Latent Dirichlet Allocation based Diversified Retrieval for E-commerce Search

Jun Yu*
Oregon State University
yuju@eecs.orst.edu

Sunil Mohan
eBay Inc.
smohan@ebay.com

Duangmanee (Pew) Putthividhya
Google
pew@google.com

Weng-Keen Wong
Oregon State University
wong@eecs.orst.edu

ABSTRACT

Diversified retrieval is a very important problem on many e-commerce sites, e.g. eBay and Amazon. Using IR approaches without optimizing for diversity results in a clutter of redundant items that belong to the same products. Most existing product taxonomies are often too noisy, with overlapping structures and non-uniform granularity, to be used directly in diversified retrieval. To address this problem, we propose a Latent Dirichlet Allocation (LDA) based diversified retrieval approach that selects diverse items based on the hidden user intents. Our approach first discovers the hidden user intents of a query using the LDA model, and then ranks the user intents by making trade-offs between their relevance and information novelty. Finally, it chooses the most representative item for each user intent to display. To evaluate the diversity in the search results on e-commerce sites, we propose a new metric, *average satisfaction*, measuring user satisfaction with the search results. Through our empirical study on eBay, we show that the LDA model discovers meaningful user intents and the LDA-based approach provides significantly higher user satisfaction than the eBay production ranker and three other diversified retrieval approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

E-commerce search; diversified retrieval; Latent Dirichlet Allocation

*This work was done when the first author was an intern at eBay Search Science.

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556215.

1. INTRODUCTION

Online marketplaces such as eBay and Amazon globally connect buyers directly to sellers. Given the large and dynamic inventory, retrieving relevant products for the users' queries is a crucial task. Due to the extremely diverse inventory of products on these sites, many inexperienced users looking to purchase different products may express their intents in the same way, leading to ambiguity [2]. For example, the query *fossil* could refer to either the brand Fossil in the Fashion category or the antique fossils in the Collectible category. A user often has a hidden purchase intent before shopping online and considers different products as relevant; without the prior knowledge of their interests, it is better to show different types of products so that at least one of them will be relevant to the user's intent [20].

Information retrieval approaches, based on the probabilistic ranking principle [19], can not satisfy users with different purchase intents. The probabilistic ranking principle ranks documents by their probability of relevance to the query and relies on the assumption that the relevance of a document is evaluated independently from the other documents. This assumption does not hold in practice because the most relevant documents often contain highly similar contents, resulting in redundancy. This problem becomes more severe in online marketplaces where there exists a large number of highly similar listings in the inventory and users are allowed to create their own product description to advertise their listings. Using product search without optimizing for diversity results in a clutter of similar products being ranked in the top slots in the search result page and wastes valuable prime real estate that could be used to satisfy other purchase intents.

To tackle this problem, we investigate methodology that introduces diversity in the search results. Diversified retrieval has been studied in the context of web search assuming the existence of a taxonomy [1, 26, 27]. In the e-commerce setting, one is tempted to follow suit and use the product taxonomy to help select a set of diverse items that satisfy different user intents. In practice, however, the existing product taxonomy on many e-commerce sites are found to be too complex and noisy due to (a) overlapping subtrees which allow similar products to reside in different parts of the taxonomy, and (b) non-uniform granularity in different subtrees, i.e. a category may include one or multiple products. For example, all iPod products are within one giant

category in the eBay product taxonomy. Thus diversified retrieval based on taxonomy can not further explore different products within the same category. While the product taxonomy works well to enhance inventory browsing experience, it is rendered useless for diversified retrieval purposes.

Unlike web search, diversified retrieval in product search must be able to accommodate dynamic inventory changes and fluctuating demand in real time. On eBay, there are hundreds of items listed and ended every second. The ability to perform fast scoring and ranking is essential in such a dynamic environment. In addition, user interests on e-commerce sites evolve quickly over time. For example, the latest generation of iPhone immediately becomes popular on eBay after its release. The model should be able to quickly adjust itself to the change of user interests over time.

To address these unique challenges in the e-commerce domain, we propose a Latent Dirichlet Allocation (LDA) [3] based diversified retrieval approach. The key idea is to choose diverse items based on the hidden user intents of a query. Our approach first discovers the user intents of a query by learning an LDA model, then ranks the uncovered user intents by making trade-offs between their popularity and information novelty, and finally selects the most representative item for each user intent to display. In addition, our approach can easily incorporate user feedback, e.g. Figure 1 shows two listings of query *keyboard* corresponding to two different user intents, "See more items like this" option can take user feedback and allow users to explore similar items within a particular user intent.

