Analysis of Hubs and Authorities Centrality Using Probabilistic Affinity Index (PAI) on directed-weighted graph in Social Network Analysis

Muhammad Thomy Farhan\textsuperscript{1}, Eko Darwiyanto\textsuperscript{2}, Ibnu Asror\textsuperscript{3}
\textsuperscript{1,2,3}School of Computing, Telkom University, Bandung, Indonesia
E-mail: \textsuperscript{1}thomyfarhan@student.telkomuniversity.ac.id, \textsuperscript{2}ekodarwiyanto@telkomuniversity.ac.id, \textsuperscript{3}iasror@telkomuniversity.ac.id

Abstract. Social media is a place for interaction that is connected to the internet network. Twitter is one of the most popular social media. In Twitter sometimes someone does not want to be left behind information related to a particular topic, so it is necessary to follow the user related to the topic so that the information is conveyed quickly. In this study, an analysis was carried out that applied the Hubs and Authorities Centrality method to determine user rankings and the Probabilistic Affinity Index method for weighting values. The results of authority centrality ranking can be used as a list of recommendations of a user who plays a role or has information about a particular topic and the results of centrality hub ranking can be used as a list of recommendations of a user who has an interest in a particular topic. From the testing in this study, changes in the number of other users that are related to the user have the largest average change in centrality value of 0.01188. While the change in the number of relations has the largest average change in the centrality value of $1.44087 \times 10^{-9}$. Based on these tests, the number of other users that are related to the user has a large influence on the results of ranking compared to the number of relationships to other users.

Keywords : Social Media, Twitter, Hubs and Authorities Centrality, Probabilistic Affinity Index, authority centrality, hub centrality

1. Introduction
Interaction between individuals can form a social network, the rapid development of social networks makes an interaction easy to do. Twitter social media is one of the ways to shape social networks. In Twitter, sometimes someone does not want to be left behind information about a particular topic, so it requires a lot of user references so that information can be obtained quickly. Therefore, a recommended list of users is needed according to the role and interest of the user on a particular topic. In this study, user ranking can be used as a user recommendation to follow, there are two ranking groups, namely the user who has wide observation of a topic and the user who has the role or has information about the topic. In social networks, users or individuals can be called nodes.

The technique for analyzing social networks is called Social Network Analysis (SNA), one of the most commonly used SNA measurements is Centrality Measurement \[5\] which consists of Degree centrality, Betweenness, Closeness, Eigenvector and others. In this study, the authors determined the most influential nodes in a network from one of the social media namely Twitter
using the Hubs and Authorities Centrality method. And applying the Probabilistic Affinity Index (PAI) method in weighting to obtain relevant results and in accordance with the actual conditions, PAI measures the ratio of the number of relations between observed value and expected value [8].

Previous researches that have carried out research related to Social Network Analysis are Pushpa Kumar and Kang Zhang where the research shows the difference in the results of centrality values by using three methods, namely Degree, Betweenness and Closeness Centrality, based on the research that the most active nodes do not always have high ranking of three Centrality calculations that have been done [6]. In another study [4] experiments conducted with the Hubs and Authorities Centrality method with a website network case study, from the research it can be concluded that the value of a good Authority will be obtained if it has a relationship with a good Hub value and vice versa.

Yasuhiro Yamashita and Yoshiko Okubo conducted a research on the use of Probabilistic Affinity Index (PAI) with a research collaboration journal between Japan and France in 24 years as data material [8], the results of this study are that each node's PAI value is affected by the total relations that occur outside the node.

2. Implementation

Overall, the system built is to provide user recommendations based on the results of ranking from the hub centrality and authority centrality using the Probabilistic Affinity Index. The results of the user's recommendations depend on certain topics and specific times on Twitter. Hub centrality and authority centrality values are calculations based on the number of relations that occur between nodes, there are only two relations in Twitter that are applied in this study namely mention and reply. After obtaining the results of ranking, data validation using a questionnaire whether the respondent agrees with the accuracy of the results of user recommendations on the topic.

