<table>
<thead>
<tr>
<th>Source</th>
<th>Title</th>
<th>Matches</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4]</td>
<td><a href="https://www.researchgate.net/publication...nformation_Retrieval">https://www.researchgate.net/publication...nformation_Retrieval</a></td>
<td>18</td>
<td>8.4%</td>
</tr>
<tr>
<td>[15]</td>
<td><a href="https://www.researchgate.net/publication...e_of_negative_issues">https://www.researchgate.net/publication...e_of_negative_issues</a></td>
<td>7</td>
<td>4.2%</td>
</tr>
<tr>
<td>[28]</td>
<td><a href="https://arxiv.org/pdf/1602.00104.pdf">https://arxiv.org/pdf/1602.00104.pdf</a></td>
<td>3</td>
<td>2.4%</td>
</tr>
<tr>
<td>[33]</td>
<td><a href="https://www.semanticscholar.org/paper/Ex...a74b74012814c5508e20">https://www.semanticscholar.org/paper/Ex...a74b74012814c5508e20</a></td>
<td>6</td>
<td>1.7%</td>
</tr>
<tr>
<td>[34]</td>
<td><a href="https://www.researchgate.net/publication...raction_perspective">https://www.researchgate.net/publication...raction_perspective</a></td>
<td>4</td>
<td>1.1%</td>
</tr>
<tr>
<td>[37]</td>
<td><a href="https://www.springer.com/gp/book/9783319676203">https://www.springer.com/gp/book/9783319676203</a></td>
<td>2</td>
<td>0.9%</td>
</tr>
<tr>
<td>[39]</td>
<td><a href="https://www.coursehero.com/file/p1q7ubi...an-coordinates-and/">https://www.coursehero.com/file/p1q7ubi...an-coordinates-and/</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>[40]</td>
<td><a href="https://www.coursehero.com/file/p6fa5am...etween-the-Phillips/">https://www.coursehero.com/file/p6fa5am...etween-the-Phillips/</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>[41]</td>
<td><a href="https://timesquare.inria.fr/simple-relation-example/">https://timesquare.inria.fr/simple-relation-example/</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>[42]</td>
<td>cdn-pubs.acs.org/doi/suppl/10.1021/acsc...ic01699_si_001.pdf</td>
<td>1</td>
<td>0.9%</td>
</tr>
<tr>
<td>[43]</td>
<td><a href="https://en.mimi.hu/finance/leakages.html">https://en.mimi.hu/finance/leakages.html</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>[44]</td>
<td><a href="https://scirate.com/arxiv/1506.07975">https://scirate.com/arxiv/1506.07975</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>[45]</td>
<td><a href="http://www.amosweb.com/cgi-bin/awb_nav.pl/?s=wpd&amp;c=average_product">www.amosweb.com/cgi-bin/awb_nav.pl/?s=wpd&amp;c=average_product</a></td>
<td>1</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

7 pages, 2214 words

PlagLevel: selected / overall
133 matches from 50 sources, of which 15 are online sources.

Settings
- Data policy: Compare with web sources, Check against my documents, Check against my documents in the organization repository, Check against organization repository, Check against the Plagiarism Prevention Pool
- Sensitivity: Medium
- Bibliography: Consider text
- Citation detection: Reduce PlagLevel
- Whitelist: --
Abstract: It is possible that a technology affects other technologies. In this paper, explored the possibility to reveal improved information retrieval performance through the extraction method of social network. Any extracted social network structurally is not a perfect graph so it is possible to build the star social networks as the optimal form of graph, which guides to model the implication of information retrieval: the formulation of recall and the precision, which generally show better performance on average over 90% and 58%, respectively.

Key words: information space, webpages, recall, precision, graph, degree.

1 Introduction

The social network as a resultant of extraction method from Web is a representation of relationship between web pages through any search engine [1]. The resultant based on occurrence and co-occurrence [2]. It depends not only on the logically relevance between the query and the webpage but also the similarity between the query and the information available, i.e. the answer to the required information as accurately as possible [3]. The latter case has become the concentration of information retrieval that is knowledge technology that focuses on the effectiveness and efficiency for retrieving information from information space like Web [4].

The method of social network extraction from the Web not only gives birth to social structure, but reveals the need for that information to be trusted [5]. Therefore, as a technology an extraction method of social networks must be equipped with evaluation tools [6]. In other words, when the method produces a resultant, another methods needs to be expressed from that result. This paper aims to reveal an IR model based on analysis of the extracted social network.
2 Problem Definition

