UNIVERSITY OF Southampton
School of Electronics and Computer Science

Content-Based Cold Starts in Recommendation System

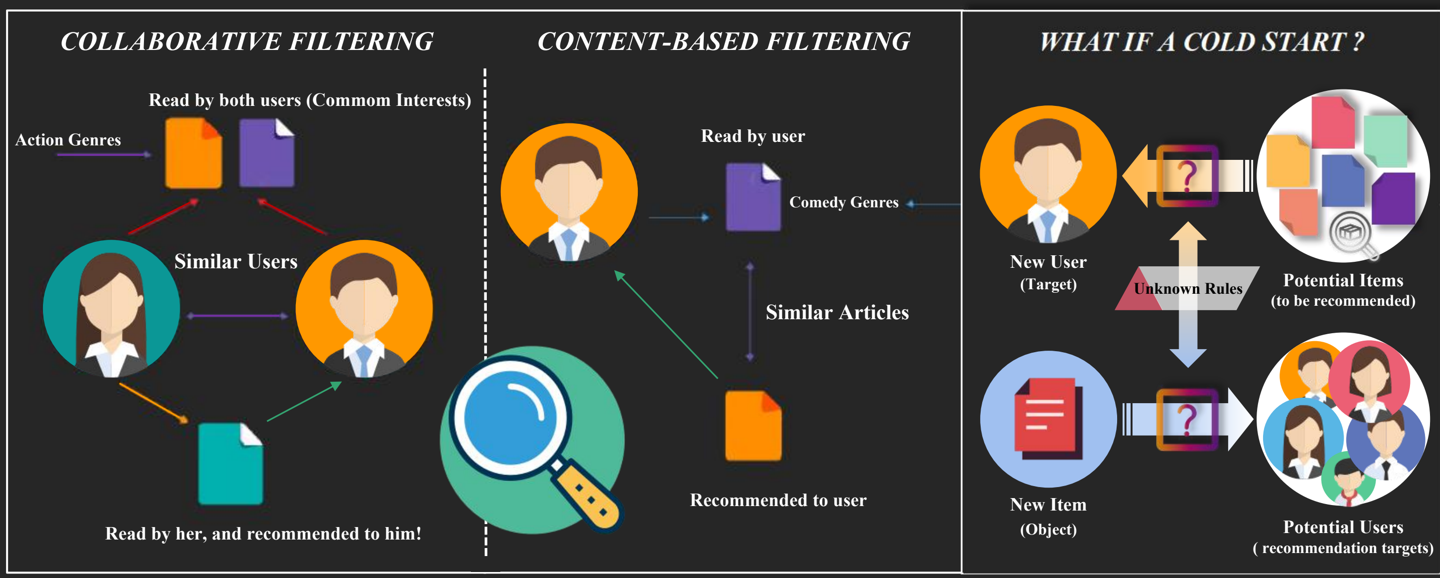
Author: Ziyi Guo (MSc Data Science)   Supervisor: Dr Nicolas Green

Background

Recommendation systems aim to identify the most relevant items from a collection to fulfil a user's search intent. By using known information about the user and items, including **interaction data**, **content data**, or both, recommendation systems are producing personalised results without requiring a query, which means the user can receive recommendations with no direct involvement.

Collaborative recommender systems aggregate user interactions to find similar users and recommend the item they liked. Common techniques include collaborative filtering, reducing the dimensionality of the sparse user-interaction matrix based on a **user's past interactions**. However, **time-sensitive items** record may **not remain valid** long enough; also it may result in **positive feedback loops**.

Content-based recommendation systems recommend items based on **similarity to content information on items**. Usually a representation for a target user is generated based on item description embeddings. Content-based systems are **less dependent on** items **with extensive interaction histories** as new items **are recommended based purely on content similarity**, but the past interaction records of users are still required for comparison.

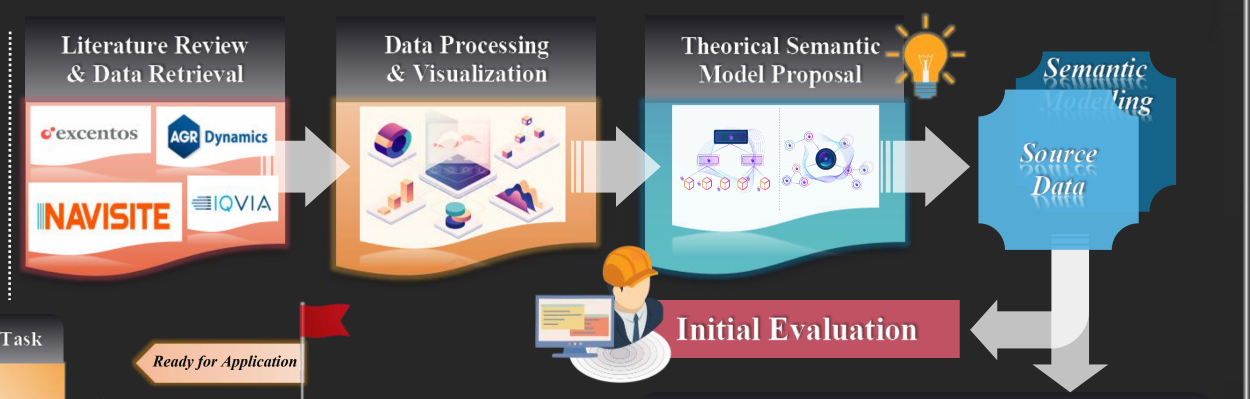


In the real-life case of recommendation system, how-ever, there is **no behaviour record** for the **new users**, which illustrates the problem of ***cold start*** scenarios. The problem is commonly divided into ***item-wise (new item)*** and ***user-wise (new user) cold starts***, which are mainly addressed by content-based and hybrid recommendation systems and can be hard relying largely on knowledge bases and ontologies, or tags and categorisations.

