IOP Conf. Series: Materials Science and Engineering **300** (2018) 012055 doi:10.1088/1757-899X/300/1/012055

Information Retrieval on social network: An Adaptive Proof

M Elveny, R Syah, M Elfida, M K M Nasution*

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

E-mail: mahyuddin@usu.ac.id

Abstract. Information Retrieval has become one of the areas for studying to get the trusty information, with which the recall and precision become the measurement form that represents it. Nevertheless, development in certain scientific fields make it possible to improve the performance of the Information Retrieval. In this case, through social networks whereby the role of social actor degrees plays a role. This is an implication of the query in which co-occurrence becomes an indication of social networks. An adaptive approach we use by involving this query in sequence to a stand-alone query, it has proven the relationship among them.

1. Introduction

Information Retrieval (IR) is a part of computer science that systematically examines the relevance of information required with information sources, which mathematically derives and proves the measurement formulation of the relevance of information resources and information required. It is from models to methods [1]. In information era with such a large and change information source dynamically [2], also continue to grow, the role of IR is crucial for generating trusty information [3]. This is based on the work of the search engine, which logically means that the document ω is relevant to the query q if it means query or $\omega \Rightarrow q$, $\omega \in \Omega$, where Ω is a space information [4].

One way of obtaining information is through the extraction of information or data mining, such as the extraction of social networks from the Web, or the mining of social structure from information sources by involving social network analysis (SNA) [5]. On the other hand, social networks obtained either by extracted (semi)-automatically or manually [6], or social networks either by document or in real terms are present in everyday life [7]. It technologically presents IR methods based on social networks [8]. Because after all the source of information such as the Web has always been a shadow of the actual state of the events [20]. However, little interest has been made to prove the existence of social network links with IR [10]. Therefore, this paper aims to reveal an IR formula adaptively based on social networks.

2. Basic Concept and Problem Definition

To be a conceptual bridge toward the problem definition, we disclose the basic concepts and related works of various literature as follows [11, 12].

Definition 1. A document d consists of the words w_k, $k = 1, \ldots, K$ or contains some vocabularies (sometimes it expressed as tokens) i.e. w_l, $l = 1, \ldots, L$ where $L \leq K$ if every word has a weight $|w_l| = \sum_{i=1}^{k_j} p(w_k) = k_i/L$ where $|w_k| = p(w_k) = 1/K$ and k_i the number of the same words for the word w_k or k_i is the word frequency of w_k.

Definition 2. A set of documents D is a collection of documents $d_i, i = 1, \ldots, I$ arranged in such a way that each document has a weight $|d|$ and every vocabulary or word has a weight $|w|$.

2.1. Information Retrieval

To model IR approach required the standard data as the comparison for results returned by the access method toward information source. Therefore, based on Definition 2, a document-set D_r serves as the comparative standard data for document generated by an access tool (search engine) whereby it depends on the query q (we call it as the evaluation document or D_e). The document comparison result will be stated in the mutual document $D_r \cap D_e$ [4, 13].

Definition 3. Document set D_r is set of documents $D = \{d_i | i = 1, \ldots, I\}$ such that each document has a unique identity $id_i = f(d_i)$ where f is a mapping that collects different addresses of the same documents, or $D_r = \langle id_i, d_i \rangle$.

Definition 4. Evaluation document D_e is a collection of documents that is accessed whereby each document has id uniquely based on the information source.

In general, to estimate the trusty information we use the measurement referred to as recall and precision as follows [14].

Definition 5. Recall (rec) is a measurement to the relevant documents is retrieved by the tool, i.e.

$$rec = \frac{|D_r \cap D_e|}{|D_r|} \tag{1}$$

where D_r is a set of relevant documents and D_e is a set of the retrieved documents, and $|D_r \cap D_e|$ is the size of $D_r \cap D_e$ and $|D_r|$ is the size of D_r.

