Title	Recommendation Cold Starts: A View of Textual Data
Student name:	Ziyi Guo (MSc Data Science)
Supervisor name:	Dr Nicolas Green

Aims/research question and Objectives

Recommendation system plays an important role in solving information overload, providing personalized services and assisting users in decision-making. There have been a wide range of applications such as E-commerce, E-tourism, social media and digital libraries, while the objects of recommendations are becoming increasingly widespread as well, including commodities, music, movies, books, tourist attractions, friends and applications. The execution of recommendation system algorithms usually relies on a **specific data base**, such as the user's historical behaviour records or user rating data. When a new user or item enters the system, the necessary data is normally not available to establish the association between the user and the item, causing the recommendation system being unable to infer the user preferences and execute recommendation mechanism, resulting in a cold-start problem (Table 1)[1]. This research intends to do surveys on such problem and explore the solutions handling the cold-start recommendations from the view of data processing.

	$Item_1$	$Item_2$	$Item_3$	$Item_4$	$Item_5$
$User_1$	3	5	-	5	4
$User_2$	-	-	-	-	-
$User_3$	2	3	-	-	-
$User_4$	4	-	-	4	3
$User_5$	-	5	-	5	5
$User_6$	3	4	-	5	-

Table 1. Sample of User-Item Ratings

Firstly, cold-start problems can be classified as complete cold-start and partial cold-start[2-3]. A complete cold-start indicates that there is no data in the recommendation system that can be used to construct a *user-item* relationship while a partial cold start indicates that there is really sparse data in the system, such as a small number of items rated by the new user. According to the research object, the cold-start problem can also be classified as new user cold-start, new item cold-start and new community/system cold-start[4]. Such classifications provide various scenarios for a wide range of researches, in which data are presenting varying distributions and all kinds of possible correlations (Fig 1)[5]. This research intends to explore the above circumstances based on data mining and highlight the key status of data in such cases.

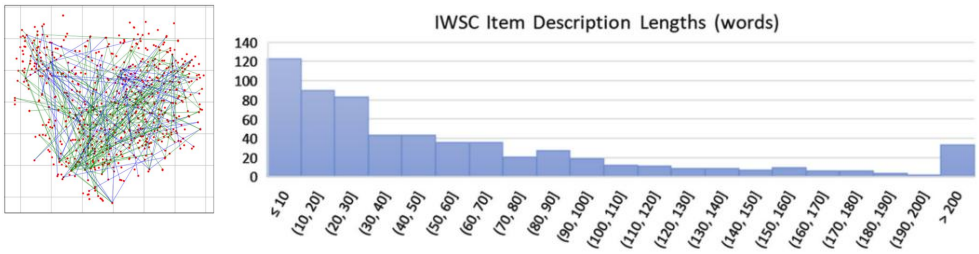

Fig 1. Study on the ISWC Dataset[5]

Secondly, as the core of recommendation systems, current recommendation algorithms are mainly divided into content-based recommendation, collaborative filtering recommendation and hybrid recommendation algorithms[6-7]. Content-based recommendations search for items following past user preference and generate similar items. Collaborative filtering algorithms classify users and items based on their similarity and then process recommendation, in which user-based and item-based collaborative filtering are the most widely used in recommendation algorithms as memory-based collaborative filtering[8]. This research intends to explore the traditional recommendation systems architectures, figure out the root causes of cold start problems and propose data processing methodologies based on specific cold-start recommendation systems.

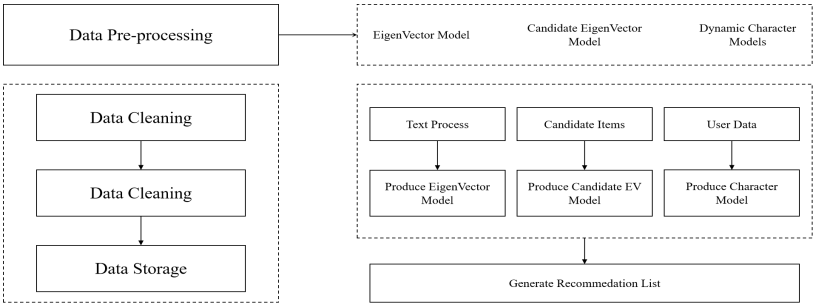

Fig 2. Content-based Recommendation System Workflow

Finally, based the previous study, this research intends to realize the proposed data processing models on various cold start tasks in multiple recommendation scenarios to test and compare the performances of the methodologies. To the best of the work, general operation structure of data processing in cold start problems is scheduled to be proposed. If possible, this research intends to make expansion on the current cold-start recommendation framework from the view of data process.