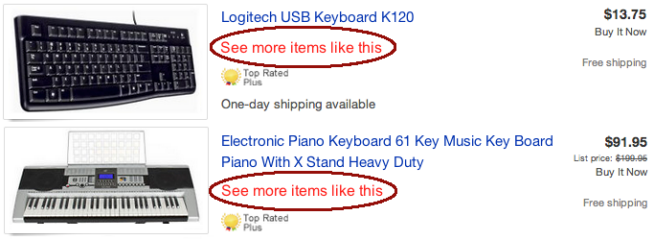

Figure 1: Listings of query *keyboard* with "See more items like this" option.

To evaluate the search results on e-commerce sites, we propose a novel offline metric, *average satisfaction*, to measure user satisfaction with the search results. Our empirical study on eBay data shows that the LDA model discovers meaningful user intents and the LDA-based approach provides significantly higher average satisfaction than the eBay production ranker and three other diversified retrieval approaches. The main contributions of this paper include:

- We identify the unique challenges for product search on e-commerce sites and propose an LDA-based diversified retrieval approach to address these challenges.
- We develop a variant of the standard LDA model, *Multivariate Bernoulli LDA*, that is more suitable for modeling short item titles without duplicated terms.
- We propose a new metric to evaluate user satisfaction with the search results and perform a detailed evaluation comparing different methods using the eBay data.

2. METHODOLOGY

To improve user satisfaction on e-commerce sites, it is important we understand different user intents associated with the same query [9]. Following this intuition, our approach consists of three steps: *discovering user intents*, *ranking user intents* and *selecting items for user intents*, which we describe in the following subsections.

2.1 Discovering user intents

The user behavior on the search result page often carries the information on a user's hidden purchase intent. For example, users with different purchase intents tend to click on different types of products on the search result page. Since most e-commerce sites collect user behavior data, this is plentiful and obtained with little extra cost. Then the challenge is how to discover the hidden user intents of a query from the user behavior data. In our approach, we uncover the hidden user intents using the Latent Dirichlet Allocation (LDA) model [3]. Latent Dirichlet Allocation is a probabilistic generative model, originally invented to uncover the underlying topics from a collection of documents. The idea behind the LDA model is that each document can be characterized by a mixture of topics where each topic defines its own probability distribution over a fixed vocabulary. The words in a document contain semantic information of its topic proportions and documents of similar topics tend to use a similar set of words.

To discover the hidden user intents of a query with the LDA model, we first generate the "corpus" for that query. Since the item title is the most descriptive, concise and relatively noise-free portion of an item on e-commerce sites, we generate the "corpus" of a query by collecting all the item titles from the user behavior data. Then we apply the LDA model on it and the discovered topics correspond to the hidden user intents for that query. Similarly, each user intent is characterized by its own multinomial distribution over the terms in the vocabulary. Terms that represent a user intent well are assigned with higher weight and the same term gets different weights in different user intents. From this point on, we use a topic and a user intent interchangeably.

Though an LDA model needs to be built for each query, we can reduce this overhead in practice. Firstly, only ambiguous queries need to be treated with diversified retrieval. Secondly, we can cluster similar queries and train a single LDA model for the group of similar queries. Finally, though training the LDA models is in general slow, it can be done offline and the hidden user intents are stored for real time scoring. In addition, recent developments in topic modeling have made inference in the LDA model very efficient and able to handle large amounts of data. For example, the parallel LDA [23] allows training the LDA model in distributed systems and the online LDA [12] allows us to easily incorporate the latest user behavior data coming in a stream and avoid the cost of retraining the model completely.

The Multivariate Bernoulli LDA model.

The LDA model assumes a word in a document is drawn from the multinomial distribution associated with its topic assignment. Under this assumption, the same term may occur multiple times in a document. In our application, however, the item titles are often short and do not contain duplicated terms. To better model this unique property of

the data on e-commerce sites, we propose a variant of the LDA model, *Multivariate Bernoulli LDA*, in Figure 2.

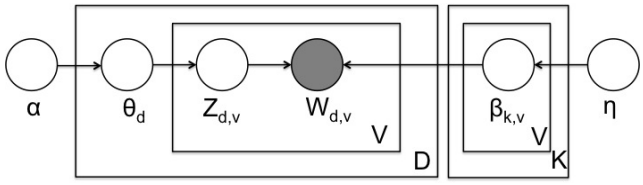

Figure 2: Graphical model representation of the Multivariate Bernoulli LDA model.

for *each topic* k **do**
  for *each term* v *in the vocabulary* **do**
    Draw probability of occurrence $\beta_{k,v} \sim Beta(\eta)$
  end
end
for *each document* d **do**
  Draw topic proportion $\theta_d \sim Dirichlet(\alpha)$ **for** *each term* v *in the vocabulary* **do**
    Draw topic assignment $Z_{d,v} \sim Multinomial(\theta_d)$
    Draw term occurrence
    $W_{d,v} \sim Bernoulli(\beta_{Z_{d,v},v})$
  end
end

Algorithm 1: Generative process of the Multivariate Bernoulli LDA model.

