Deconstructing Opinion Survey: A Case Study

Entesar Alanazi[1†]
361200365@qu.edu.sa
Department of Computer Science
College of Computer, Qassim University
Al Qassim, Saudi Arabia

Summary
Questionnaires and surveys are increasingly being used to collect information from participants of empirical software engineering studies. Usually, such data is analyzed using statistical methods to show an overall picture of participants' agreement or disagreement. In general, the whole survey population is considered as one group with some methods to extract varieties. Sometimes, there are different opinions in the same group, but they are not well discovered. In some cases of the analysis, the population may be divided into subgroups according to some data. The opinions of different segments of the population may be the same. Even though the existing approach can capture the general trends, there is a risk that the opinions of different sub-groups are lost. The problem becomes more complex in longitudinal studies where minority opinions might fade over time. Longitudinal survey data may include several interesting patterns that can be extracted using a clustering process. It can discover new information and give attention to different opinions. We suggest using a data mining approach to finding the diversity among the different groups in longitudinal studies. Our study shows that diversity can be revealed and tracked over time using the clustering approach, and the minorities have an opportunity to be heard.

Keywords:
Longitudinal Studies, Clustering, Opinion Diversity, Expert Opinion, Survey Opinion.

1. Introduction

In software engineering empirical research, there is an enlarged use of surveys and questionnaires in obtaining information from experts [1]. Because of the internet growth, it has become easier for businesses to collect experts' opinions through web practices and traditional procedures. So, that experts may have differing views and opinions. Many different types of data are generated during the software development process. Typical forms of data types [2]: Code bases, traces logs, historical code changes, fault databases. Typically, this data is analyzed according to whole descriptive statistics. Even though the processes used can help obtain general trends on the surveys, there is a risk of losing the groups' opinions or noticing differences in the opinions that the minority groups give, even as the similarity does not mean any actual difference. Therefore, the longitudinal studies clustering approach aims to get the slight differences between minority groups in surveys. Large investments have recently been made in software process automation to reduce development costs while also improving quality. Automation processes allow the storage and retrieval of new data types while still producing certain traditional data types. Some other forms of software engineering data like Test cases, System build traces, Team and personal data, and Development process data [3]. Recently, there is a large amount of survey data in most software organizations since it is easy to collect survey opinions systematically by online tools.

We notice that survey data, in general, is analyzed using statistical methods and measures (like mean, median, variance, or other) for their findings [4]. Usually, analysts consider all survey respondents as a single group and show an overall picture of participants' agreement or disagreement using some techniques to extract categories [5]. In many cases, the population is segmented into smaller groups by using available basic information. In most cases, this approach not often exposes opinion variations precisely. Sometimes, there are different opinions in the same group, but they are not well discovered [4].

The problem becomes more complex in longitudinal studies[1], where minority opinions might fade or resolute over time. In our research, we used clustering methods to analyze categorical data from a longitudinal opinion survey. Clustering divides the population into different groups of population (subgroups) that have common opinions to some level. This approach has many advantages and opportunities, including:

- Since it creates categories based on their opinions, it may reduce the manipulation of grouping. Also, by using clustering, specific information and opinions may be integrated.

[1] Longitudinal study (LS) is an observational research method in which data are collected repeatedly for the same subjects over a period of time. Longitudinal study allows researchers to study changes over time by observing individuals during the study period. In addition, with longitudinal studies, changes can be tracked over time [16].

Manuscript received April 5, 2021
Manuscript revised April 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.4.8

- It can show the difference of opinion among the population more accurately. Statistical variance [4] shows agreement or only total disagreement, while grouping by clustering can show variance in each group of agreement and disagreement within a group.
- Identifying minorities that would not be identified in any other way. If results are presented in an aggregated manner, minority groups often lose their voices. A consistent alternative opinion over the years may suggest some degree of strong conviction in a longitudinal study.
- In a longitudinal study, statistically, similar groups can be identified, making it possible to observe similar opinions over time.

