$  proves the existence of relationship between two social actors, whilst using Eq. (1) gives the strength relation as a weight. A weight of the relation  $\gamma_4$  be a behavior of same actors based on the relationship between the clusters, whereby  $\gamma_4(sim_j) * b_k$  be expression of behavior. It also can proved by the similarity of  $W_{ai} = \{(w_k, b_k) | k=1, \dots, K_i\}$  and  $W_{aj} = \{(w_k, b_k) | k=1, \dots, K_j\}$ ,  $W_{ai} \cap W_{aj} \neq \phi$ .

Theorem 1 is proved. Therefore, for exploring the behavior of the social actors and the relation between them as the social network mining [15, 16] we have algorithm as follows.

**Algorithm SN<sub>c</sub>:**

```

Declare a query for each social actor,  $q \leftarrow t_a$  for  $a$  in  $A$ ,
 $A$  is a set that consist of  $n$  social actors.
Submit a query to the search engine,  $|\Omega_{ai}| \leftarrow q$ ,  $a_i$  in  $A$ .
Collect the snippets for  $t_{ai}$ ,  $L_{ai} \leftarrow q$ ,  $L_{ai}$  is list of snippets for  $a_i$  in  $A$ .
Give a weight to every word in  $L_{ai}$ ,  $W_{ai} \leftarrow b$ ,  $W_{ai}$  is a set of words for  $a_i$  in  $A$ .
Submit a query to the search engine,  $|\Omega_{aj}| \leftarrow q$ ,  $a_j$  in  $A$ .
Collect the snippets for  $t_{aj}$ ,  $L_{aj} \leftarrow q$ ,  $L_{aj}$  is list of snippets for  $a_j$  in  $A$ .
Give a weight to every word in  $L_{aj}$ ,  $W_{bj} \leftarrow b$ ,  $W_{bj}$  is a set of words for  $a_j$  in  $A$ .
Submit a query to the search engine,  $|\Omega_{ai} \cap \Omega_{aj}| \leftarrow q$ ,  $a_i$  and  $a_j$  in  $A$ .
Collect the snippets for  $t_{ai}$  and  $t_{aj}$ ,  $L_{ai,aj} \leftarrow q$ ,  $L_{ai,aj}$  is list of snippets for
 $a_i$  and  $a_j$  in  $A$ .
Give a weight to every word in  $L_{ai,aj}$ ,  $W_{ai,aj} \leftarrow b$ ,  $W_{ai,aj}$  is a set of words for
 $a_i$  and  $a_j$  in  $A$ .
Compute the strength relations by using  $|\Omega_{ai}|$ ,  $|\Omega_{aj}|$ , and  $|\Omega_{ai} \cap \Omega_{aj}|$ .
Compute the similarity between  $W_{ai}$  and  $W_{bj}$ .
Compute the average for  $W_{ai,aj}$ .

```

**4. Extracting social network: A discussion**

For the Google search engine, whenever a query containing the search term  $t_a$  submitted, we will get hit count  $|\Omega_a|$  and a list of snippets. Each snippet consists of the URL address of the web page, the title of the web page, and the abstract of web page. For example, for two social actors obtained a number of snippets, number of words from snippets, and the hit counts like Table 2.

**Table 2.** The clusters based on search term

Search term		“Abdullah Mohd Zin”		“T Mohd T Sembok”	
Number	of	204		155	
snippets					
Number	of	5,982		4,680	
words					
Hit count		5,850		2,740	
List of words		$w_{ai}$	$p(w_{ai})$	$w_{aj}$	$p(w_{aj})$
		malaysia	0.502	malaysia	0.514
		journal	0.4099	prof	0.4652
		ismail	0.2649	university	0.2264
		nazri	0.2177	ahmad	0.1991
		university	0.1738	abdullah	0.1849
		computer	0.1043	halimah	0.1098
		science	0.0921	journal	0.1071
		prof	0.0733	badioze	0.1005
		network	0.0652	zaman	0.1005
		university	0.0405	ukm	0.0469
		information	0.0276	abu	0.0423