Assume we have D documents in the corpus, a vocabulary of size V and a predefined number of topics K. In the Multivariate Bernoulli LDA (MB-LDA) model, each topic is characterized by a multivariate Bernoulli distribution instead of a multinomial distribution. Each term v in the vocabulary within a topic k has its own Bernoulli distribution with parameter $\beta_{k,v}$. For a document d, the model first samples the topic assignment $Z_{d,v}$ for each term v in the vocabulary and then samples the term occurrence $W_{d,v}$ from the Bernoulli distribution associated with its topic assignment $Z_{d,v}$. The generative process of the MB-LDA model is given in Algorithm 1. Since the occurrence of a term in a document is modeled as a binary random variable, a term can occur at most once in a document. Like the LDA model, the inference of the MB-LDA model is also computationally intractable. In this work we use the collapsed Gibbs sampling method [10] to learn the MB-LDA model and details of the collapsed Gibbs sampler are given in Appendix A. In the experiment, we compare both LDA models in ranking the eBay listings and show that the MB-LDA model provides better performance.

2.2 Ranking user intents

We rank the uncovered user intents of a query by trading off their relevance and information novelty. Though the user intents indicate different user information needs of a query, some user intents are similar and may correspond to the same high-level user intent (e.g. *Fossil men watch* and *Fossil women watch* are part of the high-level user intent *Fossil watch*). Given some similar user intents, choosing one of them downweights the information novelty of the others. To achieve better user satisfaction, we rank all user intents by making trade-offs between their relevance and information novelty. The relevance measures the importance of a user intent to a query and the information novelty measures the amount of extra information a user intent adds onto the user intents already selected.

In this paper, we adopt the Maximal Marginal Relevance approach [5] to rank all uncovered user intents. Maximal Marginal Relevance (MMR) ranks user intents by choosing the next one that maximizes the marginal relevance in Equation 1.

$$\lambda Rel_Q(T_i) - (1-\lambda) \max_{T_j \in S} Sim(T_i, T_j) \quad (1)$$

The marginal relevance is a linear combination of relevance and information novelty with the parameter λ trading off these two factors. The first term $Rel_Q(T_i)$ measures the relevance of user intent T_i to query Q. The relevance of a user intent can be measured using its popularity in the data. User intents associated with more user behavior data (e.g. the clicked items) are more popular and thus more relevant to the query. Since each item is characterized by a mixture of user intents, we calculate the popularity of a user intent T_i by summing the proportion of that user intent over all the items in the data, i.e. $Rel_Q(T_i) = \frac{1}{D}\sum_{d=1}^{D} \theta_{d,i}$, where $\theta_{d,i}$ specifies the proportion of user intent T_i in item d. As an equivalent of measuring the information novelty of a user intent, we measure its redundancy. The second term $\max_{T_j \in S} Sim(T_i, T_j)$ measures the redundancy of user intent T_i with respect to the set of selected user intents S. The redundancy is low when the user intent is very different from any selected user intent. Since each user intent has its own distribution over the terms in the vocabulary, we measure the similarity between two user intents using the cosine similarity. The best value of parameter λ is tuned on the validation data. Again, ranking user intents can be done offline and we save the ordering of user intents for real time ranking in the last step.

2.3 Selecting items for user intents

Finally, we select the most representative item for each user intent to display. The form of a user intent in the LDA model allows for fast scoring so that we can find the most representative item for a user intent very efficiently. Since a user intent is characterized by a multivariate Bernoulli distribution in the MB-LDA model, the parameters specifying the probability of term occurrence indicate the weight of a term within that user intent. To find the most representative item for a user intent, we score all the items in the inventory using the term weights of that user intent and choose the item with the maximum score.

Given the term weights within a user intent, we score each item by summing the weights of matching terms in the title using Equation 2.

$$Score(I_i, T_k) = \frac{\sum_{j=1}^{|I_i|} \beta_{k,W_{i,j}}}{max(AvgTitleLength_Q, |I_i|)} \quad (2)$$

where $W_{i,j}$ denotes the jth word in the title of item I_i and $\beta_{k,W_{i,j}}$ specifies the weight of term $W_{i,j}$ in user intent k. In addition, we normalize the score by the max of the title length and the average title length of query Q. This normalization can help prevent title spamming[1] by penalizing listings with an unusually long title, and avoid showing

[1]The sellers put irrelevant terms in the title to get higher exposure.

listings with very short title since they are generally less informative.

3. AVERAGE SATISFACTION METRIC

Classical metrics (e.g. precision, recall and NDCG) can not be used to evaluate diversity in the search results because they do not take into account the correlation between items. In addition, studies [21, 14] have shown that accuracy metrics do not necessarily correlate with user satisfaction. Variants of classical metrics have been proposed to evaluate diversity in different settings [7, 1]. However, these metrics either require human judgement or rely on existing categories. In this section, we propose a new metric to evaluate diversity in the search results based on user satisfaction and justify its use in the e-commerce domain.