We used longitudinal opinion surveys conducted over several years as a case study to investigate the clustering approach on LS (Longitudinal Studies). They used standard statistical techniques to analyze and got some general conclusions on the population on each survey. According to statistical analysis, the overall result of requirements satisfaction was in bad shape. By applying the clustering approach, we found an important group within the participants who have different opinions from the general conclusion.

This research may help software organizations follow this approach to identify new ideas or critical opinions while conducting surveys within their respective domain. Our research analyzed opinion survey which is usually not processed in data mining (DM). According to our studies, this type of data can contain several potential patterns extracted using a clustering process. It could expose new information and give attention to different opinions.

The remaining is structured as follows: Section 2 contains related work, Section 3 presents the methodology used, Section 4 presents an overview of the LS survey used for the analysis, and section 5 shows the results of clustering. Section 6, a comparative analysis was discussed. Section 7, some issues related to our approach were discussed. Finally, in Section 8, we conclude with some future goals.

2. Related Works

After searching the literature on "data analysis of expert opinion survey" using data mining techniques, it appears that research in the software engineering field lacks in this area. One of the main research practices is to perform opinion surveys on software engineering experts. There are mostly standard guidelines based on simple statistical approaches and some rational investigation methods to analyze survey data.

Kitchenham [4] & [6] described such methods with a recommendation for using enhanced statistical methods like Bayesian analysis. They mentioned that "Bayesian methods are not usually used in software engineering studies" and suggests getting assistance from statisticians. Also, M. Mendonca and N. L. Sunderhaft [7] mention that data mining has appeared as one of the tools to analyze software engineering data. Furthermore, they said data analysts should always consider statistics-based technologies as tools that can improve data mining.

In empirical software engineering, survey research has received less focus on a methodological level than other types of research. In the survey study, Wagner and other authors [8] compiled a list of important and challenging topics from using survey research to develop and test scientific hypotheses to data analysis issues that consider quantitative and qualitative data. Recently, John Moses [9], [10]& [11] introduced a quality prediction model of software build on the experts' opinion using Markov Chain Monte Carlo (MCMC) simulation and Bayesian inference. In general, descriptive statistical procedures, with some hypothesis tests, are used to examine opinion surveys [12].

Hassan and Blom. [3] showed that using data mining on survey data can explore potential opinions that may go ignored and unreported using simple statistical analysis and traditional rationale. Moreover, the utility of Longitudinal Study has been experienced in a few software engineering research studies. The efficiency of test-driven development was examined by Maximilien and Williams [13]. They performed a year-long study with an IBM software development group.

3. Methodology

The clustering technique was applied using the Expectation-Maximization (EM) algorithm on a longitudinal survey dataset using WEKA. This algorithm was used to classify the respondents' opinion records into clusters, where each opinion belongs to a specific cluster based on the likelihood. The clustering process starts by choosing the questions to be analyzed and dismissing the others. Then, clustering begins with a low predicted number of clusters and raises the number of clusters in each stage to classify cohesive and important clusters, which are labeled. When no new significant clusters appear, the process is stopped. The size of identified groups may change in each step, which is may the same group will reappear with a small difference in size, and we extract the group when it is significant. Results after clustering could be exporting and imported for further analysis as they can be processed in other statistical tools.

