Master thesis

omernivr12

July 2017

1 Abstract

Recommendation systems has become prevalent in recent years since the boom of web. Any internet based service has the advantage of having huge collection without the need to physically present it to the customer[2]. But, even if holding millions of products, any user can only see one web page at a time. Hence, for a user not to be lost in space, the service has to organize smartly the products to its liking. Recommendation of videos is one field that gained a lot of traction since the Netflix competition[4] was presented. Though many innovations were made over the years, there is still a wide search of combining the different features of videos and users into a comprehensive network for recommendations. This paper is tackling two blocks concerning the recommendation problem. Firstly, conceptually, it offers two novel ways to think about the recommendation problem. The first one is about selecting features, while the second one concerns a different architecture. Secondly, the promising architecture of GANs is used for the first time, as far as we are aware, to train a model in the video recommendation field.

2 Introduction

Video recommendation problem is one which was widely researched in recent years. Especially since the Netflix data set was open sourced and a prize was declared. Video recommendation is also almost identical to the generic problems of matching between items and users. However, although many research has been done, it seems that most methods are touching upon one solution or the other, and not trying to combine the strength that lies in each one of the methods.
To encompass all ideas in one model seems unrealistic, but choosing some prominent ones from each technique seems like an interesting and richer. Thus we are taking the the idea of items similarity from content based methods, the idea of collaborative filtering or user similarity from its behavior or ranking, and the idea of temporal dynamics from the sequential nature of the problem. The way we choose to combine all these things is by using the RNN network as a baseline model and adding to it.

Furthermore, the basic idea of recommending items according to their rank is disappearing once we are dealing with an implicit data set. That is since whether a video was watched or not, does not imply about its liking. Despite that, most practitioners keep on predicting whether an item is active or not. If we stop for a moment to ponder about human behavior, we will see that most people would not thank or be happier if the road is clear on their way to work. The other side of the coin is when the we are facing a traffic jam and our anger levels are rising (budhist). This story is a mirror to us getting very upset when we face something we dislike, but are quite indifferent to things that we do not especially excited about. The same idea is the one directing us in this paper, where we consider a data set for dislikes along with the usual data set for likes. Both data sets are implicit, with dislikes measured as movies that were stopped after less than 25 percent of their duration and were not continued afterwards. It is definitely open to many interpretations for the reason why a user stopped watching a certain video. However, we hypothesize that by using a lower dimensional representation of the complete set of movies considered to be disliked, a shared feature space would be learned. It is with this learned feature space combined with the output from the implicit liked space that we can be more confident if this is a video that we better avoid or better recommend.
While using RNN for predicting movies seems natural from the sequential nature of the problem, its common way to infer can suffer from the likelihood trap () .
This means that once we have predicted a movie that is unlikely and that was never seen before in the data set we might continue in the same manner basing on that unlikely movie to predict and the result would be poor. Few ways to overcome this problem were suggested in literature such as scheduled sampling and ... but as these suffer from problems of .. the method that is currently looking to ease the problem is GAN, which stands for generative adverserial networks, but is just a way to learn using two networks, where one is generating

data and the other tells it if it is correct or wrong.
To our knowledge, till now GANs were applied to many fields but has not been applied yet to the movie recommendation scene. Aside from avoiding the likelihood trap, the nature of GANs and its gradients requires from us to use reinforcement learning to first: avoid the derivative of the discriminator with respect to the discrete output and, second: enable a full sequence inference using MCMC.
Further novelty comes in the way we optimize our weights. Introducing first cross entropy, then the BPR-max cost and after the mean reciprocal rank. The former is forcing a ranking such that the desired video in the next step is bigger than a sample of negative examples. The latter tries to enforce also the ordering and ranking of the video that comes second after the desired one, the video that comes third and so forth.

3 Background

This chapter will outline the framework in which recommendation systems are developed and the progress made over the years. The chapter is constructed in the way of first drawing the image of the general framework and then delving into explaining the specific popular methods used.

