# Model of Cytation Network Analysis using Sequence of Words as Structured Text Representation

IOP Conf. Series: Materials Science and Engineering **288** (2017) 012048 doi:10.1088/1757-899X/288/1/012048

# Model of Cytation Network Analysis using Sequence of Words as Structured Text Representation

**D S Maylawati[1*], A Rahman[1], M I N Saputra[1], W Darmalaksana[2] and M A Ramdhani[1]**
[1]Departement of Informatics, UIN Sunan Gunung Djati Bandung
[2]Research Center, UIN Sunan Gunung Djati Bandung

[*]diansm@uinsgd.ac.id

**Abstract.** Citation in research activity is crucial. Various of journal publication media are available as a forum for publication and a source of research reference. In this study, we propose an analytical model for cytation networks in journal publications using a network of sciences. In addition, our cited network analysis model contains a structured text representation with Sequence of Words (SOW) form in the pre-process stage. The algorithm which is used to produce text representation in this research is the algorithm that corresponds to the generated SOW form, such as sequential or frequent pattern algorithm. The conceptual validation of the cytation network model in this study was carried out at the Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung. By using Focus Group Discussion (FGD) as a conceptual validation, the cited network analysis model in this study is ready and possible to implement easily because it is described in general concept.

## 1. Introduction

Citation becomes important in research activities and publications of scientific papers. The relation between one and another citation becomes interesting to discuss about the research, ranging from finding out research ideas to the ongoing process of research. In fact, citation relationship can be an analytical tool for assessing the social network of researchers, research networks of interest based on research area, interrelated recommendation of publications for reference, and other analysis based on the results of mapping of citation networks. In this article, we create an analytical model to create a citation network by utilizing a structured text representation or Sequence of Words (SOW) because the citation data is text-shaped. Citation publication data of journals or articles need to be well prepared to obtain a good citation network as well. The frequency of occurrences of citation attributes such as author names, titles, emails, and affiliations in journals are the key in determining citation networks. Text data will be prepared in pre-process by forming a structured text representation on the basis of sequential pattern (SP) which is one of the forms of multiple of words [1] [2] [3] [4] [5]. Based on the SP concept that considers the order of occurrences of items in the transaction collection [6] [5] [3] [7], the representation of structured text in the form of SP also concerns the order of occurrences of words in the document collection [3] [1] [6] [7]. SP for ordinary text representation is called Sequence of Words.

The analysis model in this study is validated by conceptual validation and Focus Discussion Group (FDG). Conceptual validation is one of the conceptual model validation processes that has broad terminology and point of view, where the assumptions and theoretical basis are the reference [8]. The

assumptions and the theoretical foundations are made to which degree the model corresponds to the problem. The conceptual validation measures and tests the suitability between problems encountered with the concept or model being designed, after considering the provided assumptions or axioms [8]. Meanwhile, FGD is used to support the validity of models that have been tested by conceptual validation. FGD can provide assessment based on the results of discussions of groups of people who have the same background, scholarship, interest, experience, and field of research [9] [10] [11]. FGD can decide things that cannot be explained statistically by building discussions and generating opinions on issues being discussed. This becomes one of the strengths of FGD because each member of the group can either agree or reject what is being discussed in terms of perspective, experience, and practical points of view.

Furthermore, in session 2 pre-process will be discussed to prepare the text data according to the needs of the analysis model, session 3 discusses the sequence of words representation. Then in session 4 graph theory and network science model will be discussed, next session 5 discusses the citation network analysis model proposed in this research. In session 6, there will be evaluation of the proposed model that was performed using conceptual validation and FGD.

## 2. Text Data Pre-processing

Text pre-process in text mining is one of the important processes [3] [12], as well as in data mining and machine learning [13]. This is because in a good pre-process, the text data is prepared so that in the next process (mining) can also produce a good output [3] [12]. The pre-process in this study consists of extracting citation data from a collection of IEEE publication text documents, afterward tokenizing, lower case, forming a sequence of words representation are done. Pre-process in this research perform stop word removal and stemming process according to requirement. For example, when title and abstract are used as data in forming network citation, stop word removal and stemming process are certainly necessary to do first. Text data that has been prepared to do the process of citation network analysis is in the structured text representation in the form of sequence of words. The representation of sequence of words that can be formed will be discussed at session 3.

