Information Retrieval Optimization Based on Tree of Social Network

Information Retrieval Optimization Based on Tree of Social Network

M K M Nasution[1,*]

[1]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Padang Bulan 20155 USU, Medan, Indonesia

E-mail: *mahyuddin@usu.ac.id

Abstract. Information source such as the Web is a representation of social activities. Social activities create social structures that can be explored through social network concepts. Social networks not only show the structure of a social community, but prove the existence of its members and community. In that context, within the source of information, evidence of a community can be accessed through a collection of documents whose existence continues to increase, whereby the measurements can be performed using recall and precision. However, the recall and the precision as part of the measurement of information retrieval also requires technology to retrieve the related documents, an information retrieval based on network concepts. To solve it is proposed the concept of a collection of stars from trees and the degree of social actors in social networks, and produced a formula about the measurement of recall and precision with better results.

1. Introduction

The extractions about a social network from the Web [1] is directly related to documents or Web assignments [2, 3]. Social network extraction not only presents the social structure of a community, but also as a technology related to decision making [4]. Social networks as a model of the relationship between social actors [5]: personal, organization, or corporate, by which every social actor in proving his activities bring forth related documents and to measure the performance of social actors can be based on social networks extracted from the document [6].

Apart from social networks, but with regard to document management, there is a field of knowledge called information retrieval (IR), a scientific field that methodically tests the requirement of information relevance from the information sources [7, 8]. Through methods about the extraction from the Web, that is to generate social network, there is a reciprocal relationship between extracted social networks and information retrieval [9]. On the one hand, the extraction method of social networks requires IR as the validation of the results of the method [10]. On the other hand, IR development is not only related to the concept of IR itself, but also related to social networks [11]. This paper aims to develop IR that is optimally applicable on the basis of the extracted social networks.

2. Problem Definition

To a social community or a set of social actors $A = \{a_i | i = 1, \ldots, n\}$, by involving information sources containing documents or webpages will generated information about the existence of social actors and the similarities between the social actors [12, 13]. This similarity becomes the modalities to gain relationships between one social actor and another [14]. All relationships can be semantically accumulated in the co-occurrence whereas the presence of social actors is evidenced by the occurrence [1].

Occurrence and co-occurrence are mined by involving any search engine from a set of webpages by means of the assignment resulting in each singleton value $|\Omega_a|$ [15] and doubleton $|\Omega_a \cap \Omega_b|$ [16] for each $a, b \in A$ with which $\Omega = \{\omega_x\}$ is a collection of webpages or information space [17, 18]. Extraction of social network presents the strength relation between social actors by using similarity measures against two singleton and one doubleton [19, 20, 21]. Thus a collection of interconnected social actors by the method of extraction also causes the document or webpages to be in a cluster of interconnected documents.

On the one side, the basic principle of IR is comparing two sets of documents. A document that is known to be sure its existence as the standard set of documents, or D_s. A collection of documents generated from the information space as the related information that can expressed, or D_w. On the other side, all information extraction methods lift the related documents from the information space [22]. Therefore, this principle can be represented formally by the recall measurement Rec [7],

$$Rec = \frac{|D_s \cap D_w|}{|D_s|}, \tag{1}$$

and the precision measurement $Prec$ [7],

$$Rec = \frac{|D_s \cap D_w|}{|D_w|}. \tag{2}$$

In social networks, any social actor has a degree $d(a)$ or any other number of actors adjacent to a [23]. In other words, it says that $d(a_i) \geq 0$, $i = 1, \ldots, m$. Thus, Eq. (1) and Eq. (2) become [11]

$$Rec_a = \frac{|D_s \cap \cup_{i=1}^{m} D_w|}{|D_s|} \tag{3}$$

and

$$Prec_a = \frac{|D_s \cap \cup_{i=1}^{m} D_w|}{|\cup_{i=1}^{m} D_w|} \tag{4}$$

Structurally the extracted social networks constitute communities organized in such a way a collection of social actor degrees or $\{d(a_i)\}$. In other words, there are the intersection of degrees applied into Eq. (3) and Eq. (4), which can be mathematically reduced. Thus, if a set of social actors has an social network extracted from any collection of documents, then there is a set of documents that generated from the information space optimally.

