Social network extraction based on Web: A Review about Supervised Methods

Mahyuddin K. M. Nasution\textsuperscript{1,*}, Shahrul Azman Noah\textsuperscript{2,**}

\textsuperscript{1}Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, USU, Padang Bulan 20155, Medan, Sumatera Utara, Indonesia
\textsuperscript{2}Knowledge Technology Research Group, M-CAIT, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia.

E-mail: *mahyuddin@usu.ac.id,**samn@ftsm.ukm.my

Abstract. The extraction of social networks from specific sources of information is essential. It relates to the disclosure of social structures with prevailing behavior in accordance with that information source. It, of course, requires methods that are generally in a supervised stream. The method changes based on the demands of data modeling, which are generally textual, but do not rule out other types of information, such as databases or different literacy. This paper reviews the methods that have been developed and the types of information sources involved as input to the social network extraction process. This brief review follows the literature related to social network extraction involving supervised methods. Based on different information sources, there are different models in supervised stream.

1. Introduction

Social network extraction can have commented as something mapping entity relationships among all entities [1, 2], represent patterns of relationships between entities [3], measuring social capital, i.e., the value obtained by social actors individually or in groups [4], and present a variety of social structures appropriate to its importance and implementation in different domains or sources of information [5]. In this regard, the extraction of social networks is to produce technology to identify and explain especially content from information sources that are actors and their relationships and provide a way to find social structure. That is, a way to identify social actors and tie relationships between social actors.

Regarding the extraction of social networks from information sources and application for dealing with next issues [6, 7, 8], it is necessary to review as a first step to reveal various issues related to social network extraction [9, 10, 11], i.e. about paradigm, strategy, and information source. In related cases, this paper will provide a review of the interests about and supervised methods for extracting social networks from information sources based on a study [12] and expand it through next literatures, by aiming to reveal possible other model in same stream.

2. Background

Classically, the extraction of social networks is carried out through data on social relations by social scientists done with care and trust, carefully and meticulously [13]. Accordingly, the extraction involves a long time [14]. Classical methods typically entangle questionnaires,
Figure 1. A tree of the social networks where an author as central - Source: https://sinta.ristekbrin.go.id/authors/detail?id=56676&view=network

interviews, live observations, manually filtering archival records, or diverse experimental results [15]. So finding who has a relationship with whom in a large-scale information space becomes a complex problem. Furthermore, asking each individual about their relationship is impractical, given the large number of individuals involved with the dynamic change of relationships that may exist [16].

Advances in information technology in terms of natural language processing have allowed more information about individual activities to be gathered from documents [17]. In general, all these documents do not have the same structure as a database and lack semantics in their interpretation [18]. In this case, the social network extraction method requires not only a tested formulation but also items related to that formulation. Evaluation and testing reveal the trustworthy level of information.

3. An Approach
The purpose of reviewing social network extraction is to see from a broad perspective were the target of social network research that involves name recognition and the identification of relationships between social actors [19]. However, this social network extraction research only focuses on individual names in web documents, and some social network studies either involve different domains or other online information sources, which may take consideration [20]. There are also some studies from online sources of information related to famous people, such as artists and leaders or other entities in the company. For example, see Figure 1, a social network (tree) that shows one’s centrality based on online database as an information source. Any network always consists of two components as a base, namely nodes and edges. Therefore, social network extraction also bases on two tasks, namely identifying actors for nodes and the relationship between actors for edges.

The source of information contains the name of the actor ambiguity (ambiguity), which requires the process of name disambiguation [21, 22]. In this case, each actor’s name well identifies in the various patterns of name labels that may be available and take into account the appropriate pattern as a preliminary consideration for performing the obscurity process. A list
of the names of the identified actors is a basis for identifying the relationship between a couple of actors. While name disambiguation supports the trusted relationship between the actor couple [2].

Generally, information extraction is the process of filling in fields and database records (structured) from unstructured text (information sources) or loose formatting [23, 24] and involves task criteria: segmentation, classification/clustering, consolidation, normalization, and deduplication. So is the extraction of social networks, that is, assuming social networks as information to be generated from a heterogeneous source of information, then social network extraction also takes into account the five criteria of the above information extraction task.

4. Results and Discussion

In the field of artificial intelligence, there are two streams of research to extract social networks from various sources of information. One of them is supervised method [4, 5, 25, 26, 27, 28, 29, 30, 31].