Recognizing that users with different purchase intents use the same query, the goal of ranking then becomes to capture the interests of as many users as possible and as quickly as possible. This can be expressed through risk minimization framework [27, 1]. The risk of a ranker is measured by the proportion of users that can not find any relevant items. Instead of measuring the risk, we can measure the user satisfaction. **Average Satisfaction (AS)** for query Q is defined to be the user satisfaction with respect to the top N items in the ranking list R, averaged across all the users $\mathbf{U}_Q$. Average satisfaction can be written as Equation 3 where $Satisfaction(U_j, R_N)$ specifies the satisfaction of user U_j on the top N items in a ranking list R. In addition, **Mean Average Satisfaction (MAS)** is defined to be the mean of the average satisfaction over the top N rank positions shown in Equation 4.

$$AS_N(R,Q) = \frac{1}{|\mathbf{U_Q}|} \sum_{U_j \in \mathbf{U_Q}} Satisfaction(U_j, R_N) \quad (3)$$

$$MAS_N(R,Q) = \frac{1}{N} \sum_{n=1}^{N} AS_n(R,Q) \quad (4)$$

Average satisfaction takes into account both relevance and diversity based on user demands, making it a natural metric for the e-commerce domain. To show that, we use the simplified example on query *fossil*. Assume there are 10 users in which 5 users are interested in Fossil bag, 3 users are interested in Fossil watch and 2 users are interested in antique fossils. The satisfaction of a user is 1 if there exists at least one item matching his interest and 0 otherwise. Given the ranking results from three different rankers {Fossil bag, Fossil watch, antique fossils}, {Fossil watch, antique fossils, Fossil bag} and {Fossil bag, Fossil bag, Fossil watch}, the average satisfaction of these three ranking results at rank position N from 1 to 3 are: {0.5, 0.8, 1.0}, {0.3, 0.5, 1.0} and {0.5, 0.5, 0.8}, where ranker 1 has the largest mean average satisfaction of 0.767. At $N = 3$, both ranker 1 and 2 satisfy all the users while ranker 3 dissatisfies 20% of the users. Ranker 3 orders items according to their probabilities of relevance, but includes similar items, resulting in redundancy. Meanwhile, ranker 1 is better than ranker 2, because it captures the user interests more quickly by ranking them according to their probabilities of relevance.

The utility function, $Satisfaction(U_j, R_N)$, is hard to measure directly on e-commerce sites due to their rapidly changing inventory. Since the clicked items of a user indicate the user's interest, we define this utility function to be the similarity between the clicked items of user U_j and the most similar item on the ranking list R_N. The satisfaction of a user with multiple clicks can be calculated by averaging across all the clicked items. In practice, however, it is difficult to associate a clicked item with a user. Therefore, we treat each clicked item as if it is from a different user as an approximation. Finally, the average satisfaction of ranking R on query Q is shown in Equation 5 where I_j denotes the jth clicked item in the set of all clicked items C_Q for query Q and I_i denotes the ith item in the ranking list R_N.

$$AS_N(R,Q) = \frac{1}{|C_Q|} \sum_{I_j \in C_Q} \max_{I_i \in R_N} Sim(I_j, I_i) \quad (5)$$

We use the weighted cosine similarity function to measure the similarity between two items, where the weight of a term is query-dependent and reflects its relevance to the query. In our study, the weights are learned from a relevance model described in Section 4.

Given our definition of the utility function, average satisfaction and mean average satisfaction are monotonically increasing functions of N as adding more items to the ranking list never reduces the user satisfaction. Our proposed metric is similar to the metric used in [17] except that their utility function is calculated using user feedback.

4. EVALUATION

4.1 Data description

We evaluate our approach using the data at eBay, one of the largest online marketplaces. eBay manages over 90 million active users worldwide, 200 million items for sale and 8 billion URL requests daily. Unlike web search, eBay works in a dynamic environment where about 10% of the items in the inventory are listed, expired or sold every day. Given a wide variety of products in the inventory, eBay maintains a very large and complex taxonomy, with more than 18,500 leaf categories. However, the design of different subtrees in the eBay product taxonomy depends on different expert knowledge. The large-scale data, the dynamic environment and the complex and noisy product taxonomy make diversified retrieval extremely challenging at eBay.