3.1 Problem

The generic problem for recommendation systems is that of inferring whether an item is active or not - 0,1, which is commonly known as an implicit data set. This is implicit since if 'Joe' read the book 'Mein Kamf', it does not immediately imply of its liking. The explicit data set consists of some scoring of items made by the user, as an example we can have a 1-5 rating. The recommendation problem could be viewed as interpolation problem [9] - this can simply be thought of just as given two points in space, choose a third point location and connect all to a triangle. Or, it could be viewed as a prediction problem - given the points viewed so far, predict the most likely points to appear next. These two problems require different solutions, since the former is time-independent and the latter is time dependent. The simplest form of the problem is that of having a rating matrix where each user is a row and each item is a column. This matrix would be very sparse - many values are missing since an individual only viewed few movies from Netflix libraries, or bought only few items from Amazon's stock. It is possible to ask users to rate items, but it is limited to the user's patience and more worrisome it is biased. Meaning a user might not rate disliked movies. To the simple matrix form of the problem it is possible to add many other dimensions representing item attributes and/or user attributes. However, some solutions become irrelevant when doing so. As mentioned above, the implicit problem offers the challenge of differentiating between active to liked item. Most solutions concerning implicit data sets only check what is being 'liked' but put aside the 'disliked' issue. This paper will offer a way to tackle this problem. This is based on the assumption that if a user watched an item for certain period and then stopped, it might flag a dislike. So a network representing dislikes combined with a network representing activity, could be used to infer if an active item is more driven by likes or by dislikes.

3.2 Our problem

The problem we are facing with and decided to tackle is more case specific, but can be applied for many other similar cases. The problem is of predicting a full sequence cycle of videos the user should be recommended with in order to stay as long as possible in the Channel four platform.

3.3 Recommendation framework

At the top most level we can differentiate between history based recommendation to session based recommendation. History based means tracking user's behaviour through long period of time, whereas session based refers to short term behaviour of an individual that is new to the system at every visit. History based has very broad research and some of its applications are of fit to session based as well. Choosing the branch of history based recommendation we are presented with solutions under one of content based, collaborative filtering (CF)[6] or hybrid blocks [1] , which are then divided to heuristic based approaches and model based approaches. These separations fit also session based, but heuristic approaches are unstable due to sparsity. The subsections below will explore each block, its advantages and deficiencies.

3.4 Content based recommendations

The simple idea behind content based methods says: if a user watched a video about crime[7], recommend content related to crime. In order to do so, we should construct an item profile; containing any relevant features characterizing the item e.g. words that characterizes item topic. The latter example will play a major role in our novel architectures. Once a profile representing an item is constructed, it requires only to measure similarity between item' properties, and then recommending those that are most similar. To be more concrete, we calculate a matrix A where a_{ij} is the entry to i-th row and j-th column representing similarity between item i and item j. Given user U rated items i,j and k, go to rows i,j,k, order by descending value and pick top-k to recommend. Popular similarity measures in information retrieval are cosine measure and Jaccard similarity. Both similarities are intuitive, with cosine defining similarity by geometric meaning - closeness of angle between two vectors. While Jaccard is measuring similarity by proportion of overlapping items.
In a more formal way:

1. $cosine_\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$
2. $Jaccard_{A,B} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$

Some of the main arguments rising against the use of content based methods are the inability to distinguish items with same features, the inability to offer to users content different from what they have already watched, the issue of recommending something which is too similar to previous items and the problem of the new user[1]. It is worth to dwell on the new user problem [9][] since it is something that most solutions still find hard to deal with. A new user must rate/buy/use few items before she can get similarity to other items and reasonable recommendations. An alternative is to include user features which leads us to the collaborative filtering block.

3.5 Collaborative filtering

Collaborative filtering[1] is the umbrella name for all methods that use some sort of similarity among users likes or items ratings to measure their similarity. The main idea at its core is the construction of a utility matrix. A utility matrix is one where users are represented by rows and items are represented by columns. We then have many missing values, with some filled with the ratings. Our aim is to fill the matrix either in full or with K values such that recommendations will be possible.
The data inside the utility matrix is obtained by either asking the user - which is often referred as an explicit data set, or by inferring the behavior of the user - which is often referred as an implicit data set.
Asking the user to rate items involves the patience of the user, her inability to rate unknown items[10] and the bias associated with the 'not missing at random problem'[3]. This means that whether a rating is missing is missing depends on the video. In video rating this means, a user might not rate disliked videos or video that he is embarrassed sharing.
Inferring the user preferences through its activity is the problem we are facing with the channel four data at hand. So, we know that a user has watched a video or has not watched a video inside the platform. However, we do not know how a view maps to the liked/disliked space. With that in mind, even if we can reconstruct the perfect matching of the utility matrix it might still be useless in practice. Hence, one of our main ideas in the paper, is to tackle the problem of implicit information to some extent using additional information which is not less important.
Under the umbrella, the separation breaks to two major groups. The one being memory or heuristic based approach and the other being model based. The main difference is that a model based approach is to use the database to learn a model which is then used for predictions. (empirical analysis of predictive algo.).
The simplest and what was quite popular way with memory based is to find N users similar to user U, then we can average the ratings for item I. Where similarity can be measured by the cosine or Jaccard similarity mentioned above. However, this naive way suffers from many issues such as being slow, it does not work well for users with unusual preferences. In addition, the new user problem is evident with the problem of what items to show.
What items to show can be mitigated in few ways: 1. random 2. popularity 3. entropy 4. item-item. (getting to know you)
by entropy, calculate each movie using relative frequency of each of the five possible ratings and present movies with highest entropy. If user i can rate item j, it is considered as part of the users that are contributing to the entropy of user a. we then have the probability of user i(PMCF). ...
The slow computation render this user to user method impractical in real world problems. What brought upon the item-item method (Amazon). The change is to look at items in the utility matrix instead of the users. i.e we can compute cosine similarity between pairs if columns in utility matrix. all these calculations

