CHITTAGONG UNIVERSITY OF ENGINEERING AND TECHNOLOGY (CUET)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHITTAGONG-4349

(Project /Thesis Proposal)

Application for the Approval of B. Sc. Engineering Project/Thesis
(Computer Science & Engineering)

Date: 10/10/2018
Session: 2017-2018

1.	**Name of the student**	**:**	Md. Shahriar Kabir
	Student ID	**:**	1404083
2.	**Present Address**	**:**	Dr. Q.K. Hall, Room no: 457, Chittagong University of Engineering & Technology
3.	**Name of the Supervisor**	**:**	Rahma Bintey Mufiz Mukta
	Designation	**:**	Assistant Professor Computer Science & Engineering (CSE) Chittagong University of Engineering & Technology
4.	**Name of the Department**	**:**	Computer Science & Engineering (CSE)
	Program	**:**	B.Sc. Engineering
5.	**Date of First Enrolment In the Program**	**:**	18 March, 2015
6.	**Tentative Title**	**:**	Google Play Store Data Mining and Analysis

7. Introduction

With the vast popularity of smartphones, mobile application markets have been growing exponentially in terms of the number of users and downloads. As a report from Statista.com says that by 2020 mobile apps are forecast to generate around 189 billion U.S. dollars in revenues via app stores and in-app advertising [1]. The mobile app market is growing faster than before. The industry is huge and growing daily, and there is no end in sight. Google play store is one of those app stores which holds over 76% market share currently [2].

The main advantage of play store is that they aggregate vital information created by both developers and users. In the app store product pages, developers usually describe and update the features of their apps. Recent studies focused on mining app features described by developers as extracting the features from the app descriptions is essential.

In this paper, I will introduce google play store mining and analysis like software repository mining [3]. App stores usually do not provide source code. However, they do provide a variety of other information in the form of pricing, number of downloaded apps and customer rating. Therefore, I will use data mining to extract feature information, which will be combined with more readily available information to analyze apps' technical, customer and business aspects. So I will try to determine the correlation between price, rating and number of downloads based on this analysis and represent the top features of same category apps.

From Statista.com, there are more than 2.6 million Android apps present on play store as of 1st October, 2018 [1]. Almost 13% of apps are known as low-quality apps. Many apps lack core features as well as extra features that should be present on that apps. So this analysis will help to find out the correlations between different factors as well as the top existing features. This analysis will help app developers to suggest existing top quality features.

8. Background and Present State of the Problem:

Over the past years, some works were done on app store analytics [4]. Most of these works focus on mining a large amount of app store data to derive advice for analysts, developers, and users.

Mark Harman et al. [5] proposed a system which analyzed the Blackberry app store to find correlation and feature extraction. But they experimented on a limited number of apps. Besides, Blackberry app store is outdated now as it contains less than 1% market share nowadays. Timo Johann et al. [6] proposed a system to extract features from app descriptions and app reviews. But this does not show any correlation among the factors- price, download and rating for better perception.

Ning Chen et al. [7] introduced AR Miner computation framework for app review extractions. Lorenzo Villarroe, Gabriele Bavota, Barbara Russo, Rocco Oliveto, Massimiliano Di Penta introduces CLAP (Crowd Listener for Release Planning) for categorizing user reviews. They made the system based on app review. So the app description and the correlations are beyond the scope of this paper. Emitza Guzman et al. [8] proposed a system to analyze user reviews based on sentiment analysis. This gives an automated approach that helps developers filter, aggregate and analyze user reviews.

Al-Subaihin et al. [9] designed a system to measure app similarity based on mined textual features. Lorenzo Villarroel et al. [10] proposed a technique based on user reviews. Besides some more work has been done based on product descriptions in [11 ,12, 13, 14].

Related works that are done have limitations as well as scope to be developed. Very few works have been done for play store apps analyzing based on product description and correlation factors- rating, price and number of downloads. So, to remove limitations and provide an efficient way, I want to propose an app store mining technique which will ensure an efficient way to analyze the google play app store.