Definition 6. Precision ($prec$) is a measurement to the retrieved documents that is relevant to real documents based on tool, i.e.

$$prec = \frac{|D_r \cap D_e|}{|D_e|} \tag{2}$$

where D_r is a set of relevant documents and D_e is a set of the retrieved documents, and $|D_r \cap D_e|$ is the size of $D_r \cap D_e$ and $|D_e|$ is the size of D_e.

2.2. Social Network

To develop the social network, we have a set of social actors $A = \{a_i | i = 1, \ldots, I\}$ and we determine the relationship between social actors in pairs based on a set of relation clues $C = \{c_j | j = 1, \ldots, J\}$. Therefore, social network can be defined as follows [15, 16].

Definition 7. A social network (SN) is a graph $G(V, E)$, $V = \{v_i | i = 1, \ldots, I\}$ as a set of vertices in G and $E = \{e_k | k = 1, \ldots, K\}$ as a set of edges in G for representing the social relationships between social actors such that

(i) $\gamma_1 : A \overset{1:1}{\rightarrow} V$

(ii) $\gamma_2 : R \rightarrow E$

where R is a set of relation between social actors, i.e. $r_j = c(a_k, a_l)$, $r_j \in R$, $j = 1, \ldots, J$, $a_k, a_l \in A$. We notify a social network as $\langle < V, E, A, R, C, \gamma_1, \gamma_2 \rangle$.

Definition 8. A graph $G(V, E)$ is a star graph if one of vertex $v_c \in V$ has degree $d > 1$ is more than another vertices in V and other vertices have degree $d(v_i) = 1$, $v_i \neq v_c$, and v_c as center of star graph.

Lemma 1. If degree of social actors is $d(a) > 1$, then a is a center of star graph.
Proof. For some of social actors we have degrees $d(a_1) \geq d(a_2) \geq \ldots \geq d(a_m) > 1$. Based on it, there are m candidate vertices as center, there are $m - 1$ candidate vertices as leafs, and then we build a star graph by a way we eliminate degrees until $d() = 1$ of all vertices that are not center candidates. This method is done so that all $d(a_i) > 1$ alternately will be the center of the star graph.

Theorem 1. The recall and precision as a presentation of IR can be enhanced on the basis of social network if and only if the social network is optimally shaped star graph.

3. An Approach

To be get information we use cognitive structures as an approach in some of implications as follows [4, 1].

Lemma 2. If D_r and D_e each contains uniquely document id that may be the same between two sets, then the same two id are based on the iteration of $id_i \in D_r$ againts $id_j \in D_e$.
Proof. Suppose $id_i \in D_r$ and $id_j \in D_e$. $id_j \in D_e$ is generated based on the $\omega \Rightarrow q$ implication which is true value if the content q is in the Ω, in other case it is false. While D_r contains a set of documents with specified id, and $\omega \Rightarrow id$ is true if $q \Rightarrow id$. Because D_r contains a set of id_j, so $q \Rightarrow id_j$ has to round every $id_i \in D_e$, and looping $id \in D_r$ done to $id \in D_e$.

Lemma 3. If D_r and D_e each contains a document id that might be the same so as to form $D_r \cap D_e$, looping $id_i \in D_r$ against $id_j \in D_e$ generates a sequance number of D_e.
Proof. Based on the assumptions and consequences of the Lemma 1 and Lemma 2. Suppose $id_j \in D_e$ with $id_1, id_2, \ldots, id_m$, $j = 1, \ldots, m$, as a result sequence of $\omega \Rightarrow q$. Therefore, by doing iteration $id_i \in D_r$ against one by one against from $id_j \in D_e$ from the sequence 1 to m.

Proposition 1. If $D_r \cap D_e$ contains a sequence of id, then the value of $D_r \cap D_e$ is j related to id_j.
Proof. Based on the results of the Lemma 3, the size of $|D_r \cap D_e|$ is smaller than or equal to m, i.e. $1 \leq |D_r \cap D_e| \leq m$.

