PaperGalaxy: Generating Semantic Network of Academic Papers with Crowdsourcing

Hoon Kim
Korea Advanced Institute of Science and Technology
gnsrla12@kaist.ac.kr

Hyunsung Cho
Korea Advanced Institute of Science and Technology
hscho122@kaist.ac.kr

Juho Sun
Korea Advanced Institute of Science and Technology
soju@kaist.ac.kr

ABSTRACT

For students and researchers, it is essential to understand how each research is connected to one another. By doing so one can gain more insight of the field and tackle the problem they are trying to solve more easily. However, finding and understanding such connection is very difficult for individuals to deal with. Furthermore, since existing paper finding solutions, such as Semantic Scholar or Google Scholar is based on AI technology, they have limitation of lacking context of each connections. In this paper, we present PaperGalaxy – crowdsourcing platform that can generate semantic network of academic papers that represent how papers are related with one another. PaperGalaxy gathers data from academic paper readers themselves as they create a map of papers by connecting academic papers and labeling each connections with descriptions. A technical evaluation shows findings of observation and issues from subjective study on PaperGalaxy.

INTRODUCTION

There are estimated amount of 1.486 million peer-reviewed papers published within 2010 and it is becoming more and more difficult for researchers to keep up with this explosive growth of scientific literature. “Which papers are most relevant? Which are considered the highest quality? Is anyone else working on this specific or related problem? Difficulty of answering these questions is slowing down research and making it hard to solve big problems. Many attempts to solve these problems such as Semantic Scholar and Google Scholar exists. However, all of them depend on AI technology such as machine learning and computer vision to analyze papers, which leads to fundamental limitations. For example, although Semantic Scholar’s AI is very good at find connections between papers, it lacks the feature of finding the context of the connection. To address this limit, this research envisions a scenario of generating semantic network of academic papers with crowdsourcing. Our goal is to gather data from readers of academic papers through crowdsourcing and not only find the connections between papers, but also the context of each connections. For example, crowd could find a meaningful connection that this specific research is related to.

RELATED WORK

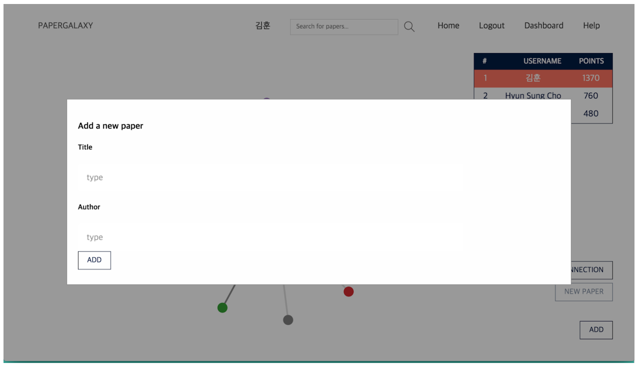
Semantic Scholar is one of a few services that users can use to search through academic papers. Semantic Scholar analyzes publications and extracts important features of academic papers to find connections between them. They use machine learning and computer vision technique to analyze the text and references to find meaningful information. PaperGalaxy, on the other hand, while aiming to achieve the same objective of finding connections between papers and effectively showing them to the users, is powered by academic paper readers themselves, instead of computers. By crowdsourcing academic readers, it is possible to find out the context of each connections and PaperGalaxy aims to create the most accurate semantic network of academic papers.

PAPERGALAXY

PaperGalaxy gathers data from academic paper readers themselves and deliver this information in the way that users can acquire meaningful information effectively. PaperGalaxy is designed to derive the maximum capability out of the crowd, and is composed of 4 parts: data collection, data visualization, paper navigation, and gamification.