9. Objective with Specific Aims and Possible Outcomes:

This proposed work will be carried out with a view to archive the following objectives:

- To design and develop a system that will analyze the correlation among apps' rating, price and number of downloads of different category apps.
- To extract the top existing features of same category apps.
- To suggest developers building more efficient apps.

10. Outline of Methodology/Experimental Design:

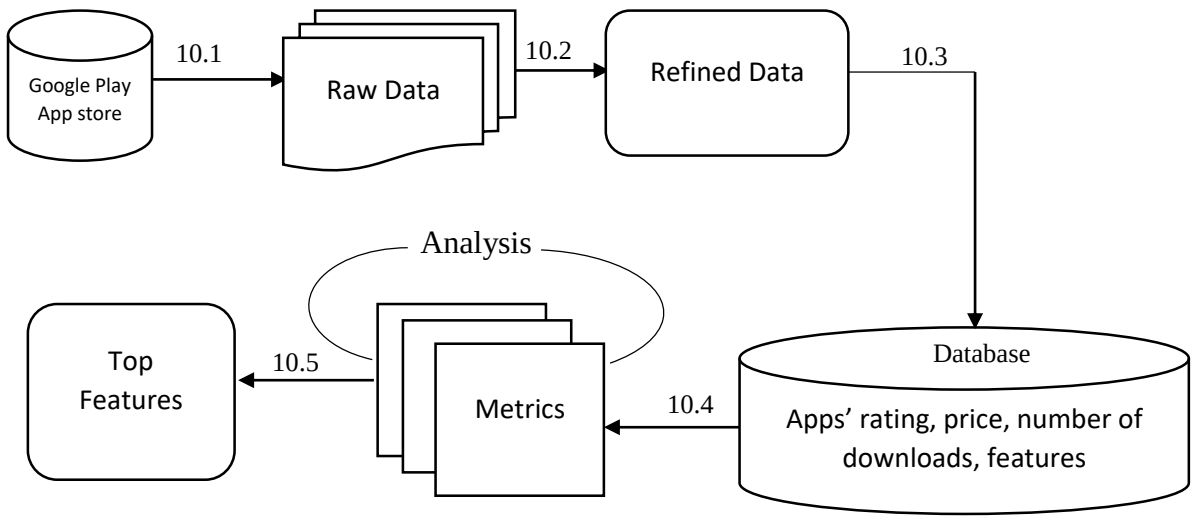

Fig 1: Schematic diagram of the proposed system.

To implement the proposed app store mining and analysis the listed procedures need to follow:

- Raw app data will be collected from google play store with the help of app crawler.
- Raw data will be parsed into app name, category, description etc.
- Features will be extracted from description.
- Refined data and features will be collected.
- Apps will be classified based on their type.
- Correlation of same category of apps will be determined.
- Finally, the top existing features of the same category of apps will be shown.

In my proposed method the entire process of app store mining and analysis will work like this:

10.1 Data Collection:

A web crawling system will be implemented to collect the raw webpage data from the app store. The crawler first collects all category information of the app store and scans each category page to find the list of addresses of all the apps in each category, then it extracts raw data of each app within each category.

10.2 Parsing Raw Data:

The raw data will be parsed according to a set of pattern templates, the attributes of which specify a unique searchable signature for each attribute of interest. For example, in my proposed system apps' name, category, description, price, customer's rating, number of downloads etc. will be extracted.

10.3 Extract Features from Description:

Natural Language Processing technique will be used to extract features from app description. Firstly, all sentences in the description will be tokenized. Then all noise will be removed. Then I need to find a pattern based on parts of speech of those words to determine the most valuable features.

10.4 Metrics for App Analysis:

Some simple metrics will be introduced to compute information about the features of an app. This captures the attributes of a feature in terms of the corresponding attributes of all app that possess the feature. The metrics will be defined with respect to an app database.