Proposition 2. If $D_r \cap D_e$ contains a sequence of id, then the value of D_e is the last number of iteration towards $id_j \in D_e$.
Proof. Based on the results of the Lemma 3, the size of $|D_e|$ is smaller than or equal to m, i.e. $1 \leq |D_e| \leq m$, $|D_e|$ is the last number of iteration towards D_e.

By involving the above systematic: Lemma 1, Lemma 2, Lemma 3, Proposition 1 and Proposition 2, we disclose the following ordinances:

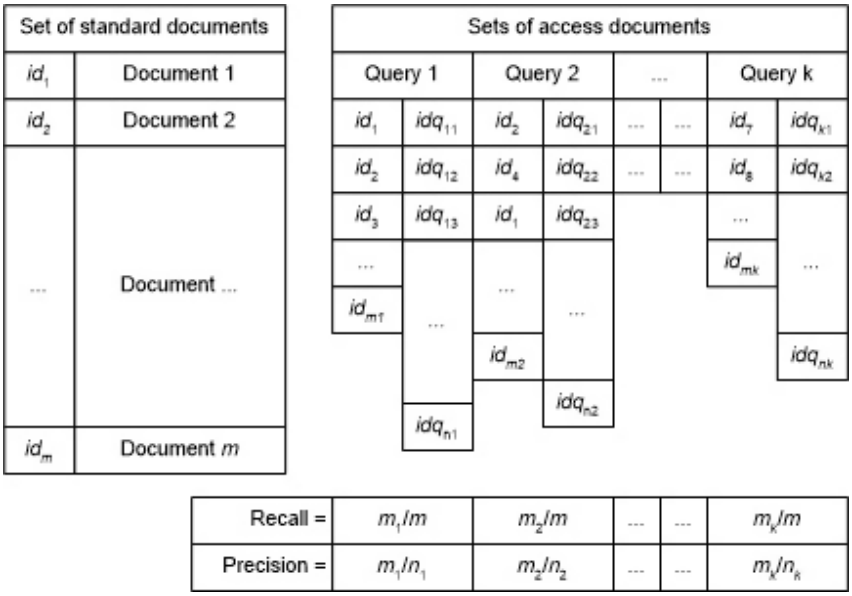

Figure 1. The recall and precision with the standard and the access documents.

Algorithm 1:
INPUT : A set of id_i from D_r
OUTPUT : $|D_r|$, $|D_e|$, $|D_r \cap D_e|$
STEPS :
(1) $id_j \leftarrow (\omega \Rightarrow q)$
(2) Set $j = 1$ to J:
Set $i = 1$ to I:
if $id_j = id_i$ then:
(a) $n = j$
(b) Collect id_j into $D_r \cap D_e$.
(3) $|D_r| \leftarrow J$, $|D_e| \leftarrow I$, $|D_r \cap D_e| \leftarrow n$.

4. Adaptive Proof

Later in this paper, we reveal the interpretive outlines involving the above approach to prove adaptively Theorem 1.

Assuming that each query stands alone, based on Algorithm 1 the recall and precision calculations can be expressed as Fig. 1, although the query contains the co-occurrence of two names of the social actor [17]. However, since any query involving co-occurrence becomes a clue of the relationship between two social actors, so it is possible that one of the social actor names is the same actor on each query [18]. Thus, the queries produce social networks that are generally the form of star graph with one social actor as the center [19].

