**A Preliminary Study on the Construction of Natural Science Resource Recommender System for High School Students**

**Pei Yu Wang, Kuo-Liang Ou**  
*Institute of Learning Sciences and Technologies  
National Tsing Hua University, Hsinchu 30013, Taiwan;*

*klou@mx.nthu.edu.tw*

**Abstract**

*This research aims to establish a recommendation system for autonomous learning resources. The research objects are high school students and teachers, and the scope is in natural sciences. Compared with general search engines, this system can search by word to find terms in the scientific field and excerpts of related articles. High school students can enhance learning and manage learning progress through the Kanban system interface for knowledge integration during SDL. The teacher's scoring results show that the recommended resource system helps learners collect learning information. In addition to the teacher questionnaire, the search results compared with commonly used search engines and the range of search results are comprehensive. However, the quality and quantity of the system training set can significantly affect query results. The results of this study can provide a reference for the development of future learning resource recommendation systems.*

**Keywords**:   
Text Mining, self-directed learning, Recommend System

Introduction

Self-directed Learning (SDL) has become a compulsory subject for high school students in Taiwan since 2019. Students need to develop their learning plans and search learning resources related to learning subjects appointed by themselves in the SDL process. Therefore, the teacher plays a scaffolder, who no longer dominates the learning subject and content just like they did with the didactic teaching strategy before. However, if students lack prior knowledge, they cannot evaluate the correctness and relevance of learning resources searched from the Internet. Notably, there was much false information on the Internet, making students misunderstand SDL.

This paper is a pilot study that uses text mining technology based on authoritative articles of popular sciences crawled from the web to provide students' SDL process. The relevant and extensive knowledge content will be recommended for students to read and explore continually to avoid students searching for learning resources on the Internet, focusing on the keywords only and facing the difficulty of discriminating the reliability. This research uses the Kanban system as the human-machine interface. It cooperates with the learning strategies and progress of autonomous learning to assist teachers in observing students' learning conditions.

Related Works

Self-directed Learning (SDL) is proposed by Knowles (1975). During the SDL process, learners need to self-diagnose learning needs, set learning goals, and find resources. Furthermore, learners need to constantly adjust the learning direction and evaluate the results. Knowles (1975) also indicates the importance of SDL. First, SDL autonomous learners learn more than passive learners. Second, the SDL process also aligns with human development, and third, learners will generate independent inquiry skills for facing problems. In addition, Kidane et al. (2020) believe that self-directed learning enables learners to understand and analyze issues in self-directed learning, which helps cultivate the ability of lifelong learning.

Iwasiw (1987) indicates that a teacher-guided model should be employed in SDL first for students to understand the learning process. High school students still need peer assistance or teacher intervention in their learning process. With the aid of using digital platforms and group learning strategies, teachers may have the opportunity to help students control their learning progress and allow learners to supervise each other to achieve the goal of SDL. In this study, the Kanban System was used as the learning platform for students to integrate the learning resources and process and assist learners in managing learning tasks and understanding their peers' learning status (Anderson, 2010).

Determining learning goals and searching learning resources are critical parts of SDL. Most high school students are highly dependent on search engines, like Google and Bing, which provide accurate keyword search results. Due to the poor use of keywords or on-authoritative articles, the query results of traditional search engines may cause students' learning myths or mislead the learning direction. Therefore, it is challenging for teachers to provide SDL learners with relevant, extensive, and authoritative information. Text mining is developed to process unstructured text data through various quantification techniques. It helps users explore the correlation between words (Shi, Chen, Han, & Childs, 2017) and information hidden in the articles (Tan, Steinbach, and Kumar, 2006). Some article recommendation systems were proposed (Zhang, Yao, Sun, & Tay, 2019 and Domingues, Sundermann, Manzato, Marcacini, & Rezende, 2014) to explore and disclose the related articles for various objectives.

Method

This study collects 34,191 authoritative popular science articles from the Internet and then uses text mining technology, including TF-IDF and Doc2Vec algorithm, to identify their correlation. Table 1 lists the data source and the article amounts.

