Page & Reel:
The Cross Book/Movie Recommender

Group 17:
Shubashree Baskar, Brian Merritt, Nirav Shah, Haritha Ramesh, Ryan Place, Irtaza Haider

Reading the document:

We have incorporated our literature survey with the 9 Heilmeier questions. Against every reference number - a,b,c (e.g. (2(a,b,c))) indicate the three questions that are answered for every paper – What is its main idea, How is it/or not useful to our project and its shortcomings respectively. Any of these three questions that are not answered along with Heilmeier questions(e.g (2(a,b))) are answered in a separate section named ‘Unanswered questions in the survey’ in the end(In this case 2. (c) <explanation>). A "-" against the number in the unanswered questions means that it has completely answered a),b),c)in the place where it was referenced.

Page & Reel

Introduction-What we are trying to do:

Recommendation algorithms are synonymous with media consumption, though nearly all only focus on one form. The goal of this project is to develop an application that when given either a book or movie recommends them other books and movies that share similarities with an emphasis on cross-recommendation. The similarities that will be evaluated by many factors including, but not limited to, genre, summary, and rating.

How is it done today? What are the limits of current practice?

Tastedive.com is a recommendation tool, constricted to a single genre type and is limited in returned recommendations. A survey[1] of content-based, collaborative and hybrid approaches to current recommendation systems gives us a brief idea of their advantages and their potential shortcomings(1(a,b)). We will take note of a variety of filtering tools when designing the output of recommendations. A tool proven useful involves clustering users and filtering neighbors close to the cluster centers[2]. To reduce mean absolute error (MAE), we will opt for maximizing the number of neighbors to a cluster(2(a,b)). Content-based and collaborative filtering is also known to perform better using a hybrid approach[3](3(a,b)).

What's new in your approach? Why will it be successful?

User are able to view information from either Goodread's or IMDB's API upon clicking a cross-media recommendation. We believe that these features will prove useful to people wanting to use our tool. Decision Tree classifier has been shown to be a more effective method of analyzing the quality than either the *k*-NN or Naïve Bayes method[4](4(a,b)). However, the Decision Tree classifier has been known to overfit in some instances, reducing overall accuracy [5]. Thus, we aim to utilize a newly defined hybrid algorithm that uses Naïve Bayes along with the Decision Tree classifier to select for only the most important subset of attributes(5(a,b),6(b)) [6].

To mitigate the effects of reducing relevance of recommendations, we seek to draw inspiration from Amazon's recommendation system:item-based collaborative filtering[7](7(a,b)) in which instead of matching users with similar interests, we match the similar user's liked books/movies.

Who cares? If you're successful, what difference and impact will it make, and how do you measure them?

Self-learning cross-media tools do not exist for recommendation services[8,9]. The android Smart Movie Recommendation system works on decision trees by evaluating genres based on an API(8,9(a,b,c)). Our tool can have a huge impact if adopted by e-commerce giants such as amazon, who currently use a combination of manual targeting and automated systems[11](11(a,b,c)). This tool can help readers, writers, and viewers discover more media based on their interests and help children become more interested in literature.

The success of this idea can be measured by:

1. Weekly Visits, Survey on recommendation quality
2. Prediction Accuracy

We plan to incorporate temporal evaluation of the same and decide on online/offline method of evaluation(12(a)) whose results may not always agree[12](12(b)). The shortcoming of online evaluation would be randomization and inherent bias(12(c)). Some metrics like serendipity[13] are highly subjective, context driven and are hence difficult to measure(13(a,b,c)).

What are the risks and payoffs?

Some movies and books in a genre are very different from others in the same genre and bias the recommender to give you media that you are not interested in. To overcome this we can implement “coverage” measures as mentioned in[14](14(a,b)). Another risk that isn’t tied to the functionality project is that of the feedback loop that recommendation algorithms can cause in its users[15](15(b)). One way of fixing the feedback loop is clustering users with similar tastes together and also introducing randomness [16](16(b))

How much will it cost?

The cost is zero as we plan to code from scratch and use open source libraries for training Machine learning models. We are currently planning to use the goodreads api and imdb api to get up-to-date information on books and movies and also plan on expanding to other free movie rating databases such as MovieLens[10].

How long will it take?

It should take 6 weeks.

What are the midterm and final "exams" to check for success? How will progress be measured.

The midterm will be a complete recommender for either book or movie with the basic visualization elements included. The final "exam" will be a fully working application that we have described in this proposal and presentation.

Author Contributions:

Every author has contributed equally in writing this proposal and plans to put in equal efforts in the work that will be done.

Visualization:

Give short lists of recommendations with information. Clicking enlarges for further information. Will be explained further in the presentation.

Proposed workflow:

Data collection from GoodReads and IMDB API - First two weeks
Data Cleaning using Open Refine - First two weeks
Data Integration using SQLite/Open Refine - Second and third weeks.
Data analysis by building a hybrid algorithm which was described above - Second, third and fourth weeks
Visualization using D3 - Fifth week
Sixth week - Buffer and overall completion.

