SAVEIT: A SMART ARCHIVE FOR IT SENIOR PROJECTS

R. Alotaibi, R. Alufi, K. Alamoudi, R. Alghamdi, S. Sami, M. Almarie

Faculty of Computing and Information Technology, Department of Information Technology, King Abdulaziz University (SAUDI ARABIA)

Abstract

The volume and diversity of senior projects are expanding each year. It is exceptionally difficult for students and advisers to keep track of each project. They face difficulty to refer to previous projects, extract some useful information and whether the proposed ideas match potential advisers and their research interests. In addition, decision-makers do not have valuable statistics about senior projects, assigned advisers, and areas of interest of senior projects. SaveIT is a web-based application that is proposed for archiving a large number of senior projects of the previous and upcoming years. It uses the power of text mining tools to generate statistics and interesting patterns from senior projects like trends of programming languages, the frequency plot of areas of interest, and the relationship among areas of interest represented in the network graph. SaveIT helps users without prior expertise in textual analysis to find answers to some important questions. Results showed that the text-mining tools were able to identify a considerable number of relevant terms from the texts analyzed, providing concise representations of projects which can support students and advisers.

Keywords: Text-mining, senior projects, smart, archive.

1 INTRODUCTION

In recent years, text mining has become more popular in the field of Education mostly because of the growing number of systems which store large data about students and their materials [5, 7]. Text mining technology provides a solution to bridge the knowledge gap between unstructured text and structured information representation [9, 11].

Text mining uses techniques such as trends and associations to automatically retrieve, extract and discover statistics information in large data [4]. It can help humans to identify and verify hidden information from the text more efficiently and it can discover relationships in the huge data. A number of text mining tools have been developed to support research [2, 6, 10, 14], yet none of them was designed for senior projects.

The volume and diversity of senior projects are expanding each year. It is exceptionally difficult for students and advisers to keep track of each project. They face difficulty to refer to the previous projects, extract some useful information and whether the proposed ideas match potential advisers and their research interests. In addition, decision-makers do not have valuable statistics about senior projects, assigned advisers, and areas of interest of senior projects. Therefore, there is a need to develop a tool that can archive and extract useful information about the archived projects automatically.

In this paper, we propose SaveIT that incorporates text-mining techniques. SaveIT extracts valuable knowledge from all the titles and abstracts of the project reports. Also, it visualizes some interesting information such as the most and least applied areas of interest in all senior projects, the correlation among areas of interest and trend analysis.

Our proposed solution is a web-based application that used for archiving a large number of senior project reports and prototypes of the previous and upcoming years. It uses the power of text mining tools to generate statistics and interesting patterns from senior projects like trends of programming languages, the frequency plot of areas of interest, and the relationship among areas of interest represented in the network graph. SaveIT helps users without prior expertise in textual analysis to find answers to some important questions.
SaveIT targets students, advisers and researchers at the Information Technology department at King Abdulaziz University. All users will be able to get some knowledge about previous projects, research areas, and the areas of interest of potential advisers. Results showed that the text-mining tools were able to identify a considerable number of relevant terms from the texts analyzed, providing concise representations of projects which can support both students and advisers.

The rest of the paper is organized as follows. Section 2 presents some related work. Next, Section 3 gives some overview of the proposed solution. Section 4 describes some experiments to validate the proposed solution while Section 5 discusses the results. Finally, Section 6 concludes the paper.

2 LITREATURE SURVEY

There are many universities spent efforts to keep an online archive for senior projects. Whitman College¹, California Polytechnic State² and University of North Carolina at Asheville³ have online archives in their websites for senior projects. The projects can be classified by the year or the department. They provide information about the coordinator and the number of projects of each year. These websites classify projects according to the year or the department, but none of them used text-mining technologies. Fig. 1 shows two examples of these systems.

On the other hand, there are some online tools that offer text analysis but they are not specific for senior projects. First, LimTox: Literature Mining for Toxicology⁴ is a web-based text-mining tool that extracts associations between compounds, drugs and genes. It incorporates many text-mining techniques that depend on classification tasks for extracting useful and interesting patterns from large textual data. It uses frequency terms and word n-gram that means co-occur of words in the same documents. These techniques are applied to PubMed abstracts [1]. A screenshot of LimTox is shown in Fig. 2(a).

RLetters⁵ is a web-based application for text analysis of journal articles. It allows users to perform textual analyses on sets of articles such as term frequency, word usage, graph dataset by publication date, and words associations (cooccurrences) as seen in Fig. 2(b).

WordItOut⁶ is a web-based text mining system that uses graphical representation of frequently used words in a text file the user uploads it to generate what is called word cloud in the text mining area. A screenshot of this tool is shown in Fig. 2(c).

