Impact of Variations of Similarity and Prediction Techniques on User-based and Item-based Collaborative Filtering Recommendations

FIRST LASTNAME1, Davidson College, USA
FIRST LASTNAME2, Davidson College, USA
FIRST LASTNAME3, Davidson College, USA
FIRST LASTNAME4, Davidson College, USA

Recommender Systems provide users with product/service recommendations in order to save time and effort maneuvering through the plethora of choices available. This work reviews various Similarity and Prediction Techniques used in the generation of Collaborative Filtering User-based and Item-based Recommendations.

Additional Key Words and Phrases: Recommender Systems, Collaborative Filtering, Similarity and Prediction techniques

ACM Reference Format:
First Lastname1, First Lastname2, First Lastname3, and First Lastname4. 2022. Impact of Variations of Similarity and Prediction Techniques on User-based and Item-based Collaborative Filtering Recommendations. In *16th ACM Conference on Recommender Systems, Seattle, WA USA, Sept. 18 - 23, 2022.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Introduce your work effort here. Explain the motivation for undertaking this work and the importance to the Recommender System field. Include a brief discussion of research questions to be pursued/answered in this study. Indicate any novel approaches.

2 RELATED WORK

This section presents the prior work that has been conducted in this area of research. Use citations and the accomplishments of the prior work.

3 THESIS

This section explains the research that will be performed in this work (WHAT), details of the motivation behind this effort (WHY), a discussion of the Hypotheses that will be tested (EXPECTED RESULTS), and a brief description of how the research effort will be conducted (HOW). Use formulae and/or equations to detail the math behind this effort.

4 EXPERIMENTAL DESIGN

This section lays out the assumptions and variations of the study. What datsaets were used. What variables will be varied and the extent to which they will be varied. Etc.

Manuscript submitted to ACM

4.1 Experimental Design/Variations Requirements (36 variations total):

- Recommender System Algorithms: User-based and Item-based
- Similarity method: Distance, Pearson (+optional=Spearman or Jaccard or Tanimoto, etc.)
- Similarity significance weighting: None, n/25, n/50 (+optional= n/75, n/100, or variance weighting)
- Similarity threshold: >0, >0.3, >0.5 (+optional= >0.1, >0.2, >0.4)
- Rating prediction normalization: weighted (+optional=deviation-from-mean weighted or z-score predictions)
- Datasets: ML-100K (run some variations with critics first for testing; must provide Stats report for ML-100K)
- Evaluation: LOOCV (+optional=k-fold CV)
- Metrics: MSE, RMSE, MAE (report all three together)
- Tests of Hypothesis: (+optional=p-values for pairs of means or ANOVA for multiple means)

Note: Any +optional(s) selected **MUST** run with all other requirements.

4.2 Here are some additional programming that will be required:

Code development does not have to be described; neither does computing configuration. I would strongly suggest that you continue to build on the code base we have been developing in class. Your code should be well documented and readable!

(1) **Dialog in main():**

(a) New RML(ead ml-100k dataset). Downloaded the ml-100k zip from GroupLens.org[1], unzip and place the contents into a "ml-100k" folder located in the "data" folder where the "critics" data are stored. Here's some code for RML ..

```
elif file_io == 'RML' or file_io == 'rml':
    print()
    file_dir = 'data/ml-100k/' # path from current directory
    datafile = 'u.data'  # ratings file
    itemfile = 'u.item'  # movie titles file
    print ('Reading "%s" dictionary from file' % datafile)
    prefs = from_file_to_dict(path, file_dir+datafile, file_dir+itemfile)
    print('Number of users: %d\nList of users [0:10]:'
          % len(prefs), list(prefs.keys())[0:10] )
```

(b) New SIMU command: similar to SIM command except that this reads/writes a User-User similarity matrix. This command will need to call the new calculateSimilarUsers() function described below.