can be made offline so it can be very quick to recommend an item in the online setting. This is in contrast to the user-user setting where a new user needs to be in the online setting calculated against many other users.
Here the problem of the new user is replaced by the new item issue. The new item does not have any ratings associated with it, so can not be recommended. Another issue is when an item is liked by all. Universally liked items are not useful in capturing similarity.
The other group under the umbrella is the model based collaborative filtering. The most renowned of these methods for many years is the dimensionality reduction and in particular Matrix Factorization in its various forms (PMCF, PMF,...). The simplest form is derived from the Principal component analysis (PCA) method.
PCA is a way to deconstruct a matrix to a lower dimension representation matrix times a basis matrix. We want to find the dimensions that are communicating the most amount of variation in the data. The idea is building upon the notion that it is many times the case that lower dimension manifold is lying inside a higher dimensional space that does not add much information at inference. Think about a data set that contains both the weight in grams and in Kg. One of the two is redundant since they communicate exactly the same information and hence can be disregarded. We would want the PCA to be able to find these so called redundancies. Formally, it is....Can be preformed more efficiently by SVD.
The idea of dimension reduction is that there are relatively small set of features of items and users that determine reaction of users to items. In the field of videos, an easy example can be made by thinking about the shared space of videos and items as finding the genres shared by the different videos. The resulting matrix would be the number of videos as rows and the number of genres as columns. the basis matrix would be the weighting of each of these genres to each user, such that the original matrix is the result for.
Fortunately, PCA can also serve to complete a partially filled matrix. The method is the same as it is for the regular PCA, just that this time we would only measure the loss in the places that values do exist (barber). We can solve this by two ways : 1. Fix U then solve linear system for V, then fix V and solve linear system for U.
2. Initialize values of the U, V matrices and preform gradient descent on the RMSE loss. As PCA has no probabilistic interpretation, it might be desired to use its probabilistic model (PMF). In this view, the rating of user i to item j is distributed around the mean UV_{ij} of a multi dimensional normal distribution. The user and video matrices are getting also a normal prior distribution with 0 mean and constant variance. the posterior is then... A problem that arises in the context of matrix completion in all its forms is, how to choose the number of features D that will constitute the lower dimension representation.
In the recent years with the rise of Neural Networks and its family, many other methods were presented for matrix completion among them RBM [**Salakhutdinov2007**], BRNN [**Schuster1997**] for filling missing values and auto-encoders [**Sedhain2015**]. With the latter being the closest to the PCA completion method and has shown

the most promising results regarding video recommendation using collaborative filtering[**Sedhain2015**]. Due to its simplicity, conceptual matching to older collaborative filtering techniques and its results, we chose to employ the auto-encoders to the disliked data. Auto-encoders [3] are a fancy name for dimensionality reduction using neural networks. We construct them by having in the minimal case a layer where we input the original matrix and then we have a lower dimensional layer where we apply a function above X. We then apply another function above the lower dimension layer to try and reconstruct X. The loss that is being measured is the reconstruction loss between output and input. This is an unsupervised method. If we use only linear functions such as f(x) = Wx at the reduction layer and output layer we get a similar reconstruction to PCA.