Summary of proposed research and analysis methodology

Review Methodology

Above all, to demonstrate complete comprehension on the proposed research, the project will start with literature review, concept justification and case study, which is not limited to academic papers but commercial recommendation frameworks from data scientists as well①. Through this review, it will be given an overview of the state-of-art recommendation techniques and the key elements in such subject as well as widely recognized cases of effective solutions on cold-start problems. Most importantly, specific objective for the following months of research, basically including the target dataset (from *Kaggle*) and applying recommendations in the project, are scheduled to be decided through review stage.

Analysis Methodology

Secondly, efforts are scheduled to be made in the analysis of chosen recommendation systems and target dataset. For the recommendation framework, clear and straight flowcharts of the systems can be drawn in *XMind*, based on which comprehensive analysis is carried out on the specific recommendation workflow. As for data visualization and analysis, tools provided with libraries in *Python*, such as *Matplotlib*, *Seaborn*, and *Pandas*, offers great functions to reveal the distributions and the correlations of data visually and numerically, which is the foundation of data processing as well as model proposal. Following the analysis above, the main body of data processing models will be proposed theoretically.

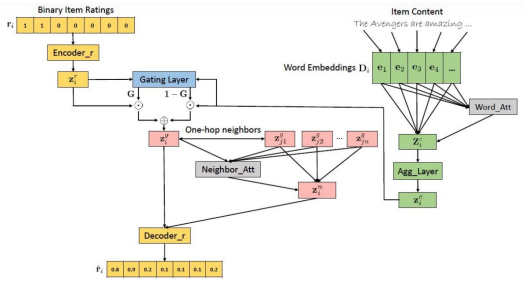

Fig 3. Recommendation System GATE[9]

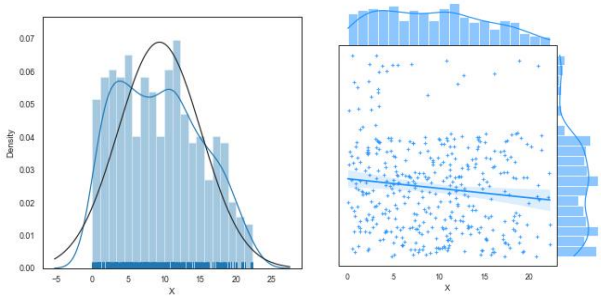

Fig 4. Visualization Samples from *Seaborn*

Implementation Methodology

Next, based on initial research and basic analysis of data and system, the project comes to the crucial stage of proposal implementation and model realization. For data processing, the major priority is to take the way of fitting proposed model into consideration and then, a series of data cleaning according to redundancy situation and data embedding based on existing mature linguistic Encoders and Decoders can be processed to transform the target data into the needed formats. For model implementation, the crucial consideration is to make corresponding predictions, namely recommendation lists, based on the processed input data. Specifically, this research intends to apply transfer-learning based pre-trained models in *Pytorch libraries*, combining with proposed theories (eg. semantic models); while for sure the frozen trained models will fit the new dataset through training with the fully-connected layers. Despite of the initial design at present, formal experiments will be processing depend on circumstances.

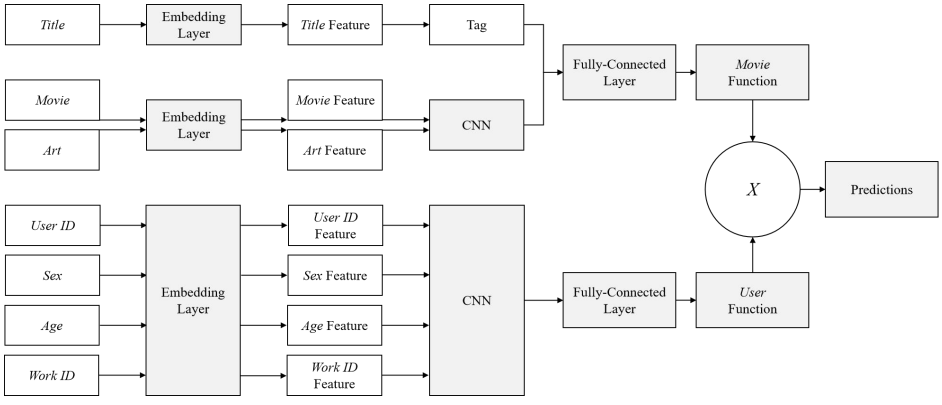

Fig 5. Deep Learning Based Recommendation Framework

Evaluation Methodology

Moreover, suitable evaluation method need to be designed to comment on the performance of proposed data processing and recommendations. Except for regular metrics, such as CTR, accuracy rate, recall rate, NDCG, MAP[10], this research intends to set more humanized and interactive evaluation mechanism, such as browse depth and long-term feedback, based on specific users and items[11] to fully describe the model performance.