3. The proposed approach

In principle mathematically in graph theory, the optimal form of the network is the tree [24]. Trees can be derived from the network, but taking into account that there is no single vertex separate from the other vertices, or at least one edge connecting one vertex with another. Based on that, the vertices with the same position are considered to be unconnected, provided that two vertices are connected to one other vertex [25].

In social networks involving the strength relation, we can reduce the weakest relations one by one to form trees, but not to cause one vertex to be alienated from the other. In other words, trees will cause all vertices to be connected to one other vertex [26]. However, not all of the weakest relationship are caused by a small co-occurrence, but the weakest strength relations is due to the high occurrence. Thus, elimination of the more weak strength relation is based on the

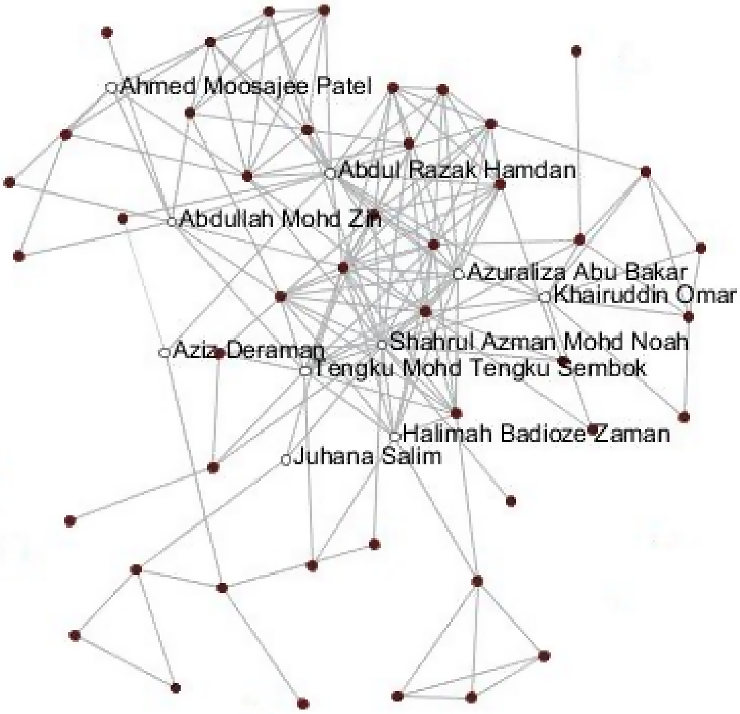

Figure 1. An extracted social network

smallest co-occurrence rankings so as not to reduce relations based on the number of documents that are more strength but have smaller strength relations than others. Therefore, there will be a sequence of the co-occurrence value: $|\Omega_a \cap \Omega_b| \geq |\Omega_a \cap \Omega_c| \cdots |\Omega_b \cap \Omega_c| \geq \cdots \geq |\Omega_c \cap \Omega_x| \geq \cdots$ [27].

4. Discussion about theory and experiments

In principle, the tree is the optimal form of network based on the graph theory. Thus, social networks extracted from information sources will provide a structure in the optimal form of trees, making it possible to minimize the process of accessing information about the community from information sources [28]. For example, there is a social network extracted from the information source based on the 10 seeds of social actors that generate 45 other social actors and a social network such as Figure 1, or $n = 55$ social actors generate $n = 55$ degrees. Each degree possible between 0 to $n-1$, or $0 \leq d(a_i) \leq n-1$, $i = 1,\ldots,n$.

Let for n social actors, it is potential that the social network becomes a tree in which one vertex (the center) has a degree equal to $n-1$. Thus, Eq. (3) and Eq. (4) are formulations that apply to $m = n-1$. A tree is called star if for n social actors may be one vertex of $n-1$ or $d(a_i) = n-1$ while the other vertices are 1 or $d(a_j) = 1$, $i \neq j$, $j = 1,\ldots,n-2$. In other words, if there is a_j having $d(a_j) \geq 1$, the vertex degree is reduced to 1 or $d(a_j) = 1$. However, if the social network has the optimal shape is a tree that is not star, there will be some vertices become the central candidates of some other vertices collection.