Same as in matrix completion methods, we can also use auto-encoder to fill in values. However, as we usually input to the first layer of our network the original matrix, an extra care should be applied when considering the missing values. If these are set to zero, we are biasing the model towards zero values which we do not want to do, especially not in our implicit data set where zeros have a meaningful interpretation as non active. Few initial values could be considered, such as the mean value of items, or the output of a matrix completion procedure. These values would still not be considered in the loss, and will not be back propagated during training, but would affect the interaction of the other weights and hence should be reasonable.

Until now we have considered models and memory based techniques that are neither time dependent nor feature rich. For example we have not considered the age of our user to affect her viewing behavior. But it is quite intuitive that a six year old would have different preferences than fifty years old. If our technique takes this feature and others into account, then it will be much easier to group the different possible videos to recommend next for users. Few methods that seem interesting ad promising in the dimension reduction sphere are inductive matrix completion, that extends the idea that we can decompose the original matrix to only a two matrices to decompose them to multiple matrices which hold also features like age, gender and postcode for the user, and features like genres and actors for the movie - this method is called inductive matrix completion [**Jain2013**]. Another method which looks promising is the Tensor decomposition. This is promising since unlike matrix decomposition methods the initial matrix can be in high dimensional space and thus much richer in features. Hence, better predictive power is assumed? (tensor-decom-paper). However, this method to date is only applicable to a very specific space of problems and does not yet have a lot of research background.

Although Tensor decomposition can be considered for modeling temporal dynamics, it is more natural to think of it as a sequence which can be modeled by the RNN which is built exactly for this reason.

3.6 Hybrid methods

Hybrid methods are exactly the ones that combine characteristics both from content based methods and from collaborative filtering. In our solution we choose to employ a hybrid method, since this can give us a richer representation and can answer more questions than the proto-typical answers concluded from filling a matrix or predicting zeros and ones just on the basis of ratings. It is clear that the nature of liking movies is temporal. Our interest can change over time, poor movies can become a nostalgia and so on. RNNs power lies not only in their temporal construction but also in that they can be input with ,any different kinds of data. including features concerning the user and/or the item.

3.7 Features

As mentioned in the content based part above, an important part is to find similarity among properties of the items or videos in our case. Some of the common features to use are genres, which we use as well. Genres are assumed to be very helpful to recommendation, since if a user is watching only teenage comedies, then it will be quite simple to construct the table of possible videos she is going to watch next. Since the number of genres is not so big, we can use one hot encoding to represent them. i.e. construct a list of genres, and place a 1 if genre is active in this specific video, otherwise place 0. example: With genres $\mathcal{G} = \{Comedy, Horror, Thriller\}$ if user i watched 'When Harry met Sally' the corresponding vector would be {1, 0, 0}.
Another feature, which is less common for video recommendation is to use the script/subtitles of the video as a characteristic. We have decided to use it with the use case that a stand-up video might be tagged in genres of comedy and stand-up, but if the text is about drug abuse, and then you move subsequently to watch 'Trainspotting' the text itself would be more telling than the genre tags. Although this idea is not very common in the video recommendation industry, the idea is prevalent in written documents classification to topics.
Several techniques are available with the widely used TF.IDF [8] from information retrieval, LDA [5] from unsupervised learning using variational infernce and Word2Vec from . The common to the three is that they are all can be used to find similarity in space of words and/documents to one another.. Our choice is the word2vec method, which was chosen since it is an inherent part of the Neural network methods that are used in our architecture. For completion we still give a short explanation regarding TF.IDF and LDA.
TF.IDF can be explained easily as a counting method. Where we count the number of times a term occurs in a document and we weight this counter by the frequency across all documents in the set. If the same word appears frequently in all documents, there is not much we can learn about the specific topic. The simplest example would be the so-called stop words like {the, on, a} which are widely used with no meaning behind. This idea is both intuitive and fits good writing practices[11], which emphasize non-redundancy and simplicity as its main properties. Once a score vector is obtained, we can pick the top K

words as characterizations. Counting however does not necessarily imply on the underlying characteristics of the text.
LDA is an unsupervised, meaning we do not have topics beforehand but we construct them from scratch. The approach is used to classify words in a document to topics. Whilst the idea is simple, its mathematical formulation is more evolved, and will not be stated here. The idea is that once we have a collection of documents each containing of words. Each of these documents has a different distribution over words. We then want to assign a word to a topic without forgetting the probability of this word under this topic. To clarify, It might be that a document consists of two topics with one 99 percent and the other 1 percent. However, the probability of the word *'qualia'* is zero under this topic. So it would still make sense to choose the topic with the one percent chance.
Word2Vec