Ethic, Health and Safety

Finally, this research may be concerning ethical issues with data privacy and commercial confidential. Therefore, the experiments will be carried out based on public datasets and common recommendation architectures, which are allowed to be used in academic researches, in order to avoid constituting an infringement. Basically, all the researches and experiments will be done online and remotely and thus having no healthy and safety issues.

Research plan – Gantt chart or Pert chart

Gantt Charts

Firstly, this research intends to carry out works including both implementation and report following the timescales as below.

	Project Stage & Tasks \ Timeline	Week 1 13th June	Week 2 20th June	Week 3 27th June	Week 4 4th July	Week 5 11th July	Week 6 18th July	Week 7 25th July	Week 8 1st August	Week 9 8th August	Week 10 15th August	Week 11 22nd August	Week 12 29th August
	Theoretical Study												
1	Literature Review												
2	Cold-Start Case Study												
3	Target Datasets Preparation												
4	Recommendation Scenario												
	Data Analysis and Visualization												
5	Recommendation Framework												
6	Data Visualization												
7	Data-processing Model Proposal												
8	Data-processing Model Pretest												
	Recommendation Implementation												
9	Architecture Implementation												
10	Pre-processed Data Embedding												
11	Recommendation Model Test												
	Recommendation Evaluation												
12	Performance Evaluation												
13	Justification and Retraining												
14	Recommendation Retest												
	Model Function Expansion												
15	Reflection and Critical Review												
16	Generation to Wider Tasks												
	Project Completion												
17	Experimental Data Collection												
18	Dissertation Organization												

Table 2. Initial Project Outline of Gantt Chart

Pert Charts

In addition, different objectives are mutually relevant as below.