Suppose that there is a query sequence $q_1, \ldots, q_k$ whereby each query contains $q_1 \leftarrow a_1, a_2$; $q_2 \leftarrow a_1, a_3$; $\ldots$; $q_k \leftarrow a_1, a_k$. So we get a list of $D_{e_1}, D_{e_2}, \ldots, D_{e_k}$, or a sequence of $D_r \cap D_{e_1}, D_r \cap D_{e_2}, \ldots, D_r \cap D_{e_k}$, and a set of documents is a collection of evaluation documents as follows

$$D_{es} = \cup_{l=1}^{k} D_{e_l}, \tag{3}$$

whereby $|D_{es}| \leq |D_{e_1}| + |E_{e_2}| + \ldots + |D_{e_k}|$. Whereas, a collection of documents comes from the comparison between id of 2 sets of documents is $D_r \cap D_{es} = \cup_{l=1}^{k} D_r \cap D_{e_l}$, and based on Eq.

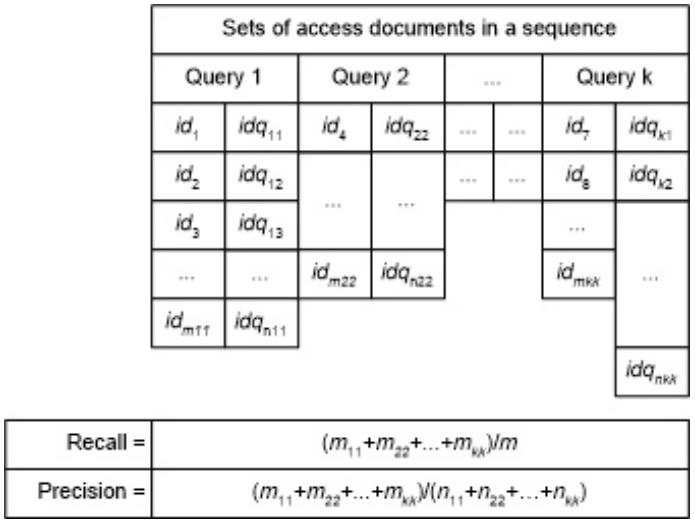

Figure 2. Adaptive approach to the recall and precision.

(3) it be

$$D_r \cap D_{es} = D_r \cap \cup_{l=1}^{k} D_{e_l} \tag{4}$$

whereby $|D_r \cap D_{es}| \leq |D_r \cap D_{e_1}| + |D_r \cap D_{e_2}| + \cdots + |D_r \cap D_{e_k}|$. In general, in Eqs. (1) and (2), the value of $|D_{e_1}| + |D_{e_2}| + \cdots + |D_{e_k}|$ and the value of $|D_r \cap D_{e_1}| + |D_r \cap D_{e_2}| + \cdots + |D_r \cap D_{e_k}|$ each has been reduced to $|\cup_{l=1}^{k} D_{e_l}|$ and $|D_r \cup_{l=1}^{k} D_{e_l}|$. This reduction as a result of merging the set of documents where the same documents to be listed once so that the value of $|D_{es}|$ close to the value of $|D_r|$ or $|D_r \cap D_{es}| \leq |D_r|$, but the number of documents in D_{es} has potential to exceed the number of documents in D_r. This causes a low precision value even if recall value is high. Taking into account that the keyword can reduce unsuitable documents, each query with a form of co-occurrence (one of the social actor names being the keyword for the other) lifts the appropriate document up to the surface [20]. Randomly assigned queries such that every D_{e_l}, $l \neq k$, $|D_{e_l}|$ has the highest value in accordance with $|D_r \cap D_{e_l}|$ in sequence, where in the next sequence the value of $|D_{e_l}|$ does not come from the same document in the previous query, while in the last query or for $|D_{e_k}|$ involves all possible documents, see Fig. 2. Thus by involving the same query as Fig. 1, taking into account the precise measurements consecutively. The query results, except the last query, are considered only to the extent that the last document is appropriate. In other words, if Eqs. (1) and (2) are restated based on Eqs. (3) and (4) as follows [1]

$$rec = \frac{|D_r \cap \cup_{l=1}^{k} D_{e_l}|}{|D_r|} \tag{5}$$

and

$$prec = \frac{|D_r \cap \cup_{l=1}^{k} D_{e_l}|}{|\cup_{l=1}^{k} D_{e_l}|} = \frac{|D_r \cap \cup_{l=1}^{k} D_{e_l}|}{|D_r \cap \cup_{l=1}^{k-1} D_{e_l}| + |D_{e_k}|} \tag{6}$$

As the implementation of the Eqs. (5) and (6) can be seen in Fig. 2. Theorem is proven.