3.8 Loss

In machine learning, the choice of your loss function can make a big difference for whether your weights will learn to find an optimum or it will struggle to learn and provide no meaningful results. The common loss in recent years for training classification problems is defaulted as cross-entropy. Why not to use the generic mean squared error for classification problems? the answer to that question will become evident when we discuss the cross entropy. However, problems arise also when dealing with large scale tasks; Since the cross entropy computes the probability for all possible labels, when the number of labels increases we can encounter instability rising from underflow and overflow and heavy computation times (insert paper here). Solutions for that are based upon looking at only part of the labels, consider them as undesired examples and compare to the desired one. Such methods are the Bayesian personalized ranking [**Rendle2009**] loss and the mean reciprocal rank [**Chapelle2009**] approaches. The former tries to maximize the desired one, while the latter tries to enforce the correct ordering of the first K videos. We will explain here under the intuitive use of each one of these losses and the advantages and pitfalls of each of them.

3.9 Cross-entropy

To understand the fit of cross entropy to the classification problem, we would need to go through some Information theory concepts as the communication of code words and entropy.
As explained very clearly in (insert paper). If we try to look at it in the video field, we can consider we have a budget to spend for making three types of movies: comedy, old people that like sweaters or drugs for East Asian mammals. The basic idea of entropy says that if the comedy genre comprises of 95 percent of the market, the optimal thing to do is to invest 95 percent of our budget in producing these type of movies. Now imagine we can only communicate the genres to our director by zeros and ones, and we pay for each number we send - then, our goal would be to sending the message with shortest length to

pay as little as possible. The cost for each such message is $\frac{1}{2^(messagelength)}$. Since we have only two options $0, 1$ for each message letter if we send a message starting with 1 we can't use 1 any more as a starting letter and half of our options are gone. Otherwise, we would not know the meaning of the message 110 is it 11 and then 0 or is it 1 and 10. To further clarify the cost, lets look at an example where we use length two message such as 10. in this case we will not be able to start any other message with 10 however, 11, 01 and 00 are still available, so only every fourth option is blocked. i.e $\frac{1}{4}$. Now to conclude how much we will pay for writing a lengthy twitter message we will have to multiply the cost times the probability that this message will be read. From the cost defined we can calculate the message length as $\log_2 \frac{1}{cost}$. As discussed if the optimal spending is p(x), then the average message length using the optimal spending is $H(p) = \sum_x p(x) \log_2 \frac{1}{p(x)}$ which is the definition beyond the entropy term.

We give here only a simple explanation for why p(x) is the optimal spending. Consider that we spend 55 percent of our budget on a message that 5 percent read and 45 percent on a message that is widely read. If our investor sees this behavior, he will ask why aren't we spending more on a widely read message? If we then decide to spend all on the widely read message then we will lose 5 percent of the budget although it was not required. The mathematical proof follows the same line of thought.

Since p(x) is the probability that the message will be read, if we plug into the entropy term a probability, the more concentrated the probability the shorter my messages can be to convey the same idea.

Think that if all my followers know what I am talking about i can say 'we won' but if I have followers from many sectors then I would need to explain who won? where? and when?. Now if I don't know the distribution and still try to write 'we won', this can potentially be very confusing to my followers.

This confusion is the idea of cross entropy. How far are we from the entropy, which is the optimal spending for our real distribution. Formally, $H_p(q) = \sum_x q(x) \log(\frac{1}{p(x)})$ is the cross entropy of P with respect to Q.

In classification problem, we exchange P with the predicted probability and Q with the real label. i.e. Q will be probability concentrated at one X_i value with probability 1. Using cross entropy with softmax output layer if many labels are involved, as in the video recommendation task, then the gradient can stay stable (insert paper) - this part will be explained further in the gradient section. The gradient calculation is also the main driver to use cross entropy and not mean squared error loss (see below).

