

# Recommending Functions in Spreadsheets from the Fuse Corpus

Shaown Sarker, Emerson Murphy-Hill  
Department of Computer Science  
NC State University  
Raleigh, North Carolina  
Email: {ssarker, emurph3}@ncsu.edu

**Abstract**—The most common form of end-user programming software is spreadsheets. Despite spreadsheets having a large array of functions built-in, most of these functions are underused. We look into modern recommender system technology to address this problem and present a collaborative filtering based function recommender system for spreadsheets. Our evaluation of the recommender system shows that the suggested functions are indeed helpful to real world spreadsheet users compared to a most popular algorithm based system. In this paper, we detail the methodology used to implement the system and the evaluation process. In addition, we outline a road-map to integrate the system as an effective tool in MS Excel.

## I. INTRODUCTION

End-user programmers range from children to professionals including teachers, accountants, administrators, managers, and research scientists [1]. While many of these users are not formally trained software developers, the tasks they perform on a frequent basis are quite similar to that of a professional software engineer's.

Spreadsheets are the most common form of end-user programming software. According to a study in 2005 [2], nearly 23 million Americans use spreadsheets, constituting 30% of the workforce. The use of spreadsheets is also very common in industry for analytical purposes. Winston [3] estimates that around 90% of all analysts in industry perform calculations in spreadsheets. From an end-users perspective, spreadsheet formulas can be viewed as fragments of source code as spreadsheet formulas can contain programming constructs like constants, variables, conditional statements, and references to other parts of the spreadsheet. It is important that the analysts using spreadsheet makes efficient use of the functionality available to them.

In Microsoft Excel, there are almost 350 unique functions, (472 in Microsoft Excel 2013 including the compatibility functions<sup>1</sup>). But, most of these are rarely used. To facilitate these very large number of spreadsheet analysts to use these functions, it is viable that we explore effective function recommendation in spreadsheets.

Not sure how to tie in Fuse here.

Researchers have been assembling and preserving spreadsheet corpora for the purpose of better understanding of end-user activities and designing tools to assist them [4], [5]. In this

paper, we look into Fuse [6], the largest reproducible spreadsheet corpus known till date. We applied a slightly modified version of the user based collaborative filtering algorithm on the spreadsheets in Fuse to get function recommendations for an input spreadsheet. We evaluated the performance of our system against another commonly used recommender algorithm and found out that our collaborative filtering based recommendation system recommends functions in spreadsheet by insert future calculated percentage here.

We first review previous work on recommendation systems in softwares and related work on making spreadsheet formula usage efficient (Section II). We then provide details of the spreadsheet corpus used for our system and the details of the algorithm used (Section III) before describing our evaluation measures (Section IV) and their outcome (Section V). We discuss the various aspects of our system (Section VI) along with the limitations and future works (Section VII) based on our contribution in this paper afterwards.

## II. RELATED WORK

## III. METHODOLOGY

### A. Fuse Spreadsheet Corpus

### B. Modified Collaborative Filtering

## IV. EVALUATION

## V. RESULTS

## VI. DISCUSSION

## VII. LIMITATIONS & FUTURE WORK

## ACKNOWLEDGMENT

## REFERENCES

- [1] A. J. Ko, R. Abraham, L. Beckwith, A. Blackwell, M. Burnett, M. Erwig, C. Scaffidi, J. Lawrance, H. Lieberman, B. Myers *et al.*, "The state of the art in end-user software engineering," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 21, 2011.
- [2] C. Scaffidi, M. Shaw, and B. Myers, "Estimating the numbers of end users and end user programmers," in *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*. IEEE, 2005, pp. 207–214.
- [3] W. Winston, "Executive education opportunities millions of analysts need training in spreadsheet modeling, optimization, monte carlo simulation and data analysis," *OR MS TODAY*, vol. 28, no. 4, pp. 36–39, 2001.
- [4] M. Fisher and G. Rothermel, "The euses spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4. ACM, 2005, pp. 1–5.

<sup>1</sup><https://support.office.com/en-us/article/Excel-functions-by-category-5f91f4e9-7b42-46d2-9bd1-63f26a86c0eb>

- [5] F. Hermans and E. Murphy-Hill, "Enrons spreadsheets and related emails: A dataset and analysis," Delft University of Technology, Software Engineering Research Group, Tech. Rep., 2014.
- [6] T. Barik, K. Lubick, J. Smith, J. Slankas, and E. Murphy-Hill, "Fuse: A reproducible, extendable, internet-scale corpus of spreadsheets," 2015.