(c) Extensions to LCVSIM command:

- Check for a larger len(prefs) and name accordingly (for printing)
- Have the user enter either User-based or Item-based recommendations. Be sure to set algo to correspond to the input .. algo = getRecommendationsSim or algo = getRecommendedItems
- Asking for error metric no longer required since your loo_cv_sim() code will calc/print all three at the end of each run.

[1] https://grouplens.org/datasets/movielens/100k/

- Be sure to run SIM or SIMU commands **before** executing this command.

(2) **Base Recommender code:**

(a) New calculateSimilarUsers() function: Similar to calculateSimilarItems() function but for users instead of items so no transpose required. This function is called by SIMU. **Be sure to set n=100 in the parameter list for this function and for calculateSimilarItems() as well!**

(b) New getRecommendationsSim() function: Similar to getRecommendations() function but uses the user-user sim matrix (created by the calculateSimilarUsers() function) to calculate recommendations (much faster, but not as precise perhaps).

(c) Extensions to loo_cv_sim() function:

- Calls the getRecommendationsSim() function for user-based recommendations and getRecommendedItems for item-based recommendations **via parameter**
- Calc /report MSE, RMSE, and MAE at the end of each run
- Remove the metric parameter since your code will calc all three.
- Note: Keep the **loo_cv()** function as-is.

(d) Extension to data_stats() function:

- Average number of ratings per user, standard dev (over all users)
- Min, Max, Median number of ratings per user

(e) **Optional (extra credit) but recommended:** The getRecommendationsSim() and getRecommendedItems() functions return a rankings list that may or may not contain the item removed by the LOOCV process. To improve performance consider creating new copies of these functions that calc/return a rankings list with ONLY the item removed by the LOOCV process, or an empty list if no recommendation calculated. A change to loo_cv_sim() will also be needed.

4.3 Additional Requirements:

(1) Required sections: Abstract, Introduction, Related Work, Thesis, Experimental Design, Results, Discussion, Conclusion, References

(2) Documentation: Use of LaTeX including bib, **MUST use ACM template provided**. Use Overleaf, TeXShop, MiKTeX„ etc. to build manuscript

(3) Figures: Use of Excel/matplotlib/etc., png/jpg/gif/etc.

(4) Code: Python v3 – Well documented and readable! Use Anaconda3 framework. Use Spyder, VScode, etc., IDE.

(5) Moodle Deliverable (zipped file):

(a) Technical paper pdf file

(b) Folder 1: LaTeX tex file, and everything required to generate the pdf document such as bib file (if used), figures, etc.

(c) Folder 2: Code py file(s) with instructions on how to run the program(s)

(d) Folder 3: Data files used/generated

(6) Honor Code: This is a pledged activity. No sharing of work/content between teams is allowed. Do not represent other?s work as your own.

(7) Teaming: Teams will be assigned

(8) Grading Rubric:

(a) Grade reflects the degree to which the criteria below are satisfied: (A/A-: Exceeded/All/most, B+/B/B-: Many/some, C+/C/C-: Few/lacking, D+/D: Needs major re-work, F: no meaningful content delivered. High quality optional work will be used to potentially improve the final grade)
  (i) Completeness (40%): Meets requirements and specifications
  (ii) Quality of content (60%): Clarity of expression (text, formulae, figures, etc.), methodical approach, meaningful content in all required sections, analytical discussion of results
(b) Your grade will be based on how well the project met the above criteria AND your individual contributions documented (by you) upon project submission to Moodle

5 RESULTS

Provide text, charts, and tables that indicate/explain the results that were obtained by executing the Experimental Design. **You do not need to report on critics dataset results!** These are the experimental variations with Dataset=ML-100k, Rating Prediction=Weighted:

(1) Recommendation Algorithm: User-based Collaborative Filtering, Item-based Collaborative Filtering
(2) Similarity method: Euclidean distance for RS, Pearson Correlation
(3) Similarity significance weighting: None, n/25, n/50
(4) Similarity threshold: > 0.0, >0.3, >0.5
(5) Evaluation metric: MSE, RMSE, MAE