Unanswered Questions in Survey:

1. (c)Only some of the extensions suggested in this paper are relevant to us.
2. (c)-Still relies on neighborhood sizes to cluster.
3. (c)The conclusions are drawn from evaluating a system that predicts only movies based on a specific dataset and may not be generalizable.
4. (a)- Performed 10-fold cross validation on overall quality of movies.
   (b) Professional critics provide a higher accuracy on overall quality of a movie
   (c)-Decision Trees can overfit, reducing accuracy.
5. (c)-Sparseness of awards is common if using those as a filtering metric.
6. (a)-show that hybrid algorithms can work just as well or better than non-hybrid algorithms and also explain how these algorithms work with a focus on looking at how the differences of the multiple algorithms work together in the hybrid. (c)-not cross classifying datasets.
7. (c) Algorithm is highly effective in practice and does not mention any shortcomings.

8. -
9. -
10. (a)-Movie Dataset
    (b)-Uses dataset to provide Movie Recommendations
    (c)-Only focuses on movies
11. -
12. -
13. -
14. (a)Not all movies have “predictive” power. (c) Uses a 3-way-split in the choices in the initial feedback on user’s tastes but does not include a “neutral” rating, which we can incorporate.
15. (a)Used simulations to discuss how recommendation systems can institute a feedback loop with the user. (c) only used training data and not direct user interface
16. (a)gives a general overview of the multiple algorithms Netflix uses, how they are being improved up, and the financial outcomes of different decisions. (c) They are only looking at one type of media.
17. Presents a technique-Modified EM to learn a new user profile using information from existing profiles.However, this only works when ratio of related to unrelated user-feature pairs is small.
18. Presents a Bayesian network based approach, to use context for making movie recommendations. However, the performance was tested on sparse data.
19. Presents an approach that classifies short texts in a set of generic classes. This approach is specialized for Twitter ecosystem and works only if context is taken into account, which might not be our case.

References:

1. Gediminas Adomavicius, Tuzhilin A: **Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions**. *IEEE Transactions on Knowledge and Data Engineering* 2005, **17**(6):734-749.

2. Gong S: **A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering**. *JSW* 2010, **5**(7):745-752.

3. George Lekakos PC: **A hybrid approach for movie recommendation**. *Multimedia Tools and Applications* 2008, **36**(1-2):55-70.

4. Johann Schaible ZC, Oliver Hopt, and Benjamin Zapilko: **Utilizing the Open Movie Database API for Predicting the Review Class of Movies**. *KNOW@LOD*, **2015**.

5. Dewan Md.Farid LZ, Chowdhury MofizurRahman, M.A. Hossain, Rebecca Strachana: **Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks**. *Expert Systems with Applications* 2014, **41**(4-2):1937-1946.

6. Hsu K-W: **A Theoretical Analysis of Why Hybrid Ensembles Work**. *Computational Intelligence and Neuroscience* 2017.

7. Greg Linden BS, Jeremy York: **Amazon.com recommendations: item-to-item collaborative filtering**. *IEEE Internet Computing* 2003, **7**(1):76-80.
8. Sang-Ki Ko S-MC, Hae-Sung Eom,Jeong-Won Cha,Hyunchul Cho,Laehyum Kim,Yo-Sub Han: **A Smart Movie Recommendation System**: Springer, Berlin, Heidelberg; 2011.
9. Rahul Katarya OPV: **An effective collaborative movie recommender system with cuckoo search**. *Egyptian Informatics Journal* 2017, **18**(2):105-112.
10. F. Maxwell Harper JAK: **The MovieLens Datasets: History and Context**. In., vol. 2016: ACM Digital Library; 2016.
11. Olivier Chapelle TJ, Filip Radlinski, Yisong Yue: **Large-scale validation and analysis of interleaved search evaluation**. *ACM Transactions on Information Systems* 2012, **30**(1).
12. Joeran Beel MG, Stefan Langer, Andreas Nurnberger, Bela Gipp: **A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation**. *RepSys '13 Proceedings* 2013:7-14.
13. Mouzhi Ge CD-B, Dietmar Jannach: **Beyond accuracy: evaluating recommender systems by coverage and serendipity**. *RecSys '10* 257-260.
14. Nadav Golbandi YK, Ronny Lempel: **Adaptive bootstrapping of recommender systems using decision trees**. *WSDM'11* 2011.
15. Allison J.B. Chaney BMS, Barbara E. Engelhardt: **How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility**. *Cornell University Library* 2017.
16. Carlos A. Gomez-Uribe NH: **The Netflix Recommender System: Algorithms, Business Value, and Innovation**. *ACM Transactions on Management Information Systems* 2016, **6**(4).
17. Yi Zhang JK: **Efficient Bayesian Hierarchical User Modeling for Recommendation Systems. SIGIR.** 2007.
18. Ono C, Kurokawa, M., Motomura, Y., & Asoh, H.: **A Context-Aware Movie Preference Model Using a Bayesian Network for Recommendation and Promotion**. *International Conference on User Modeling* 2007.
19. Sriram B, Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: **Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval**. *ACM* 2010:841-842.