**Table 1**. *Data Sources*

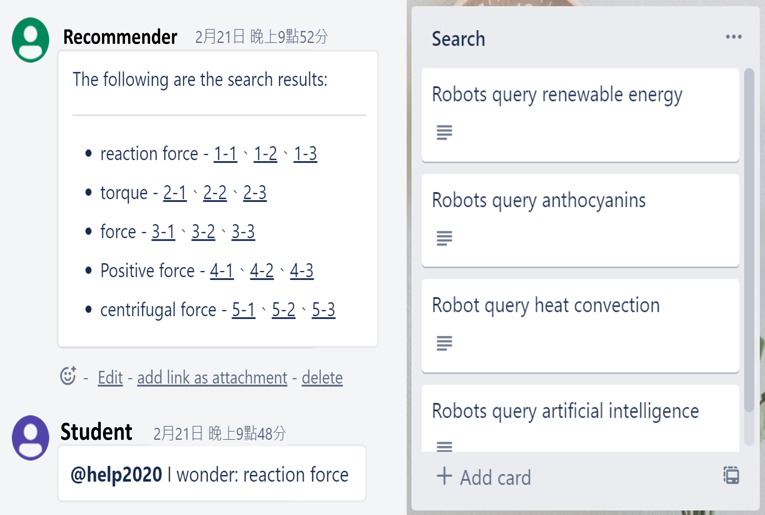
|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Title** | **URL** | **Article Numbers** |
| 1 | Environmental Information Center | https://e-info.org.tw | 11322 |
| 2 | PanSci | https://pansci.asia | 9817 |
| 3 | Sci-Tech Vista | https://scitechvista.nat.gov.tw | 6180 |
| 4 | Science Online | https://highscope.ch.ntu.edu.tw | 6104 |
| 5 | Scientific American | https://sa.ylib.com | 2013 |
| 6 | The News Lens | https://www.thenewslens.com | 1232 |
| 7 | National Geographic | https://www.natgeomedia.com/ | 787 |
| 8 | NTU CASE | https://case.ntu.edu.tw/blog/ | 719 |
| 9 | Physics Bimonthly | http://pb.ps-taiwan.org/ | 533 |
| 10 | Science Monthly | https://www.scimonth.com.tw/ | 418 |
| 11 | NTCU science lab | http://scigame.ntcu.edu.tw/ | 330 |
| 12 | Lifechem | https://www.lifechem.tw/ | 302 |
| 13 | phycat | https://hfdavidyu.com/ | 174 |
| 14 | Science Buds - Science Games | http://www.bud.org.tw/ | 56 |
| Total |  |  | 34191 |

The first step of text mining is segmentation and conversion. The Jieba, an open-source word segmentation tool, is employed in this study. Then, the text vectors among each word are established by the Doc2Vec algorithm (Le & Mikolov, 2014). These vectors represent the similarity between each word. The closer the semantics are, the higher the score. For example, as shown in Table 2, when students inquire about the word "physical," the system responds that the most relevant character is "physics" with a similarity of 0.71, followed by condensed matter with a similarity of 0.63. The top 10 most similar terms are mechanics, theoretical physics, elementary particles, chemistry, quantum optics, quantum mechanics, celestial mechanics, and mathematics.