In our study, we randomly sample 120 queries from all the queries submitted to eBay in two weeks (excluding those extremely rare queries). These queries cover a wide spectrum of query frequencies with some queries having as many as 30,000 daily submissions and as few as 300 daily submissions. For each query, we collect two months of user clicked data where items with multiple clicks have multiple copies in the data. Then, we randomly sample 10,000 items for training, 10,000 items for evaluation and another 2,000 items as validation data for parameter tuning. To leverage the category of an item, we add the item category as an extra term to the item title. For example, the item *iPhone 4 black ATT 16GB* in category *Cell Phones and Smartphones* is converted into a bag of words with "cat-Cell Phones and Smartphones" added in as an extra term. Items in more than one category have multiple categorical terms added into the bag of words.

4.2 Experiment

One way to evaluate a model on e-commerce sites is to run the A/B testing and calculate the online metrics, e.g. click through rate (CTR). However, the online metrics in the A/B

fossil		basketball		iPod	
Fossil bag and purse	"cat-Handbags and Purses" purse handbag bag leather tote shoulder key cross	*basketball*	"cat-Basketball" official size spalding ball 29.5 street	*iPod touch*	"cat-Portable Audio & Headphones" touch apple gb 8th generation
Fossil men watch	"cat-Wristwatches" watch men chronograph mens ch stainless steel fs	*basketball shoes*	"cat-Men's shoes" shoes nike size mens air black adidas	*iPod nano*	nano "cat-Portable Audio & Headphones" gb generation 8th model 4th
Fossil women watch	"cat-Wristwatches" watch es women stella relic gold dial by	*basketball card*	"cat-cards" card jordan lot michael auto rookie topps	*iPod case*	for case usb iPhone touch 4th new apple
Fossil wallet	wallet "cat-Wallets" leather clutch nwt brown new coin	*basketball shirt*	nike shirt shorts xl "cat-Men's Clothing" new jersey	*iPod charger*	for charger usb iPhone cable 4th mp 3rd
antique fossils	ammonite shark "cat-Shark Teeth" dinosaur "cat-Amphibian,Reptile and Dinosaur" "cat-Ammonites" tooth	*basketball hoop*	hoop "cat-Basketball" backboard rim nba portable in ground	*iPod classic*	"cat-Portable Audio & Headphones" classic apple 5th 30 gb black generation

Table 1: The user intents of query *fossil*, *basketball*, and *iPod* using the MB-LDA model with 5 topics.

testing may not accurately capture the aspect of user satisfaction and the A/B testing itself requires many resources. Alternatively we evaluate our models using the offline metric *average satisfaction* defined in Section 3. In our experiment, we compare six different approaches in ranking the eBay inventory, including the eBay production ranker, Maximal Marginal Relevance (MMR), Probabilistic Latent Maximal Marginal Relevance (PLMMR), the category-based approach and two LDA-based approaches (LDA and MB-LDA).

The eBay ranker is a machine learning ranker with many features built in to improve sales on eBay. Since the eBay ranker scores each item independently, the diversity in the search results is introduced as a side-effect. We use the eBay ranker as our baseline and compare the other approaches against it.

MMR [5] takes the ranking from a relevance model and reranks the items by trading off their relevance and novelty. The relevance model used in eBay makes the bag-of-word assumption and learns the weight for each term in the vocabulary based on both clicked and skipped user data. In general, the weight of a term is proportional to the number of occurrences in the clicked data. However, a term may get a negative weight if it rarely occurs in the clicked data, but frequently in the skipped data[2]. The relevance model scores an item by summing the weight of the terms in its title and normalizing it by the average title length of the query. Given the output from the eBay relevance model, MMR reranks 500 most relevant items by greedily choosing the next item that maximizes the marginal relevance. The parameter λ, that makes trade-offs between relevance and novelty, is chosen from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and the best value of λ is tuned on the validation data.

PLMMR [11] is a probabilistic latent version of MMR and takes into account the hidden user intents. Like our approach, PLMMR learns an LDA model for each query and then represents the document and the query as probability vectors of latent topics. The relevance term and novelty term are then calculated in the space of latent topics. In the experiment, the same LDA models learned in our approach are used in PLMMR.

[2]The skipped data is defined to be the items above the first clicked item on the search result page.

The category-based approach selects diverse items based on the eBay product taxonomy. First we estimate the weight of each leaf category for a query based on the user demands, sort them in decreasing order, and choose the top-ranked item by the relevance model within a category.

The LDA-based approach For each query, we first construct its vocabulary by removing the rare terms whose frequency is below 1% in the training data. For both the LDA model and the MB-LDA model, we set the number of topics to be 10, use the symmetric hyper parameter α and η with value 0.1, and run collapsed Gibbs sampling for 5,000 iterations and check on the convergence. The trade-off parameter λ is chosen from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and the best value of λ is tuned on the validation data.