5.1 Required layout of dataset statistics

```
Descriptive analytics data/Chart for ml-100K:
-- Total number of users, items, ratings
-- Overall average rating, standard dev (all users, all items)
-- Average item rating, standard dev (all users)
-- Average user rating, standard dev (all items)
-- Matrix ratings sparsity
-->> Average number of ratings per user, standard dev (all users)
-->> Min, Max, Median number of ratings per user
-- Ratings distribution histogram (all users, all items) figure.

Popular items analytics data/Chart for ml-100K:
-- popular items: most rated (sorted by # ratings)
-- popular items: highest rated (sorted by avg rating)
-- popular items: highest rated items that have at least a "threshold" number of ratings
```

5.2 Required Results Charts

Note: Calc MSE, RMSE, and MAE for each run so you don't have to repeat the run just for a different error metric.

```
1. User-based Recommendation, y-axis: MSE, x-axis: simWtg (none, n/25, n/50)
      Curve1: Pearson, >0 sim
      Curve2: Pearson, 0.3 sim
```

Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim

2. User-based Recommendation, y-axis: RMSE, x-axis: simWtg (none, n/25, n/50)
Same as chart 1 but with RMSE
Curve1: Pearson, >0 sim
Curve2: Pearson, 0.3 sim
Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim

3. User-based Recommendation, y-axis: MAE, x-axis: simWtg (none, n/25, n/50)
Same as chart 1 but with MAE
Curve1: Pearson, >0 sim
Curve2: Pearson, 0.3 sim
Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim

4. Item-based Recommendation, y-axis: MSE, x-axis: simWtg (none, n/25, n/50)
Curve1: Pearson, >0 sim
Curve2: Pearson, 0.3 sim
Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim

5. Item-based Recommendation, y-axis: RMSE, x-axis: simWtg (none, n/25, n/50)
Same as chart 4 but with RMSE
Curve1: Pearson, >0 sim
Curve2: Pearson, 0.3 sim
Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim

6. Item-based Recommendation, y-axis: MAE, x-axis: simWtg (none, n/25, n/50)

```
Same as chart 4 but with MAE
Curve1: Pearson, >0 sim
Curve2: Pearson, 0.3 sim
Curve3: Pearson, 0.5 sim
Curve4: Distance, >0 sim
Curve5: Distance, 0.3sim
Curve6: Distance, 0.5sim
```

6 DISCUSSION

This section analyzes the results obtained and indicates the extent to which the Hypotheses were accepted or rejected.

7 CONCLUSION

Summary of the work completed and the findings that were made. Future work is also included in this section.

8 REFERENCES

This is the bibliography of citations used in this paper. The citations can be made part of this document or kept in a separate file and "imported" (this is the more typical approach for seasoned researchers).

9 ACKNOWLEDGMENTS

Identification of funding sources and other support, and thanks to individuals and groups that assisted in the research and the preparation of the work should be included in an acknowledgment section, which is placed just before the reference section in your document.

This section has a special environment:

```
\begin{acks}
...
\end{acks}
```

so that the information contained therein can be more easily collected during the article metadata extraction phase, and to ensure consistency in the spelling of the section heading.

Authors should not prepare this section as a numbered or unnumbered `\section`; please use the "`acks`" environment.

Example ..

ACKNOWLEDGMENTS

Motto: Alenda Lux Ubi Orta Libertas. Translates to "Let Learning Be Cherished Where Liberty Has Arisen."

10 APPENDICES

If your work needs an appendix, add it before the "`\end{document}`" command at the conclusion of your source document. **==» NOTE: The information in the Appendices below is FYI and does NOT need to be included in your manuscript!!**

A CITATIONS AND BIBLIOGRAPHIES

The use of BibTeX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTeX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before "`\begin{document}`") of your LaTeX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only output the series if the volume number is given [**?**] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible

Table 1. Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Table 2. Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**], a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**? ? ?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], 'YYYYb'-test for prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Boris / Barbara Beeton: multi-volume works as books [**?**] and [**?**]. A couple of citations with DOIs: [**? ?**]. Online citations: [**? ? ?**]. Artifacts: [**?**] and [**?**].

