On Application of Bayesian Parametric and Non-parametric Methods for User Cohorting in Product Search

Shashank Gupta
Flipkart Internet Private Limited
India
shashank.gupta@flipkart.com

Abstract

In this paper, we study the applicability of Bayesian Parametric and Non-parametric methods for user clustering in an E-commerce search setting. To the best of our knowledge, this is the first work which presents a comparative study of various Bayesian clustering methods in the context of product search. Specifically, we cluster users based on their topical patterns from their respective product search queries. To evaluate the quality of the clusters formed, we perform collaborative query recommendation task. Our findings indicate that simple parametric model like Latent Dirichlet Allocation (LDA) outperforms more sophisticated non-parametric methods like Distance Dependent Chinese Restaurant Process and Dirichlet Process based clustering in both tasks.

1 Introduction

Online retail business has been on an unprecedented growth since last few years, with a market share of 600 Billion dollars last year in the United States [1]. To cope up with such an epochal growth, product discovery becomes an important aspect for any e-commerce platform. It is an important aspect to help customers navigate an ever increasing inventory of products.

Product search is an important aspect of E-Commerce discovery, where displaying relevant and personalized products in answer to an user query is directly tied to customer satisfaction (Su et al., 2018; Moe, 2003).

Personalizing search results entails mining search behavior logs to represent individual user's search intent. However, due to the large variance in individual user's preference, finding a good representation is often difficult. In the context of web
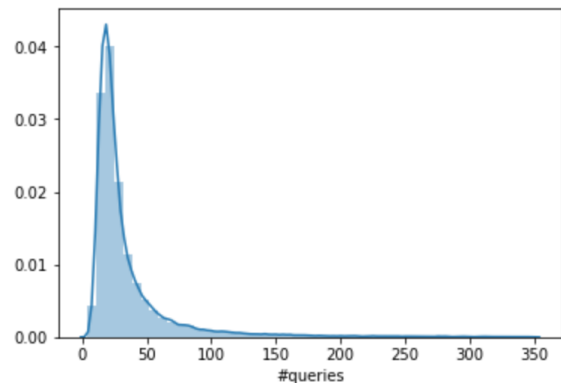

Figure 1: Query Distribution across users.

search, search intent mining has been extensively studied (Teevan et al., 2008; White and Drucker, 2007; Joachims et al., 2017).

However, despite it's apparent importance, personalization in product search is not very well studied. There has been some attempts to model user's search intent based on her past product interactions. Qingyao et al. (Ai et al., 2017) jointly learn representations of user, query and users' product reviews for personalization. In a follow-up work, Qingyao et al. (Ai et al., 2019), learn a 'zero-attention' model for representation learning to account to cold-start problem in new product categories. However, it should be noted that both work attempt to model search interest at the granularity of individual users, which are highly spare in nature.

Fig. 1 is the distribution of queries across users in the Amazon Product Search Dataset (Ai et al., 2017). The distribution is skewed towards left side of the distribution. It shows that query distribution across users is sparse, with very few users have high number of queries and subsequently few interactions as-well. This makes it challenging to model search interests at the granularity of individual users.

By modelling each user separately, we are also

---

[1] https://www.digitalcommerce360.com/article/us-ecommerce-sales/

missing on the chance to capture the heterogeneity in the user's search interests. Modelling user's jointly can aid in learning individual user's interest, as-well as aid in handling sparsity through sharing statistical strength within the group. Towards this end, we study various Bayesian clustering methods to model latent user cohorts by mining user's search query patterns.

To the best of our knowledge, this is the first work in the domain of E-Commerce search to study the application of bayesian methods for modelling user cohorts. The aim of this paper is to present a comparative study of different bayesian parametric and non-parametric methods in the domain of product search. Product search system designers can benifit from this study by making an informed decision when designing E-commerce personalization system.

2 Related Work

There has been plenty of work in web search for modelling user cohorts for learning an effective ranking function. Previous studies (Bian et al., 2010; Giannopoulos et al., 2011) propose a K-means clustering of queries and then a learn a ranking function within each sub-group to capture query patterns in each group.

However, the focal unit of the previous work is queries, whereas we are interested in modelling user. The closest to our work is by Yan et al. (Yan et al., 2014), where they model user cohorts based on Click-Through Rates (CTR) and Open Directory Project (ODP) based topics. However, instead of relying on human curated ODP based topics, we use probabilistic topic models to identify topics from user's log data and further use it for clustering.

Towards this end, in this work, we study the problem of user cohorting by clustering user's queries in the probabilistic topic space. Specifically, we study application of various probabilistic topic models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Hierarchical Dirichlet Process (HDP) (Wang et al., 2011; Teh, 2006), Dirichlet Process Gaussian Mixture Model (DPGMM) (Teh, 2010) and the more recent Distance Dependent Chinese Restaurant Process (ddCRP) (Blei and Frazier, 2011). To the best of our knowledge, this is the first work in product search which studies application of probabilistic topic models for user cohorting.

3 User Cohorting using Probabilistic Topic Models

In this section, we will formally define the problem statement followed by a brief description of the non-parametric ddCRP user clustering method.

3.1 Problem Statement:

We define the search log D as : $D = \{u, q, p\}$, where u is the individual user identifier, q is the product search query and p is the unique id of the product purchased by the user in the context of the query q fired. As part of pre-processing, we aggregate the queries fired by user as basic representation unit. Each user u_i is represented as, $u_i = \{q_1, q2, ..., q_n\}$, where each query q_j is represented by bag-of-words of of basic search tokens. Here, the user can be treated as a document of search tokens. The goal is to cluster users' based on topical representation.

3.2 ddCRP for user clustering

In this section, we will describe the ddCRP method for user clustering.