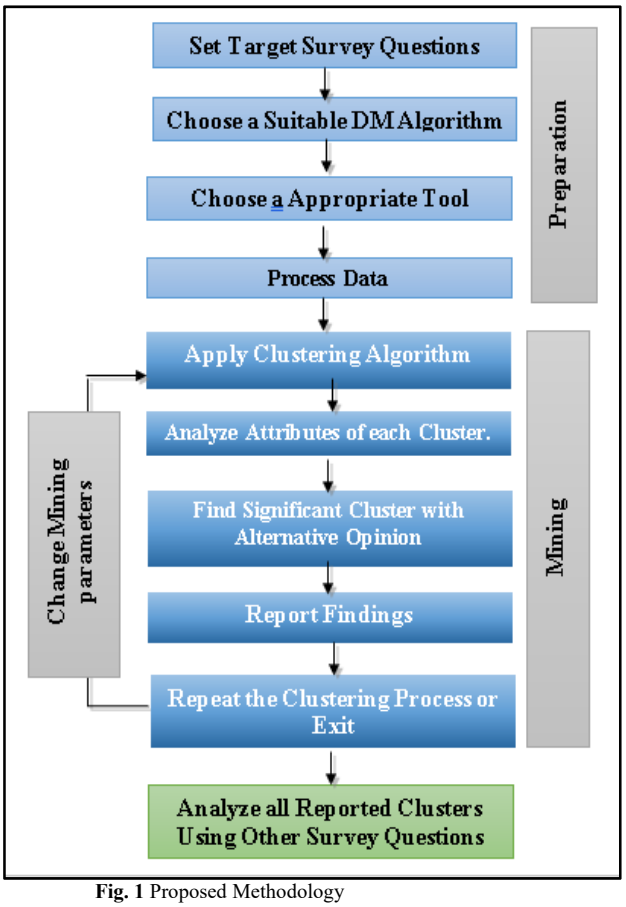

Fig. 1 Proposed Methodology

Fig.. 1 illustrates the proposed methodology. The following explained each step briefly, with the decision taken at each step while applying this approach to those surveys. Findings are presented and explained in the next section.

A. Set Target Survey Questions

Attribute selection is a difficult problem in DM. Usually, each survey attribute can be viewed as a clustering attribute (in our case survey question). By considering only interesting attributes, we can decrease computation cost and improve the extracted data [14]. The number of clustering attributes seems to have no limit; instead, it is determined by the data's nature. Attribute selection methods such as CFS (Correlation-based Feature Selection) and Wrapper Subset Evaluation [14] may help in this respect.

B. Select the Appropriate Data Mining Algorithm

The choice of an appropriate algorithm may influence the results and their analysis. When choosing an algorithm, should be considered the data form. In our research, categorical data is used. We used the Expectation-Maximization (EM) algorithm. Each instance is given a probability distribution that shows the possibility of belonging to one of the clusters.

C. Select the Appropriate Tool

Several open-source data mining platforms include clustering algorithms and visualization of results to applying the clustering algorithm. We used WEKA, a free data mining application.

D. Process Data

Standardize or Normalize the data is a significant step forward to eliminate arbitrary effect because of unit variation of different attributes [15]. To making the dataset consistent with tools, data cleaning or formatting may be required. We had to pre-process the data to be compatible by remove characters such as the quotation mark, comma, and semicolon from CSV files to be opened on WEKA.

E. Apply Clustering Algorithm

The dataset is ready for clustering after standardization. At First, the Clustering Parameter, the number of expected clusters, needs to set, then start the clustering process.
Initially, clustering starts with a low expected number of clusters (we set the value of expected cluster N =2) and afterward increases the number of clusters to identify cohesive and significant clusters in each step and labeled them.

F. Analyze Attributes

We are investigating the interesting clusters at first and comparing them with one another and at the same time with the overall population using the statistics for each survey question. During this process, we could also observe some relationships between the various attributes. A correlation between the attributes used in clustering versus attributes that are not used in clustering may suggest the reason for opinion diversity.

G. Find the Alternative Cluster

When observing the differences between clusters, the choice of indicators affects. Moreover, according to the traditional survey analysis, the average is used to show participants' overall tendency.

H. Report Findings

After analyzing each critical question, a list of findings was generated based on the clustered data and the question's attributes. Also, compared with general conclusions. It may describe a new suggestion or reveal contradictions. Also, it could describe the nature of minority groups.

I. Repeat the Clustering Process or Exit

Those findings can be checked by running the same DM algorithm with different parameter values and observing those findings' consistency. Across the case study,

at some N's value, the clustering process stopped because there were no significant groups that introduce a new opinion.

J. Analyze All Reported Clusters

For the longitudinal analysis, the interesting groups in the first year are used as the main groups. Therefore, identify those groups in the following years by applying the same steps year by year and observing clusters with similar characteristics to the main groups. Analyzed them and studied the change over the study period.