In the case, the extracted social network depicts a community, such as Figure 1, and the tree of the social network also becomes the optimal form of the community structure concerned. To obtain IR measurement formulations based on the Eq. (1) and Eq. (2), one of the largest degrees of vertices in the social network is chosen to be the center of one star ($d(a_i) = n_i - 1$, and then form a leaf from star ($d(a_j) = 1$) by eliminating all edges connected to the third actor. Then reselect from the remaining social network, to get the center of the second star and so on, whenever a star is formed eliminating the relationship with the non-leaf forward vertices, so resulting in a set of k star. Suppose there are 5 social actors forming each star with the number of documents 103, 105, 160, 210 and 70. Each document has its own identity based on the address URL of the publisher, which is collected in a set as the group of documents for the community. Based on the star of the social network, documents related to social actors can be achieved as follows: For author A1 = `Abdullah Mohd Zin`, generated 96.12% and 51.03%; A2 = `Azuraliza Abu Bakar`, generated 96.19% and 33.72%; A3 = `Shahrul Azman Mohd Noah`, generated 96.88% and 69.82%; A4 = `Tengku Mohd Tengku Sembok`, generated 95.24% and 77.52%; A5 = `Ahmed Moosajee Patel`, generated 97.14% and 41.46% [11]. Respectively the results are recall and precision for documents already stated. These results are much better than involving measurement based on disambiguation only [29]. Therefore, for k stars are obtained $\sum Rec = \sum_{i=1}^{k} Rec_{a_i}$ and $\sum Prec = \sum_{i=1}^{k} Prec_{a_i}$ or

$$Rec = \sum_{i=1}^{k} \frac{|D_s \cap \cup_{j=1}^{n_i-1} D_w|}{|D_s|} \tag{5}$$

and

$$Prec = \sum_{i=1}^{k} \frac{|D_s \cap \cup_{j=1}^{n_i-1} D_w|}{|\cup_{j=1}^{n_i-1} D_w|} \tag{6}$$

Thus, the mean of Rec and $Prec$ for Eq. (5) and Eq. (6) illustrates Rec and $Prec$ as a whole if each star is independent of each other [30, 8].

In graph theory, if the leaf candidate has a high enough degree, the initial step of edge elimination is still performed, but the next step the edges that has been eliminated is returned to the candidate leaf unless it is connected to the first star center, in order to the candidate of the leaf to become the next star center candidate. Thus, although there are two vertices within the social network having the same degree and each being the center of the star, but the degree of center will not be the same. In other words, the connecting edge between two stars becomes the constraint for equation Eq. (5) and Eq. (6), so as to lift all relevant documents of a community from information sources can be optimized by reducing the constraints, i.e.

$$Rec_c = \frac{|D_s \cap \cup_{i=1}^{k} \cup_{j=1}^{n_i-\ell} D_w|}{|D_s|} \tag{7}$$

and

$$Prec_c = \frac{|D_s \cap \cup_{i=1}^{k} \cup_{j=1}^{n_i-\ell} D_w|}{|\cup_{i=1}^{k} \cup_{j=1}^{n_i-\ell} D_w|} \tag{8}$$

where ℓ is a deduction to optimize the results of constraints. For example, for an average of Rec and $Prec$ for five social actors are 96.31% and 54.71%, respectively. By involving Eq. (5) and Eq. (6) we generate Rec = 97.07% and $Prec$ = 55.40% for optimizing the recall and the precision.

5. Conclusion

The extraction of social networks requires information retrieval to validate the truth of information, while information retrieval involving social structures can be used to improve information retrieval performance. The mutual interest relationship between IR and social networks has been revealed through the extraction of social networks so that recall and precision formulations as measurements of IRs can be adjusted. This adjustment results in an optimal form of recall and precision by the implementing in accessing information about a community.

References

[1] Nasution M K M, Sitompul O S, and Noah S A 2018 Social network extraction based on Web: 3. the integrated superficial method *Journal of Physics: Conference Series* **978(1)**.