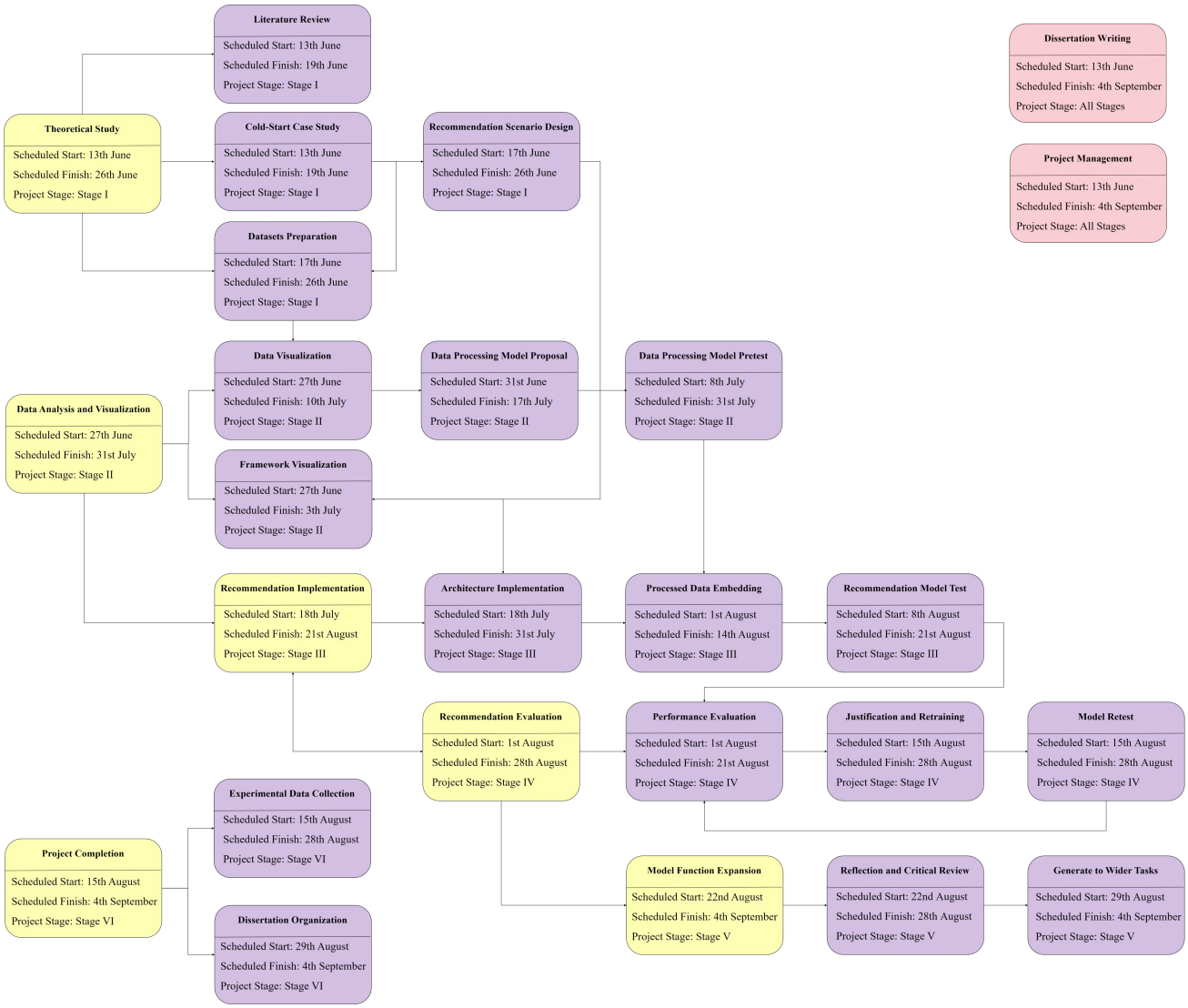

Fig 6. Pert Chart of Project Objectives

Ethical Statement and Data Management Plan

Ethical Statement

Recommendation systems are designed based on human demand and user data and widely applied in human lives. It is stated that this research will cause no user privacy breach and the technical research results proposed in the work will be informative and voluntary-optional to all users included in the built systems.

Health and Safety Statement

As the all the study and experiments are implemented online and remotely, this research presents no obvious healthy and safety issues. Due to the possible demand of GPU resource, it is stated that experiments carried out in the Building 16 will follow the Laboratory Safety Regulations on electric facilities and pay attention to the right use of computing resources.

Environmental Impact Statement

As far as recommendation systems are concerning virtual user and items on the Internet, this research has no obvious environmental impact.

Data Management Plan

As is mentioned before, recommendation system architectures are built based on large amount of user and item data. It is stated that all the data used in this research will be collected from public datasets available for academic research and stored for research aims only. Moreover, the recommendation models and relevant test data will be uploaded to the university at the end of the project.

Ethical aspects

Generally, this research aims to optimize recommendation systems based on handling with the cold start problem caused by data sparsity from the view of data. As the research subject, recommendation systems are designed highly related to human demand of detecting useful information. However, with the rapid development of recommendation systems, related ethical issues abound constantly.

Firstly, recommendation systems collect user behaviour and recommends objects of interest to users based on their behaviour. If the user browses content that is private and not willing to be known, then the systems still gain access to these particular hobbies and therefore recommends similar content, which is a privacy breach. As a sort of user privacy, the user's preferences are acquired by the recommendation system, through which enterprises may use these preferences for unethical commercialization (aka big data killing), and thus harming users.

Secondly, issue named *Filter Bubble* exists widely in personalized recommendations, which is described that recommendation algorithms suggest possible content of interest to users based on a variety of information, but trap users in their own cultural and cognitive bubbles as receiving no contrary views in the long run. As a result, users are constantly recommended with their favorites while quality content not exposed to users are excluded and gradually marginalized, leading to the cost reduction of user decision making and the appearance of *Filter Bubble*.

To implement mature techniques of modern recommendation systems, researches need to pay more attention to 1) the human factor, to make recommendations more warm and emotional with better understand of social needs, 2) the regulations on data quality and protections of user privacy, 3) the ability to access users to differentiated information for learning and growth, 4) the rights of self-control of user and autonomy over their choices. The points illustrated here will always be important considerations during this project.

Commercial aspects

The value of recommendation systems is reflected in both improving user satisfaction and making business profits. For most enterprises, the ultimate goal of improving the user experience is also to capture business value. Given the huge commercial value of recommendation systems, almost all companies define the value of recommendation systems in terms of making more commercial profits, with the short-term commercial value of recommendation systems as the most important goal. The efforts of current researches optimizing recommendation functions are expanding its commercial value constantly.