The ddCRP is an extension of the popular Bayesian Non-parametric method Chinese Restaurant Process (CRP). Like CRP, it is defined as a distribution over all possible partitions of a set (users in our case). Customer (user u_i in our case) enter the restaurant in a sequence and select a table z_i following another customer (user) u_j based on some probability and sits on that table. This is unlike CRP, where the probability of a customer choosing a seat is proportional to number of customers already seated on that seat.

Formally, the ddCRP process can be summarized by the following mathematical notation:

$$p(z_i = j | D, \alpha) \propto \begin{cases} f(d_{ij}), & \text{if } i \neq j \\ \alpha, & otherwise \end{cases} \quad (1)$$

where z_i is the latent table assignment for the customer i. $z_i = j$ represents that the customer i is linked to customer j with probability defined above. D is the matrix of all pairwise distances (defined in the next section) and alpha is the hyper-parameter of the model. f is the decay function and d_{ij} is the pre-defined distance between user u_i and the user u_j. $R(u_{1:N})$ defines the clustering structure by traversing the final user linking generated by the model.

The generative story of the ddCRP model is defined as below:

1. For $i \in [1:N]$, sample $z_i \sim \text{ddCRP}(\alpha, f, D)$.

2. For $i \in [1:N]$

    (a) Draw parameter from the base distribution, $\theta_i \sim G_0$

    (b) For each of the M words in the user query chain, draw $w_i \sim F(\theta_i)$.

Here, since we are dealing with bag-of-words model of text, we choose G_0 as a dirichlet distribution and F to be a Multinomial distribution, which can generate the words in the user's query chain.

3.3 Defining the Distance Function

To fully specify the ddCRP model, we need to define the distance function. In our case, to define the distance between two users as Hellinger distance between their LDA topic distribution.

4 Experiments

In this section, we will describe the experimental results from our study.

4.1 Dataset

For experiments, we use the product search dataset described by Qingyao et. al (Ai et al., 2017). They use the Amazon product review dataset and extract queries from the user's product reviews and use the purchased product as the relevant product. The dataset is generated for multiple product categories, we choose two of them for our experiments: 1) Compact Discs (CDs) and Vinyls. 2) Kindle Dataset. We removed users with atleast 200 queries, which leaves us with 1300 users in both datasets.

We select the following two tasks to evaluate performance of the clustering method: 1) Collaborative Query Recommendation and 2) Collaborative Document Recommendation.

For each user, we select a random sample of 5% queries as test queries and 95% for training. Similarly for products. We train LDA model on query chain constructed from training queries of each user. Specifically, we treat a simple concatenation of queries in user's training queries as a document and the collective query chain of all users as the corpus.

4.2 Collaborative Query Recommendation:

Once the users are segmented into various groups, we compute frequency of queries in each group. Each user is recommended top-K queries computed from his respective group. We treat user's test query as relevant queries and compute the performance of recommended queries against the relevance set.

Method	Prec@5	Prec@10
LDA	0.071934	0.088187
HDP	0.044093	0.058691
ddCRP	0.040030	0.056057
LDA + IGMM	0.007073	0.008202
LDA + GMM	0.017156	0.020767

Table 1: Query Recommendation Performance for CDs data

Method	Prec@5	Prec@10
LDA	0.072863	0.088133
HDP	0.069046	0.079751
ddCRP	0.054440	0.063734
LDA + IGMM	0.031701	0.027718
LDA + GMM	0.009627	0.011535

Table 2: Query Recommendation Performance for the Kindle data

5 Results

We present some preliminary results on the query recommendation task on the two datasets mentioned above. We compare the following methods:

1) LDA: We use the MAP estimates from the posterior over topics for each user to get the cluster assignment.

2) HDP: Similar to LDA, we use MAP estimates from the posterior to get the cluster assignment.

3) LDA + GMM: We perform Gaussian Mixture Model (GMM) clustering with LDA topic distribution as feature vector.

4) LDA + IGMM: We use Infinite Gaussian Mixture model, which is a non-parameteric model to perform clustering over user's topic distributions.

From the results, it is clear that LDA outperforms more sophisticated Non-parameteric methods in terms of Precision metric.

6 Conclusion

This paper can serve as a starting point of discussion around the use of Bayesian Non-parametric models for user cohorting. The preliminary results weigh in favor of simple parametric model like LDA, however, we believe that with further investigation, ddCRP method's performance can be improved. More specifically, if we perform clustering in the space of recently proposed neural embedding methods, instead of conventional topic model space, it's performance can be improved. We hope this paper can start that discussion.

References

Qingyao Ai, Daniel N Hill, SVN Vishwanathan, and W Bruce Croft. 2019. A zero attention model for personalized product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 379–388.

Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. 2017. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

Jiang Bian, Xin Li, Fan Li, Zhaohui Zheng, and Hongyuan Zha. 2010. Ranking specialization for web search: a divide-and-conquer approach by using topical ranksvm. In *Proceedings of the 19th international conference on World wide web*, pages 131–140.

David M Blei and Peter I Frazier. 2011. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461–2488.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Giorgos Giannopoulos, Ulf Brefeld, Theodore Dalamagas, and Timos Sellis. 2011. Learning to rank user intent. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 195–200.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. Acm New York, NY, USA.

Wendy W Moe. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1-2):29–39.

Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 547–555.

Jaime Teevan, Susan T Dumais, and Daniel J Liebling. 2008. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 163–170.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, Sydney, Australia. Association for Computational Linguistics.

Yee Whye Teh. 2010. Dirichlet process.

Chong Wang, John Paisley, and David Blei. 2011. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760.

Ryen W White and Steven M Drucker. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 21–30.

Jinyun Yan, Wei Chu, and Ryen W White. 2014. Cohort modeling for enhanced personalized search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 505–514.