Enhancing to method for extracting Social network by the relation existence

Maria Elfida, M K Matyuso Nasution*, O S Sitompul

Technical Information, Fasilkom-TI, Universitas Sumatera Utara, Padang Bulan 20155 USU Medan Indonesia

E-mail: mahyuddin@usu.ac.id

Abstract. To get the trusty information about the social network extracted from the Web requires a reliable method, but for optimal resultant required the method that can overcome the complexity of information resources. This paper intends to reveal ways to overcome the constraints of social network extraction leading to high complexity by identifying relationships among social actors. By changing the treatment of the procedure used, we obtain the complexity is smaller than the previous procedure. This has also been demonstrated in an experiment by using the denial sample.

1. Introduction

Future of social network is not depend only on its roles in decision making [1], but information source as a resource for generating social network, whereby the resources of social network today not only consist of: vertices, edges, social actors, traditional relations like family, office, society, nation, and cultural, but also included occurrence, co-occurrence, personal names, text, graphic, image, semantic, corpus, media, and Web [2]. Therefore, the social networks extracted from the Web to be closely related to the treatment of the last mentioned collection of resources for generating trusty information [3]. In this case, the trusty information not just related to extracting social network completely from Web, but identifying correctly the components of a social network builder like the clues of relations, and the labels of vertex and edges [4, 5].

Extracting the social network is a relative approach which is formed through modal relations that depends heavily on the co-occurrence for representing relations between individuals, groups or organization [6]. Some of methods, named with superficial methods, have led to many advances in other fields and natural language processing (NLP) fields involving artificial intelligence (AI) [7]. However, the development of information extraction methods, especially with regard to social networks, is not comparable to the growth of webpages and the dynamics of their change in content. While the method of extraction is always collided with limited access to the overall information available in the source information [8]. Therefore, this paper will reveal a way of reducing the complexity of a method by identifying information related to the relationship between social actors, in addition to overcoming the limitations of access in social network extraction from the Web.

2. Related Works and Motivation

In general, to extract social network from Web formally is as follows [9, 11, 12, 10]

A1 $\gamma_1 : A \overset{1:1}{\rightarrow} V$, whereby $A = \{a_i|i = 1, \dots, I\}$ is a set of social actors, and $V = \{v_i|i = 1, \dots, I\}$ is a set of vertices in graph G.

A2 $\gamma_2 : A \times A \rightarrow E$, $E = \{e_j|j = 1, \dots, J\}$ is a set of edges in graph G.

A1 means that, there is information about social actors coming from the Web. This is interpreted as submitting a query q to search engine whereby the query contains the name of a social actor, or $q_i \leftarrow a_i$ (the occurrences) [13].

A2 indirectly explains that one of clues about relation between social actors is derived from the application of co-occurrence, i.e. submitting a query q to search engine whereby the query contains a pair of social actor names, or $q_j \leftarrow a_i, a_k$ [14]. However, Cartesian product on A2 states with certainty that for n social actors, in the procedure of social network extraction there is the potential of $n(n-1)$ iterations computationally. In other words, there n times to submit query (occurrence) to search engines and $n(n-1)$ times to submit query (co-occurrence) to search engines, or $n + n(n-1) = n^2$ computationally. This is symmetrically expressed as $n + n(n-1)/2 = n/2 + n^2/2$ times for submitting query to search engine [15].

On the one hand, each search engine has the limitations of receiving query submissions, such as 1000 submissions/day/computer [10]. Meanwhile, as social media the Web constantly changing and so dynamic [16]. Thus, it may be that the old social actors still exist, but the new social actors must have arisen from everyday social activities in society [2]. On the other hand, information about social actors and their relationships must be trusty [12]. For that purpose, in step A1 the method involve keywords to uncover ambiguity or use well defined name in the form of patterns to reduce the bias, nevertheless this is not to reduce the complexity of procedures and the information space like Web [8].

In principle, the method of social network extraction from the Web depend on strength relations sr through the concept of similarity. Generally, the similarity uses three vectors based on two social actors [18], i.e.

$$sr = sim(\Omega_{a_i}, \Omega_{a_k}, \Omega_{a_i} \cap \Omega_{a_k}) \tag{1}$$

whereby $\Omega_{a_i} = q_i \leftarrow a_i$, $\Omega_{a_k} = q_k \leftarrow a_k$, $\Omega_{a_i} \cap \Omega_{a_k} = q_{i,k} \leftarrow a_i, a_k$.
Meanwhile, the last query is a form that involves symmetry of mutual keywords between a social actor and another social actor, and when a name pattern is applied, it will simultaneously be done the disambiguation and the bias reduction [14, 16, 17].

