The 5th ICIBA 2016, International Conference on
Information Technology and Engineering Application
Palembang-Indonesia, 19-20 February 2016

Extracted Social Network Mining

Mahyuddin K. M. Nasution

Fakultas Ilmu Komputer dan Teknologi Informasi (Fasilkom-TI)
Universitas Sumatera Utara, Padang Bulan, Medan 20155, Sumatera Utara, Indonesia
e-mail: mahyuddin@usu.ac.id, nasutionmahyu2012@gmail.com

Abstract

In this paper we study the relationship between the resources of social networks by exploring the Web as big data based on a simple search engine. We have used set theory by utilizing the occurrence and co-occurrence for defining the singleton or doubleton spaces of event in a search engine model, and then provided them as representation of social actors and their relationship in clusters. Thus, there are behaviors of social actors and their relation based on Web.

Keywords : *Singleton, doubleton, cluster, behavior*

1 INTRODUCTION

An extracted social network is a resultant from the methods of extracting social network from information sources (web pages, documents, or corpus) [1] or the transformation of the raw data into a social network (pre-processing) [2]. However, Web not only dealing with everything changed dynamically [3], but Web as social media represent all members of social (population) [4, 5], or containing big data as big picture of world. Thus, extraction of social networks always based on parts of social (communities) [6], and then to analyze it so that enable to generate useful information, for example, in the decision making [7]. This needs the sample that can represent population. Therefore, for getting significance of information source and trust, in the extracted social networks need the suitable approaches [8].

In other side, the resources of social network such as vertices/actors, edges/relations, and Web/documents have the relations between one to another [9]. A lot of relations between vertices and edges for expressing some of social structures in Social Network Analysis (SNA) [10], but still a little formula to get information about relations among first two resources and Web [11]. Therefore, this needs a formalism study about resources. In all sides of social network, this paper aimed to provide a basic means of discovering knowledge formally about the extracted social network, we call it social network mining.

2 RELATED WORK AND MOTIVATION

The social networks can be modeled naturally by the graph $G < V, E >$ where $V = \{vi|i = 1, ..., n\}$ is a set of vertices, and $\{ej|j = 1, ..., m\}$ is a set of edges and e_j in E if two vertices vk in V and v_l in V are adjacent, or $e_j = v_k v_l = v_l v_k$ [12]. In pre-processing of social network mining, extracting the social network from the information sources is the relatively approaches which is formed through modal relations [1]. One of extraction methods is the superficial method that depends heavily on the occurrence and the co-occurrence [3].

Let a word "Web" or a phrase "World Wide Web" is representation an object according to what we think [13]: the computer network is a social network [14]. In expressing the behavior of social, Natural language processing (NLP) as basic layer of social network mining, and we define the term related it as follows.

Definition 1. A term tx consists of at least one or a set of words in a pattern, or $tx = (w_1, ..., w_l), l \leq k, k$ is a number of words $w(s), l$ is number of vocabularies (tokens) in $t_x, |t_x| = k$ is size of t_x.

In NLP, 'Shahrul Azman Noah' and 'Opim Salim Sitompul' as terms, for example, are well-defined names of social actor. We have defined a dynamic space based on concept of NLP application as follows [15].

Definition 2. Let a set of web pages indexed by search engine by Ω. For each search term t_x, where t_x in $\sum$, i.e. a set of singleton search term of search engine. There are a dynamic space Ω containing the ordered pair of the term $t_{xi} i = 1,, I$ and web pages $\omega_{xj} j = 1, ..., J : (t_{xi}, \omega_{xj}) = (t_x, \omega_x)_{ij}$, or a vector space $\Omega_x = (t_x, \omega_x)_{ij}$ (is subset of or equal to Ω) is a singleton search engine event of web pages (singleton event) that contain an occurrence (event) of t_x in ω.

Definition 3. Suppose t_x in q and q is a query. Clustering web pages based on query is an implication, i.e if $\omega \rightarrow t_x$ is TRUE then a web page ω in Ω is relevant to q or $\Omega_x = 1$ if t_x is true at all ω in Ω, 0 otherwise, and Ω_x as the cluster of t_x.

Classically, a logical implication associated with inference [16].

Lemma 1. If $|\Omega|$ is the cardinality of Ω and $|\Omega_x| \leq |\Omega|$ then probability of a singleton event Ω_x is

$$P(t_x) = |\Omega_x|/|\Omega| \; in \; [0, 1]$$

Proof. For any term t_x in q, each web page ω in Ω is relevant to a query q has a probability to other web pages in $\Omega, 0 \leq p(\omega) = 1/|\Omega| \leq 1$. Probability of all web pages that relevant to a query in Ω is $0 \leq p(\Omega_x) = \sum p(\omega) = |\Omega_x|/|\Omega| \leq 1$, or $P(t_x) = p(\Omega_x)$.