Data Collection

All the data used in PaperGalaxy are sourced from the crowd. In PaperGalax, the crowd can add connections between papers. They first select two papers they want to connect. Then they choose a type of the relationship of selected papers from three options: similar motivation, similar technique, and similar workflow. They finally write a description explaining how the two papers are related as shown in Figure 2 If papers that users want to connect are not available in the platform yet, they can supplement missing papers easily with minimal information of the title and author of the paper. (Figure 1) The addition of paper is immediately updated on the platform, so the crowd can check whether their contributions were successfully made or not.



#	USERNAME	POINTS
1	김훈	1370
2	Hyun Sung Cho	760
		480

Add a new paper

Title

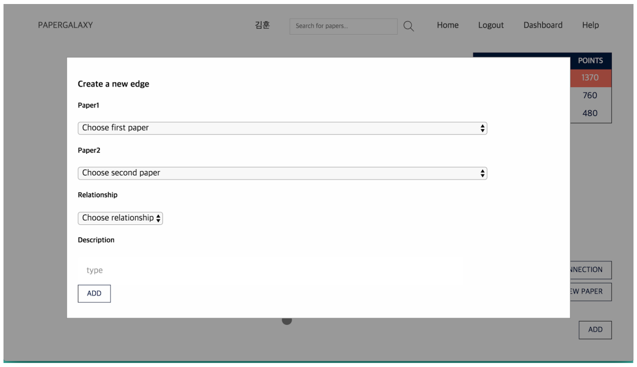
type

Author

type

ADD

Figure 1. Addition of new academic paper



POINTS
1370
760
480

Create a new edge

Paper1

Choose first paper

Paper2

Choose second paper

Relationship

Choose relationship

Description

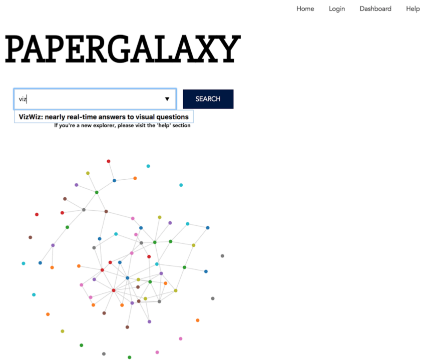
type

ADD

Figure 2. Addition of new connection between papers

Data Visualization

To display the semantic network of academic papers in an intuitive and attractive way, PaperGalaxy uses an interactive data visualization library, D3.js, for the graph. Figure 3 and Figure 4 below show a screenshot from PaperGalaxy that captures the visualization of academic papers as nodes and connections between them as edges. The graph on the main page includes all the papers and connections in the platform. Users can get an overview of entire data from this graph. (Figure 3) The graph on the result page shows searched paper and related papers in depth 1. (Figure 4) Users can check the information of papers by hovering over the node. They can also check the information of connections by clicking the edge. (Figure 5) The descriptions on an edge are aggregated in a table format that involves the relation type, text description, up-vote, and down-vote. Edges in the graph have different colors according to the number of descriptions added to the edge. The more descriptions an edge has, the darker the edge is.

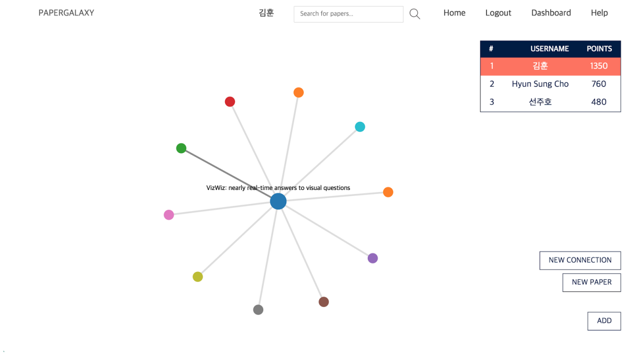


viz

VizWiz: nearly real-time answers to visual questions

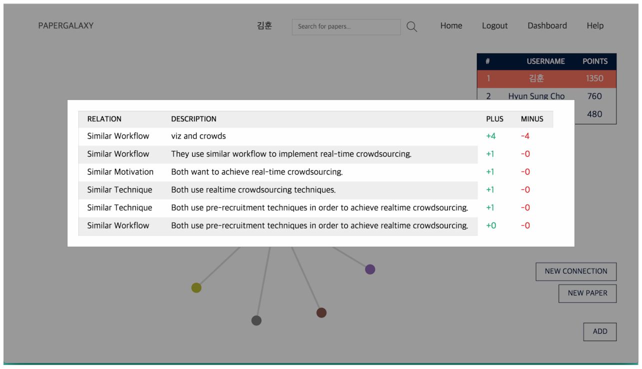
If you're a new explorer, please visit the 'help' section