5. RESULTS AND DISCUSSION

5.1 Discovered user intents

First, we examine whether the LDA models discover semantically meaningful user intents for a query. We use queries *fossil*, *basketball* and *iPod* as our illustrative examples and more examples can be found in Appendix B. We run the MB-LDA model with 5 topics, and show the discovered user intents in Table 1.[3] The left column shows the manually interpreted labels of the user intents and the right column shows the terms of the largest probability of occurrence. The MB-LDA model indeed discovers meaningful user intents. Firstly, the MB-LDA model is able to distinguish different categories of a query and associate the terms with the corresponding category. For example in query *basketball*, the model is able to associate *spalding* and *ball* with category "Basketball", and *nike* and *shoes* with category "Men's shoes". Secondly, the discovered user intents are more expressive than the eBay product taxonomy. In query *iPod*, the MB-LDA model is able to further split all iPod products in one giant category "Portable Audio & Headphones" into more detailed user intents *iPod touch*, *iPod nano* and *iPod classic*, which better captures users' different information needs. Finally, the MB-LDA model is able

[3]We only show the results from the MB-LDA model due to space limitations, but the user intents discovered by the LDA model are similar.

to combine similar categories adaptively. The user intent *antique fossils* of query *fossil* includes three different categories of antique fossils. Given the relatively small user demands on antique fossils, combining different categories of antique fossils into one user intent helps save the space to show items for other popular user intents.

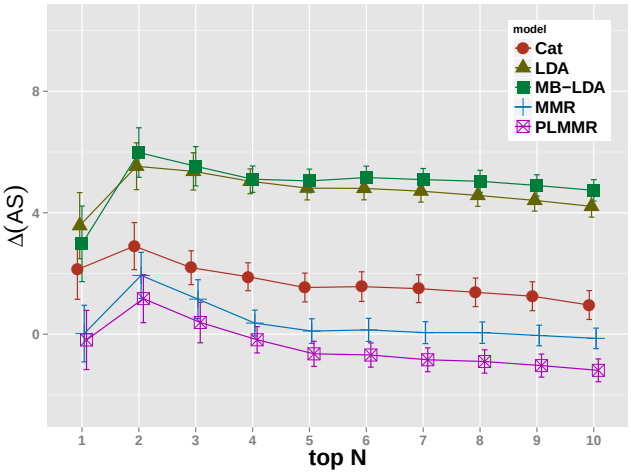

Figure 3: The mean and 95% CI of the difference of average satisfaction for different approaches against the eBay ranker on all testing queries. The X axis denotes the rank position N and the Y axis denotes the difference of average satisfaction.

5.2 Performance against the eBay ranker

Next, we compare five different approaches against the eBay ranker. For each query, we rank the eBay inventory with different approaches, evaluate on the top 10 diverse items, and compute the difference of average satisfaction against the eBay ranker. In Figure 3, we show the mean and 95% confidence interval of the difference of average satisfaction on all testing queries. The paired t-test indicates only the LDA-based approaches and the category-based approach are statistically better than the eBay ranker across all top 10 rank positions and they improve the baseline by as much as 6% and 3% respectively. The MB-LDA model is better than the LDA model even though the difference is not statistically significant. In Table 2, we show the mean average satisfaction on top 10 items in the first row. The MB-LDA model shows the best overall performance and is more than 3% better than the category-based approach. Since both LDA models perform similarly with the MB-LDA model being slightly better. For the sake of clarity, we use the results of the MB-LDA model in the following analyses.

Both MMR and PLMMR do not work well. MMR fails to consider the user intents, thus the selected items may not represent user intents well and certainly can not cover all different user intents of a query. Though PLMMR considers the hidden user intents, it does not consider the length of an item title in computing the relevance and diversity, therefore it may choose items with either very short or very long titles, which are less informative. Since MMR and PLMMR are obvious less effective in improving user satisfaction, we focus on the comparison between the LDA-based approach and the category-based approach.

Table 2: The mean average satisfaction at rank 10 on different types of queries. The approaches highlighted in bold indicate the best performer in that type of queries.

Query Type	eBay ranker	MMR	PL-MMR	Cat	LDA	MB-LDA
All	68.1%	68.7%	68.2%	69.8%	73.2%	**73.7%**
High	60.3%	60.6%	60.1%	66.0%	66.5%	**67.4%**
Low	74.1%	75.1%	74.8%	72.7%	**78.3%**	**78.3%**

Table 3: The queries that the MB-LDA based approach outperforms and underperforms the category-based approach the most. + indicates the MB-LDA model is better and - otherwise.