The supervised method uses a function $\lambda$ to classify $Z$, namely $\lambda : Z \rightarrow C$ is like $\lambda(z) = c$, i.e., $z$ in $Z$, and $c$ in $C$ is the class label, i.e., $C = \{c_1, c_2, \ldots, c_{|C|}\}$ is the data set as a target attribute, $|C| \geq 2$ is the number of classes, and $Z \cap C = \emptyset$. This method involves corpus analysis to identify labels from relationships in social networks, in which each actor takes and gives various labels extracted from sources of information [32, 33]. Formally the corpus, latent, and triplet variables are defined as follows.

T1 Corpus:
Corpus $D$ is a collection of $n$ documents represented by $D = \{d_k|k = 1, \ldots, n\}$, where $d$ is the document [34].

T2 Latent variables:
The relationship between actor names, documents, and words are interpreted by a set of latent variables $Z = \{z_1, \ldots, z_l\}$, $l$ is size of $Z$ and each element of $Z$ represents a latent topic [35].

T3 Triplet:
Triplet $\{A, D, W\}$ is an aspect model called a generative model, through $a \times d \times w$, or $P(a, d, w) = P(d)P(a, w|d)$, where $P(a, w|d) = \sum_{z \in Z} P(a, w|z)P(a|D)$, $P(d)$ is the probability of document $d$ in corpus $D$, latent class $z$ with probability $P(z|d)$, word $w$ with probability $P(w|z)$, and the name $a$ with the probability $P(a|z)$ [36].

There are many supervised method strategies that can be used in information extraction such as Bayesian, decision tree, induction tips, and support vector machine (SVM). However, most social network extraction methods involve a generative probabilistic model (GPM) strategy [37].

In the supervised research stream, each actor matches a variety of labels extracted from sources of information related to the actor, and appropriate labels will classify the relationship between the actors. Statistical language processing is statistically an analysis to master the richness of language content from interactions involving words, topics, and other dimensions that derive from interactions between actors [38]. Statistically, probability models have been constructed and used on End-to-End systems [5] to analyze corpus and group actors based on attribute and connection parameters. The corpus built from various sources of information such as email [4, 5, 25, 26, 27, 39, 40], Web [5, 41, 42, 43], working paper group [29, 44] or documents [28, 30, 31], databases [29, 44], different services [41, 42, 43] or homepage [5]. The researchers propose methods. It involves strategies for using machine learning and training conditional random field (CRF). It is to recursively recognize the actor’s name and produce a network with the concept of "friend of friend": That is a system of defining seed actors from email and subsequently generating other actors and relationships between actors from the seed homepage.
Table 1. Summary of reviews on supervised research papers/articles

| References | Paradigm | Strategy | Information Source |
|------------|----------|----------|--------------------|
| [5]        | End to Eng system | Probability: email, Web, homepage |                     |
| [25, 29]   | ART      | Bayesian network-Monte Carlo Gibbs sampling | email |
| [4, 23, 26, 29] | RART    | Bayesian network-Stochastic block-structures | email |
| [30, 31]   | GT       | Bayesian network | Text US senate dataset |
| [40]       | DART     | Bayesian network-CRF, MMR, and TF.IDF | email |
| [28]       | GOT      | Bayesian network-GSI | Text US senate dataset |
| [29, 44]   | APT      | Bayesian network-Rexa database, papers |                      |
| [41, 42, 43] | ArnetMiner system and ACT | Bayesian network-CRF, EM and HMMF | Web, Web Services |

In other words, using parameters generally is to gain knowledge from the corpus, which is a modality of information that can be combined to discover hidden structures based on triplet definitions, involving Bayesian probabilities, for which such a network is called a Bayesian network. Based on those parameters, generative probabilistic models (GPM) can determine the relationship between actors in pairs. The author recipient topic (ART) model, for example, from GPM, is specially designed to capture the meaning of the language used to generate a network of direction by invoking the Monte Carlo Gibbs sampling strategy to obtain a sequence of observations more or less as defined [25]. As an extension of the ART model, role-ART (RART) model groups actors with similar roles using the Stochastic Blockstructures strategy [23]. While, the group topic (GT) model takes into account the characteristics of the relationship text and enables group discovery through new topics that appear in the text [30, 31]. As a modification of the RART model, the discriminative ART (DART) model uses mean reciprocal rank (MRR) for defining latent variables. It involves term frequency-inverse document frequency (TF.IDF) to obtain document vectors and trains CRF to represent hidden predictors of performance for some users [40]. In specific cases combining time and text modalities is to identify topic trends over time by involving group switch index (GSI) indexing strategies, and this is known as the group over time (GOT) model [28]. While exploring the expertise modality through the author-persona-topic (APT) model only is to match the reviewers of the research paper to be submitted to the appropriate person as an implicit form of social network [44].

