Creative Fusion Seminar

Network analysis of KPOP idol twitter hashtags

Team 2 - 김기명(2018-12153), 김지수(2021-15600),
김태윤(2021-15148), 최동혁(2016-12214)

1 Introduction

Twitter is one of the most famous social networking services. About 5.17 million Korean users visited Twitter in January, 2021[1] . It means that formidable texts are floating on the Twitter, and the texts can be treated as a data in the aspect of digital humanity.

'Web Crawling' , or 'Web Scraping' is a technique used for text mining. For instance, assume the situation that we are trying to make a list of texts in the web which includes 'dog' . If we use Web Crawling, then we can easily make the list without searching every text in the search engine like google. This technique allows us to handle millions of texts in short time. 'Twitter Crawling' that we use in this project is a technique to find the texts in the Twitter.

To analyze the group, it is important to first find out the characteristic of the group. We cannot substitute the result of analysis to general groups. One of the significant properties of Twitter users is strong interest in Idol. We postulated that the rate of people who are fan of one of the Idol groups are higher than general people.

In the selection of Idol group, we tried to find the group with proper size of fandom to control the size of database. Idols like 'BTS', or 'Blackpink' have too big size of fandom. If we cut the data to the controllable size, then the result of analysis might b e changed from the first. Selection of Idols with too small fandom leads to decrease in size of database, then the analysis becomes meaningless. Our selection is 'IVE'. The group has proper size of fandom, and had been working recently, so that we can analyze current Twitter users.

Our final goal of this project is to build a database of texts about 'IVE' in twitter and investigate the correlation with the hashtags that are used.

[1] '2021 소셜미디어 시장 및 현황 분석' , DMC미디어

2 Process of building database

We used ‘snscrape’ is one of a python library used for scraping social network services. Using the library, we restricted the location of the post to Korea and searched the texts that were posted during the first working period of ’IVE’, from December 1st, 2021 to January 9th, 2022. We saved the list of hashtags of each posts into csv file.

The csv file only includes the list of concatenated hashtags with each post, and to analyze the hashtags in Gephi(one of a tool for network analysis), pre-processing is required. First, we need to split the concatenated hashtags individually. We decomposed the hashtags into 672 pieces of hashtag. To regulate the size of database, we arranged the hashtags according to the number of each hashtag’s usage. We made a node with 37 hashtags that are used the most.

We also need edge file for analysis in Gephi. Based on the node file, we saved the count how many times target hashtag is posted with source hashtag into weight.

3 Analysis

3.1 Centrality

3.1.1 Betweenness Centrality

Betweenness centrality scores the nodes with the number of times shortest path passing by each node. The 3 nodes with the largest betweenness centrality are ’안유진’, ’아이브’, ’IVE’.

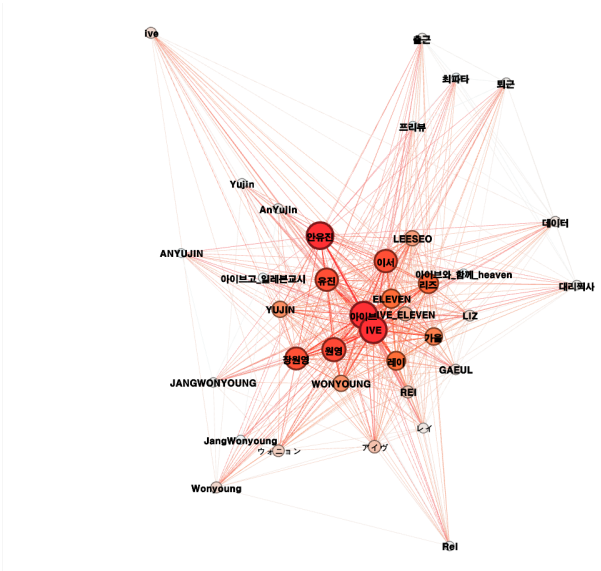

Figure 1: Ranked by betweenness centrality

'아이브', and 'IVE' is a hastag of the name of the group, so it is reasonable that they got the highest score. On the other hand, '안유진' is a node of a specific member's name. We can guess the cause of this aspect into three point of view. First is that she is the most popular member in the group. Second, she undertakes the most important role in the group. Third, she works individually from the group.