**Table 1.** The structured text representation in the form of SOW [2] [3] [5]

| Structure | Structure Illustration | Order Rules | Meaning |
|---|---|---|---|
| **Representasi FWS:** | | | |
| Documents are seen as the set of Frequent Word Sequence (FWS). | < (w1, w2), (w3, w4),>, where (w1, w2) is FWS I, (w3, w4) is FWS i+1 and so on. | a. The order of occurrence of FWS = sequence of occurrences in the document.<br>b. The order of occurrences of the FWS element = the order of occurrences of words in the document. | In the collection of documents FWS I often appear followed by FWS i + 1 and so on. Elements or items in FWS I, w1 always appear followed by w2, as well as the appearance of elements on FWS i + 1 and so on. |
| **Set of FWS Representation** | | | |
| Documents are seen as the set of the Frequent Word Sequence Set | <<(w1,w2)>,<(w3,w4), …>>, where (w1,w2) is FWS I is in the first sentence, (w3, w4) is FWS i + 1 fst is in the second sentence. | a. A. Sentence sequence = sequence of occurrences in the document.<br>b. B. The order of occurrence of FWS = sequence of occurrences in the document.<br>c. C. The order of occurrences of the FWS element = the order of occurrences of words in the document. | In the collection of documents other than the occurrence of FWS I often appear followed FWS i + 1 and so on, the appearance of the first sentence often appears followed by the second sentence and so on. Elements or items in FWS I, w1 always appear followed by w2, as well as the appearance of elements on FWS i + 1 and so on. |
| **FWI Representation:** | | | |

| Structure | Structure Illustration | Order Rules | Meaning |
|---|---|---|---|
| Documents are seen as the set of Sequential Pattern Item sets | <(w1,w2),(w3,w4),…>, where (w1,w2) is FWI I, (w3,w4) is FWI i+1 and so on. | a. A. The order of occurrences of FWI = sequence of occurrences in the document.<br>b. B. The order of occurrences of FWI elements does not have to be the order of occurrences of words in the document. | In the collection of documents FWI I often appear followed by FWI i + 1 and so on. Elements or items in FWI I, w1 always appear simultaneously with w2 without having to be sequentially w1 followed by w2, if w2 appears earlier than w1 is considered the same FWI, as is the appearance of elements on FWS i + 1 and so on. |
| **Set of FWI Representation:** | | | |
| Documents are viewed as the set of sets of sets Sequential Pattern | <<(w1,w2),(w3,w4),… >>, where (w1,w2) is FWI I is in the first sentence, (w3, w4) is FWI i + 1 dst is in the second sentence. | a. Sentence sequence = sequence of occurrences in the document.<br>b. The order of occurrences of FWI = sequence of occurrences in the document.<br>c. The order of occurrences of FWI elements does not have to be the order of occurrences of words in the document | In the collection of documents other than the occurrence of FWI I often appear followed by FWI i + 1 and so on, the appearance of the first sentence often appears followed by the second sentence and so on. Elements or items in FWI I, w1 always appear simultaneously with w2 without having to be sequentially w1 followed by w2, if w2 appears earlier than w1 is considered the same FWI, as is the appearance of elements on FWS i + 1 and so on. |

## 3. Sequence of Word Representation Analysis

Text is unstructured data that needs to be structured before mining can be done. A well-structured text representation is a form of multiple of words [2] [3]. Sequence of Words (SOW) is one form of multiple words that is adapted from the sequential pattern [5] [3] [1] [2] [7]. Basically, representations of structured text have the form of sequential pattern (SP) and frequent pattern (FP) [5]. SP represents text data into SOW by observing the order of occurrences of words in text data. While the FP does not pay attention to the order of occurrences of the word, but the representation of the text formed consisted of words that always appear together. The representation of structured text in the form of SP is called Frequent Word Sequence (FWS) [5] [14] [2] [15], while the structured text representation in FP forms is called Frequent Word Itemset (FWI) [5] [3] [4]. There is also a structured text representation developed from FWS and FWI such as the Set of FWS [1] [2], Set of FWI [3] [4], and Cord FWI [1]. Set of FWS and Set of FWI have a syntax by taking note of the order of occurrence of sentences in the text data so that the meaning of the text is safer. More detailed explanations relating to FWS, FWI, Set of FWS, and Set of FWI are listed in Table 1.

## 4. Graph Theory and Network Model Concept Analysis

Basically, the network model formed is a graph. The graph is a representation of discrete objects and the relationship among these objects [16] [17] [18]. In which, discrete objects are represented by vertices and edges. In this study, the node represents the citation and the side represents the relationship between one citation and another citation. The network model adapts the graph theory by combining the type of graph and terminology graph, especially the degree of vertex of a graph. The degree of a node is the number of sides adjacent to that node or connecting between vertices [16] [17]. The degree of vertices in the network model can show how much and in the connection between the nodes representing citation [18].

Network model used in this research is related to network science. Network science is a network model that represents real-world networks, such as neural networks, communications, control, social,

economic, etc. that are complex and need to be modeled so that they are easily understood. [19] [20] Various network models can be implemented, including the widely used Barabasi-Albert [18] network, the Random network, Coevolving Multigraph built using multi-layer analysis with multiple multigraphs inspired by social networks [19], and other network models that can be implemented for analyzing the citation network. Collaboration on the network model is fundamental [18] [19] on the network science model. It is necessary to measure the robustness of network science models. Measurements can be done in many ways, one of them is by measuring the robustness of the network correlation degree.