5. Conclusion

The involvement of social actors can be used to improve the performance of recall and precision through effective approaches. An effective approach is made to the use of queries in sequence via a stand-alone query. However, implementation needs to be done by involving data and search

engines, in addition to providing more definitive proof of the relation existence between the extracted social networks and IR.

References

[1] M K M Nasution, R Syah and M Elfida 2018 Information retrieval based on the extracted social network *Applied Computational Intelligence and Mathematical Methods*, Advances in Intelligent Systems and Computing **662**.
[2] M K M Nasution, M Elveny, R Syah, and S A Noah 2015 Behaviour of the resoures in the growth of social network *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015, 7352551.*
[3] L Kirchhoff, K Stanoevska-Slabeva, T Nicolai, and M Fleck 2008 Using social network analysis to enhance information retrieval systems. *Social Networks Applicatons Conference.*
[4] M K M Nasution and S A Noah 2012 Information retrieval model: A social network extraction perspective *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management*, (CAMP'12), 6204999.
[5] M K M Nasution 2016 Social network mining: A definition of relation between the resources and SNA, *International Journal on Advanced Science, Engineering and Information Technology* **6(6)**.
[6] M Hamasaki, Y Matsuo, K Ishida, T Hope, T Nishimura, and H Takeda 2006 An integrated method for social network extraction *Proceeding of the 15th International Conference on World Wide Web* (WWW 2006).
[7] F Benhawi, N M Ali and H M Judi 2012 User engagement attributes and levels in facebook *Journal of Theoretical and Applied Information Technology* **41(1)**.
[8] P Mika 2007 *Social Networks and the Semantic Web* Springer-Verlag: Berlin.
[9] 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)* **8397** LNAI (PART 1).
[10] M K M Nasution, and O S Sitompul 2017 Enhancing extraction method for aggregating strength relation between social actors *Advances in Intelligent Systems and Computing* **573**.
[11] M K M Nasution, and S A Noah 2011 Extraction of academic social network from online database *2011 International Conference on Semantic Technology and Information Retrieval* (STAIR), 5995766.
[12] 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), (arXiv:1601.02904v1 [cs.IR] 12 Jan 2016).
[13] M K M Nasution 2015 Extracting keyword for disambiguating name based on the overlap principle *International Conference on Information Technology and Engineering Application* (4-th ICIBA), Book 1. (arXiv: 1602.00104v1 [cs:IR] 30 Jan 2016).
[14] W B Croft, D Metzler, T Strohman 2010 *Search Engines Information Retrieval in Practice* (New York: Addison Wesley).
[15] M K M Nasution 2016 Social network mining (SNM): A definition of relation between the resources and SNA *International Journal on Advanced Science Engineering Information Technology* **6(6)**.
[16] M K M Nasution, M Hardi, R Syah 2017 Mining of the social network extraction *Journal of Physics: Conference Series* **801 (1)**.
[17] M K M Nasution 2012 Simple search engine model: Adaptive properties for doubleton *Cornell University Library* (arXiv:1212.4702v1 [cs.IR] 19 Dec 2012).
[18] M K M Nasution 2013 Simple search engine model: Selective properties *Cornell University Library* (arXiv:1303.3964v1 [cs.IR] 16 Mar 2013).
[19] M K M Nasution, O S Sitompul, E P Sinulingga, and S A Noah 2016 An extracted social network mining *Proceedings of 2016 SAI Computing Conference* (SAI).
[20] M K M Nasution 2014 New method for extracting keyword for the social actor *Source of the Document Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8397** LNAI (PART 1).