However, a number of enterprises ignore laws, regulations and moral bottom line and use recommendation systems as tools for enrichment and benefit by recommending obscene content to increase user stickiness. Also, item providers exploit the features and loopholes of the recommendation algorithm to attack the system and boost their product rankings they offer, maliciously increasing the weight of their subject matter and gaining more attentions. As the platform that provides the recommendation algorithm, it needs to constantly study the countermeasures and adjust the algorithm by correcting the data and optimizing the recommendation algorithm to assess the real value and true popularity of the subject matter, which is a long-term technical confrontation process.

As the root of a commercial phenomenon, scientists working on related technologies have the obligation to be aware of such issues ahead of enterprises, in which 1) the pursuit of legitimate commercial value with normal commercialization to obtain profits is the basis of a company's operation but maliciously harming users and causing negative social impact are unacceptable, 2) recommendation system should be user-oriented and could provide quality recommendations and relaxing interaction experiences, avoiding user being immersed in it at the meantime, 3) the long-term development of recommendation systems concerning users, merchants and platforms need to be taken into consideration to build positive-sum game system and achieve win-win situations.

Legal aspects

According to the previous discussion, in the applications of recommendation systems, issues are existing as the misuse of commercial recommendations to affect normal user behavior and commercial ecology, the basic functions of recommendation systems curing user thinking and violating user privacy, as well as the technical limitations to handle data quality problems. All these issues are indicating the necessity of the legal framework and restriction regulations on the use of commercial recommendation systems, which is a wake-up call for relevant academia researches.

The operation of recommendation systems should obey the Market Regulation Act, providing items and services of good quality to users, obey the Information Protection Act to prevent malicious commercial attacks and protect user privacy, and obey the Bill of Rights to give users the right of self-selection and wide exploration space based on diversified recommendations.

Recommendation systems are sought after by the business community because of their huge business value, but business value is only a part. The practitioners of recommendation systems could learn to be more humanistic. Recommendation systems do not have values per se, but humans have given them lives and value proposition, and need to overcome many problems and promote more mainstream social values, which requires humans to better integrate their own values into the systems through rules, algorithms and even human intervention.

Specifically for this project, experiments concerning data will be carried out based on the respect for user privacy and promote general researches; while experiments concerning recommendation architectures will fully consider the modern legal framework for such systems. This research will adhere to a responsible attitude towards humanity and society.

References

[1] Gonzalez Camacho L A, Alves-Souza S N. Social Network Data to Alleviate Cold-Start in Recommender System: A Systematic Review [J]. Information Processing & Management, 2018, 54 (4)：529-544.

[2] Wei J, He J H, Chen K, et al. Collaborative Filtering and Deep Learning Based Recommendation System for Cold Start Items [J]. Expert Systems With Applications, 2017, 69：29-39.

[3] Hernando A, Bobadilla J, Ortega F, et al. A Probabilistic Model for Recommending to New Cold-Start Non-Registered Users [J] . Information Sciences, 2017, 376：216-232.

[4] Bobadilla J, Ortega F, Hernando A, et al. Recommender Systems Survey [J]. Knowledge-Based Systems, 2013, 46：109-132.

[5] Ralph, D., Li, Y., Wills, G. et al. Recommendations from cold starts in big data. Computing 102, 1323–1344 (2020). https://doi.org/10.1007/s00607-020-00792-y.

[6] Ye X, Yuan P,Guo X,et al. Collaborative filtering recommendation algorithm based on user interest and project cycle [J]. Journal of Nanjing University of Science and Technology, 2018, 42 (4)：392.

[7] Fu M,Qu H,Yi Z. A novel deep learning-based collaborative filtering model for recommendation system [J]. IEEE Transactions on Cybemetics, 2018：1-13 ,doi：10.1109/TCYB. 2018.2795041.

[8] Tseng G, Lee W. An enhanced memory-based collaborative filtering approach for context-aware recommendation [C]//Proceedings of the World Congress on Engineering(WCE 2) ,2015,1：1-5.

[9] Ma, Chen, et al. "Gated attentive-autoencoder for content-aware recommendation." Proceedings of the twelfth ACM international conference on web search and data mining. 2019.

[10] ZHU Y X, LV L Y. Evaluation Metrics for Recommender Systems[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-176.

[11] ZOU L X, XIA L, DING Z Y, et al. Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems [C] // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 2810-2818.

Related Resources

① https://data-flair.training/blogs/data-science-at-netflix/