Completeness of social networks is generally expressed in graph \( G(V, E) \). The set of vertices \( V = \{v_i | i = 1, \ldots , I\} \) as visual representations of a set of social actors \( A = \{a_i | i = 1, \ldots , I\} \); \( a_i \) are the names of possible social actors from information space \( \Omega \). hit count \( |\Omega_{a_j}| \) as the cardinality of \( a_j \). The set of edges \( E = \{e_i | i = 1, \ldots , J\} \) as a visual representation of relation between two social actors \( a_i, a_j \in A \) in the information space \( \Omega \), hit counts \( |\Omega_{a_i} \cap \Omega_{a_j}| \) as the cardinality of \( a_i \) and \( a_j \) [9]. Hit count of two occurrences co-occurrence computationally within similarity distance [10], e.g. using

\[
sim_{text}(a_i, a_j) = \frac{2|\Omega_{a_i} \cap \Omega_{a_j}|}{|\Omega_{a_i}| + |\Omega_{a_j}|}
\]  

(1)

is to give weight to each relationship. The weight not only determines the rank of any relation between social actors, but encourages the growth of social networks. Moreover, semantically the weights indicate the proximity of webpages that are not actually interlinked with each other [11]. Thus, social structures indirectly form other structures of documents scattered within the information space, the different structures than links built into webpages or within the server in which the webpage resides.

In general, if a webpage is accessed then another webpage that is interlinked to it will indirectly impacted by the click event. Likewise, if one of social actor’s name becomes the keyword of another social actor’s name in the co-occurrence form, through the extracted social network, then the different webpages although having one of the social actor names will be clustered by search engine into appropriate cluster.

Proposition 1. If a vertex is representation of a social actor in the extracted social network, then a cluster of webpages is representation of a social actor in Web.

Proof. Based on Eq. (1) to get information about a social actor \( a \in A \) from the Web is assigned a query \( q \), and it generates a collection of webpages \( \Omega_{a} = \{\omega_i | k = 1, \ldots , K_{a}\} \) as a composition of information: a hit count \( |\Omega_{a}| \) and list of snippets \( L_{a} \). Formally, \( \Omega_{a} \leftarrow q = a \) or \( L_{a} \leftarrow q = a \). In other words, for all \( v_i \in V \) we have

\[
v_i = q(a_i) = \Omega_{a_i} = \{\omega_i | i = 1, \ldots , m_i\}
\]  

(2)

where \( \Omega_{a_i} \) is a cluster of webpages.

Proposition 2. If an edge is representation of the relation between two social actors in the extracted social network, then a cluster of webpages is representation of the relation in Web.

Proof. Eq. (1) have shown for getting information of the relation between two social actors \( a_i, a_j \in A \) from the Web is assigned a query \( q \), and it produces a collection of webpages \( \Omega_{a_i} = \{\omega_i | k = 1, \ldots , K_{a_i}\} \) and \( \Omega_{a_j} = \{\omega_j | k = 1, \ldots , K_{a_j}\} \) as a
Information Retrieval Based on the Extracted Social Network

Compotision of information: a hit count \( |\Omega_a \cap \Omega_{a_i}| \) and list of snippets \( L_{a_i(a_j)} \).

In other words, for all \( e_i \in E \) we have

\[
e_i = q(a_{a_i}) = \Omega_{a_i} = \Omega_a \cap \Omega_{a_i} = \{\omega_k | k = 1, \ldots, m\}
\]

where \( \Omega_{a_i} \) is a cluster of webpages.

The cluster of webpages in either occurrence or co-occurrence consists of webpages in general conical to intersection between a cluster based on \( a_i \) and a cluster based on \( a_j \). Therefore, to obtain a reliable technology it is necessary to assess the performance of social network extraction methods. Assessments done on the side of getting trustworthy information and getting side of technology associated with it. Assume \( D_r \) is the set of documents relevant to the query and \( D_c \) is an accomplished documents, then the commonly used the assessment measures are recall \( \text{Rec} \) and precision \( \text{Pr} \) as follows.

1. To measure the permissibility (approximate ability) of reaching approach to all relevant documents based on query,

\[
\text{Rec} = \frac{|D_r \cap D_c|}{|D_r|}
\] (4)

2. To measure the ability of the approach to achieve only documents relevant to the query and may reject irrelevant documents,

\[
\text{Pr} = \frac{|D_r \cap D_c|}{|D_c|}
\] (5)

Theorem 1. If a social network is representation of social structure, then resultant of method for extracting social network from Webis representation of webpages structure.

3 An Approach

If the social network extracted from the information source consists of \( n \) social actors, then in general the social structure is formed can be expressed through the degree of social actors. In this case, the social network theoretically consist of two categories as defined below.

Definition 1. A social network is completely shaped or abbreviated a complete social network consists \( n \) vertices and \( n(n - 1) \) edges. Hereinafter it denoted as \( SN(n, n(n - 1)) \).

Definition 2. A star-shaped social network or abbreviated a star social network consists \( n \) vertices and \( n - 1 \) edges. Hereinafter it denoted as \( SN(n, n - 1) \).

The comparison between the complete social network and the star social network based on degree of all vertices is to reveal that the extracted social network will be between two categories, see Fig. 1. Thus, the predictably extracted social network may consist of parts, i.e. the sub of social network are composed of social actors as the center and the social actors who become the leaf.
Lemma 1. If social actor has the highest degree in social network, then the social actor become candidate of center in social network.