[2] Soussi R, Aufaure M-A, and Baazaoul H 2010 Towards social network extraction using a graph database *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*

[3] Nasution M K M and Noah S A 2017 Social network extraction based on Web. A comparison of superficial methods *Procedia Computer Science* **124**.

[4] Kijkuit B and Ende J V D 2007 The organizational life of an idea: Integrating social network, creativity and decision-making perspectives *Journal of Management Studies* **44(6)**.

[5] Nasution M K M 2018 Social network extraction based on Web: 1. Related superficial methods *IOP Conference Series: Materials Science and Engineering* **300(1)**.

[6] Nasution M K M 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)**

[7] Nasution M K M, Noah S A M and Saad S 2011 Social network extraction: Superficial method and information retrieval *Proceeding of International Conference on Informatics for Development* (ICID'11).

[8] Elveny M, Syah R, Elfida M and Nasution M K M 2018 Information retrieval on social network: An adaptive proof *IOP Conference Series: Materials Science and Engineering* **300(1)**

[9] Nasution M K M and Noah S A 2012 Information retrieval model: A social network extraction perspective *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management* (CAMP'12).

[10] Nasution M K M, Hardi M and Sitepu R 2016 Using social network to assess forensic of negative issues *Proceedings of 2016 4th International Conference on Cyber and IT Service Management*, CITSM.

[11] Nasution M K M, Syah R, and Elfida M 2018 Information Retrieval Based on the Extracted Social Network *Advances in Intelligent Systems and Computing* **662**.

[12] H Kautz, B Selman, and M Shah 1997 ReferralWeb: Combining social networks and collaborative filtering *Communications of the ACM* **40(3)**.

[13] Nasution M K M, and Noah S A 2011 Extraction of academic social network from online database *2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011.*

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

[15] Nasution M K M 2018 Singleton: A role of the search engine to reveal the existence of something in information space *IOP Conference Series: Materials Science and Engineering* **420(1)**

[16] Nasution M K M 2018 Doubleton: A role of the search engine to reveal the existence of relation in information space *IOP Conference Series: Materials Science and Engineering* **420(1)**

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

[18] Dridi A, and Slimani Y 2017 Leveraging social information for personalized search *Social Network Analysis and Mining* **7(1)**.

[19] Nasution M K M 2012 Simple search engine model: Adaptive properties *Cornell University Library* arXiv:1212.3906 [cs.IR].

[20] Nasution M K M 2012 Simple search engine model: Adaptive properties for doubleton *Cornell University Library* arXiv:1212.4702 [cs.IR].

[21] Nasution M K M 2018 Semantic interpretation of search engine resultant *IOP Conference Series: Materials Science and Engineering* **300(1)**.

[22] Nasution M K M 2013 Superficial Method for Extracting Academic Social Network from the Web, *Ph. D. Thesis*, FTSM UKM Bangi, Malaysia.

[23] Nasution M K M, Elveny M, Syah R, and Noah S A 2015 Behavior of the resources in the growth of social network *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community* (ICEEI).

[24] Boettcher S, and Percus A G 2001 Extremal optimization for graph partitioning *Phys. Rev. E.* **64**.

[25] Manik E, Suwilo S, Tulus, and Sitompul O S 2018 On the 5-local profile of trees *IOP Conference Series: Materials Science and Engineering* **300(1)**.

[26] Kawaguchi N, Azuma Y, Ueda, S, Shigeno, H and Okada K 2006 ACTM: Anomaly connection tree method to detect silent worms *20th International Conference on Advanced Information Networking and Applications* Volume **1** (AINA'06).

[27] Nasution M K M 2017 Modelling and simulation of search engine *Journal of Physics: Conference Series* **801(1)**.

[28] Nasution M K M, Syah R and Elveny M 2017 Studies on behaviour of information to extract the meaning behind the behaviour *Journal of Physics: Conference Series* **801(1)**

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

[30] Nasution M K M and Sitompul O S 2017 Enhancing extraction method for aggregating strength relation between social actors *Advances in Intelligent Systems and Computing* **573**.