**Table 2.** *The top 10 terms and similarity scores of the term physics*

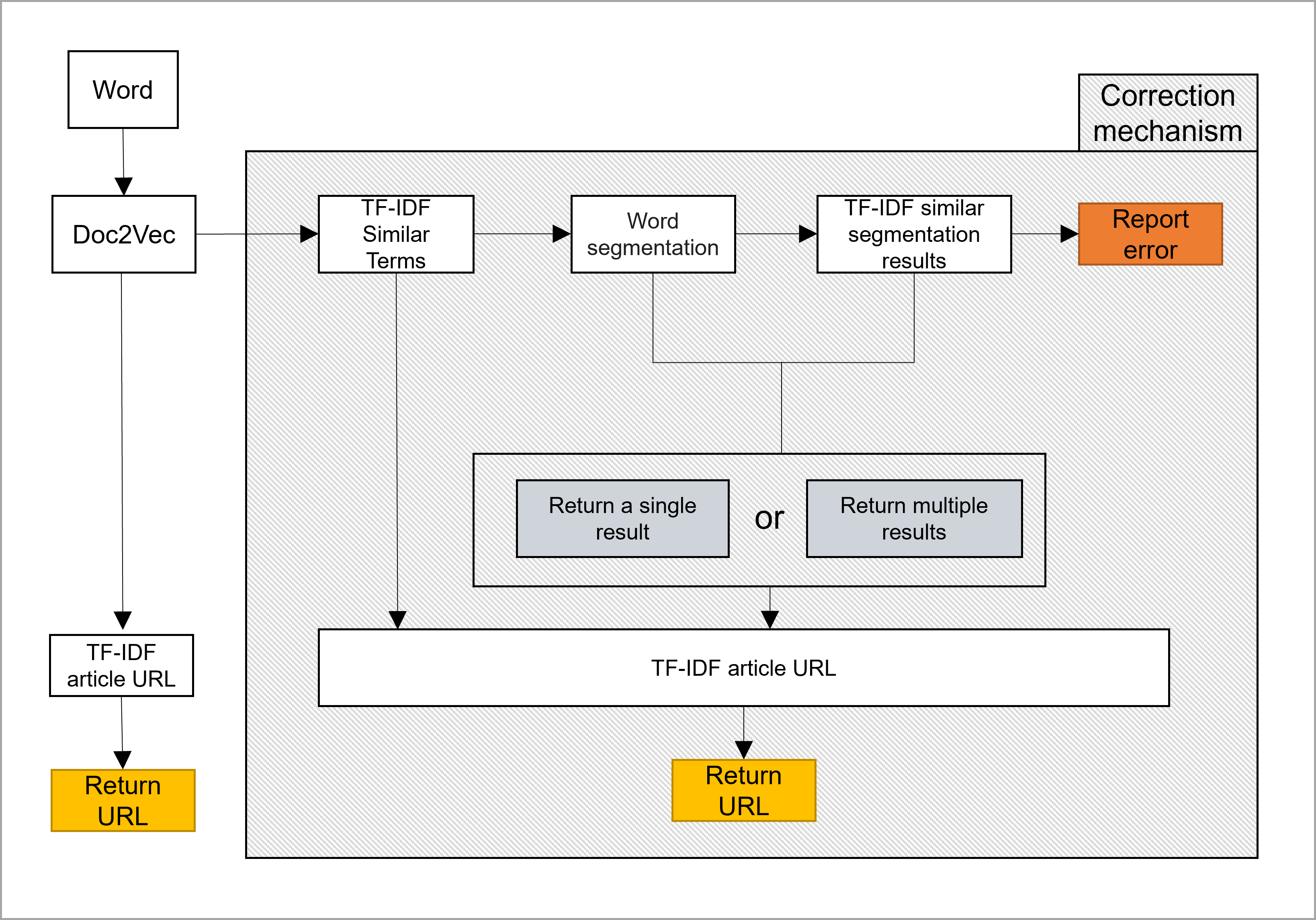
|  |  |  |
| --- | --- | --- |
| **Inquiry term: physical** | | |
| order | Related terms | Similarity scores |
| 1 | physics | 0.71 |
| 2 | condensed matter | 0.63 |
| 3 | mechanics | 0.63 |
| 4 | theoretical physics | 0.61 |
| 5 | elementary particle | 0.61 |
| 6 | chemistry | 0.61 |
| 7 | quantum optics | 0.60 |
| 8 | quantum mechanics | 0.60 |
| 9 | celestial mechanics | 0.60 |
| 10 | mathematics | 0.58 |

After the top 10 relevant characters are pulled from the dataset, the three most relevant articles corresponding to each character will respond to the Kanban system for students reading using the TF-IDF algorithm. For example, in Figure 1, a student queried the science term “reaction force” in the Kanban card. The system responded to the student with five related terms including “reaction force” itself and “torque”, “force”, positive force”, and “centrifugal force” above the card.