4. Case Study Overview (State-of-Practice in IT-Industry in QTEMA)

QTEMA, a Swedish IT consulting firm, surveyed to analyze the Swedish IT industry's state of practice. The questionnaires form included 21 questions on different aspects of working in the IT sector. It contained background questions and other questions about technical aspects related to the participants' development processes. The study was repeated annually, and in this study, the questionnaires from 2010 to 2013 were used. On average, it was answered by 150 respondents each year.

4.1 Questions Used in Clustering

We focused on questions that assess the requirements process at agile teams. The questions used in our study are listed below:

Q1: Which of the following statements best describes the development method you use most often?

Answer Choices: - Agile / Blend / Traditional / Other

Q2: Has your company/entity a functioning organization and process for working with Requirements?

Answer Choices: - Very low / Low / High / Very high

Q3: Which of the following techniques do you typically use in your development projects? -Review of requirements?

Answer Choices: - Yes / No

5. Applying Clustering in Longitudinal Study Approach

From a single-year survey in the previous study [3], authors have defined an approach using clustering to identify and analyze interesting and minority groups with diverse opinions. The clustering process starts with a low expected number of clusters and then increases the number of clusters. Then identify cohesive and significant clusters in each step and labeled them. The process stops when no new significant groups appear. Because the size of identified groups may change in each step, authors recognize the groups based on their statistical closeness.

In this paper, we used the same approach on longitudinal survey data to analyze and detect opinion differences over a period and reveal the minority whose voice may disappear when analyzing the survey using statistical analysis methods. We applied clustering to a case study to investigate the effectiveness of clustering in extracting groups that represent the minority and have a different opinion and voice.

5.1 Analyzing Case Study

As their analysis, the overall result of requirements satisfaction was in bad shape, but techniques used to review requirements were in good shape, as shown in Fig.2. After applying the proposed approach, we found a significant group inside the participants with different opinions discussed in the following section. This group is used as the longitudinal analysis group, and Fig.3 illustrates the group size over the years.

Fig. 2 Requirements Satisfaction for Year 2010

5.2 The Diverse Group

This group appears in 2010, 2011, 2012, and 2013. Most of the population in this group are confident regarding requirements and techniques used to review requirements. Some distinctive properties of this group:

- Across the years, 87% to 100% of the population suggests higher confidence in their requirement process.
- In 2010, 68% of members followed the traditional development process.

- In 2011,100% of members follow the traditional development process. In 2013, 69% of members followed a blend of the traditional and agile development process.
- 26% of members follow the agile development process in 2010, 29% in 2011, and 27% in 2013.

6. Comparative Analysis

To confirm the effectiveness of the approach used, we compared them with previous statistics in this chapter. Table 1represnents data distribution over the survey years. In terms of requirements, it shows higher confidence compared to the total population - over 75% compared to 40% -45%.

We observed that there are respondents who express satisfaction with the techniques used to review requirements ranging from 53% to 65% during the years of the survey. The alternative group seems to be dissatisfied through the clustering process, which appeared in 2012 by 94%. Table 2 illustrates the group distribution over the years with the percentage for each year.