Although a surge of importance and complexity in the models that reveal the role-based relationship of the actors. Nevertheless, the stream of supervised research involves corpus analysis to identify the descriptions of relationships given in social networks with increased and pre-determined parameters. Nevertheless, the stream of supervised research involves corpus analysis to identify the descriptions of relationships given in social networks with increased and
pre-determined parameters. The use of the corpus in identifying the relationship between
the actors causes disambiguation in parallel can do during the collection of documents. Such as
in the specific case, social networks of the academic, some models from GPM are proposed by
adding and modifying some parameters for the ArnetMiner system [41], but only concerning
topic and conference model implementation strategies. The model is known as the author
conference-topic (ACT) model, and it has a supporter by the use of a hidden Markov random
field (HMRF) in name disambiguation and expectation maximization (EM) as a framework for
assessing parameters in the form of triplets [42]. Nevertheless, the ArnetMiner system uses
the ACT model as the basis for the generation of social networks, where the mining of social
networks of academics is enhanced by involving unregulated methods [43].

The ACT model and the ART model generally have the same concept. The fundamental
difference between these two models is that ART involves a group of documents with a log
system, i.e., email [25]. While the ACT base on a list of papers presented at a scientific conference
(not involving a log system). Thus, it makes possible sources of information come directly
from the Web, although it involves the limited representation of documents [42, 43]. Another
difference between the two is that ART models sometimes produce asymmetric relationships
when messages do not respond. While ACT models always reveal symmetrical relationships
because of the existence of relationships bases on co-occurrence.

A summary of the existed research work on the supervised research stream likes in Table
1. A table that describes, in general, the method of social network extraction with contact
naming involving document groups (corpus), and to further serve as a guide to generate labels
for contact naming involves different methods.

The sources of information involved in Table 1 contain social actors with different roles. In
a specific case, emails as the information source where actors can become the recipients and the
senders of messages or both at once, and they are connected by the existence of communication
via email [4, 5, 25, 26, 27, 29, 40]. At Web information sources, social actors act as authors, and
it will generate relationships through the concept of the existence of two or more actor names in
a document or co-occurrence [5, 41, 42, 43]. Similarly, by involving the source of information on
the homepage (personal site), each actor is represented by his/her homepage where the concept
of relationship as generation from the record defects of other actors that exist on the homepage
or hyperlinks to other Web sites or homepage [5]. Using the concept of co-occurrence also in
the list of working papers (databases) representing any scientific conferences [29, 44] or other
sources of information [41, 42, 43], with which generating relationships through the presence
of actors or co-authors of working papers. The concept of similarity of discussion topics will
connect senators as social actors or group them into different groups [28, 30, 31].

Also, on different occasions, using the same stream of methods, social network extraction is
based on a strategy by marking the link of the event-participant community [45, 46]. The use of
relational databases has also become a target source of information from social networks, namely
the graph operation strategy [47]. In a different research, the development of a methodology
to analyze several social data individually into a tree structure and then collecting it into a
social network [48]. Using novel sources, literature, social network extraction using character
name truncation, and detecting conversations as determinants of relationships between actors
in novels, which reveal different things about influential characters [49]. The adjustment of the
method to information sources for extracting social networks also applies to data sources for
literary works that vary in style and writing every era. Adjustments, in this case, involve the
approach of artificial intelligence [50].

By considering the historical and literary literature, social network extraction involving
supervised stream has its challenges, especially when expressing the identity of each community,
social structure, and the ups and downs of a culture.
5. Conclusion

The supervised method is one of the streams in extracting social networks, which has its own characteristics, especially with regard to the concept of network shared probability. The methods are developed by involving one parameter after another to reveal a social structure that is richer with meaning. The modeling approach is to deal with data models from information sources, especially with regard to tracing entities that are social actors. Each method of forming a social community is different according to the information revealed. Obviously, it is through this the development of sustainable methods when dealing with literature that does not behave in the same way.