**Figure 1.** *The user interface of the Kanban system*

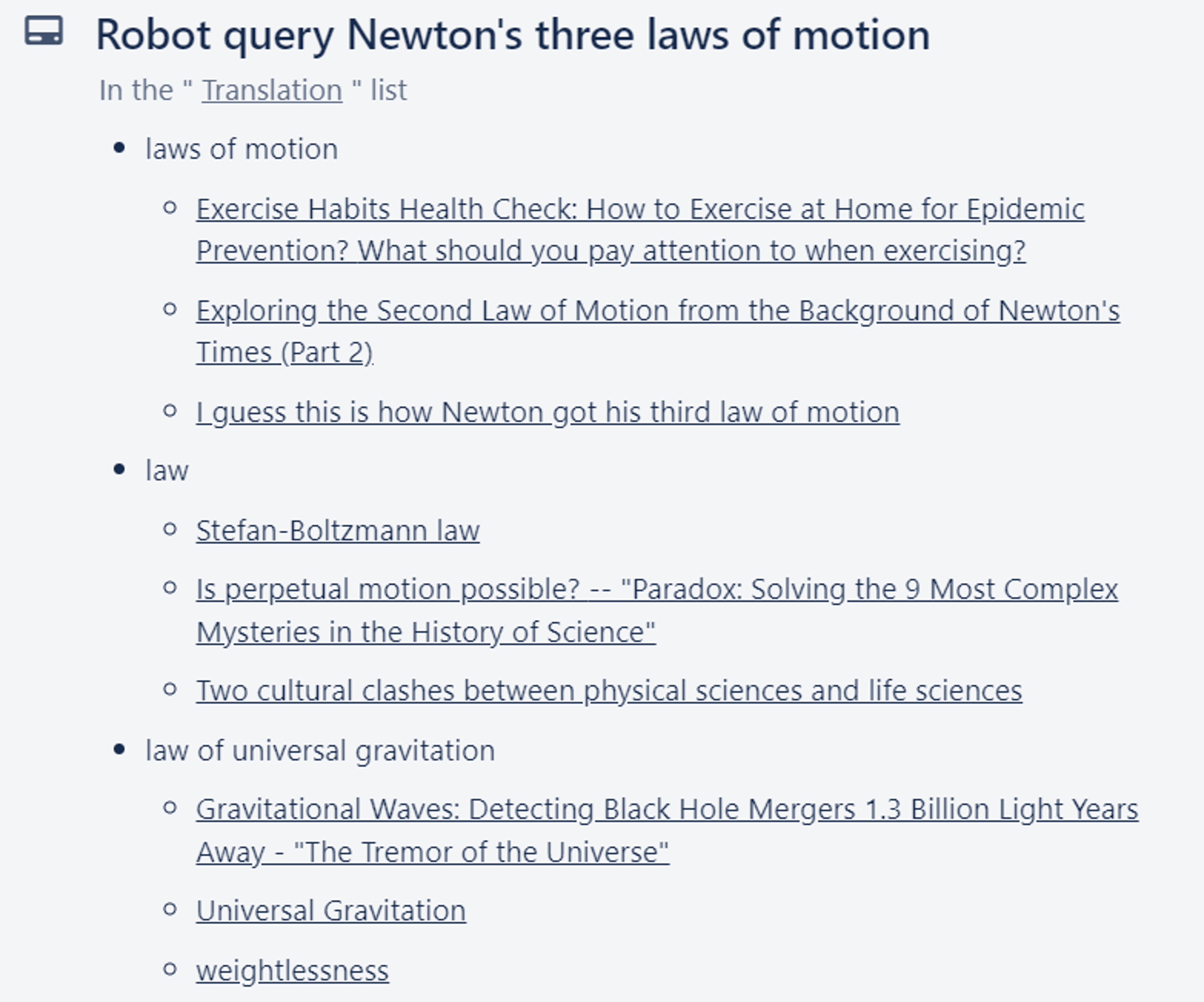
When students ask a question in the SDL resource recommender system, some related terms and articles will be replied to the groups' space where the students belong in the Kanban system. In the case of Figure 1, three related articles are listed after each term for students’ reference when learning. Therefore, students can explore the related knowledge after reading these related articles in the SDL process. The Kanban system helps group members to learn through the cards and the lists on web browsers. The system architecture is illustrated as follows.