B TABLES

The "`acmart`" document class includes the "`booktabs`" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

C MATH EQUATIONS

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

C.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

C.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

D FIGURES

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Your figures should contain a caption which describes the figure to the reader. Figure captions go below the figure. Your figures should **also** include a description suitable for screen readers, to assist the visually-challenged to better understand your work.

Figure captions are placed *below* the figure.

D.1 The "Teaser Figure"

A "teaser figure" is an image, or set of images in one figure, that are placed after all author and affiliation information, and before the body of the article, spanning the page. If you wish to have such a figure in your article, place the command immediately before the `\maketitle` command:

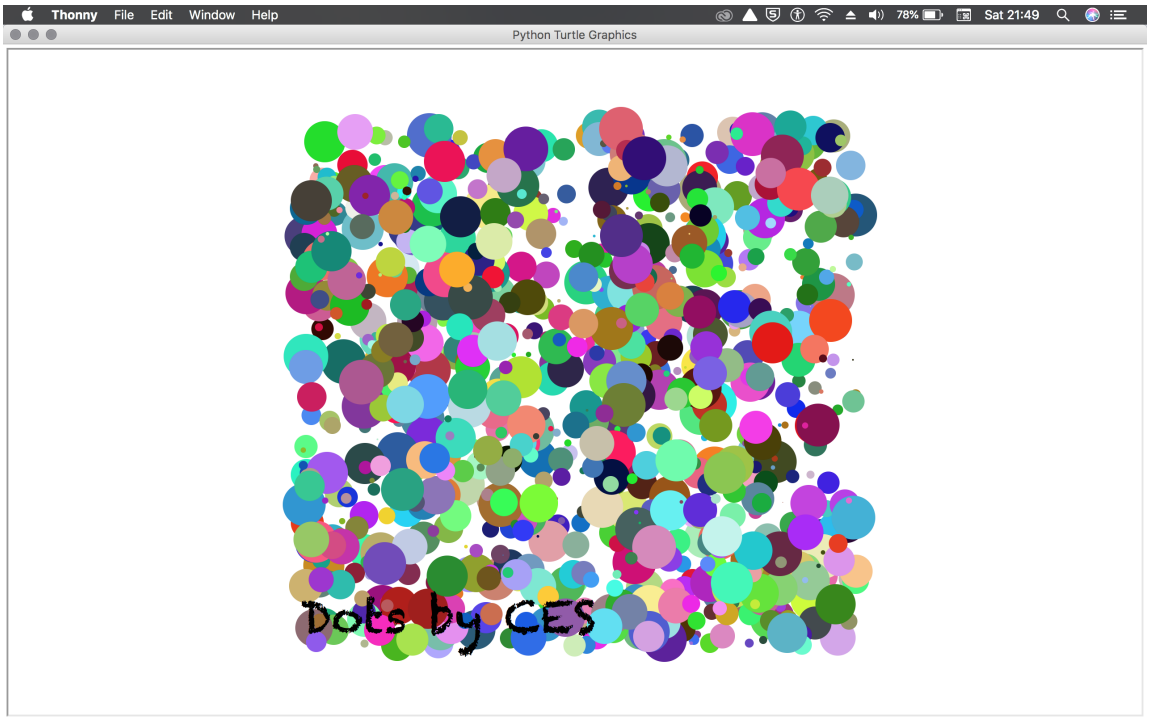

Fig. 1. 1000 Dots, created with Turtle on Python v3. Randomized dot size, color, and location.

```
\begin{teaserfigure}
  \includegraphics[width=\textwidth]{sampleteaser}
  \caption{figure caption}
  \Description{figure description}
\end{teaserfigure}
```

E RESEARCH METHODS

E.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

E.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

F ONLINE RESOURCES

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.