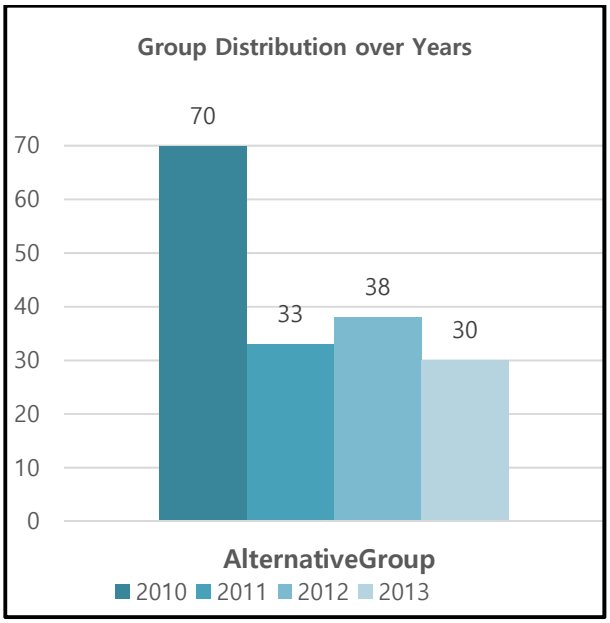

Fig. 3 Group Distribution

Table 1: Data Distribution in General Population- 2010 to 2013

SQ	Answer	2010	2011	2012	2013	2010 (%)	2011 (%)	2012 (%)	2013 (%)
Development method used	**Agile**	21	25	26	27	13.7%	17.6%	18.7%	18%
	Blend	79	71	76	86	51.6%	50%	54.7%	57.3%
	Traditional	46	38	31	31	29.4%	26.8%	22.3%	20.7%
	Other	7	8	5	5	4.6%	5.6%	3.6%	3.3%
Requirement process	**Very low**	16	26	20	19	10.5%	18.3%	14.4%	12.7%
	Low	69	60	54	72	45.1%	42.3%	38.9%	48%
	High	54	45	56	51	35.3%	31.7%	40.3%	34%
	Very high	12	11	8	7	7.8%	7.8%	5.8%	4.7%
Technique to Review Requirements	**No**	55	64	64	66	36%	45.1%	46%	44%
	Yes	95	76	74	82	62%	53.5%	53.2%	55%
	No answer	3	2	2	2	2%	1.4%	1%	1%
Total population		**153**	**142**	**139**	**150**				

Table 2: Data Distirbution in Group 1

SQ	Answer	2010	2011	2012	2013	2010 (%)	2011 (%)	2012 (%)	2013 (%)
Development method used	**Agile**	12	4	14	7	17.14%	12.12%	34.15%	23.33%
	Blend	37	26	25	18	52.86%	75.75%	70%	60%
	Traditional	16	3	1	5	22.86%	9.09%	2.44%	16.67%
	Other	5	0	1	0	7.14%	0	2.44%	0
Requirement process	**Very low**	3	0	5	0	4.29%	0	13.16%	0
	Low	0	0	0	0	0	0	0	0
	High	54	25	25	30	77.4%	75.76%	65.78%	100%
	Very high	11	8	8	0	15.71%	24.24%	21.05%	0
	No Answer	2	0	0	0	2.86%	0	0	0
Technique to Review Requirements	**No**	11	17	36	0	15.71%	51.52%	94.73%	0
	Yes	57	16	2	29	81.43%	48.48%	5.26%	96.67%
	No answer	2	0	0	1	2.86%	0	0	3.33%
Total population		**70**	**33**	**38**	**30**				

7. Discussion

Anderberg showed that, by simple human ability, it is challenging to understand possible partitions from a dataset. He gave an example that to group 25 observations into 5 groups can be huge (exactly 2,436,684,974,110,751) [17]. It is very difficult to divide the population manually and explore their features for even small surveys. The problem becomes more complex in the case of longitudinal studies where the additional data. On the other hand, similar problems in other domains can be solved by clustering.

Clustering methods were used in the current research in a systematic approach to segment the survey population, then recognize important groups with different opinions and analyze them.

The process with examples already was discussed in the previous study [3]. In this chapter, we discuss some of the important factors that may constrain the clustering. Data preparation is a significant step before starting the mining process because some clustering algorithms are not configured to process the missing data, so empty records must be eliminated or filled with appropriate data to differentiate them from others.

Certainly, some variations in questions at each survey are expected in the longitudinal study, which may affect the analysis process. In our research, we focused on a standard set of questions across surveys at each case study.

8. Conclusion & Future Work

Opinion-based surveys in software engineering usually analyzed using descriptive statistical tools which have overall conclusions. The small number of participants may lead to a researcher being excluded from using data mining as an analysis tool, that's why it is rare to use data mining tools in this kind of data.