Query	Daily submissions	Ambiguity score	$\Delta(AS_3)$
ipad 1st generation	100	9.4	+15.5%
verizon cell phones	11K	0.6	+12.8%
nike 60	600	14.8	+12.7%
iphone 4	1.3M	7.2	+12.2%
unlocked cell phone touch screen	3.6K	1.1	+12%
ps 3	11K	12.7	+11.3%
porcelain doll	800	21.9	-4.2%
lego	800	21.8	-2.9%
coach	23K	25.9	-2.7%

To better understand the different behaviors of the LDA-based and the category-based approaches, we show the queries where these two models differ the most in ranking the top 3 items and their statistics in Table 3. The ambiguity score of a query describes how ambiguous the query is based on the distribution of user demands over different categories [16]. The LDA-based approach provides 10-15% more average satisfaction in the best case and only about 4% less in the worst case. Since the category-based approach heavily depends on the existing taxonomy, it becomes less effective when the granularity of the part in the eBay taxonomy is coarse. The LDA-based approach, however, discovers user intents based on item titles and is not affected much by the quality of the eBay taxonomy. As a result, the LDA-based approach works better on queries of low ambiguity. In queries where the LDA-based approach does worse, the item titles are often noisy and less informative, thus the LDA model can not discover meaningful and stable user intents. For example in query *lego*, the sellers use very different words in the title to describe similar products and there is a large variety of different lego products on the market. On the other hand, the eBay product taxonomy is well designed for this type of product and the category-based approach is able to achieve better performance on this type of queries.

5.3 Queries of different ambiguity

To further investigate the model performance, we run the same analysis on two groups of queries, one including queries of high ambiguity and the other one including queries of low ambiguity. In particular, we take 25% most and 25% least ambiguous testing queries based on their ambiguity scores for this analysis. The results are shown in Figure 4. The

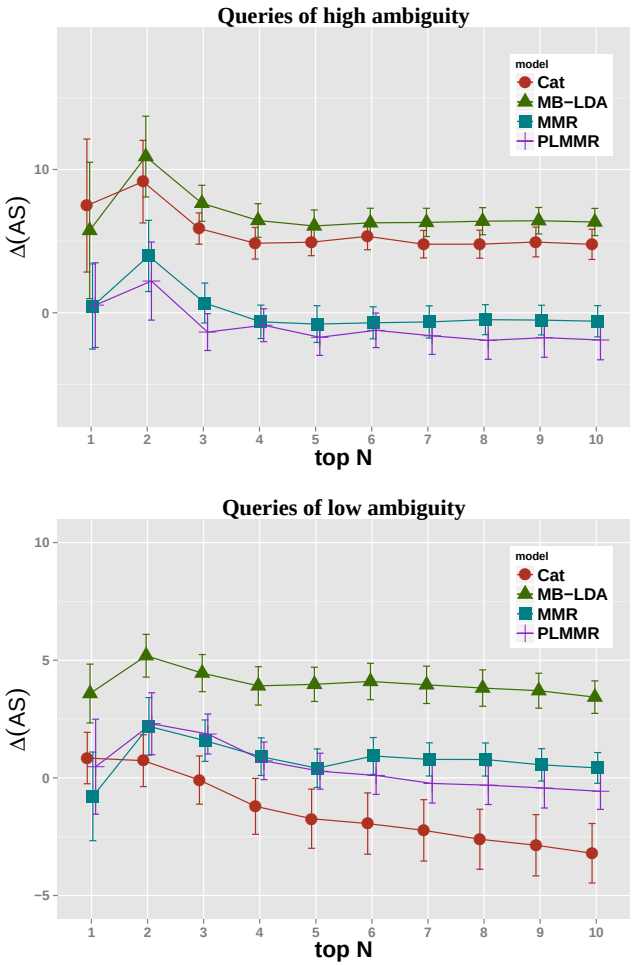

Figure 4: The mean and 95% CI of the difference of average satisfaction for different approaches against the eBay ranker on queries of high ambiguity (top) and queries of low ambiguity (bottom).

figure on the top shows the difference of average satisfaction on queries of high ambiguity. Both the LDA-based and the category-based approaches perform statistically better than the eBay ranker. Since there are often more different user intents associated with queries of high ambiguity, more items needs to be selected to cover all different user intents, resulting in lower mean average satisfaction (the second row in Table 2). Again, the MB-LDA model shows the best performance and is able to improve the eBay ranker by as much as 7% on mean average satisfaction.

The difference between the LDA-based and the category-based approaches becomes smaller on queries of high ambiguity. Since the granularity of the eBay taxonomy is often more refined on this type of queries, the category-based approach is able to capture most of the user intents based on the category information. However, the LDA-based approach is able to handle the correlations between categories and the overlapping subtrees in the taxonomy. For example, we show the top three items of query *fossil* selected by the eBay ranker, the category-based approach and the LDA-based approach in Table 4. All three items selected by the eBay ranker are bags and purses, showing very limited diversity. The category-based approach shows a bag, a watch and a purse. Though the bag and the purse are from different categories, they are actually similar due to the strong correlation between the categories *Handbags and Purses* and *Wallets*. The LDA-based approach selects a bag, a watch and a set of fossil teeth, and clearly shows the most diversity in the search results.

	eBay ranker	Cat	MB-LDA
1			
2			
3			

Table 4: The top three items selected by the eBay ranker, the category-based approach and the LDA based approach for query *fossil*.