**Figure 2.** *The system architecture*

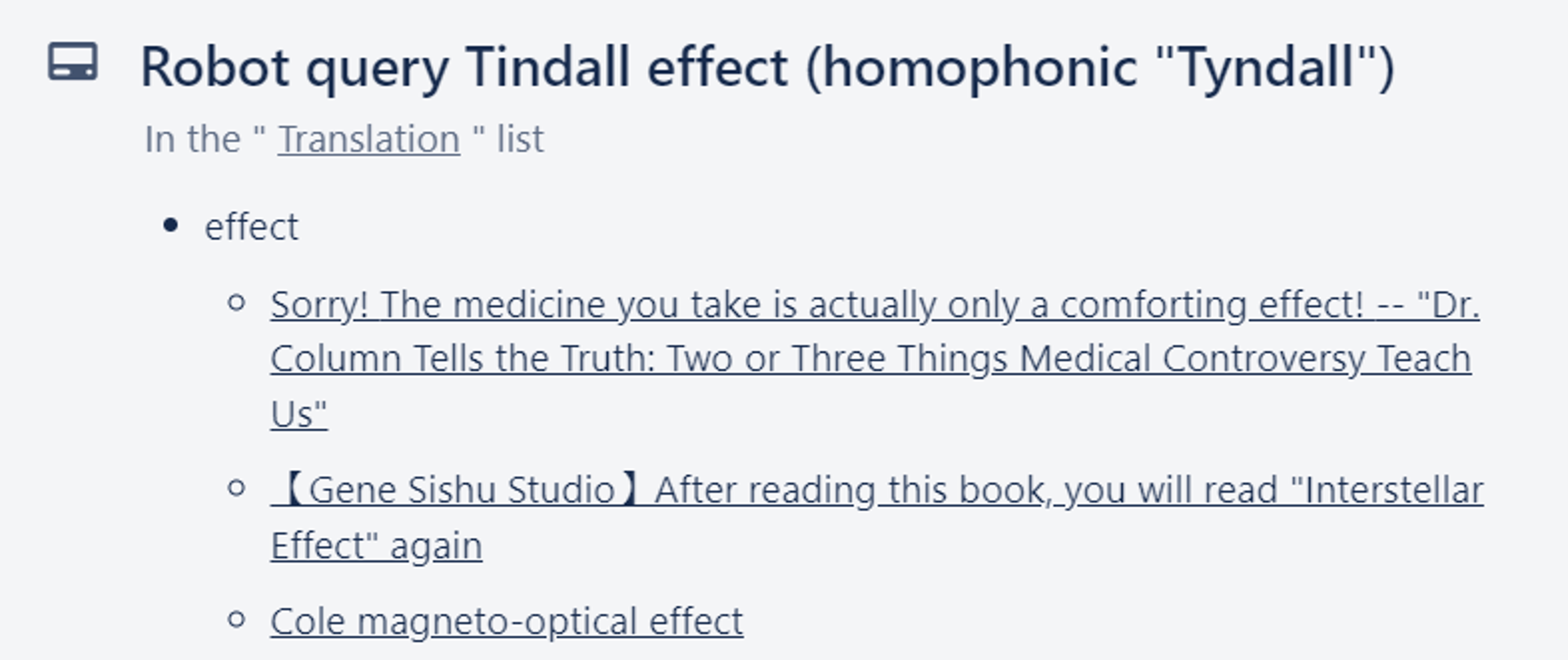
# Results

Figure3 illustrates the results of the student's inquiry of Newton's Laws of Motion, which are the most popular science term in high school physics.



**Figure 3**. *The result of the student's inquiry of Newton's Laws of Motion*

Figure4 illustrates the results of the student's inquiry of the Tyndall Effect, which is the advanced science term in high school physics.



**Figure 4.** *The result of student's inquiry of the Tyndall Effect*

A questionnaire with eight items was employed in this pilot study to investigate teachers’ views of their students using this SDL learning resource recommend system in the future. Twenty-seven science teachers in Taiwan were recruited in this study. The following table shows the statistical results.

**Table 3**. *The result of the questionnaire of system performance*

|  |  |  |  |
| --- | --- | --- | --- |
| **Q ID** | **Item** | **Mean** | **SD** |
| Q 1 | I am satisfied with the content of this system feedback. | 4.47 | 0.62 |
| Q 2 | The system responses are related to student’s inquiries. | 3.82 | 1.38 |
| Q 3 | The system responses are helpful to students’ inquiries. | 4.59 | 0.62 |
| Q 4 | The system responses are helpful to students’ dispersing thinking ability. | 4.65 | 0.49 |
| Q 5 | The system responses are under my expectations. | 3.47 | 1.33 |
| Q 6 | The system responses are helpful to students’ data collecting ability. | 4.59 | 0.62 |
| Q 7 | The system responses are helpful to students’ SDL process. | 4.38 | 0.72 |
| Q 8 | The system responses are helpful to teachers to assist students’ SDL process. | 4.41 | 0.71 |
|  | Mean | 4.30 | 0.81 |