![Diagram](image-url)

**Figure 1.** Description System

In this study, the dataset used is a list of users who tweeted with the hashtag #AsianGames2018 to other users and a list of relationships made between users. In the dataset
there are 1176 number of nodes, number of mention relations as many as 1590 and number of reply relations as many as 23, taken from 4 to 5 August 2018. This research process is shown in Figure 1.

3. Preprocessing
In preprocessing, the relation value is weighted into the matrix, where the system will calculate every relation that occurs between nodes. The weighted graph concept is applied in this weighting, where the number of relation weights from a node 'A' to node 'B' with node 'B' to node 'A' will not be the same ($W_{AB} \neq W_{BA}$). For example, node 'A' makes a reply relation two times to node 'B' then $W_{AB} = 2$ and if node 'B' makes a mention relation three times to node 'A' then $W_{BA} = 3$.

Table 1. Example of a Matrix

|   | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 0 | 1 | 0 |
| B | 1 | 0 | 3 | 0 |
| C | 2 | 0 | 0 | 0 |
| D | 0 | 0 | 2 | 0 |

Table 1 shows an example of a matrix $n \times n$, there are four nodes and each node has the value of the number of relations with other nodes. Indegree is the number of relations received by the node, while outdegree is the number of relations initialized by the node for example node 'C' has 2 outdegree and 6 indegree. Node In is the number of nodes pointing to a node, while Node Out is the number of nodes designated by a node for example node 'C' has 1 Node Out and 3 Node In.

4. Probabilistic Affinity Index
After the data is completed in the preprocessing stage, the weight value in the matrix will be processed in the calculation and manufacturing stage of the PAI matrix. In this process, the observed value and the expected value of each weight value will be searched and the ratio of the two values will be searched. Formula 1 is the calculation of PAI [8].

\[
PAI_i = \frac{n_{ij}}{E[n_{ij}]} = \frac{n_{..} \times n_{ij}}{n_i \times n_j} 
\]

$n_{ij}$ value is the observed value which is the number of relations that occur on a network, while $E[n_{ij}]$ is the expected value which is the total condition of the relation of the node. The variables $n_i$ and $n_j$ represent the number of nodes that are related to the node 'i' and the node 'j', $n_{..}$ is the value that contains the size of the matrix.

5. Hubs and Authorities Centrality
After the PAI matrix is obtained, hubs centrality and authority centrality are calculated to get the results of the ranking. Hub value is the centrality value of a node in its ability to make a relation with other nodes, while authority value is the centrality value of a node based on the number of relations to the node. Formula 2 is a hub centrality calculation and Formula 3 is an authority centrality calculation [7].

\[
AA^Ty = \lambda y 
\]

\[
A^Tx = \lambda x 
\]
Variable $A$ is matrix $n \times n$ and variable $A^T$ is the transpose of matrix $A$. The variable $\lambda$ is the largest eigenvalue of the matrix $A$ and $y$ is a vector that holds the value of the hub while $x$ is a vector that holds the value of authority.

6. Evaluation

The Testing Phase is divided into three scenarios where the results of the ranking and changes in centrality values will be analyzed, each scenario has its own objectives. In the first scenario the weighting is not modified or using a pure dataset, while in the second and third scenarios an additional weight is added to see the effect. After these three scenarios were tested, validation of the results of the first scenario was made by making a questionnaire. After these three scenarios were tested, the results of the first scenario were validated by making a questionnaire.