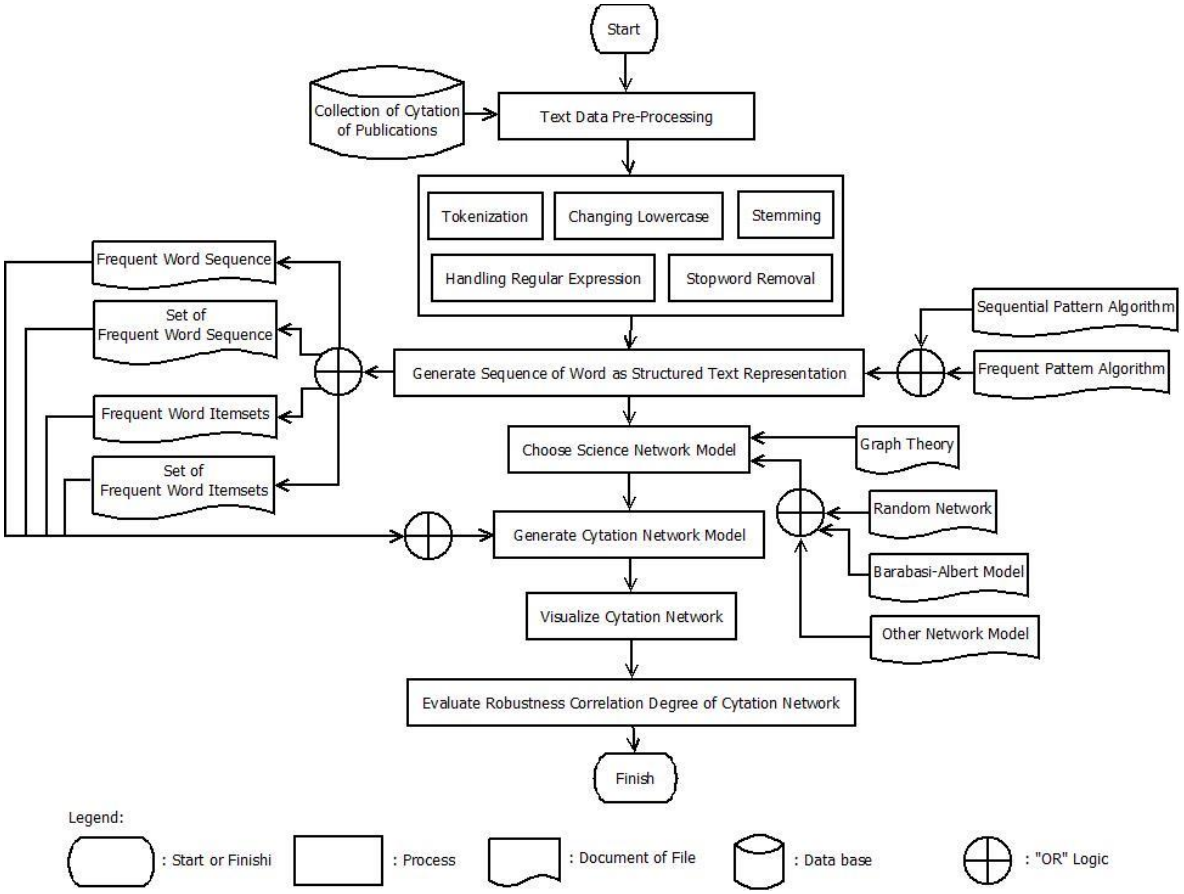


**Figure 1.** Publication Citation Network Analysis Model with SOW Structured Text Representation

## 5. Results and discussion

Based on the discussion of sessions 2, 3, and 4, we proposed a model for the analysis of the cited network presented in Figure 1. Starting from collecting the publication citation documents the then pre-process to prepare text data into a structured representation in the form of SOW. The formation of SOW representation can use SP algorithm or FP algorithm [5]. The SOW formed can be the representation of FWS, Set of FWS, FWI, or Set of FWI [2] [3]. Furthermore, after determining the model of network science that can be used for network citation model based on graph theory, network citation model according to network science model and applied SOW representation will be formed. The models of science network that can be applied are as described in session 4, including Barabasi-Albert, Coevolving Multigraph, Random, or other network science models [18] [19] [20]. The network model is visualized and evaluated robustness of the resulted citation network correlation degree.