In the first point of view, '장원영' is the most popular member in the group. The number of usage in hashtags about '장원영' such as '원영', and '장원영' is 6571, which is much bigger than the number of usage in hashtags about '안유진', 4504. Also, in the third point of view, '장원영' also appeared in many TV programs apart from the work with groups. Since she is a leader of the group and undertaking a role of main vocal in the group, this shows that she is playing an important role in the group.

In addition, the popularity of '안유진', and '장원영' seems to be different a lot. Hashtag '안유진' were used a lot obviously than other hashtags of '안유진'. On the other hand, hashtags '원영', '장원영', 'WONYOUNG' are used evenly. We can guess that there are many celebrities having '유진' in the name, so people used '안유진' hashtag to prevent the confusion with other celebrities. This also leads to increase of '안유진' score in betweenness centrality.

3.1.2 Closeness Centrality

Closeness centrality scores scores each node with division of the number of other nodes into the sum of the distance of shortest path with other nodes. '아이브', 'IVE', '안유진' got the biggest score and '유진', '이서', '리즈', '원영', '장원영' were ranked after them. //

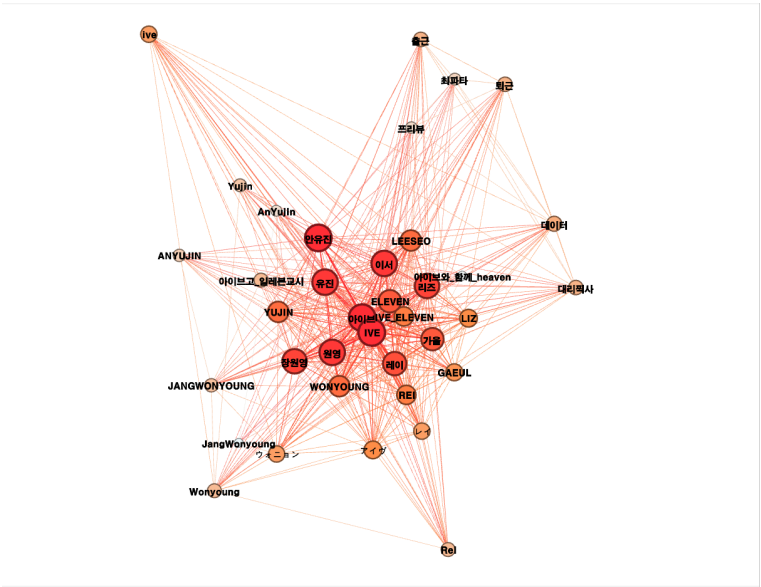

Figure 2: Ranked by closeness centrality

We can guess that the three nodes, ranked at the top got highest score with the same reason

with betweenness centrality. Four nodes followed by can be guessed as score based on the popularity. '원영', '장원영' are the hashtags of same member, and they are just divided with existence or nonexistence of last name. Hashtag '유진' got lower score than '안유진' with the same reason mentioned in betweenness centrality. We found out that the four members with the highest score are mentioned more often than other two members in 'IVE'.

3.1.3 Eigenvector Centrality

Eigenvector centrality scores the nodes based on how many influential nodes are connected the node. '아이브', 'IVE', 'ELEVEN', '유진', '안유진', '원영', '장원영', '이서', 'Leeseo', '리즈', '가을', '레이' are ranked at the top in scores of eigenvector centrality.