6.1. First Scenario Testing

In this test, the comparison of the results of ranking in hub centrality and authority centrality by applying PAI or without PAI is analyzed which aims to determine the cause of the ranking order differences.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
No. & Node & Hub Centrality & Node Out & Outdegree \\
\hline
1 & sushimaccheese & 0.352488488 & 1 & 8 \\
2 & destikaambar & 0.265204957 & 2 & 7 \\
3 & ilovecenon & 0.264366366 & 1 & 6 \\
4 & memethachai & 0.176244244 & 1 & 4 \\
5 & isminauldyati & 0.176244244 & 1 & 4 \\
6 & jonachlers & 0.176244244 & 1 & 4 \\
7 & aminah2 & 0.138937451 & 2 & 6 \\
8 & citradiania & 0.132183183 & 1 & 3 \\
9 & juakhoos2010 & 0.04757027 & 10 & 11 \\
10 & martavye & 0.000499958 & 9 & 12 \\
\hline
\end{tabular}
\caption{Hub Centrality Results without using PAI}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
No. & Node & Hub Centrality & Node Out & Outdegree \\
\hline
1 & juakhoos2010 & 0.074411906 & 10 & 11 \\
2 & lalalaaaf39 & 0.07065396 & 3 & 5 \\
3 & aminahc2 & 0.070305126 & 2 & 6 \\
4 & mulyawanagr & 0.070305126 & 2 & 3 \\
5 & andikagd101 & 0.070305126 & 2 & 3 \\
6 & ayubioscela & 0.070305126 & 2 & 3 \\
7 & bimoerpras & 0.070305126 & 2 & 3 \\
8 & serha77 & 0.070305126 & 2 & 2 \\
9 & sushimaccheese & 0.067330561 & 1 & 8 \\
10 & martavy & 0.0001226467 & 9 & 12 \\
\hline
\end{tabular}
\caption{Hub Centrality Results using PAI}
\end{table}

Based on Table 2, there are several nodes with Hub Centrality, Node Out and Outdegree values sorted from the highest Hub Centrality values. In testing without using PAI, the node 'sushimaccheese' has the highest centrality value of 0.3524. While in testing using PAI, the node 'sellkaoos2010' has the highest centrality value of 0.0744. Both of these values have a far difference because the measurement without PAI the number of outdegree to the node that has a high ranking has a strong influence to determine the centrality value. While the measurement with PAI the number of Node Out to nodes that have a high ranking has a strong influence. This reason also makes a list of the highest node sequences to be significantly different between the use of PAI and not. Although the node 'martavye' has a large Node Out and outdegree value but has a small hub value, this is because the node that has a relation with the node 'martavye' does not have a high authority value.
Table 3. Authority Centrality Results

(a) Authority Centrality results without using PAI

| No. | Node         | Authority Centrality | Node In | Indegree |
|-----|--------------|-----------------------|---------|----------|
| 1   | inabadminton | 0.998340514           | 219     | 280      |
| 2   | antoagustin  | 0.051012989           | 36      | 57       |
| 3   | asiangames2018 | 0.019000892         | 64      | 73       |
| 4   | badminton2017 | 0.01412427           | 12      | 13       |
| 5   | tniu         | 0.004745219           | 19      | 21       |
| 6   | footballina_eng | 0.004373252      | 6       | 7        |
| 7   | mbahuyok     | 0.004143381           | 31      | 31       |
| 8   | jaktvcom     | 0.003896799           | 1       | 1        |

(b) Authority Centrality results using PAI

| No. | Node         | Authority Centrality | Node In | Indegree |
|-----|--------------|-----------------------|---------|----------|
| 1   | inabadminton | 0.998913911           | 219     | 280      |
| 2   | antoagustin  | 0.044909079           | 36      | 57       |
| 3   | badminton2017 | 0.029041402         | 12      | 13       |
| 4   | asiangames2018 | 0.019764138         | 64      | 73       |
| 5   | mbahuyok     | 0.016064846           | 31      | 31       |
| 6   | yamur266210  | 0.017067926           | 83      | 104      |
| 7   | tniu         | 0.007719989           | 19      | 21       |
| 8   | tniu         | 0.0076842             | 15      | 15       |

Based on Table 3, there are several nodes with Authority Centrality, Node In and Indegree values sorted from the highest Authority Centrality values. In both tables, there are nodes that have the same order but have different authority values, this is because the differences in ranking in Table 2 affect the values of each node in Table 3.

6.2. Second Scenario Testing

This test aims to see the effect of adding relations to the value of authority centrality using either PAI or without PAI. In this test the mention relation was added to the weight of 1 to 8 which was carried out during the 8 times of testing, this test was only tested for 8 times because the value change pattern was clearly seen.