3. An Approach

In many ways, the similarity span plays a role, especially when it involves measurements related to co-occurrence [16, 18] Whether directly or indirectly, the extraction of social networks actually rests on the role of co-occurrence, which generally involves the value of $\Omega_{a_i} \cap \Omega_{a_k}$, i.e. the hit count $|\Omega_{a_i} \cap \Omega_{a_k}|$, which meets the following criteria

$$|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}| \tag{2}$$

and

$$|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_k}| \tag{3}$$

whereby $|\Omega_{a_i}|$ and $|\Omega_{a_k}|$ are the hit counts based on occurrences for $a_i, a_k \in A$. It means that the existence of the relationship between two actors depends on the value of $|\Omega_{a_i} \cap \Omega_{a_k}|$ for all a_i and a_k in A. Therefore, based Eq. 1 we obtain a rule as follows.

Rule 1. $\forall v_i, v_k \in V, \exists e_j \in E$ if and only if $sr > 0$ between a_i and a_k, $a_i, a_k \in A$.

Table 1. Scale of Iterations

Modification	oc	cs	sr
Before	n	$n(n-1)/2$	$n(n-1)/2$
After	$\leq n$	$n(n-1)/2$	$\leq n(n-1)/2$

Computationally, for example sr depends on formulation of hit counts as follows [18].

$$sr = \frac{2|\Omega_{a_i} \cap \Omega_{a_k}|}{|\Omega_{a_i}| + |\Omega_{a_k}|}. \tag{4}$$

So, let $\mathcal{A}$ as a space for all occurrence for A we have a procedure as follows.

```
Social_Network($A$)
  for $i \leftarrow 1$ to $n$
       $\mathcal{A} = [q_i \leftarrow a_i]$
  for $i \leftarrow 1$ to $n$
      for $k \leftarrow i+1$ to $n-1$
           function.sr($a_i, a_k, |\Omega_{a_i}|, |\Omega_{a_k}|$)

function.sr($a_i, a_k, |\Omega_{a_i}|, |\Omega_{a_k}|$)
  $|\Omega_{a_i} \cap \Omega_{a_k}| \approx q_{i,k} \leftarrow a_i, a_k$
  if $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}|$ AND $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_k}|$:
      return $sr(|\Omega_{a_i}|, |\Omega_{a_k}|, |\Omega_{a_i} \cap \Omega_{a_k}|)$
  else:
      return 0.
```

Or, briefly the procedure contain steps as follows

(i) Get $|\Omega_{a_i}|$, $i = 1, \ldots, I$.

(ii) Get $|\Omega_{a_i} \cap \Omega_{a_k}|$ in pairs for all social actors.

(iii) If $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}|$ AND $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_k}|$, then compute sr.

Thus, we get a rule for identifying the relationship between two social actors.
Rule 2. $sr > 0$ if and only if $|\Omega_{a_i} \cap \Omega_{a_k}| > 0$.

Under Rule 2, it can be stated that if one social actor has not the hit count about co-occurrence > 0 with all the other social actors, then occurrence should not be disclosed.

Therefore, two first steps in procedure of extracting social network from Web by using search engine change be as follows:

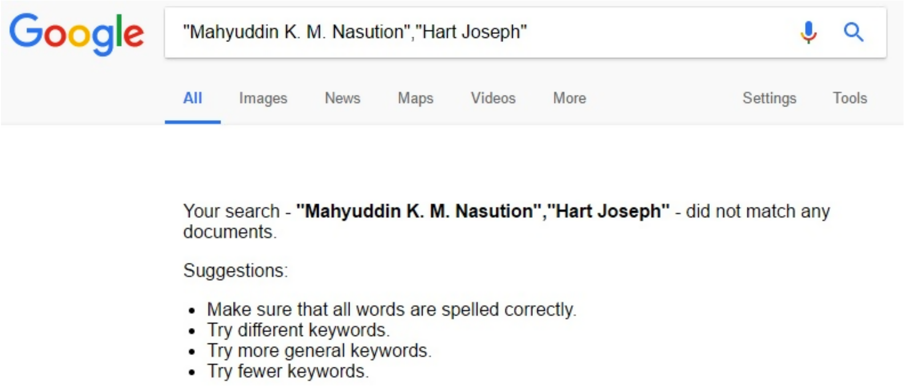
(i) Get $|\Omega_{a_i} \cap \Omega_{a_k}|$ in pairs for all social actors.