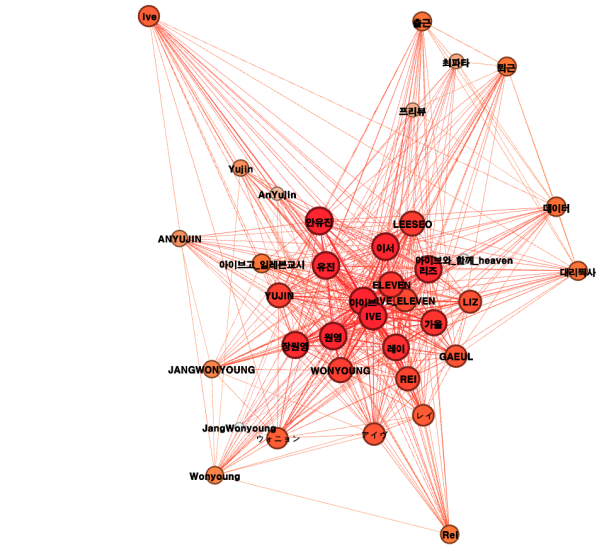

Figure 3: Ranked by eigenvector centrality

The nodes are about name of the team, name of the members, title of the song. They are influential nodes and they are connected with other influential nodes. Among the nodes above, some members are included in two or more nodes and some members are not. It seems to be the difference between popularity of the members.

3.2 Modularity

We visualized the nodes and edges in Gephi with Fruchterman layout. The Fruchterman layout organizes the nodes by communities, clustering nodes that are densely connected to each other. In the case of our network, there doesn' t seem to exist multiples communities. There is just one single community as it can be seen: there is only one cluster.

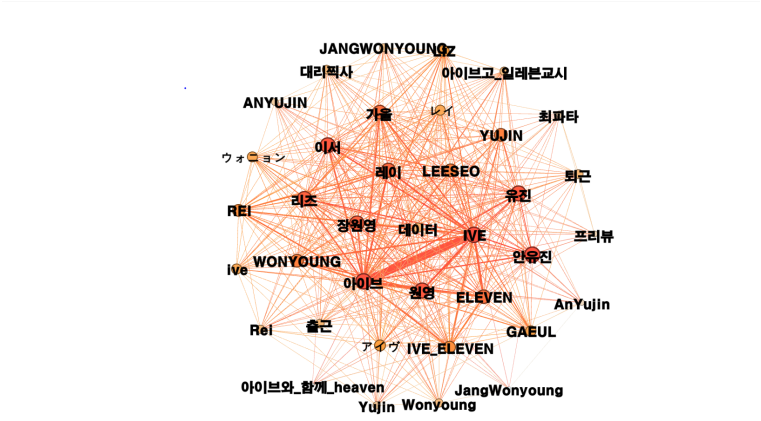

Figure 4: Organized using Fruchterman layout

However, the modularity calculation revealed that there are three communities.

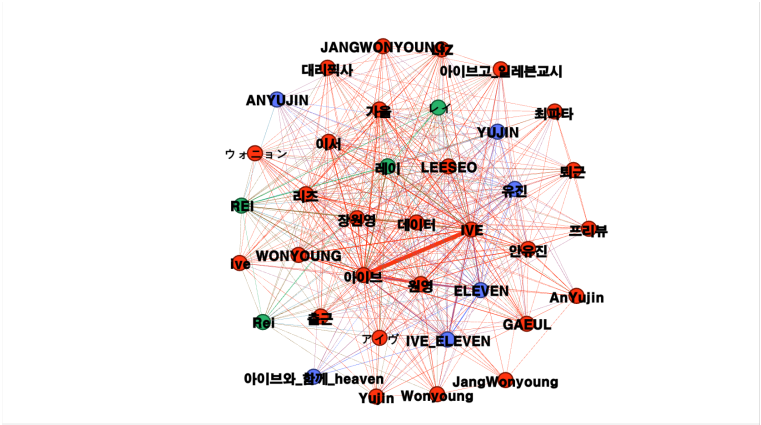

Figure 5: Divided by modularity class

Following table is the list of the nodes belong to each class.

