

Recommendation Systems on the Web

Jordan Simmons
Indiana University Bloomington
jomsimm@iu.edu

ABSTRACT

Recommendation Systems are being used all over the web. There are different popular techniques that are being used in modern systems. Some of the larger well know companies are using this technology very well. This material is an overview of some techniques, state of the art systems, and challenges and limitations of Recommendation Systems.

KEYWORDS

i523, hid336, Recommendation Systems, Big Data

1 INTRODUCTION

Recommendation systems (RS) leverage big data in order to create value for both businesses and customers. “The goal of a recommender system is to generate meaningful recommendations to a collection of users for items or products that might interest them.” [7]. RS are effective for a variety of industries and products which can range from a product in a store, a news article on a site, or a search query. RS is beneficial to businesses and customers by increasing metrics such as revenue and customer satisfaction [2]. Many online platforms are starting to use RS to analyze their data. In order to gain a better understanding of RS, general analysis of modern techniques, companies currently using RS, and challenges and limitations within the field will be covered.

2 RECOMMENDATION TECHNIQUES

Three common RS techniques would include content-based, collaborative, and hybrid recommendations [1]. Other techniques exist, but these three are the most widely used today. In order to determine which technique is best depends on the recommendations to be made, and the data used to make them. Many times, the hybrid approach is used because there can be limitations with other approaches [1]. Overall, it is best to understand a little bit about each technique before choosing which is best.

2.1 Content-Based

Content-Based RS recommend items to users by using descriptions of items and how the user is profiled based on their interest [8]. Items are classified by different characteristics, attributes, or variables [8]. Once items are classified, they can be grouped together based on the classifications. Users are classified by data they provide to the system, and/or the data collected by interacting with the system.

Content-Based RS are commonly seen on web applications and E-commerce sites. These types of systems readily track and monitor almost all user activities. Typically a user has an account with the system, which is where data was voluntarily provided. With this data, users can be classified easier compared to a customer walking into a brick and mortar business.

2.2 Collaborative Filtering

“Collaborative Filtering is the process of filtering or evaluating items using the opinions of other people” [10]. This type of RS is commonly seen on systems where an item can be rated by a user. With this technique, user rating are collected and stored from a user for an item that they have used or purchased. The ratings from the user are then compared to other users that have rated the same item. For example, person A buys items 1 and 2 and rates each item highly. Then, person B buys item 1 and rates it highly. Since person A and B both bought and rated item 1 highly, the system would likely recommend item 2 to person B. On the contrary, if person B gave item 1 a low rating, the system would not likely recommend item 2 to person B. This concept uses the assumption that “people with similar tastes will rate things similarly” [10]. This assumption may not be true in all cases, but it is a good base for RS to start learning users interests, and recommend items based on those interest. With this technique, the more ratings that the systems has collected per item, and the more ratings given by the user, the easier it is for that system to make recommendations to that specific user.

2.3 Hybrid

Hybrid RS takes two or more techniques and combines them to improve performance and reduce limitations that a single technique might have [3]. In most cases, collaborative filtering is used with one or more of the other techniques to improve performance. Other techniques that are used and not discussed include Demographic, Utility-Based, and Knowledge-based recommendations [3]. The hybrid approach narrows down items with one technique, and then uses another technique on that subset of items to make a more accurate recommendation. Determining the best hybrid system depends on the specific business case, and the data used to make the recommendation.

An example of a hybrid approach would use collaborative filtering and the content-based methods described above. For example, if User A is interested in baseball. The system would use the content-based approach to narrow down all items that are classified as baseball items. From this subset of baseball items, the system could then use the collaborative-filtering approach to find the items with ratings from other users which will be user group B. The system would then find all item ratings from user group B and compare those item ratings to person A. If there are any users in group B that have similar likes to person A, the system would likely recommend the baseball items to person A that person B has previously rated highly. This is a high-level example of how a hybrid RS would work. Real world examples are more complex than this example, and use large amounts of data.

3 MODERN SYSTEMS

Two well known companies that are currently using RS are Netflix and Amazon. These two companies have huge customer bases, in which they collect data on. The data collected within these sites and how they utilize it to generate suggestions to their users is what makes these companies have successful advanced recommendation systems.

3.1 Netflix

Netflix is an internet based company that offers a variety of movies and television shows. Netflix had a problem of customers sorting through its large selection of movies and shows, and eventually losing interest which resulted in abandonment of their services [5]. Over the years, Netflix has created and continually developed new RS algorithms which they claim saves them more than one billion dollars per year and a monthly turnover in the low double digits [5].

Netflix does very well at recommending movies and shows to its users. They have incorporated different strategies to collect data from users which is the base of their RS. Data is collected in the form of customized search, video ratings, continue watching feature, amount of time spent watching and other user activities [5]. Using the data collected from these features, Netflix can recommend top rated, now trending, and videos based on user interest, which is very appealing to the user when there are so many selections to choose from.

