Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 124 (2017) 86–92

Procedia Computer Science

www.elsevier.com/locate/procedia

4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

Social Network Extraction Based on Web. A Comparison of Superficial Methods

Mahyuddin K. M. Nasution[a,*], Shahrul Azman Noah[b]

[a] *Information Technology Department, Universitas Sumatera Utara, Medan 20155, Sumatera Utara, Indonesia*

[b] *Knowledge Technology Research Group, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi 43600 UKM Selangor, Malaysia*

Abstract

The Web as a source of information has many potentials which allow to use the different treatments on social network extraction methods. The approaches generally we identified as superficial methods in unsupervised stream. However, the same resources of social networks, i.e. based on a community of social actors, reveal many of different approaches to produce social networks. Therefore, based on a treatment to another treatments, from the given treatments until the different social networks generated and it has been declared different methods. It requires comparison to reveal the properties of social networks and their methods in this paper. It is revealed that there is a core social network has similarity with other social networks is more than 1% as general property of the extracted social networks, whereby there is a social network for different methods has the common edges in graph.

Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017.

Keywords: Superficial method, occurrence, co-occurrence, hit count, snippet, URL address, Web, vertices, social actors, edges, strength relations

1. Introduction

The Web as an information source has a lot of potential to be extracted into consideration for decision making [1]. Extracting social network from Web not only considers the available potential only but systematically is also the way to gain structural social behavior [2]. Therefore, the methods applied for extraction also vary, and each involves the different potentialities of information source, although it involves the same social actors. However, since the method externally is outside the search engine system [3], although the methods in unsupervised stream heavily depends on the search engine [4], consequently the method can't fully utilize the information resources optimally [5], moreover the methods in supervised stream [6] such as the use of: Latent Dirichlet Allocation (LDA) [7] or Hidden Markov Model (HMM) [8], for example.

* Corresponding author. Tel.: +62-061-821-1633 ; fax: +62-061-821-9411.
E-mail address: mahyuddin@usu.ac.id

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017
10.1016/j.procs.2017.12.133

The different information involvement of the Web in disclosing social networks has resulted in the different approaches for extracting social network from Web [9, 10]. However, in the same potential, it has not been revealed that there is a difference between these approaches. Therefore, this paper will reveal the performance of each approach in the superficial methods involving the experiment of social network extraction.

2. A Review

In the formal definition of social networks, expressed in graph theory $G(E, V)$, it has been disclosed that there is a set of vertices V and a set of edges E, with which $v_i \in V\ i = 1, \ldots, n$ denotes entities in social networks and $v_i v_j \in E$ denotes relationship between entities $v_i, v_j \in V$ in a social network [11, 12]. This definition reveals that in the extraction of social networks using the basic superficial method (BSM) occurs the process (a) determines the social actors and (b) builds relationships between them [13, 14]. Next we consider some basic characters related to the superficial methods.

In the first process, the use of the social name a_i (without quotes) in the query q is to represent a social actor or

$$\mathbf{a}_i = q \leftarrow a_i, \tag{1}$$

then generally search engines generate ambiguous information about the actor social [15]. However, with the addition of the keyword kw, in general it can reduce naturally the default property of used social name [13, 16], i.e. consequent of

$$\mathbf{aw}_i = q \leftarrow a_i, kw \tag{2}$$

happen reduction ambiguity, with which $|aw_i| \leq |a_i|$, $|a_i| \in \mathbf{a}_i$ is a cardinality of a_i and $|aw_i| \in \mathbf{aw}_i$ is a cardinality of a_i, kw [17]. While using the well-defined name of social actor (in quotes) in the query q will raises the entire social actors related information or

$$\mathbf{a}_{"i"} = q \leftarrow "a_i". \tag{3}$$

In last case, $|a_{"i"}| \leq |a_i|$, and $|a_{"i"}| \in \mathbf{a}_{"i"}$ is a cardinalty of "a_i" [18]. As well as with

$$\mathbf{aw}_{"i"} = q \leftarrow "a_i", "kw". \tag{4}$$

is about one of information concentrations of a social actor, $|aw_{"i"}| \leq |a_{"i"}|$, and $|aw_{"i"}| \in \mathbf{aw}_{"i"}$ is a cardinalty of "a_i", "kw" [19, 20, 17].

In the second process, the relationship between two social actors is based on the concept of co-occurrence [21, 22, 23]. Thus,

$$\mathbf{a}_i\mathbf{a}_j = q \leftarrow a_i, a_j, \tag{5}$$

is a process to elevate the clue of relation be relationship between two actors, with which $|a_i \cap a_j| \leq |a_i|$ and $|a_i \cap a_j| \leq |a_j|$, and $|a_i \cap a_j| \in \mathbf{a}_i\mathbf{a}_j$ is a cardinality of a_i, a_j. The addition of a keyword towards the co-occurrence will usually reduce the number of information presented, that is

$$\mathbf{aw}_i\mathbf{aw}_j = q \leftarrow a_i, a_i, kw, \tag{6}$$

but it should meet that $|aw_i \cap aw_j| \leq |a_i \cap a_j|$, $|aw_i \cap aw_j| \in \mathbf{aw}_i\mathbf{aw}_j$ is a cardinality of a_i, a_i, kw [10, 23]. Likewise, the use of the well-defined name of social actor in the query will reveal the relationoship between two social actors appropriately, that is

$$\mathbf{a}_{"i"}\mathbf{a}_{"j"} = q \leftarrow "a_i", "a_j". \tag{7}$$

whereby $|a_{"i"} \cap a_{"j"}| \leq |a_{"i"}|$ and $|a_{"i"} \cap a_{"j"}| \leq |a_{"j"}|$, $|a_{"i"} \cap a_{"j"}| \in \mathbf{a}_{"i"}\mathbf{a}_{"j"}$ is a cardinality of "a_i", "a_j". Thus, if $|a_{"i"} \cap a_{"j"}| = 0$, then it means exactly there is no relationship between two social actors [14]. Whereas the addition of a keyword towards the co-occurrence based on pattern as follows

$$\mathbf{wa}_{"i"}\mathbf{wa}_{"j"} = q \leftarrow "a_i", "a_j", "kw", \tag{8}$$