Figure 3. The main page of PaperGalaxy displays the entire graph of papers that crowds have generated.



#	USERNAME	POINTS
1	김훈	1350
2	Hyun Sung Cho	760
3	선주호	480

Figure 4. Search result page of the paper "VizWiz"



#	USERNAME	POINTS
1	김훈	1350
2	Hyun Sung Cho	760
		480

RELATION	DESCRIPTION	PLUS	MINUS
Similar Workflow	viz and crowds	+4	-4
Similar Workflow	They use similar workflow to implement real-time crowdsourcing.	+1	-0
Similar Motivation	Both want to achieve real-time crowdsourcing.	+1	-0
Similar Technique	Both use realtime crowdsourcing techniques.	+1	-0
Similar Technique	Both use pre-recruitment techniques in order to achieve realtime crowdsourcing.	+1	-0
Similar Workflow	Both use pre-recruitment techniques in order to achieve realtime crowdsourcing.	+0	-0

Figure 5. Detailed information of a certain connection between two papers.

Paper Navigation

There are two ways to search a paper in PaperGalaxy. One way is to type in a title of paper in a search box that also provides autocomplete function. Users can also find papers by clicking a node in a graph provided on the main page. After they find the paper in either way, it leads to the result page where selected paper and other papers that have connections with it are displayed.

Gamification

We leverage gamification to motivate users to make contributions in the platform. One element of gamification is the points system. When users add papers or connections, they get a certain amount of points. They can also gain points if their inserted connections are up-voted by other users meaning the connections had good descriptions. On the other hand, they lose points if their connections are down-voted meaning low quality description. (Figure 5) This also acts as a quality control mechanism as it blocks users from adding meaningless data because they might lose points for generating low-quality data. The score and name of users are displayed on every page. The leaderboard shows top three rankers and player right below and right above the logged-in user. The user is also marked by special color.

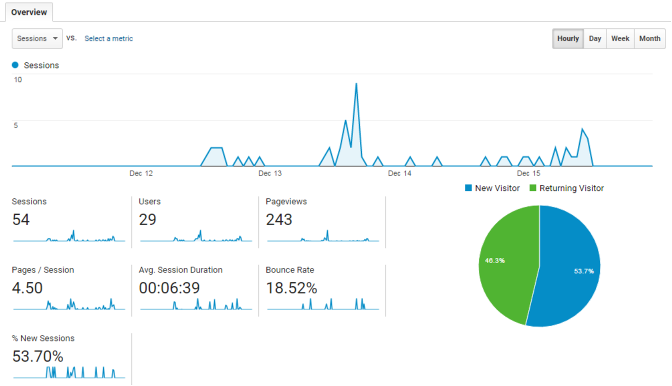
USER STUDY

We conducted an open, voluntary user study to analyze usage patterns of users in a natural setting. The goal of the user study was to observe if the current prototype alone was attractive enough to motivate users voluntarily make contributions to the system.

Procedure

Before launching user study, we inserted academic papers that were discussed in CS492 Crowdsourcing course, and some sample connections among the papers. We posted the link to PaperGalaxy prototype with a short description on Piazza (for CS492 Crowdsourcing course at KAIST) and Facebook to recruit prospective users—researchers—as voluntary participants. We did not give any specific instructions for what tasks the participants have to accomplish. The participants were allowed to freely explore the system.