In the same concept we have to define also the co-occurrence based on NLP [17].

Definition 4. Let t_x and t_y are two different search terms, $t_x \neq t_y, t_x, t_y$ in $\sum$, where $\sum$ is a set of singleton term of search engine. There are a dynamic space Ω containing the ordered pair of two terms $\{t_{xi}, t_{yi}\} i = 1, ..., I$ and web pages $\omega_{xj} j = 1, ..., J : (\{t_{xi}, t_{yi}\}, \omega_{xyj}) = (\{t_x, t_y\}, \omega_{xy})_{ij}$, or a vector space $\Omega_x \cap \Omega_y = (\{t_x, t_y\}, \omega_{xy})_{ij}$ (is a subset of or equal to Ω) is a doubleton search engine event of web pages (doubleton event) that contain a co-occurrence (event) of t_x, t_y in ω.

Lemma 2. If $|\Omega|$ is the cardinality of Ω and $|\Omega_x \cap \Omega_y| \leq |\Omega|$ then probability of a doubleton event $\Omega_x \cap \Omega_y$ is

$$P(\{t_x, t_y\}) = |\Omega_x \cap \Omega_y|/|\Omega| \; in \; [0, 1]$$

Proof. As direct consequence of: Definition 4 and Lemma 1.

At the time conducting the extraction for getting occurrences and co-occurrence, we submitted the queries containing the name to Google search engine, we have the hit count = 20,000 for 'Shahrul Azman Noah' (as occurrence) and = 3,000 for 'Opim Salim Sitompul' (as occurrence), while the hit count for 'Shahrul Azman Noah,Opim Salim Sitompul' (as co-occurrence) is 218. However, if the query contains names that are enclosed in quotation marks, produced the hit count = 2,680 for "Shahrul Azman Noah" (as occurrence) and the hit count = 5,650 for "Opim Salim Sitompul" (as occurrence), while the hit count for '"Shahrul Azman Noah","Opim Salim Sitompul"' (as co-occurrence) is 61. Therefore, information about social actors in occurrence and social networks in co-occurrences are different in behavior, and we have an assumption [18, 19]: Each probability of forming its own distribution. Different data distribution gives different behavior. In this case, we have the problem.

Theorem 1. The behavior of clusters describes the behavior of a social actor, then the behavior of other actors expressed by the relationships between the clusters.

3 MODEL AND APPROACH

Literally, we can identify social actor based on Named-Entity Recognition (NER) in web pages or any document as follow.

Definition 5. Suppose there are the well-defined actors, then there is $A = \{\alpha_i | i = 1, ..., n\}$ as a set of social actors.

Each actor literally also has attributes, thus we can define it as follow [20].

Definition 6. Suppose there be the well-identified attributes, then there is $B = \{b_j | j = 1, ..., m\}$ as a set of attributes of actors.

Definition 7. For all pairs (dyads) of n social actors, a set of relationships $R = \{r_p | p = 1, ..., m\}$ where a relationship between two actors there are a tie connect them by one or more relations, or $r_p(\alpha_k, \alpha_l) = B_{\alpha k} \cap B_{al}$.

Definition 8. An extracted social network, i.e. $SN =< V, E, A, R, \gamma_1, \gamma_2 >$ satisfies the conditions as follow:

1. $\gamma_1(1:1)A \rightarrow V,$ *and*
2. $\gamma_2 : R \rightarrow E$

As an approach to formalize the relationship between resources of social networks, and for exploring the behavior, we use the association rule.

Definition 9. Let $B = \{b_1, b_2, ..., b_m\}$ is a set of attributes. Let M_i is a set of transactions are subsets of attributes or M_i are the subset of or equal to B. The implication $\Omega_{bk} \rightarrow \Omega_{bl}$ with two possible value TRUE or FALSE as an *association rule* if Ω_{bk}, Ω_{bl} are subset of B and $\Omega_{bk} \cap \Omega_{bl} = \phi$.

4 FORMULATION OF BEHAVIOR

Each cluster represents an actor based on the extraction of social networks.

Lemma 3. If for a cluster Ω_x of a search term t_x there exist other cluster Ω_y of a search term t_y where $t_x \neq t_y$, then Ω_x is a stand-alone cluster.

Proof. Based on *Definition 2* and *Definition 9*, we have $t_x \rightarrow t_y$ literally or $\Omega_x \rightarrow \Omega_y$, but $t_x \neq t_y$ such that $\Omega_x \cap \Omega_y = \Box$. Therefore, Ω_x is a stand-alone cluster.