References

[1] YingZi Jin, Y Matsuo, M Ishizuka 2006 Extracting a social network among entities by web mining. Workshop on Web Content Mining with Human Language, ISWC.
[2] YingZi Jin, Y Matsuo, M Ishizuka 2007 Extracting social networks among various entities on the Web ESWC 2007, LNCS bf 4519.
[3] Y Matsuo, J Mori, M Hamasaki, T Nishimura, T Takeda, K Hasida, M Ishizuka 2007 POLYPHONET: An advanced social networks extraction system from the Web Journal of Web SEManetics: Science, Services and Agents on the World Wide Web 5.
[4] A McCallum, A Corrada-Emmanuel, X. Wang 2005 The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and Academic Email Computer Science Department Faculty Publication Series. Papeer 44.
[5] A Cullota, R. Bekkerman, A McCallum 2004 Extracting social networks and contact information from email and the Web Proceedings of the 1st Conference on Email and Anti-Spam. CEAS.
[6] N Nursalam, M Octavia, R D Tristiana, F Efendi 2019 Association between insomnia and social network site use in Indonesian adolescents Nursing Forum 54(2).
[7] R Pannyiwi, A Agustang, T Kasnawi, A T Pada, A Yani, S Syam 2020 Social network for drug circulation in sidenreng rappang regency, Indonesia Systematic Reviews in Pharmacy 11(9).
[8] H Pravowo, R Bramulya, Yuniar 2020 Student purchase intention in higher education sector: The role of social network marketing and student engagement Management Science Letters 10(1).
[9] S Escalera, X Baró, J Vitià, P Radeva, B Raducanu 2012 Social network extraction and analysis based on multimodal dyadic interaction Sensors 12(2).
[10] B Turnbull, S Randhawa 2015 Automated event and social network extraction from digital evidence sources with ontological mapping Digital Investigation 13.
[11] R M Aliguliyev, R M Aliguliyev, G Y Niftaliyeva 2019 A method for social network extraction from e-government International Journal of Information Systems in the Service Sector 11(3).
[12] M K Nasution 2013 Kaedah dangkal bagi pengekstrakan rangkaian sosial akademik dari Web Disertasi - Universiti Kebangsaan Malaysia.
[13] F L Gaol, B Widjaja 2008 Framework of regression-based graph matrix analysis approach in multi-relational social network problem Journal of Mathematics and Statistics 4(1).
[14] K S Morrow, H Glanz, P O Ngakan, E P Riley 2019 Interactions with humans are jointly influenced by life history stage and social network factors and reduce group cohesions in moor macaques (Macaca maura) Scientific Reports 9(1).
[15] E Bott 1955 Urban families: Conjugal roles and social networks Human Relations 8(4).
[16] P Mika 2007b Social Networks and the Semantic Web Berlin: Springer-Verlag.
[17] Erlin, N Yusof, A A Rahman 2010 Messages segmentation of asynchronous text-based discussion in social network analysis ICEICT 2010 - 2010 2nd International Conference on Education Technology and Computer 3.
[18] G T I Tawakkal, R Damayanti 2020 Social networks and brokerage behavior in Indonesian elections: Evidence from Central Java Asian Affairs(UK) 47(3).
[19] M S Shaharudin, Y Fernando, C J Chiappetta Jabbour, R Sroufe, M F A Jasmi 2019 Past, present, and future low carbon supply chain management: A content review using social network analysis Journal of Cleaner Production 218.
[20] Y-C Lin, R Padliansyah, T-C Lin 2019 The relationship and development trend of corporate social responsibility (CSR) literature: Utilizing bibliographic coupling analysis and social network analysis Management Decision 58(4).
[21] Y Song, J Huang, I G Councell, J Li, C L Giles 2007 Generative models for name disambiguation WWW 2007.
[22] Y Song, J Huang, I G Council, J Li, C L Giles 2007 Efficient topic-based unsupervised name disambiguation JCDL’07.
[23] A McCallum 2005 Extraction: distilling structured data from unstructured text ACM Queue.
[24] M R Alifi, S H Supangkat 2016 Information extraction for traffic congestion in social network: Case study: Bekasi city 2016 International Conference on ICT for Smart Society, ICISIS.
[25] A McCallum, A Coorada-Emmanuel, X Wang 2004 The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and Academic Email Technical Report, University of Massachusetts, Amherst.