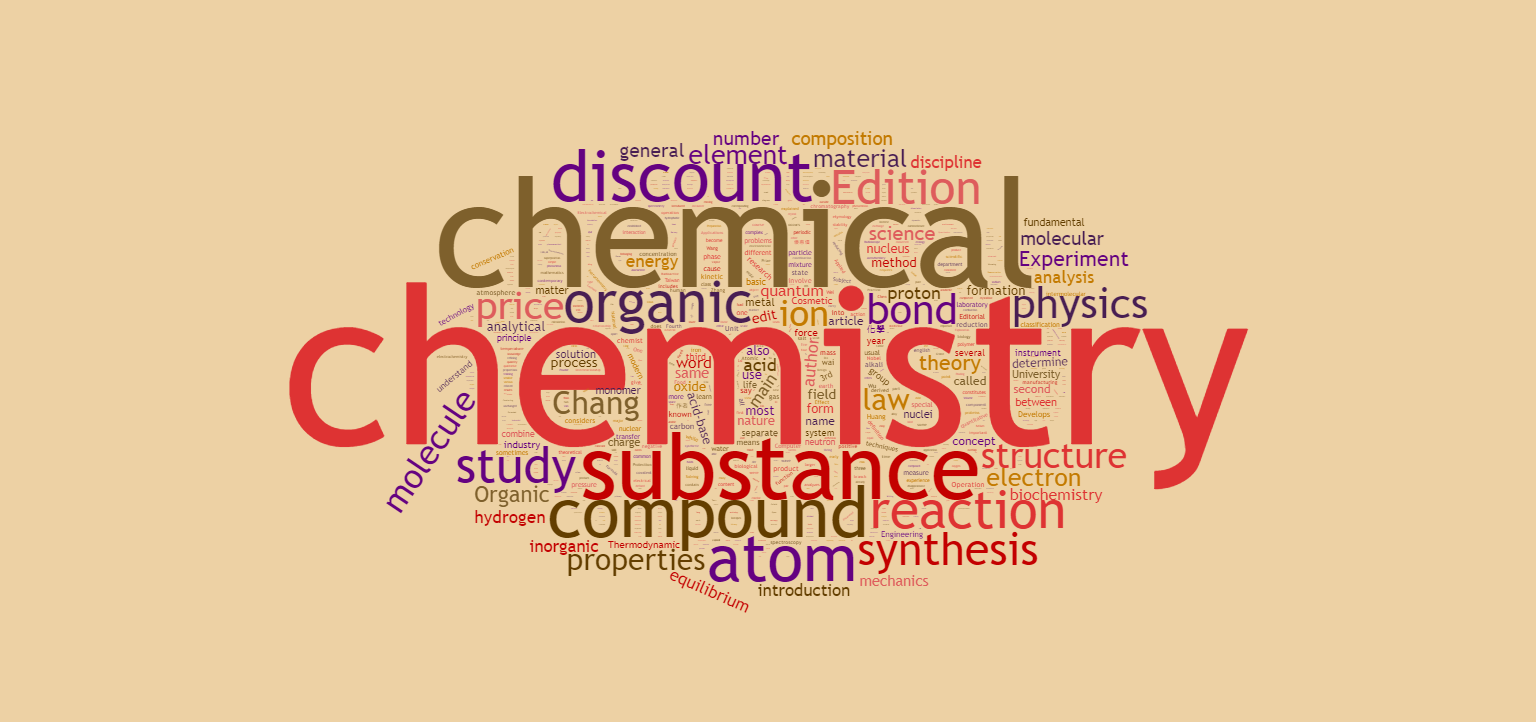
The system satisfaction score is 4.47 (Q1), which shows that teachers regard this recommender system as benefits students' SDL process and assist teachers in teaching(Q8). In system capability scores (Q2 and Q3) show that this system is helpful for students' term inquiry and helps students SDL process (Q7). To help students' diverse thinking, the Q4, Q5, and Q6 reflect that the system is suitable for assisting students' dispersing thinking ability. Furthermore, the standard deviation of Q5, "The system responses are under my expectations." reached 1.33. This high value shows that the SDL recommender response results are pretty unexpected, unlike traditional search engines response ever.