10.5 Correlation Analysis:

From the metrics I will analyze the correlation among app's rating, price and number of downloads within same category of apps. Spearman's correlation formula will be used. The formula is-

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

where n is the number of data pairs, and D = rank x_i - rank y_i. Then top features of same category apps will be shown. Thus I will suggest the top features based on this analysis.

11. Resources Required Completing the Work:

Following resources will be required to complete the task:

- Personal Computer
- Windows 10
- Python

12. Cost Estimation

The costs that will occur to implement out proposed system are given below:

No.	Materials/Purpose with specification	Price
01	Pen drive	Tk. 1200
02	Paper	Tk. 500
03	Internet cost	Tk. 4000
04	Drafting	Tk. 500

Total cost: Tk. 6200

13. References

[1] Worldwide mobile app revenues in 2015, "Mobile app revenues 2015-2020 | Statistic", *Statista*, 2018. [Online]. Available: https://www.statista.com/statistics/269025/ worldwide -mobile-app-revenue-forecast/.

[2] "Mobile Operating System Market Share Worldwide | StatCounter Global Stats", *StatCounter Global Stats*, 2018. [Online]. Available: http://gs.statcounter.com/os-market-share/mobile/worldwide.

[3] A. Hassan, "The road ahead for Mining Software Repositories", *2008 Frontiers of Software Maintenance*, 2008.

[4] W. Martin, F. Sarro, Y. Jia, Y. Zhang and M. Harman, "A Survey of App Store Analysis for Software Engineering", *IEEE Transactions on Software Engineering*, vol. 43, no. 9, pp. 817-847, 2017.

[5] Mark Harman, Yue Jia and Yuanyuan Zhang (2012). App store mining and analysis: MSR for app stores. *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*.

[6] T. Johann, C. Stanik, A. B. and W. Maalej, "SAFE: A Simple Approach for Feature Extraction from App Descriptions and App Reviews", *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017.

[7] N. Chen, J. Lin, S. Hoi, X. Xiao and B. Zhang, "AR-miner: mining informative reviews for developers from mobile app marketplace", *Proceedings of the 36th International Conference on Software Engineering - ICSE 2014*, 2014.

[8] E. Guzman and W. Maalej, "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews", *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 2014.

[9] A. Al-Subaihin, F. Sarro, S. Black, L. Capra, M. Harman, Y. Jia and Y. Zhang, "Clustering Mobile Apps Based on Mined Textual Features", *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '16*, 2016.

[10] L. Villarroel, G. Bavota, B. Russo, R. Oliveto and M. Di Penta, "Release planning of mobile apps based on user reviews", *Proceedings of the 38th International Conference on Software Engineering - ICSE '16*, 2016.

[11] J. Davril, E. Delfosse, N. Hariri, M. Acher, J. Cleland-Huang and P. Heymans, "Feature model extraction from large collections of informal product descriptions", *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2013*, 2013.

[12] M. Acher, A. Cleve, G. Perrouin, P. Heymans, C. Vanbeneden, P. Collet and P. Lahire, "On extracting feature models from product descriptions", *Proceedings of the Sixth International Workshop on Variability Modeling of Software-Intensive Systems - VaMoS '12*, 2012.

[13] H. Dumitru, M. Gibiec, N. Hariri, J. Cleland-Huang, B. Mobasher, C. Castro-Herrera and M. Mirakhorli, "On-demand feature recommendations derived from mining public product descriptions", *Proceeding of the 33rd international conference on Software engineering - ICSE '11*, 2011.

[14] H. Yu, Y. Lian, S. Yang, L. Tian and X. Zhao, "Recommending Features of Mobile Applications for Developer", *Advanced Data Mining and Applications*, pp. 361-373, 2016.

14. CSE Undergraduate Student (CUGS) Committee reference:

Meeting No: **Resolution No:** **Date:**

15. Number of Under-Graduate Student(s) working with the Supervisor at present:

_____

Signature of the Student

_____

Signature of the Supervisor

_____

Signature of the Head of the Department