Table 4. Authority Centrality results without using PAI and number of Relations

| No | Node          | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|----|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1  | inabadminton  | 0.0043486 | 0.9983394 | 0.9983357 | 0.9983342 | 0.9983328 | 0.9983322 | 0.9983322 |
| 2  | antoagustin   | 0.0511054 | 0.05112759 | 0.05112599 | 0.05113136 | 0.05113286 | 0.05113377 | 0.05113436 | 0.05113476 |
| 3  | asiangames2018 | 0.01903462 | 0.01904085 | 0.01904036 | 0.01904048 | 0.01904086 | 0.01904096 | 0.01904157 | 0.01904522 |
| 4  | badminton2017 | 0.01412431 | 0.01412432 | 0.01412432 | 0.01412432 | 0.01412432 | 0.01412433 | 0.01412433 | 0.01412433 |
| 5  | _tniu         | 0.00473348 | 0.00473552 | 0.00475576 | 0.00475589 | 0.00475596 | 0.00475601 | 0.00475605 | 0.00475605 |
| 6  | footballina_eng | 0.0038044 | 0.00438777 | 0.00438223 | 0.00438245 | 0.00438256 | 0.00438263 | 0.00438268 | 0.00438271 |
| 7  | mbahuyok     | 0.0041498 | 0.00415099 | 0.0041514 | 0.0041516 | 0.00415176 | 0.00415181 | 0.00415183 | 0.00415183 |
| 8  | jaktvcom     | 0.003591247 | 0.003590323 | 0.003590389 | 0.003590406 | 0.003590416 | 0.003590421 | 0.003590425 | 0.003590427 |

Based on Table 4, the red color is a decrease in the centrality value, the green color is the increase in the centrality value and there is a number of relations located below the centrality value at each node. Node ‘inabadminton’ has decreased the authority value because the node has a lot of relations with nodes that experience a decrease in hub value, but it is not offset by an increase in the value of the hub on the node that is related to the node ‘inabadminton’. The average value of the biggest authority change is owned by the ‘antoagustin’ node with a value of 3.66456x10^-6. The eigenvalues have an average change in value of 4931.
Table 5. Authority Centrality results using PAI and number of Relations

| No | Node                | Testing 1 | Testing 2 | Testing 3 | Testing 4 | Testing 5 | Testing 6 | Testing 7 | Testing 8 |
|----|---------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1  | inabadminton        | 0.99801391| 0.99801291| 0.99801391| 0.99801391| 0.99801391| 0.99801391| 0.99801391| 0.99801391|
| 2  | antoaugustian       | 0.044091  | 0.044091  | 0.044091  | 0.044091  | 0.044091  | 0.044091  | 0.044091  | 0.044091  |
| 3  | badminton talk      | 0.2950414 | 0.2950414 | 0.2950414 | 0.2950414 | 0.2950414 | 0.2950414 | 0.2950414 | 0.2950414 |
| 4  | asianganes2018      | 0.1976414 | 0.1976414 | 0.1976414 | 0.1976414 | 0.1976414 | 0.1976414 | 0.1976414 | 0.1976414 |
| 5  | mnbahyok            | 0.0106449 | 0.0106449 | 0.0106449 | 0.0106449 | 0.0106449 | 0.0106449 | 0.0106449 | 0.0106449 |
| 6  | yanurr266210        | 0.0107696 | 0.0107697 | 0.0107697 | 0.0107697 | 0.0107697 | 0.0107697 | 0.0107697 | 0.0107697 |
| 7  | _tnia               | 0.00772002| 0.00772002| 0.00772002| 0.00772002| 0.00772002| 0.00772002| 0.00772002| 0.00772002|
| 8  | _trial              | 0.00768423| 0.00768423| 0.00768423| 0.00768423| 0.00768423| 0.00768423| 0.00768423| 0.00768423|

Based on Table 5, the red color is a decrease in the centrality value, the green color is the increase in the centrality value and there is a number of relations located below the centrality value at each node. The 'yanurr266210' node has the highest average change in the authority value which is 1.44087 x 10^{-9} and the eigen value has an average change in value of 1.12618 x 10^{-6}. The increase in relations for each test on node ‘yanurr266210’ is large, but the change in authority on the node is very small and the change in eigenvalues of each test in Table 5 has a far difference compared to Table 4. This proves that the number of relations has a small effect on the results of the centrality value by applying PAI, even though the number of relations owned by each node experiences a large increase. This test is only analyzed for authority centrality because the pattern of changing values in the centrality hub has similarities with authority centrality, so the analysis for this authority has described the analysis for the hub.