JANGWONYOUNG	LIZ	아이브고_일레븐교시	최파타	퇴근
프리뷰	An Yujin	JangWonyoung	Wonyoung	Yujin
Ive	ウォニョン	대리찍사	가을	이서
리즈	WONYOUNG	출근	アイヴ	GAEUL
안유진	데이터	IVE	원영	아이브
장원영	LEESEO			

Table 1: Class 0 (Red)

ANYUJIN	YUJIN	유진
ELEVEN	IVE_ELEVEN	아이브와_함께_heaven

Table 2: Class 1 (Blue)

레이(Japanese)	레이	REI	Rei

Table 3: Class 2 (Green)

We can analyze this class classification in some points. We can see that class 1 is Yujin's class, and class 2 is Rei's class. Only two members of the team has their own class.

First we can see that 'ELEVEN', 'IVE_ELEVEN', '아이브와_함께_heaven' are included in Yujin' s class. When twitter users talked about Ive' s new song 'Eleven' , they mentioned Yujin more frequently than other members. Or it could be the other way around: When people talked about Yujin, they mentioned 'Eleven' . Whatever the direction, it doesn' t matter because both mean that Yujin was the spotlight of the new song 'Eleven' . If Yujin became the spotlight because of a particular reason, then we would have seen another hashtag within the class. However, that is not the case. So, finding the reason by just looking at the network is limited, and we had to look outside the network. One possible reason we came up with is the fact that Yujin is the main vocal of the team. However, we cannot confidently assert this because we haven' t looked into other networks of other time periods when IVE released a different song. If Yujin also had a separate class with the title of that song included, and that pattern is seen multiple times, then we would be more confident in saying that it is because Yujin is the main vocal of the team. However, we do realize that the fact that she is the main vocal of the team is not the 'only' reason. There must be other reasons which should be studied with other datasets.

Second, Rei has a separate class but she is not included in class 0 which includes the rest of the members' names and other hashtags such as '프리뷰', '데이터', '대리찍사', '최파타', '퇴근', '출근'. Rei is the only member who is not included in class 0. We first looked at whether if Rei' s hashtags had a strong connection to each other. We compared it to Wonyoung' s connection because she had similar number of nodes with that of Rei (6 compared to 4) and her nodes were included in class 0 but she did not have a separate class. We calculated strength by the following equation

$$\textbf{(strength of connection)} = \frac{\text{(Sum of weight of all edges)}}{\text{(Number of all possible edges)}}$$

This calculation method is questionable whether it can fully model "the strength of connection of a group of nodes" but we used this equation for simplicity. Rei, and Wonyoung got 522.33, and 199.33 as the result of the strength. According this this calculation, Rei' s hashtags had stronger connection with each other. Thus, partially explaining why only Rei has a separate class. Now

we have to compare the strength of connection of Rei' s hashtags to the nodes in class 0 and that of Wonyoung' s hashtags to the nodes in class 0. However, due to lack of coding knowledge, we couldn' t compute. We would have to manually compute 4*20 + 6*20 number of edges. The takeaway is that Rei' s hashtags had strong connection to each other, which partially explains the reason why she has a separate class and is not included in class 0. Finding the reason behind Rei' s strong connection, was very limited. There could be multiples reasons, but we lacked concrete evidence. This also should be studied from the research followed by this research.

4 Conclusion

With the database of hashtags of 'IVE' in Twitter, we revealed that the hashtags are related to each other in some points of view. The hashtags are used following some rules macroscopically. However we could not find out the causes of whole aspects of result.

First, database was too small for analysis. Regulating the size of database was required for us to handle, but to get better results, we need larger dataset and study based on it.

Also, we only researched with one working period. To make reliability higher and demonstrate the social phenomenon in Twitter, we have to take several working periods with one group and generalize the hypothesis assumed in this report.

We chose a group that made their debut recently. It means that the group's characteristic or each members' characteristics are not fixed. Also the influence of the group is not big enough, so the apsects of Twits about 'IVE' can be differed everyday, which means that the tendency is not stable.

To make a better result, larger dataset with various time period should be based on the research followed by.