[26] A McCallum, A Coorada-Emmanuel, X Wang 2005 The author-recipient-topic model for topic and role discovery in social networks, with application to Enron and Academic email Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security (in conj. with SIAM International Conference on Data Mining).
[27] A McCallum, A Coorada-Emmanuel, X Wang 2005 Topic and role discovery in social networks Proceedings of the 19th International Joint Conference on Artificial Intelligence.
[28] A McCallum, X Wang, A Mohanty 2007 Joint group and topic discovery from relations and text Statistics Networks Analysis: Model, Issues, and New Directions (ICML), LNCS 4503.
[29] A McCallum, X Wang, A Corrada-Emmanuel 2007 Topic and role discovery in social networks with experiments on Enron and Academic Email Journal of Artificial Intelligence Research 30.
[30] X Wang, N Mohanty, A McCallum 2005 Group and topic discovery from relations and text Proceeding of the 3rd Workshop on Link Discovery International Conference on Knowledge Discovery and Data Mining LinkKDD.
[31] X Wang, N Mohanty, A McCallum 2006 Group and topic discovery from relations and their attributes Advances in Neural Information Processing Systems 18.
[32] Oded Maimon, Lior Rokach 2005 Data Mining and Knowledge Discovery Handbook Berlin: Springer-Verlag.
[33] Oleg Okun, Giorgio Valentini 2009 Applications of Supervised and Unsupervised Ensemble Methods Berlin: Springer-Verlag.
[34] S Fitzmaurice 2007 Questions of standardization and representativeness in the development of social networks-based corpora: The story of the network of eighteenth-century English texts Creating and Digitizing Language Corpora: Volume 2: Diachronic Databases.
[35] B Dabbs, S Adhikari, T Sweet 2020 Conditionally Independent Dyads (CID) network models: A latent variable approach to statistical social network analysis Social Networks 63.
[36] Y Zhang, B Wu, N Ning, C Song, J Lv 2019 Dynamic Topical Community Detection in Social Network: A Generative Model Approach IEEE Access 7.
[37] B Liu 2007 Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data Berlin: Springer-Verlag.
[38] D M Blei, A Y Ng, M I Jordan 2003 Latent dirichlet allocation J. Mach. Learn. Res. 3.
[39] A McCallum, X Wang, A Corrada-Emmanuel 2007 Topic and role discovery in social networks with experiments on Enron and Academic Email Journal of Artificial Intelligence Research 30.
[40] C Pal, X Wang, A McCallum 2007 Transfer learning for enhancing information flow in organizations and social networks DTIC - Online Information for the Defence Community.
[41] J Tang, D Zhang, L Yao 2007 Social network extraction of academic researchers Proceedings of Seventh IEEE International Conference on Data Mining.
[42] J Tang, D Zhang, L Yao, J Li 2008 Extraction and mining of an academic social network Proceedings of WWW.
[43] J Tang, D Zhang, L Yao, J Li, L Zhang, Z Su 2008 ArnetMiner: Extraction and mining of academic social networks Proceedings of KDD’08.
[44] D Minno, A McCallum 2007 Expertise modelling for matching papers with reviewers Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[45] M Hamasaki, Y Matsuoi, K Ishida, Y Nakamura, T Nishimura, H Takeda 2005 Community focused social network extraction Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) LNCS 4185.
[46] M Hamasaki, Y Matsuoi, K Ishida, T Hope, T Nishimura, H Takeda 2006 An integrated method for social network extraction Proceedings of the 15th International Conference on World Wide Web.
[47] R Soussi, M-A Auffare, H Baazouzi 2010 Towards social network extraction using a graph database 2nd International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA.
[48] K-Y Wang, J-H Ting, H-J Wu, P-S Chang 2010 A dynamic and task-oriented social network extraction system based on analyzing personal social data Proceedings - 2010 International Conference on Advances in Social Network Analysis and Mining, ASONAM.
[49] D L Elson, N Dames, K R McKeown 2010 Extracting social networks from literary fiction ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.
[50] N Dekker, T Kuhn, M van Erp 2019 Evaluating named entity recognition tools for extracting social networks from novels PeerJ Computer Science 4.