6.3. Third Scenario testing

This test aims to see the effect of adding Node In and Node Out values to the centrality value, this test is carried out 8 times. In each test, Node Out from the 8 highest hub nodes in Table 2 (b) added 1 and Node In from the 8 highest authority nodes in Table 3 (b) added 1.

Table 6. Hub Centrality results using PAI and number of Node Out

| No | Node               | Testing 1 | Testing 2 | Testing 3 | Testing 4 | Testing 5 | Testing 6 | Testing 7 | Testing 8 |
|----|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1  | jualkao2010        | 0.0739661 | 0.0819603 | 0.0853084 | 0.0905991 | 0.0952527 | 0.1007817 | 0.1067871 | 0.1134955 |
| 2  | lalazaazaf9        | 0.07396602| 0.07579055| 0.08154776| 0.08587875| 0.09062902| 0.09558326| 0.10160629| 0.10795138|
| 3  | anniahe2           | 0.07361842| 0.0772694 | 0.0816679 | 0.08547891| 0.09029055| 0.09541052| 0.10133914| 0.10743762|
| 4  | mulyawananagaris   | 0.07361839| 0.0772689 | 0.0816672 | 0.08547881| 0.09029041| 0.09541033| 0.10133914| 0.10743734|
| 5  | andikag101          | 0.07361839| 0.0772689 | 0.0816672 | 0.08547881| 0.09029041| 0.09541033| 0.10133914| 0.10743734|
| 6  | azubahjosepha      | 0.07361839| 0.0772689 | 0.0816672 | 0.08547881| 0.09029041| 0.09541033| 0.10133914| 0.10743734|
| 7  | bismoopras         | 0.07361839| 0.0772689 | 0.0816672 | 0.08547881| 0.09029041| 0.09541033| 0.10133914| 0.10743734|
| 8  | aerha77            | 0.07361839| 0.0772689 | 0.0816672 | 0.08547881| 0.09029041| 0.09541033| 0.10133914| 0.10743734|

Based on Table 6, the red color is the decrease in the centrality value, the green color is the
increase in the centrality value and there is a number of Node Out located under the centrality value at each node. All nodes experience an increase in the hub value, this is because the node has more relations with the node that has increased the authority value than the node that has decreased the authority value. The ‘jualkaos2010’ node has the largest average change in the hub value, which is 0.00444 and the eigen value has an average change in value of 0.70013. The addition of the number of Node Out has a significant impact on the centrality value compared to the addition of the number of relations in Table 5, this proves that the change in Node Out affects the centrality value more than the change in relation.

Table 7. Authority Centrality results using PAI and number of Node In

| No | Node   | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|----|--------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1  | inabadminton | 0.09670197  | 0.99465039  | 0.9916655  | 0.98751909  | 0.98195517  | 0.97438328  | 0.96543297  | 0.93935123  |
| 2  | antoaugustine | 0.05061455  | 0.057765715  | 0.06559509  | 0.07412875  | 0.08342698  | 0.09354583  | 0.104463038 | 0.11545705  |
| 3  | badmintonTalk | 0.03642395  | 0.04076837  | 0.0475233  | 0.05493098  | 0.06021111  | 0.07180301  | 0.08125662  | 0.09132392  |
| 4  | asianGames2018 | 0.02795954  | 0.03617902  | 0.04552179  | 0.0558218  | 0.06464466  | 0.07384249  | 0.0917619  | 0.10528134  |
| 5  | nhabuyok | 0.01687256  | 0.02375743  | 0.03136861  | 0.0397006  | 0.04880957  | 0.05869504  | 0.06934651  | 0.08060443  |
| 6  | yanur266210 | 0.01983771  | 0.02989373  | 0.04231213  | 0.0547866  | 0.06539936  | 0.08305263  | 0.09857728  | 0.11498598  |
| 7  | _nianu | 0.01569196  | 0.02443383  | 0.03400569  | 0.04445201  | 0.05580265  | 0.06869516  | 0.08114989  | 0.09499274  |
| 8  | _nianz | 0.01594949  | 0.02426945  | 0.03376791  | 0.04413734  | 0.05543046  | 0.0675541  | 0.08057126  | 0.09431889  |