In the case of longitudinal studies, where minority opinions might fade or resolute over time, the problem becomes more complex. We suggest using a data mining approach to finding the diversity among the different groups in longitudinal studies. Longitudinal survey data can contain potential patterns that a clustering process can identify. It can discover new information and draw attention to alternative opinions. Our main objective in this research is to demonstrate that there are strong alternate opinions in longitudinal studies that can be revealed and tracked over time, and the clustering approach can expose them. The

experimental results of our proposed approach to reveal alternative opinions were provided. In the future, we will propose a systematic process structure that can be used to analyze empirical software engineering data using clustering techniques

Acknowledgment: I gratefully acknowledge Qassim University, represented by the Deanship of Scientific Research, on the material support for this research under the number (3984-coc-2018-1-14-S) during the academic year 1439 AH/2018 AD. The author would like to thank Dr. Mohammad Mahdi Hassan, Department of Computer Science, Qassim University for his supervision of this study.

References

[1] R. C. Henry and J. D. Zivick, "Principles of survey research.," *Fam. Pract. Res. J.*, vol. 5, no. 3, pp. 145–157, 1986.

[2] T. Xie, J. Pei, and A. E. Hassan, "Mining software engineering data," *Proc. - Int. Conf. Softw. Eng.*, no. May, pp. 172–173, 2007.

[3] M. Blom, "Applying clustering to analyze opinion diversity," 2015.

[4] B. Kitchenham and S. L. Pfleeger, "Principles of survey research part 6," *ACM SIGSOFT Softw. Eng. Notes*, vol. 28, no. 2, pp. 24–27, 2003.

[5] B. Kitchenham and S. L. Pfleeger, "Principles of Survey Research Part 5: Populations and Samples," *ACM SIGSOFT Softw. Eng. Notes*, vol. 27, no. 5, p. 17, 2002.

[6] B. A. Kitchenham *et al.*, "Preliminary guidelines for empirical research in software engineering," *IEEE Trans. Softw. Eng.*, vol. 28, no. 8, pp. 721–734, 2002.

[7] M. Mendonca and N. L. Sunderhaft, *Mining Software Engineering Data: A Survey A DACS State-of-the-Art Report*, vol. 4000. .

[8] S. Wagner, D. M. Fernández, M. Felderer, D. Graziotin, and M. Kalinowski, “Challenges in survey research,” *arXiv*, 2019.

[9] J. Moses, "Benchmarking quality measurement," *Softw. Qual. J.*, vol. 15, no. 4, pp. 449–462, 2007.

[10] J. Moses and M. Farrow, "Tests for consistent measurement of external subjective software quality attributes," *Empir. Softw. Eng.*, vol. 13, no. 3, pp. 261–287, 2008.

[11] J. Moses, "Should we try to measure software quality attributes directly?," *Softw. Qual. J.*, vol. 17, no. 2, pp. 203–213, 2009.

[12] T. Gorschek, E. Tempero, and L. Angelis, "On the use of software design models in software development practice: An empirical investigation," *J. Syst. Softw.*, vol. 95, pp. 176–193, 2014.

[13] E. M. Maximilien and L. Williams, "Assessing Test-Driven Development at IBM 5505 Six Forks Road Department of Computer Science," *Proc. 25th Int. Conf. Softw. Eng.*, vol. 6, 2003.

[14] M. Hall and G. Holmes, "Uow-Cs-Wp-2002-02.Pdf," no. April, 2002.

[15] K. Tanioka and H. Yadohisa, "Effect of data standardization on the result of k-means clustering," *Stud. Classif. Data Anal. Knowl. Organ.*, no. October, pp. 59–67, 2012.

[16] T. Hall, "Longitudinal studies in evidence-based software engineering," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4336 LNCS, no. 3, p. 41+, 2007.

[17] M. R. Anderberg, \Cluster analysis for applications," DTIC Document, Tech. Rep., 1973.

Ms Entesar Alanazi received the B.Sc. degree in Information Technology from the Information Technology Department, Qassim University, Saudi Arabia, in 2014. Currently, she is a Computer Science Master student at the Department of Computer Science, College of Computer, Qassim University.