State Islamic University (UIN) Sunan Gunung Djati Bandung is one of the educational institutions that run Tri Dharma Perguruan Tinggi, one of which is research. Department of Informatics is considered appropriate to be the environment to evaluate the citation network model in this study. This is because the conceptual and theoretical basis for network citation model in this study is owned by Department of Informatics Engineering. Department of Informatics UIN Bandung has several Group of Expertise (KK), among others KK Programming and Software Engineering and KK Computer Vision and Intelligent System. The two KK are the object of conceptual validation with FGD as it related to the research area in this study. Based on FGD evaluation results on the evaluation environment, approximately 84% of the 16 second members of both KK plus 3 other KK members with similar research concluded that the citation network analysis model in this study could be implemented.

## 6. Conclusion

The citation network analysis model in this study was constructed in a structured way with the flow of text data collection up to evaluation. The citation network analysis uses a science network based on graph theory. Various types of network science that can be applied are among others; Barabasi-Albert network, random, multigraph coevolving, and other science networks. In addition, prior to establishing a citation network, in the pre-process of preparing text data using a structured representation of the SOW form, it can be FWS, FWI, Set of FWS, and Set of FWI. The SOW representation can be formed using various SP or FP algorithms.

Based on the FGD results as a conceptual validation process on two KK in Informatics Engineering Department, UIN Sunan Gunung Djati Bandung, concluded that the citation network analysis model with structured text representation in this study can be implemented with various SOW and applied science network options. For further research, development can be done over the citation network analysis model with more detailed method and flow. In addition, implementation needs to be done until the formation of citation networks in various media publications journal using the model in this study. The implementation of the citation network model is being a more measurable way to test the model.

## Acknowledgments

This research is funded by Center for Research and Publishing, Institute of Research and Community Service UIN Sunan Gunung Djati Bandung.

## References

[1] A. Kania 2009 *Aplicating CFWS* Algorithm for Document Grouping with Set of Frequent Word Sequence and Qord Frequent Itemset as Text Representation* Bandung Institute Of Technology, Bandung

[2] R. A. Nalistia 2008 *Feature-based Clustering untuk Pengelompokan Dokumen dengan Representasi Teks Terstruktur* Bandung Institute of Technology Bandung

[3] D. S. Maylawati 2015 *Pembangunan Library Pre-processing untuk Text Mining dengan Representasi Himpunan Frequent Word Itemset* Studi Kasus Bahasa Gaul Indonesia Bandung Institute of Technology Bandung

[4] D. S. Maylawati and P. Saptawati 2016 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang in *Accepted and Presented in 1st International Conference on Computing and Applied Informatics (ICCAI)* Medan

[5] R. Agrawal 1999 Mapping Words, Phrases using Sequential Pattern to Find User Specific Trends in a Text Database - Patent Document *International Business Machine Corporation*

[6] R. Agrawal and R. Srikant 1995 Mining Sequential Pattern *IBM Research Center*

[7] A. Honloor 2011 S*equential Pattern and Temporal Patterns for Text MIning* Graduate Faculty of Rensselaer Polytechnic Intitute, Major Subject: Computer Science, New York

[8] M. A. Ramdhani 2001 *Perancangan Sistem Pendukung Keputusan Kriteria Majemuk Pada Pengambilan Keputusan Kelompok* Bandung Institute of Technology

[9] R. A. Krueger 1988 *Focus Groups: A Practical Guide for Applied Research* United Kingdom: Sage

[10] D. L. Morgan 1988 *Focus Group as Qualitative Research* United Kingdom: Sage

[11] D. W. Stewart and P. N. Shamdasani 1990 *Focus Groups: Theory and Practices* Sage United Kingdom

[12] C. Slamet, A. Rahman, M. A. Ramdhani and W. Darmalaksana 2016 Clustering the Verses of the Holy Qur'an Using K-Means Algorithm," *Journal of Information Technology* **15** 54 5159-5162

[13] G. Sandi, S. H. Supangkat and C. Slamet 2016 Health Risk Prediction for Treatment of Hypertension in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management (CITSM)*, Bandung

[14] H. Ahonen-Myk 2002 Discovery of Frequent Word Sequence in Text *Pattern Detection and Discovery* pp. 180-189

[15] A. Helena 1999 Knowledge Discovery in Document by Extracting Frequent Word Sequence," *Library Trends* **48** 1 160-181

[16] R. Munir 2005 *Matematika Diskrit* 3rd ed. Informatika

[17] K. H. Rosen 2007 *Discrete Mathematics and Application to Computer Science*, 6th ed. Mc Graw-Hill

[18] A.-L. Barabasi 2017 *Network Science* [Online]. Available: http://barabasi.com/networksciencebook/. [Accessed 26 04 2017].

[19] J. S. Baras 2014 A fresh look at network science: Interdependent multigraphs models inspired from statistical physics in *6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*

[20] V. C. F. Lima and C. J. A. Bastos-Filho 2016 An approach based on network science to detect communities in Social Networks in *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*