(ii) If $|\Omega_{a_i} \cap \Omega_{a_k}| > 0$, then get $|\Omega_{a_i}|$ and $|\Omega_{a_k}|$.

(iii) If $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}|$ AND $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_k}|$, then campute sr.

For which the number of submitting query to search engine as the occurrence may be less than or equal to n social actors. Thus, the complexity of social network extraction $< n/2 + n^2/2$. Computationally, the number of iterations for sr less than or equal to $n(n-1)/2$ with considering that $|\Omega_{a_i} \cap \Omega_{a_k}| = 0$ is not count in computation.

Let us redefine the number of queries for occurrence for n social actors as the occurrence scale (os) and the number of queries for co-occurrence as co-occurrence scale (cs), we obtain the comparison like Table 1 (Scale of Iterations).



Your search - **"Mahyuddin K. M. Nasution","Hart Joseph"** - did not match any documents.

Suggestions:

- Make sure that all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Figure 1. One of 7 names in denial sample: "Hart Joseph"

4. Experiment

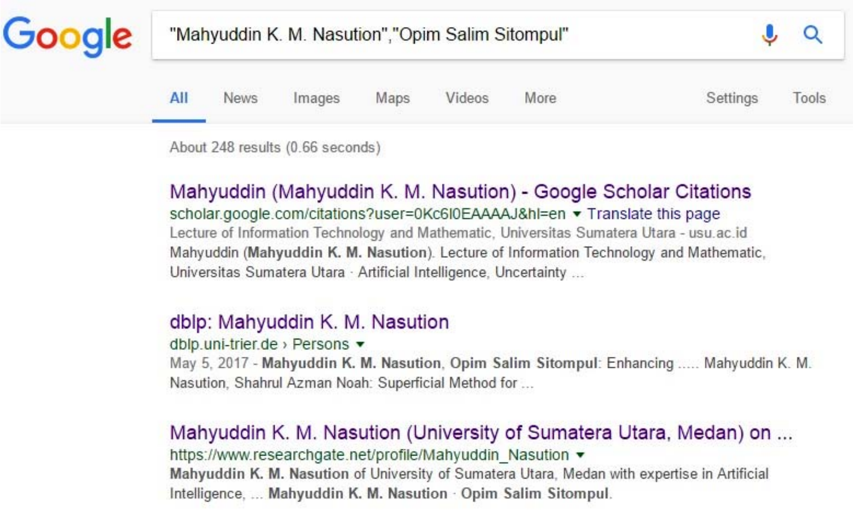
Let us implement the last procedure as follows

```
Social_Network($A$)
  identify($A$)
  if $\mathcal{I} > 0$:
      for $i \leftarrow 1$ to $n$
          $\mathcal{A} = [q_i \leftarrow a_i]$
  for $i \leftarrow 1$ to $n_{\mathcal{I}}$
      for $k \leftarrow i+1$ to $n_{\mathcal{I}} - 1$
          function.sr($\mathcal{A}, \mathcal{A}_{cs}$)

identify($A$)
  $\mathcal{I} = [0]$
  for $i \leftarrow 1$ to $n$
      for $k \leftarrow i+1$ to $n-1$
          $|\Omega_{a_i} \cap \Omega_{a_k}| \approx q_{i,k} \leftarrow a_i, a_k$
          if $|\Omega_{a_i} \cap \Omega_{a_k}| > 0$:
              $\mathcal{I} = [1]$
              $\mathcal{A}_{cs} = [|\Omega_{a_i} \cap \Omega_{a_k}|]$

function.sr($\mathcal{A}, \mathcal{A}_{cs}$)
  if $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}|$ AND $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_k}|$ AND
    $|\Omega_{a_i} \cap \Omega_{a_k}| \leq |\Omega_{a_i}| > 0$:
      return $sr(|\Omega_{a_i}|, |\Omega_{a_k}|, |\Omega_{a_i} \cap \Omega_{a_k}|)$
```

where $\mathcal{I}$ contains binary number $\{0,1\}$, and the relations between two social actors we call as the binary relationship.