Proposition 1. If $t_{\alpha i}$ in $qi = 1, ..., n$ and $\Omega_{\alpha i}$ are a stand-alone cluster for each of $\{\alpha_1, \alpha_2, ..., \alpha_n\} = A$, then $\Omega_{\alpha i}$ represent the behavior of α_i in A, respectively.

Proof. Based on *Definition 9*, we have ω in $\Omega \rightarrow t_\alpha$in q and ω is representation of actor α in A, and because of each ω in Ω has a probability then ω in Ω be the behavior of actor

α in A, but based on *Definition 2* $\Omega\alpha = \{(t_a, \omega_a)_{ij}\}, \Omega_a$ is representation of α in A. Let there be t_{ak}, t_{al} in $qt_{ak} \neq t_{al}$, we have $\Omega_{ak} \to \Omega$ and $\Omega_{al} \to \Omega$: Even though $\Omega_{ak} \to \Omega_{al}$ or $\Omega_{al} \to \Omega_{ak}$, but $\Omega_{ak} cap \Omega_{al} = \Omega_{al} \cap \Omega_{ak} = \phi$. Each of $\Omega_{\alpha i}, i = 1, ..., n$ is a stand-alone cluster that represent the behavior of an actor.

Lemma 4. Let $t_{ak} \neq t_{al}$ is the different search terms represent two social actors. If t_{ak}, t_{al} in q, then Ω_{akl} is a stand-alone cluster for a pair of social actors.

Proof. As applicable in *Lemma 3* to *Definition 3* and *Definition 2*, ω in $\Omega \to \{t_{ak}, t_{al}\}$ in$q$ or ω in $\Omega \to \{t_{ak} \square t_{al}\}$ in q and (ω in $\Omega \to t_{ak}$ *in* $q)\square(\omega$ *in* $\Omega \to t_{al}$ *in* q) and we have $\Omega_{ak} \to \Omega_{al}$ and $t_{ak} \neq t_{al}$, but $\Omega_{ak} \cap \Omega_{al} \neq \phi$ then $\Omega_{al} \to \Omega_{ak}$. However, based on Definition 9 we have $((\Omega_{ak} \to \Omega_{al}) \to \Omega) = ((\Omega_{al} \to \Omega_{ak}) \to \Omega)$. In other word, $\Omega_{akl} = \{(t_{akl}, \omega_{akl})_{ij}\} = \{(t_{ak} \square t_{al}, \omega_{ak} \square \omega_{al})_{ij}\} = \{(t_a \square t_a, \omega_a \square \omega_a)_{ij}\}(i, j = k \square l) = \{(t_a, \omega_a)_{ij}\}(i, j = k \square_l) = \Omega_a$. Thus, akl is a stand-alone cluster of a pair of social actors.

Proposition 2. If Ω_{akl} is a stand-alone cluster for a pair of $\{a_1, a_2, ..., a_n\} = A$, then Ω_{akl} represent the behavior of relationship between $a)i$ in $A, i = 1, ..., n$.

Proof. Based on *Lemma 4* we have $\Omega_{akl} = \Omega_a$, and $\Omega_a = \{(t_a, \omega_a)_{ij}\} = \{(t_a \square t_a, \omega_a \square \omega_a)_{ij}\} = \{(t_{ak} \square t_{al}, \omega_{ak} \square \omega_{al)}\}(i, j = kl) = \{(t_{ak} \square t_{al}, \omega_{ak} \square \omega_{al})\} = \{(t_{ak}, \omega_{ak}) \square (t_{al}, \omega_{al})\} = \{(t_{ak}, \omega_{ak})\} \cap \{(t_{al}, \omega_{al})\} = \Omega_{ak} \cap \Omega_{al}$. Or because name also can be an attribute of social actor, then Based on Definition 7 we have $\Omega_{ak} \cap \Omega_{al} = B_{ak} \cap B_{al} = r_p(a_k, a_l)$.

Definition 8 has set the existence of a social actor by means of γ_1 and behavior of a social actor based on the result clusters (Proposition 1, while the behavior of relationship between social actors refers to the cluster based on dyad (Proposition 2) and this behavior based on γ_2 also become behavior of an edge in social network. Specially, in superficial methods r_p in R means the strength relation between two actors ak and al in A by involving one or more of the similarity measurements: mutual information, Dice coefficient, overlap coefficient, cosine, or for example Jaccard coefficient

$$J_c = |\Omega_{ak} \cap \Omega_{al}| / |\Omega_{ak}| + |\Omega_{al}| - \Omega_{ak} \cap \Omega_{al} \ in\ [0, 1]$$