3.10 BPR

BPR-max (insert paper) is a loss that tries to circumvent the problem that is introduced with many labels to predict. Using a softmax as an output layer, to calculate the probability of all labels such that they sum to one can be a big computational hurdle. Think of a case with one million labels. As we start our prediction, we assign probabilities in more or less random manner such that it can be thought as distributing the different probabilities uniformly. In that case each label gets a probability of $\frac{1}{1,000,000}$ which is already quite an extensive representation. The further we continue our training, some probabilities will increase, while other decrease to preserve the sum to one rule. i.e. a situation of underflow is easily achieved. A solution that comes into mind is not to use the softmax to calculate the full output. Instead, we can use the sigmoid function applied over two labels at a time, forcing one to be greater than the other. Furthermore, as it is too expensive to use the 1m labels we can try to smartly sample only a fraction of labels without losing much of the information.
Formally, we define the loss function to be $BPR{-}max = -Mean(\sum_j \log \sigma(y_i - y_j))$ where $j \subset negativesamples$, i is the ground truth position of the label and $\sigma(y_i - y_j) = \frac{1}{1+\exp^{-(y_i - y_j)}}$. To understand sigmoid term let us consider few values:
$sigmoid(x) = f$ which means that when $(y_i - y_j) \longrightarrow \infty$ closer the sigmoid expression is to 1. Applyig log on 1 yields zero. If on the other hand, the term $(y_i - y_j) \longrightarrow -\infty$ then the sigmoid term is going to 0 which means the log argument will yield a big negative number. Our goal is to minimize the loss and hence what we ask is the constellation in which the predicted value at the ground truth label position is bigger than all the sampled predictions y_j.
The question to rise is how to choose the samples smartly? and what are negative samples?.

3.11 future development

On a bigger outlook, you would wonder when a user would be the owner of his own data and companies would offer a deal for that user based on his willingness to share data with them and which data it will be. Privacy issues are of a big concern and further so many efforts are made to learn an implicit behavior of a user that might be willing to communicate her personal data if she would know for what it serves and how it helps her to get better personalized data.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*. 2005. DOI: 10.1109/TKDE.2005.99. arXiv: 3.

[2] Chris Anderson. "The Long Tail". In: *WIRED magazine* 12.10 (2004), pp. 170–177. ISSN: 1580979X. DOI: 10.3359/oz0912041. URL: http://www.geeknewscentral.com/2010/08/13/the-long-tail/.

[3] David Barber. "Bayesian Reasoning and Machine Learning". In: *Machine Learning* (2011), p. 646. ISSN: 9780521518147. DOI: 10.1017/CBO9780511804779. arXiv: arXiv:1011.1669v3. URL: http://eprints.pascal-network.org/archive/00007920/%7B%5C%%7D5Cnhttp://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Bayesian+Reasoning+and+Machine+Learning%7B%5C#%7D0.

[4] James Bennett and Stan Lanning. "The Netflix Prize". In: *KDD Cup and Workshop* (2007), pp. 3–6. ISSN: 1554351X. DOI: 10.1145/1562764.1562769.

[5] David M Blei et al. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022. ISSN: 15324435. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1.

[6] Jonathan L Herlocker et al. "Evaluating collaborative filtering recommender systems". In: *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), pp. 5–53. ISSN: 1046-8188. DOI: 10.1145/963770.963772. arXiv: 50. URL: http://portal.acm.org/citation.cfm?id=963770.963772.

[7] Anand Rajaraman and Jeffrey D Ullman. "Mining of Massive Datasets". In: *Lecture Notes for Stanford CS345A Web Mining* 67 (2011), p. 328. ISSN: 01420615. DOI: 10.1017/CBO9781139058452. arXiv: arXiv:1011.1669v3. URL: http://ebooks.cambridge.org/ref/id/CBO9781139058452.

[8] Amit Singhal. "Modern information retrieval: A brief overview". In: *IEEE Data Engineering Bulletin* (2001), pp. 1–9. ISSN: 00218979. DOI: citeulike-article-id:1726446. arXiv: 9780201398298. URL: http://ilps.science.uva.nl/Teaching/0405/AR/part2/ir%7B%5C_%7Doverview.pdf.

[9] Chao-yuan Wu et al. "Recurrent Recommender Networks". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17*. 2017, pp. 495–503. ISBN: 9781450346757. DOI: 10.1145/3018661.3018689. URL: http://dl.acm.org/citation.cfm?doid=3018661.3018689.

[10] Kai Yu et al. "Probabilistic Memory-Based Collaborative Filtering". In: *IEEE Transactions on Knowledge and Data Engineering* 16.1 (2004), pp. 56–69. ISSN: 10414347. DOI: 10.1109/TKDE.2004.1264822.

[11] William Zinsser. "On Writing Well". In: *Quill* (2001), pp. 1–308. ISSN: 0060891548.