Naturally a social network is built with the social actors. The number of social actors who may be extracted from the Web cannot be predicted, but for resolving this constraint always the extraction of social network based on the community of social actors [4]. In the superficial method, the social network for two social actors formed from two singleton and one doubleton, or formed from three clusters of Web. The clusters have differences and similarities. Each difference causes their behavior

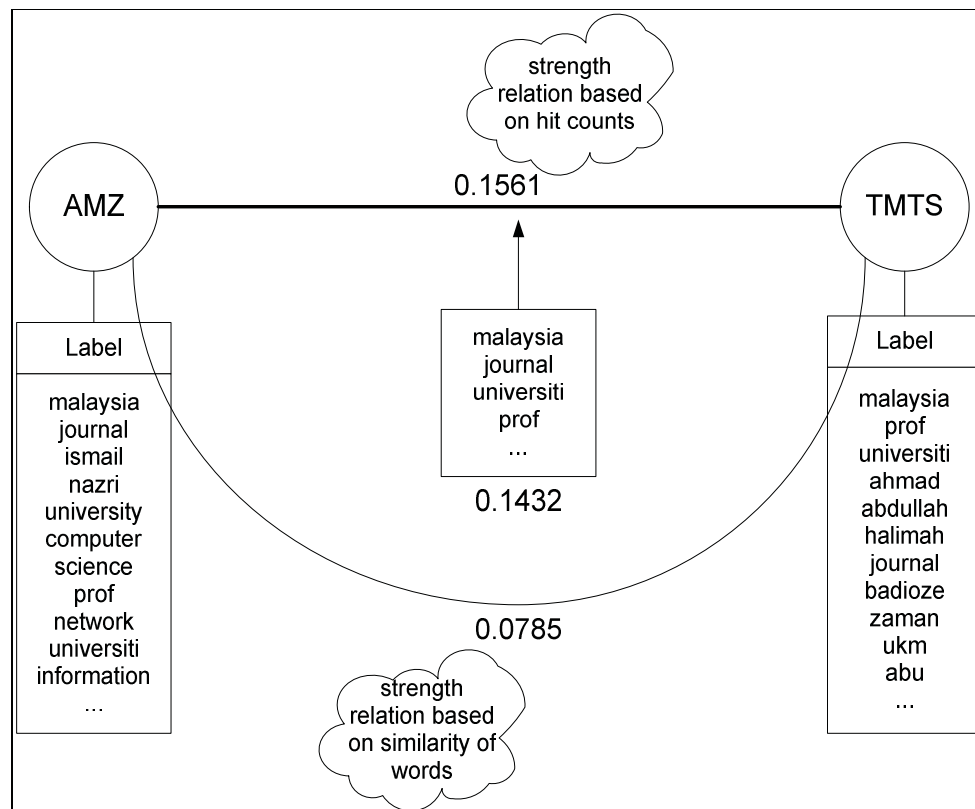
differs, for example, if the word "computer" does not exist in the list of words for "T Mohd T Sembok", then it became an expression of behavior differentiator for clusters based on term search "Abdullah Mohd Zin" [5], and this cluster has the weight as follows

$$|\Omega_{\text{"Abdullah Mohd Zin" and "computer"}}| = |\Omega_a| * p(w_i) = 5,850 * 0.1043 = 610$$

or vice versa

$$|\Omega_{\text{"T Mohd T Sembok" and "zaman"}}| = |\Omega_a| * p(w_i) = 2,740 * 0.1005 = 275$$

Thus, the behavior of any social actor be formed by the expression. It defined by a number of words with its weight that affects the cluster  $|\Omega_a|$  [9].