In an experiment involving 469 social actors, there were 7 (seven) names of social actors who at the time of the experiment were not on the Web. 7 names are used as a denial and endorsement for the identification part of procedure, to identify the relationship between social actors. For example, by using Google search engine for query q ="Mahyuddin K. M. Nasution","Hart Joseph", we have

$$|\Omega_{\text{"Mahyuddin K. M. Nasution"}} \cap \Omega_{\text{"Hard Joseph"}}| = 0,$$



"Mahyuddin K. M. Nasution","Opim Salim Sitompul"

About 248 results (0.66 seconds)

Mahyuddin (Mahyuddin K. M. Nasution) - Google Scholar Citations
scholar.google.com/citations?user=0Kc6I0EAAAAJ&hl=en ▾ Translate this page
Lecture of Information Technology and Mathematic, Universitas Sumatera Utara - usu.ac.id
Mahyuddin (**Mahyuddin K. M. Nasution**). Lecture of Information Technology and Mathematic, Universitas Sumatera Utara · Artificial Intelligence, Uncertainty ...

dblp: Mahyuddin K. M. Nasution
dblp.uni-trier.de › Persons ▾
May 5, 2017 - **Mahyuddin K. M. Nasution**, **Opim Salim Sitompul**: Enhancing ..... Mahyuddin K. M. Nasution, Shahrul Azman Noah: Superficial Method for ...

Mahyuddin K. M. Nasution (University of Sumatera Utara, Medan) on ...
https://www.researchgate.net/profile/Mahyuddin_Nasution ▾
Mahyuddin K. M. Nasution of University of Sumatera Utara, Medan with expertise in Artificial Intelligence, ... **Mahyuddin K. M. Nasution** · **Opim Salim Sitompul**.

Figure 2. Relation between "Mahyuddin K. M. Nasution" and "Opim Salim Sitompul"

see Fig. 1, while for query $q =$"Mahyuddin K. M. Nasution","Opim Salim Sitompul", we have

$$|\Omega_{\text{"Mahyuddin K. M. Nasution"}} \cap \Omega_{\text{"Opim Salim Sitompul"}}| = 248,$$

see Fig. 2. Tautologically, the method of identification by involving a denial sample behaves correctly if and only the method behaves in a general manner in the population. In other words, if the existence of 7 social actors is not on the Web, then any relationship between one of 7 social actors and one of 462 other social actors does not exist. Thus, there are as many as 3,276 binary relationships cannot be expressed through the information source. Meanwhile, involving 462 names of social actors in the pattern form we obtained 22,836 binary relationships or 21.44% of 106,491 potential relationships. Therefore, the complexity computationally requires 22,836 iterations and 23,988 times generating information for producing the strength relations.

5. Conclusion

In this study have disclosed the key that determines the complexity of social network extraction methods. By changing the treatment of the procedure based on identifying the existence of the relationship between the social actors, the number of queries submissions into the search engine can be reduced, as well as computationally, the iterations number is less than $n/2 + n^2/2$.

References

[1] 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)**.

[2] M K M Nasution, M Elveny, R Syah, and S A M Noah 2015 Behavior of the resources in the growth of social network *Proceedings of the 5th International Conference on Electrical Engineering and Informatics* (ICEEI).

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

[4] M K M Nasution and Shaurul Azman Noah 2011 Extraction of academic social network from online database *Proceeding of 2011 International Conference on Semantic Technology and Information Retrieval* (STAIRS11)).

[5] 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(1)**.

[6] 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**.

[7] 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).

[8] M K M Nasution 2017 Modelling and Simulation of Search Engine *Journal of Physics: Conference Series* **801(1)**.

[9] M K M Nasution and Shahrul Azman Noah 2012 Information retrieval model: A social network extraction perspective *International Conference on Information Retrieval & Knowledge Management* (CAMP12).

[10] Y Matsuo, J Mori, M Hamasaki, T Nishimura, H Takeda, K Hasida, M Ishizuka 2007 POLYPHONET: An advanced social network extraction system from the Web *Journal of Web Semantics* **5**.

[11] P Mika 2005 Flink: Semantic web technology for the extraction and analysis of social networks *Journal of Web Semantics* **3**.

[12] P Mika 2007 *Social Networks and the Semantic Web* Heidelberg, Berlin: Springer.

[13] M K M Nasution 2012 Simple Search Engine Model: Adaptive Properties *Cornell University Library* arXiv:1212.3906 [cs.IR].

[14] M K M Nasution 2012 Simple Search Engine Model: Adaptive Properties for Doubleton *Cornell University Library* arXiv:1212.4702 [cs.IR].

[15] M K M Nasution 2013 Superficial method for extracting academic social network from the Web *Ph.D. Dissertation* Universiti Kebangsaan Malaysia.

[16] M K M Nasution 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)* **LNAI 8397 (PART 1)**.

[17] M K M Nasution 2015 Extracting keyword for disambiguating name based on the overlap principle *Proceeding of International Conference on Information Technology and Engineering Application* (4-th ICIBA) **Book 1**.

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