Due to the high barrier to entry recommendation system, researchers study time sensitive cases and propose high performance techniques **pairing labelled and unlabelled items based on the similarity of their contents**, which can be generalised to cold-start problems.

Project Aims & Process

Facing the large labelled data requirement of most existing recommendation techniques, techniques that **function well with few labels** are highly desirable, as **obtaining well labelled data is actually problematic**. In this project, based on initial investigations on data labelling problems and the limitations of existing recommendation systems, considerable work is scheduled to explore new techniques for **providing personalised recommendations** for **new users** in highly chanllenging datasets where **few labels** are available for most **items**, and **no labels** for new **users**. Based on user and item information in the form of **natural language descriptive text** to expand the **sparse relationship distribution**, the project intends to address both ***user-wise*** and ***item-wise cold starts***.



Feasibility Discussion & Conclusion

All in all, the project schedules the following workflow:

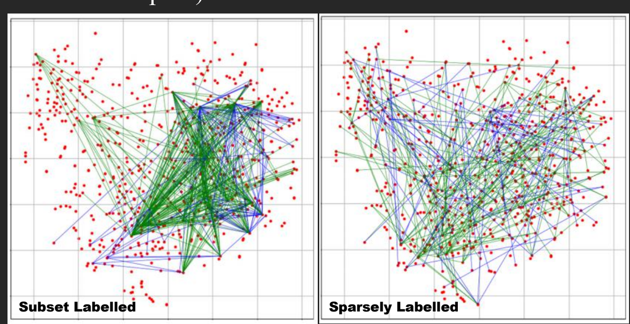
- study current recommendation systems and extremely sparse data labelling problem
- process practical data visualization on to representative datasets to **explore the data structures and features**
- propose several semantic models to redefine **content-based correlations of textual data**

Furthermore,based on the proposed semantic models:

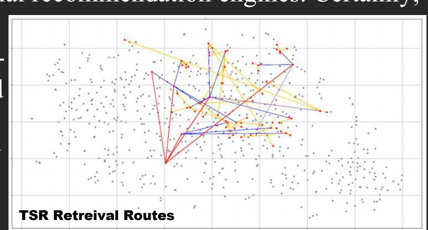
- carry out pre-experiment on subset of data to check the availability of the models as the initial evaluation
- apply the semantic model to the dataset
- make use of **current linguistic encoders** as upstream embedding model and **test a number** of them
- evaluate performance in various tasks and systems.

Before this project proposal came out, related work has been proposed on the case of supply chain on the Isle of Wight, based on the ***Isle of Wight Supply Chain (IWSC) dataset***, which consists of varying length text descriptions of 630 companies on the Isle of Wight taken via web scraping from the websites, containing general textual descriptions of enterprises' market role, activities and products, etc.

According to the initial study on the ISWC dataset, researchers work out the data structure and draws the text-label reflection (data correlation path) in different tasks.



Further, on applying a proposed ***Transitive Semantic Relationships (TSR)*** model, researchers show the retreival routes for any possible query in potential recommendation engines. Certainly, the TSR model is tested on recommendation system cold starts tasks and achieves considerable results on both accuracy scores and programing interpretability.



In spite of the achievements and the feasibility that past researches have provided, there still remains problems to be solved, which is the start point and expected outcome of the project:

- Tasks concerning ***multi-user scenario*** can be expressed as ***multi-class classification*** because of the multi-layer structure of users network, which is more general in recommendation systems, and yet to be discussed.
- ***Content-based recognition*** of similar nodes and target nodes, can be wide in approach (eg. Clustering), and need to be applied to improve performance.
- The choice of ***upstream embedding models*** influence the performance of proposed semantic models, and is thus worth abundant experiment.

In conclusion, starting with ***IWSC dataset*** and ***TSR model***, this project will go on to other databases for new data-sets, study the provenance and structure of data, propose new content-based semantic model and carry out sufficient evaluation, making attempts to provide considerable suggestions for cold recommendation systems.

References

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

Fuzheng Zhang, et al. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 353–362. DOI:https://doi.org/10.1145/2939672.2939673

J. Yuan, W. Shalaby, M. Korayem, D. Lin, K. AlJadda and J. Luo, "Solving cold-start problem in large-scale recommendation engines: A deep learning approach," 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1901-1910, doi: 10.1109/BigData.2016.7840810.

Musto, C., Semeraro, G., de Gemmis, M., Lops, P. (2016). Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems. In: , et al. Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science(), vol 9626. Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_60

https://www.navisite.com/blog/open-source-vs-commercial-database-systems/

https://news.cision.com/agr-dynamics

https://www.iqvia.com/?utm_medium=cpc&utm_source=google&utm_campaign=2021_NextBestGoogleAds_GBU_TCS_MA&utm_content=126031534540&utm_term=content%20based%20recommender

https://www.excentos.com/de/