**Table 4.** *Resource recommender system compared with the general search engine.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Inquiry term: biology** |  |  |  |
| **Resource Recommender System** | **Related terms** | **Similarity scores** | **Related terms** |
| metabolic similarity | 0.866 | biological | 0.962 |
| submarine volcano | 0.854 | ecology | 0.957 |
| medical environment | 0.836 | overlapping | 0.957 |
| pedigree | 0.827 | cell | 0.956 |
| cancer treatment | 0.819 | genetics | 0.954 |
| Ediacaran Biota | 0.785 | life | 0.953 |
| monoclonal antibody | 0.778 | zoology | 0.952 |
| predator | 0.774 | biogeography | 0.950 |
| Diversity Indicator | 0.773 | evolutionary biology | 0.949 |
| Geographical location | 0.769 | Cell Biology | 0.947 |

The system compares search results with Google search engine results. After obtaining the same number of texts from the term biology, segment the words separately, establish the word vectors by the Doc2vec and inquire "biology" again. The results are as shown in Table 4 above. The system's value is more scattered, the value is between 0.7 and 0.8, and the words are extensive; the value of Google is basically in the range of 0.9, and the word "biology" appears more frequently in the related terms results the words are precise.

On the other hand, the general search engine is more precise than the system, and the result of the system is extensive.



**Figure 5.** *The word cloud of the general search engine's inquiry of the Chemistry*

As shown from Figure 5 above, Google search results contain words such as "price" and "discount," acceptable for a commercial website, but as a source of SDL resources, it will inevitably affect learners' information collection Process.

Due to the lack of the total number of texts, the inquiry results of general search engines are more accurate than the system. However, for a system built with limited sources, the scope of the query results is sufficient to provide learners with data collection and use in the early stage of SDL's idea of ​​the topic, and there will be no results such as shopping websites and advertisements. If more texts collect more in the future, the accuracy can be further improved.

# Conclusion

This research constructs a learning resource recommender system for high school students' SDL. The relationships between the popular science terminologies are analyzed by text mining technology on 34,191 authoritative popular science articles crawled from the Internet. When students inquiry about a keyword in the Kanban system, the top 10 related terms, and their related articles will be responded to and shown in the cards of Kanban. This research is a pilot study, and a questionnaire is enrolled for 27 high school natural science teachers to evaluate the availability for students' SDL process. The results show that teachers have high scores of satisfaction with this system. Teachers also believe that it can assist students in keywords inquiry and enhance students' thinking divergence. Therefore, the system is more suitable for use in the early stage of SDL, such as expanding cognition and theme development.

**Acknowledgment**

This research was funded by the Ministry of Science and Technology (MOST), Taiwan, under the grant number MOST 109-2511-H-007-006-MY3.

**References**

Anderson, David. (2010). Kanban - Successful Evolutionary Change for your Technology Business. Blue Hole Press.

Domingues, M. A., Sundermann, C. V., Manzato, M. G., Marcacini, R. M., & Rezende, S. O. (2014). *Exploiting Text Mining Techniques for Contextual Recommendations.* Paper presented at the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).

Kidane, H. A.-O., Roebertsen, H., & van der Vleuten, C. P. M. Students' perceptions towards self-directed learning in Ethiopian medical schools with new innovative curriculum: a mixed-method study. (1472-6920 (Electronic)).

Knowles, M. S. (1975). Self-directed learning: A guide for learners and teachers. New York: Association Press.

Iwasiw, C. L. (1987). The role of the teacher in self-directed learning. Nurse education today, 7(5), 222-227.

Shi, F., Chen, L., Han, J., & Childs, P. (2017). A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *Journal of Mechanical Design, 139*(11)*.* doi:10.1115/1.4037649

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to data mining.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR), 52*(1), 1-38.

Zheng, Saijing & Rosson, Mary Beth & Shih, Patrick & Carroll, John. (2015). Designing MOOCs as Interactive Places for Collaborative Learning. 10.1145/2724660.2728689.