Proof. Suppose there are m social actors within social networks within each degree can be sorted as follow \( d(v_1) \geq d(v_2) \geq \ldots \geq d(v_m) \). It is clear that the social actor with the highest degree \( d(v_1) \) is center candidate in the social network.

Proposition 3. If there is more than one social actor is of the highest degree and likely to have multiple leaves, the social actors are the candidates of center in subs of social network.

Proof. Based on Lemma 1, if vertices with the highest degree in social network are eliminated but each vertex still has leaves, it is clear that the social actors represented by the vertices become the candidates of center in subs of social network.

In graph theory, a tree is optimal-shaped of graph. The star is one of tree forms.

Lemma 2. If there are vertices with the highest degree in social network, then the subs of social network with social actors as the center are the optimal form of social networks.

Proof. Based on Proposition 3, the subs of social networks have a center of candidate that makes them be a sub social network independently. Suppose that on each sub social network there is \( m_i \) vertices, by eliminating the edges that does not cover the center and leaf vertices, or eliminating the leaf-linking edges, it produces \( m_i - 1 \) edges within the sub social network. A sub social network is with a center has a degree is \( m_i - 1 \) while other vertices are of degree 1.
This social network is denoted by $SN_i(m_i, m_i - 1)$ be subs of star-shaped social network or the star social networks.

Proposition 4. If there are the star social networks in a social network, then the collection of information spaces about social actors is conjoined into the social actors’ information space as the center of star social networks.

Proof. In formally, $SN_i(m_i, m_i - 1) \subset SN(n, n(n-1))$ are star social networks, and $m_i \leq n, i = 1, \ldots, I$. Based on Lemma 2 and Proposition 2, for each star social networks there are $m_i - 1 \Omega a \cap \Omega a j - 1$, where $a \cap a j - 1 = \{\omega k | k = 1, \ldots, m l i - 1\}$. In other words, for vertex as center of the star social network, we have $m_i - 1$ edges or number of $|\Omega a \cap \Omega a j - 1|$ is $m_i - 1$, and

$$\prod_{j=2}^{m_i-1} |\Omega a \cap \Omega a j - 1| = \prod_{j=2}^{m_i-1} \Omega a \cap \Omega a j - 1 = \Omega a \quad (6)$$

or

$$\prod_{j=1}^{m_l - 1} \Omega a \cap \Omega a j - 1 = \Omega a \quad (7)$$

So based on Eq. 1, whereby Proposition 1, 2 and Proposition 4 take the appropriate role to structure the webpages, then Theorem 1 is proved.

Table 1: Statistic of Dataset $D_i$

<table>
<thead>
<tr>
<th>Personal Name</th>
<th>Position</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdul RaZak Hamdan</td>
<td>Professor</td>
<td>103</td>
</tr>
<tr>
<td>Abdullah Mohd Zin</td>
<td>Professor</td>
<td>105</td>
</tr>
<tr>
<td>Shahrul Azman Mohd Noah</td>
<td>Professor</td>
<td>160</td>
</tr>
<tr>
<td>Tengku Mohd Tengku Sembok</td>
<td>Professor</td>
<td>210</td>
</tr>
<tr>
<td>Md Jan Nordin</td>
<td>Professor</td>
<td>70</td>
</tr>
</tbody>
</table>

4  Information Retrieval

As the application of Theorem 1, we modelled the structure of webpages for information retrieval in social network extraction perspective based on Eq. 4 and Eq. 5. The documents set $D_i$ is modeled as a collection of all the collected webpages based on the academic actor (professor), in which each document has the unique identity that is URL address. From a collection of academic actors, there are the star social networks in which each professor as a center and an other academic actors (who is not a professor) as a leaf, and each query is built from a pair of such social actors. The documents $D_i$ was obtained based on the query. For evaluation of the approaches, we have gathered and labeled a dataset.
of 648 webpages like Table 1. By using concept the name disambiguation, Eq. 4 and Eq. 5 produce Table 2.

In general, there is an increase in the performance of the access process by means of involving social networks whereby the webpages of influential social actors will be dug up by other social actors as keywords rather than keywords that are not the names of social actors. Thus, the formulation of recall and precision based on Proposition 4 are as follows:

\[ \text{Rec}_a = \frac{|D_r \cap S_m|}{|D_r|} \]  

(8)

and

\[ \text{Pres}_a = \frac{|D_r \cap S_m|}{|S_m|} \]  

(9)

where \( m \) is degree of center in star social network.

5 Conclusion and Future Work

By studying the principle of social network extraction, from the extraction method, there is a change in formulation on the implication of information retrieval that is recall and precision, which generally have better performance based on computation. Furthermore, to test this formula will be built larger datasets.

References


Information Retrieval Based on the Extracted Social Network