**Figure 1.** The strength relation between AMZ (“Abdulah Mohd Zin”) and TMTS (“T Mohd T Sembok”) in the extracted social network

The computing toward three clusters using Eq. (1) yield a strength relation between "Abdullah Mohd Zin" and "T Mohd T Sembok", that is 0.1561 where  $|\Omega_{ai} \cap \Omega_{aj}| = 1160$ , whereas the relationship behavior affects the behavior of social actor personally or the social actors have the relationship with others. This relation behavior are defined by an expression through a number of words in the same cluster. It is also indicated by the similarity of expression between social actors although from the different clusters, see Figure 1. By using Eq. (1) for two set of words with their weights we obtain the strength relation between "Abdullah Mohd Zin" and "T Mohd T Sembok" based on similarity of words, i.e. 0.0785, while the average of word weights based on list of snippets (of  $\Omega_{ai} \cap \Omega_{aj}$ ) is 0.1432. Therefore, the social network that is extracted through two singletons and a doubleton, can be mined by involving the snippets to give meaning to the strength relation that have been generated.

## 5. Conclusion

In particular, the behavior of social actors and the relationships between them based on the Web depends on the expression and the behavior of clusters. The superficial method can be developed to express these behaviors and to mine it by following the extraction of social networks. Further study is to reveal the properties associated with the data, information, and analysis of social networks.

## References

- [1] M K M Nasution, S A M Noah and S. Saad 2011 Social network extraction: Superficial method and information retrieval, *Proceeding of International Conference on Informatics for Development* (ICID'11).
- [2] M K M Nasution 2016 Extracted Social Network Mining *Proceeding of International Conference on Information Technology and Engineering Application* (5-th ICIBA), Palembang, February 19-20.
- [3] Y Masunaga, K Ito, Y Miyama, N Oyama, C Watanabe and K Tachi 2010 SERPWatcher: A sophisticated mining tool utilizing search engine results pages (SERPs) for social change discovery *IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust*.
- [4] 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**.
- [5] M K M Nasution 2013 Superficial method for extracting academic sosial network from the Web *Ph.D Dissertation*, Universiti Kebangsaan Malaysia.
- [6] M K M Nasution and S A Noah 2012 Information retrieval model: a social network extraction perspective *IEEE International Conference on Information Retrieval & Knowledge Management*.
- [7] M K M Nasution and S A M Noah 2012 Keyword extraction for identifying social actors arXiv:1212.3023v1 [cs.IR] 13 Dec 2012.
- [8] M K M Nasution 2014 New method for extracting keyword for the social actor *Intelligent Information and Database Systems* LNAI **8397**, Heidelberg, Springer.
- [9] M K M Nasution and S A Noah 2011 Extraction of academic social network from online database *Proceeding of 2011 International Conference on Semantic Technology and Information Retrieval*, Putrajaya, Malaysia.
- [10] M K M Nasution and S A M Noah 2010 Superficial method for extracting social network for academic using web snippets *Rough Set and Knowledge Technology* LNAI **6401**, Heidelberg, Springer.
- [11] M K M Nasution, R Sitepu and M Hardi 2016 Using social networks to assess forensic of negative issues, *IEEE CITSM*, Bandung.
- [12] M K M Nasution 2012 Simple search engine model: Adaptive properties, arXiv:1212.3906v1 [cs.IR] 17 Dec 2012.
- [13] M K M Nasution 2012 Simple search engine model: Adaptive properties for doubleton, arXiv:1212.4702v1 [cs.IR] 19 Dec 2012
- [14] M K M Nasution, M Elveny, R Syah, and S A M Noah, 2015, Behaviour of the resources in the growth of social network *IEEE Proceeding of the 5th International Conference on Electrical Engineering and Informatics*, Bali.
- [15] M K M Nasution, O S Sitompul, E P Sinulingga, and S A N 2016 An extracted social network mining, *IEEE SAI Computing Conference*, London, UK.
- [16] S Vaithyanathan 1999 Introduction: Data mining on the Internet *Artificial Intelligence Review* **13**