**EIGENVALUE**  
221.090597  222.639309  224.380884  226.34402  228.560699  231.065695  233.89557  237.087014

Based on Table 7, the red color is the decrease in the centrality value, the green color is the increase in the centrality value and there is a number of Node In located under the centrality value at each node. The ‘yanur266210’ node has the highest average change in the authority value of 0.011188 and the eigen value has an average change in value of 1.9995. The addition of the number of Node In has a greater effect than the addition of the relation in Table 5 to the centrality value. It can be concluded that the addition of Node Out or Node In to a node is enough to affect changes in centrality values with the application of PAI.

6.4. Validation of Ranking Results

(a) Hub Centrality questionnaire

(b) Authority Centrality questionnaire

Figure 2. Match ranking results

The questionnaire is a list of questions to be filled in by the respondent, which is used as a communication tool between the researcher and the respondent. In the study [2], validation
was carried out using a questionnaire to determine whether the recommendation system was better or not based on the user’s subjective evaluation. In this study the results of the questionnaire are used as validation whether the results of centrality ranking can be used as a user recommendation, the higher the validation value, the recommendations are in accordance with what the user wants.

At this stage, questionnaires were given to 30 people. In Figure 2(a) as many as 77 percent of respondents agree that the results of the hub ranking are users who are related to the topic and have extensive observation. In Figure 2(b) as many as 83 percent of respondents agree that the results of the ranking of authority are users who are related and have information about the topic. Both questionnaires get positive results from respondents, this proves that ranking results can be applied as a user recommendation on Twitter.

7. Conclusion

Based on the results of the testing and analysis obtained, the effect of changes in weight using the Probabilistic Affinity Index (PAI) on the results of measurements of Hubs and Authorities Centrality is the change in the centrality value and the change in the highest user order. Based on the results of measurements in this study, changes in the number of nodes that are related have the highest average change in the centrality value of 0.01188. While the change in the number of relations has the highest average change in the centrality value of $1.44087 \times 10^{-9}$.

From the test results, obtained the factors that determine the order of a node in testing Hubs and Authorities Centrality using PAI are sorted from the most important is the state of the nodes that are related, the number of nodes that are related and the number of relations that occur from each node. Based on the results of the questionnaire, the results of ranking authority can be used as recommendations for users who play a role or have information about a particular topic and the results of hub ranking can be used as recommendations for users who have an interest in a particular topic.

References

[1] Anggarwal C C 2011 Social Network Data Analytics (New York: Springer)
[2] Baizal Z K A, Murti Y R and Adiwijaya 2017 Evaluating Functional Requirements-Based Compound Critiquing on Conversational Recommender System 2017 Fifth International Conference on Information and Communication Technology pp 256–268
[3] Bathul S 2010 Special Functions and Complex Variable (Delhi: PHI Learning Private Limited)
[4] Ding C H Q, Zha H, Husbands P and Simon H D 2001 Link Analysis: Hubs and Authorities on The World Wide Web Society for Industrial and Applied Mathematics pp 256–268
[5] Kretschmer H and Kretschmer T 2007 Application of a New Centrality Measure for Social Network Analysis to Bibliometric and Webometric Data 2006 1st International Conference on Digital Information Management pp 199–204
[6] Kumar P and Zhang K 2007 Social Networks Analysis of Online Marketplaces IEEE International Conference on e-Business Engineering pp 363–367
[7] Zhu X 2010 Link Analysis on Graph CS769 Spring 2010 Advanced Natural Language Processing pp 1–2
[8] Yamashita Y and Okubo Y 2005 Patterns of scientific collaboration between Japan and France: Inter-sectoral analysis using probabilistic Partnership Index (PPI) ISSI 2005: Proceedings of The 10th International Conference of The International Society for Scientometrics and Informetrics, Vols 1 and 2 pp 517–526
[9] Ya-ting L and Jing-min C 2011 The Social Network Analysis of Political Blogs in People Based on Centrality International Conference on Consumer Electronics, Communications and Networks pp 5441–5444