We used Google Analytics to observe and analyze the behavior patterns of users in our system and collected feedbacks using Google Forms.



Sessions
54

Users
29

Pageviews
243

Pages / Session
4.50

Avg. Session Duration
00:06:39

Bounce Rate
18.52%

% New Sessions
53.70%

Figure 6. Google Analytics screenshot of PaperGalaxy

Results

For 4 days after advertisement, total of 29 users visited PaperGalaxy as shown in Figure 4 from Google Analytics. The 29 users visited the website 54 times in total, resulting in 46.3% revisit rate. Out of the 29, 9 of them were active users who signed up and made contributions to our database.

The active users added 42 edges and 6 papers in total. Each active user contributed an average of 4.7 edges and 0.7 papers. However, two of the users who added 25 edges turned out to be scammers. They discovered a bug in the system where they could add a connection from one paper to the same paper. Because this connection does not appear on the visualization, they created 25 of meaningless edges. Excluding these scammed edges, the average of edge addition falls to 2.6 edges per user.

Three users left feedbacks for improvement, and they will be discussed in the next section.

DISCUSSION

Even though PaperGalaxy showed success to gather meaningful data from users, there are some clear limitations. First of all, there was a high entry barrier to making contribution. The crowd can add meaningful data only if they read an academic paper. This task requires time and at least certain amounts of knowledge to make valuable contributions. Also, PaperGalaxy currently contains papers from crowdsourcing research area. This only allows users who have knowledge on crowdsourcing to participate in the system, blocking users from other areas. The results of user study also showed that the number of edges one user made (4.7 edges including scammers, 2.6 edges without scammers) was significantly higher than the number of papers one user made (0.7 papers).

To resolve the issue of high entry barrier, we will flourish the database more with a huge amount of academic papers in a wider range of field, for example by combining PaperGalaxy's database with an open database of academic papers. Furthermore, the system currently provides only title and author of a paper as the paper's information. According to feedbacks from user study, users felt uncomfortable that they could not read the paper immediately after they have found a good paper using PaperGalaxy. They have make additional search in the search engine and find a valid source to read the paper. Therefore, PaperGalaxy will also need to include a link as part of the paper's information or a paper preview function in the future. We also expect that this will lower the high entry barrier as first-time readers are enabled to read the papers immediately.

Moreover, there is lack of motivation for first-time users. As we discussed before, it is hard for newcomers to make contributions to PaperGalaxy. Also, what users can achieve from the platform is not straightforward. This makes users keep questioning why they should make contributions. We should add more instructions explaining what is the benefits. Targeted advertisement to reading groups of researchers or new authors of academic papers would be another solution as well.

For the long-term motivation for users, gamification was used. The results from user study partially proved its success in terms of motivation. Out of the small number of active users in user study, there were scammers who added meaningless data merely to gain points and higher rank in the leaderboard. However, the increase in the number of scammers will lead to unreliable data, so the quality control system needs further improvements to block scammers from gaming the system.

Lastly, some technical limitations also exist. By analyzing user data from Google Analytics, we found that nearly quarter of users access our platform with a mobile device. PaperGalaxy, however, does not support mobile-friendly interface yet. It is also fully functional only in Chrome, not in other browsers like Safari or Internet Explorer. These technical issues reduce user accessibility.

CONCLUSION

As each field of science becomes more and more specialized, it is becoming more and more difficult for researchers to keep up with this explosive growth of scientific literature. In this paper, we demonstrated generating semantic network of academic papers with crowdsourcing that can help researchers to navigate through galaxy of scientific literatures. Although it is impossible for each individual to understand the entire field of science, our study of PaperGalaxy demonstrated that collective intelligence of academic paper readers can achieve such goal.

REFERENCES

1. Butler, A., Izadi, S., and Hodges, S. SideSight: multi-"touch" interaction around small devices. In Proc. UIST ’08, 201-204.
2. Discriminating Different Ways of Grasping and Holding a Tangible User Interface. TEI ’09, 359-362.