3 TEXT ANALYSIS TECHNIQUES

A wide variety of analysis methods are implemented in SaveIT. The text-mining tasks are applied to the whole senior projects database and the results are generated automatically. It is possible for the admin to add new projects and the system will include them during the analysis. In addition, the user can define a specific period of years as shown in Fig. 4 and Fig. 6. The most common text mining analyses are described below.

3.1 Word Frequency

Word frequency is applied on areas of senior projects to explore the frequency of each area. It shows the frequency of areas of each adviser. Fig. 3 shows the graphical representation of the number of projects in each research area.

¹https://www.whitman.edu/academics/departments-and-programs/mathematics-and-computer-science/resources-for-students/senior-project-archive
²http://digitalcommons.calpoly.edu/seniorprojects/
³https://csci.unca.edu/projects/connecting-guardian-ad-item-association-buncombe-county
⁴http://limtox.bioinfo.cnio.es/
⁵https://www.rletters.net/
⁶https://worditout.com/
3.2 Word Cloud

Word cloud is applied to all abstracts of senior projects to extract the most frequent words appear in the abstracts. Fig. 4 shows the word cloud generated from the abstract of all senior projects from 2014 to 2018. It represents the most frequent words mentioned in the cloud representation such as application, mobile, patient and system.

3.3 Relationships Between Words (Network of Bigrams)

The network models the co-occurrence or correlations among keywords within the same documents [8]. It shows the relationship between areas of interests. For example, Fig. 5 shows a strong relationship between games and Virtual Reality (VR) research areas.
3.4 Topic Clustering

A topic is defined by a cluster of words where each word in the cluster has a probability of occurrence for the given topic, and different topics have their respective clusters of words along with corresponding probabilities. Different topics may share some words and a document can have more than one topic associated with it [13]. It is applied to areas of interest over the years to show the trends of areas. For example, Fig. 6 shows the most widely adopted programming languages in the senior projects from 2014 to 2018.

Moreover, SaveIT has other functions such as the navigation bar that contains a clickable list of the following pages: senior projects, research areas, and advisers. These lists allow the users to search senior projects by the name or the area of interest or the adviser name. Furthermore, users can extract some useful statistics about senior projects. Also, the system offers communication between former and new students.
4 IMPLEMENTATION

SaVeIT web-application is developed by ASP.NET MVC with bootstrap to be responsive with all devices [3]. The text-mining tasks are achieved by R console that gets data from SQL Server database. The main architecture and the data flow of the system are shown in Fig. 7.

Senior projects reports of IT students at King Abdulaziz University are uploaded as PDF and stored in the database. Then, the titles and abstracts will be extracted electronically into Report Summary tab with other information about the projects such as title, research area, abstract, adviser name, students names, video prototype, etc as shown in Fig. 8.

Figure 7. SaveIT Main Architecture
4.1 Preprocessing Documents

Preprocessing step plays an important role in text mining techniques and applications. It is the first step in the text mining process. It includes eliminate and remove stop words and many other functions.

SaVeIT is developed using tidy text mining approach in R. In general, tidy data as described by Hadley Wickham [12], has a specific structure. Tidy text format is a table with one-token-per-row. A token is a meaningful unit of text, such as a word. The table is described as follows:

- Each table is a type of observational unit.
- Each column is a variable.
- Each row is an observation.

Therefore, we convert the abstracts of senior projects to tidy format using libraries in R. Then, the outputs will be the inputs to libraries that deal with the text in the tidy format then integrates R with visual studio to visualizing the results of text-mining tasks on web-application pages.

5 RESULTS AND DISCUSSION

The efficiency and accuracy of the tool were evaluated intrinsically as well as extrinsically by case studies. The correlations identified by the tool show that it offers a great potential to organize and correctly classify topics. Furthermore, the tool can be useful, for example, in identifying areas of interest and their correlations.

Moreover, we used the System Usability Scale (SUS)\(^2\) which provides a quick and reliable tool for measuring the usability of the proposed solution. It consists of a 10-item questionnaire with five response options for respondents; from Strongly agree to strongly disagree. A SUS score above 68 would be considered above average and anything below 68 is below average. Fig. 9 shows the 10 questions. We got 92.5% SUS Score which is a good score and means that the system fulfilled the user's expectations and was clear and easy to use.

6 CONCLUSION

This paper presents SaVeIT web-based application for archiving senior projects at IT department, King Abdulaziz University with use of text-mining techniques to help researchers to find answers for some important questions. It uses the tidy text format approach and visualizes the text-mining results that applied on the whole senior projects database.

Results showed that the text-mining tools were able to identify a considerable number of relevant terms from the texts analyzed, providing concise representations of projects which can support both students and advisers.

REFERENCES