In this concept of similarity, $B)ak \cap B_{al} = |\Omega_{ak} \cap \Omega_{al}| / (|\Omega_{ak}| + |\Omega_{al}| - |\Omega_{ak} \cap \Omega_{al}|) = J_c$ such that e_j in E if $r_p > 0$. However the behavior of $r_p (0 \leq r_p \leq 1)$ depends on the behavior of Ω_{ak} is a subset of Ω, Ω_{al} is subset of Ω and $\Omega_{ak} \cap \Omega_{al}$ is a subset of Ω : $|\Omega_{ak}| \leq |\Omega_{al}|$ *or* $|\Omega_{ak}| \geq |\Omega_{al}|, |\Omega_{ak} \cap \Omega_{al}| \leq |\Omega_{ak}|$, and $|\Omega_{ak} \cap \Omega_{al}| \leq |\Omega_{al}|$. If another measurement concept is similar to J_c, then Theorem 1 is proved. Therefore, we have

Corollary 1. If the behavior of social actors behaves in clusters (of big data) then the behavior of the clusters (of big data) can be represented by the extracted social network.

5 CONCLUSIONS

In this social network study we have presented an analysis for formulating the behavior of resources of social network as a social network mining. Formulation based on a search engine model and the clustering model, and we have obtained an explanation that there are relations between social actors/vertices, relationships/edges, and documents/web based on the clusters are formed. The future work will involve the extraction of a social network to describe the research collaboration for exploring the behavior of social actors and their relationships.

References

[1] M. K. Nasution and S. A. Noah, "Superficial method for extracting social network for academics using web snippets," in *Rough Set and Knowledge Technology, ed: Springer*, 2010, pp. 483-490.

[2] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 337-348.

[16] M. K. Nasution and S. A. Noah, *"A Methodology to Extract Social Network from the Web Snippet,"* arXiv preprint arXiv:1211.5877, 2012.

[4] L. Bent, et al., "Characterization of a large web site population with implications for content delivery," *World Wide Web*, vol. *9*, pp. 505-536, 2006.

[5] L. A. Abdillah, "Indonesian's presidential social media campaigns," in *Seminar Nasional Sistem Informasi Indonesia (SESINDO2014), ITS*, Surabaya, 2014.

[6] M. K. Nasution and S. A. Noah, "Extraction of academic social network from online database," in *Semantic Technology and Information Retrieval (STAIR)*, 2011 International Conference on, 2011, pp. 64-69.

[7] S. Azman, "Efficient identity matching using static pruning q-gram indexing approach," *Decision Support Systems*, vol. *73*, pp. 97-108, 2015.

[8] Y. Yao, et al., "Subgraph extraction for trust inference in social networks," in *Encyclopedia of Social Network Analysis and Mining, ed: Springer*, 2014, pp. 2084-2098.

[9] K. Mahyuddin, et al., "Behavior of the resources in the growth of social network," in *Electrical Engineering and Informatics (ICEEI), 2015 International Conference* on, 2015, pp. 496-499.

[10] J. P. Scott, *Social Network Analysis: A Handbook*, 2nd ed. London: Sage Publications, 2000.

[11] N. Memon, et al., "Social network data mining: Research questions, techniques, and applications," in *Data Mining for Social Network Data, ed: Springer*, 2010, pp. 1-7.

[16] M. K. Nasution and S. A. Noah, *"Probabilistic Generative Model of Social Network Based on Web Features,"* arXiv preprint arXiv:1207.3894, 2012.

[13] M. K. Nasution, "Kolmogorov Complexity : Clustering Objects and Similarity," *Bulletin of Mathematics*, vol. *3*, pp. 1-16, 2011.

[14] B. Wellman, et al., "Computer networks as social networks: Collaborative work, telework, and virtual community," *Annual review of sociology*, pp. 213-238, 1996.

[15] M. K. Nasution, *"Simple search engine model: Adaptive properties,"* arXiv preprint arXiv:1212.3906, 2012.

[16] M. K. Nasution and S. A. Noah, *"Information retrieval model: A social network extraction perspective,"* arXiv preprint arXiv:1207.3583, 2012.

[17] M. K. M. Nasution, *"Simple search engine model: Adaptive properties for doubleton,"* arXiv:1212.4702v1, 2012.

[18] L.-C. Chen and A. Sakai, "Critical behavior and the limit distribution for long-range oriented percolation. I," *Probability Theory and Related Fields*, vol. *142*, pp. 151-188, 2008.

[19] X. Wang, et al., "Measurements on movie distribution behavior in Peer-to-Peer networks," in Integrated Network Management (IM), *2011 IFIP/IEEE International Symposium on*, 2011, pp. 618-621.

[20] M. K. Nasution, "Extracting Keyword for Disambiguating Name Based on the Overlap Principle," in *The 4th International Conference on Information Technology and Engineering Application (ICIBA2015)*, Palembang, 2015.

[]