3.2 Amazon

Amazon is an online store that sell a large variety of products. Amazons RS provides recommendations for millions of customers from a catalog that has millions of products. [11]. Instead of comparing customers to customers, amazon uses an item-based collaborative filtering approach. This process finds items that were bought together with unusually high frequencies, and uses these relationships to recommend products to customers based on what they have purchased in the past [11]. With this algorithm, Amazon is providing a unique experience to every user and helping them find products they may not have found. Since the initial launch of this algorithm, it has "been tweaked to help people find videos to watch or news to read, been challenged by other algorithms and other techniques, and been adapted to improve diversity and discovery, recency, time-sensitive or sequential items, and many other problems." [11]

4 CHALLENGES AND LIMITATIONS

As with most technologies, RS has its challenges and limitations. It is hard to speak of this topic without speaking about the questions "more data usually beats better algorithms" [9]. This quote has raised controversy about which of the two actually produce better results. In most cases, there are many different variables to consider when answering this question.

4.1 Limitations

With complex systems, there can be many variables that cause issues that limit full capabilities of that system. Specifically, in RS, some of these limitations include cold start problems, data sparsity,

limited content analysis, and latency problems [6]. These limitations seem to be more data related rather than the actual techniques and approaches of the technology being used to analyze that data. When there is no data for a new user, it is hard for RS to create suggestions for this user. The system has no data on the users activities or what interests that user has. When a new item is added to a system, there are no reviews and no data collected with the interaction of user for this particular item. On the other hand, too much data can become redundant. At this point gathering more data will have limited gains.

4.2 Cross-Domain Recommendations

Cross-Domain recommendations aim to "leverage all the available user data provided in various systems and domains, in order to generate more encompassing user models and better recommendations" [4]. Every day the amount of data being collected increases. This data is being collected from different sources. Cross-Domain RS could use data from different sources, which could make up for some of the data caused problems. An example of a Cross-Domain recommendation would be Netflix using data from Facebook to help recommend movies to a new user. Using data from various systems like this would bring up new issues like privacy and security, but if systems started working together and sharing data there could be benefits for both systems.

Cross-Domain Recommendations help with domain specific data issues. Two different systems may have different ways of collecting and organizing data. If system 1 collects variables A, B and C, and system 2 collects variables A, B, and D, each system has information that the other system does not have. This is where sharing the data between systems could have benefits for both systems. In doing this, each system is not only benefiting from more data, but different and perhaps better data. This would also require using better algorithms to analyze the different sets of data. Depending on the system, more data can be more beneficial than better algorithms. In terms of scalability, gathering more data that is different from what is currently being collected, and using better algorithms along with the different data could potentially maximize recommendations for that system.

5 CONCLUSION

With a base understanding of RS, it is easy to see how this technology can be very beneficial in online platforms. RS has different techniques that can be used in a variety of online systems. Many large companies are creating custom RS and are benefiting greatly from them. As the massive amount of data grows from day to day, the ways in which RS is used will continue to evolve. It will be interesting to see how Cross-Domain Recommendations are used in the future, and if companies start to adopt this concept of sharing data. Data being analyzed from various systems could unlock hidden information that a single system may not be capable of producing.

ACKNOWLEDGMENTS

The author would like to thank course instructors for organizing setup of the latex format used in this paper.

REFERENCES

- [1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17, 6 (June 2005), 734–749. <https://doi.org/10.1109/TKDE.2005.99>
- [2] Xavier Amatriain and Justin Basilico. 2016. Past, Present, and Future of Recommender Systems: An Industry Perspective. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 211–214. <https://doi.org/10.1145/2959100.2959144>
- [3] Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction* 12, 4 (01 Nov 2002), 331–370. <https://doi.org/10.1023/A:1021240730564>
- [4] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. 2015. *Cross-Domain Recommender Systems*. Springer US, Boston, MA, 919–959. https://doi.org/10.1007/978-1-4899-7637-6_27
- [5] Carlos A. Gomez-Urbe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (Dec. 2015), 19 pages. <https://doi.org/10.1145/2843948>
- [6] Shah Khushro, Zafar Ali, and Irfan Ullah. 2016. *Recommender Systems: Issues, Challenges, and Research Opportunities*. Springer Singapore, Singapore, 1179–1189. https://doi.org/10.1007/978-981-10-0557-2_112
- [7] Prem Melville and Vikas Sindhwani. 2010. *Recommender Systems*. Springer US, Boston, MA, 829–838. https://doi.org/10.1007/978-0-387-30164-8_705
- [8] Michael J. Pazzani and Daniel Billsus. 2007. *Content-Based Recommendation Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 325–341. https://doi.org/10.1007/978-3-540-72079-9_10
- [9] Anand Rajaraman. 2008. More Data Usually Beats Better Algorithms. (03 2008). <http://anand.typepad.com/datawocky/2008/03/more-data-usual.html>
- [10] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 291–324. https://doi.org/10.1007/978-3-540-72079-9_9
- [11] Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.Com. *IEEE Internet Computing* 21, 3 (May 2017), 12–18. <https://doi.org/10.1109/MIC.2017.72>

6 BIBTEX ISSUES

7 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

7.1 Uncaught Bibliography Errors

DONE:

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

DONE:

I copied my content and did not see any [?]. Not sure if this is a general issue being address to all, or if i am doing something wrong?

7.2 Writing Errors

DONE:

Spelling and grammatical errors, inconsistent capitalization of terms

DONE:

Do not use the phrase *In this paper/report we show* instead use *We show*. It is not important if this is a paper or a report and does not need to be mentioned

DONE:

Use a space after . , :

7.3 Citation Issues and Plagiarism

DONE:

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material

7.4 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”