The figure on the bottom in Figure 4 shows the difference of average satisfaction on queries of low ambiguity. As we expected, the difference of average satisfaction on this type of queries is smaller because they are in general less ambiguous. The category-based approach performs badly on this type of queries due to the coarse granularity of the subtrees in the eBay product taxonomy. The LDA-based approaches are still very effective in showing diverse items in the search results and is the only method that is statistically better than the eBay ranker. The third row in Table 2 shows the mean average satisfaction of different approaches on queries of low ambiguity, both the LDA model and the MB-LDA model perform the best and improve the eBay ranker by around 4% on mean average satisfaction.

5.4 The number of user intents

In our last analysis, we investigate how the number of user intents affects the performance of the LDA-based approach. We run the MB-LDA model with 10 and 20 topics and plot the difference of average satisfaction against the eBay ranker in Figure 5. The MB-LDA (K=10) model outperforms the MB-LDA (K=20) model early in the ranking, but becomes worse afterwards. The MB-LDA model with more topics discovers more specific user intents, but these user intents are sometimes too specific to users and are not as effective as more abstract user intents. For example in query *fossil*, it would be less effective to display a men watch and a women watch if only two slots are allowed to show diverse items. However, as we have more slots to show the diverse items in the search result page, more specific user intents become better in satisfying users with a particular interest. Depending on the number of slots we have to show the diverse items in the search result page, we can adjust the number of

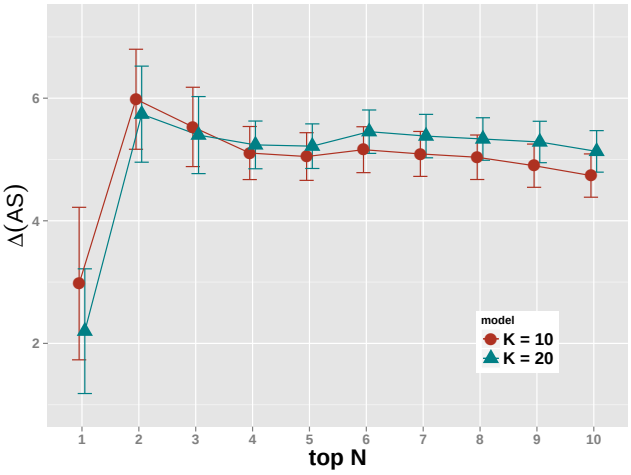

Figure 5: The mean and 95% CI of the difference of average satisfaction for the MB-LDA model with 10 and 20 user intents on all testing queries.

topics in the LDA model accordingly to achieve better user satisfaction.

Given the analyses we have done so far, we try to give some insights on when users benefit the most from introducing the diversity in the search results and which diversified retrieval approach we should apply. On queries of high ambiguity, introducing diversity in the search result page can improve user satisfaction significantly, while the gain is generally less on queries of low ambiguity. Out of all the queries we examine, there are about 15% queries where the LDA-based approach provides 10% or more average satisfaction than the eBay ranker, and there are only 5% queries where the LDA-based approach does worse. The LDA-based approach is preferred when the item titles of the query are descriptive and relatively noisy-free and the granularity of the subtrees in the product taxonomy associated with the query is coarse. On the other hand, the category-based approach is preferred when the product taxonomy is well designed and its granularity is fine-grained.

6. RELATED WORK

Carbonell and Goldstein [5] first looked into the problem of diversity in document summarization. They proposed the Maximal Marginal Relevance (MMR) approach where items are ranked by a linear combination of their relevance and novelty. Along this line of research, Zhang et al. [29] proposed different measures to compute redundancy in adaptive information filtering system. Guo and Sanner [11] derived a probabilistic latent view of the MMR that can balance relevance and diversity automatically. Instead of ranking items by the marginal function, [4, 6] maintain a probability of relevance conditioned on the previously selected documents and update it whenever a new document is selected. Bookstein [4] updated this probability of relevance using user feedback. Chen and Karger [6] proposed the *1-greedy* approach which updates the probability of relevance by treating all previously selected documents as irrelevant.

Another line of research used the Markov walk on a graph to select the diverse items. This approach first converts the problem into a graph where each item is a node in the graph and the edge, measuring the similarity between two items, defines the transition probability. Then the diverse items are selected based on different criterions, including the information richness [28], a reinforced random walk [15] and an absorbing random walk [30]. Instead of calculating the weight of the edge with some specific function, Dubey et al. [8] learned the edge weights from data and then selected the center nodes that maximize the entropy of the conductance between the center nodes and the rest of the graph.

The above approaches work without assuming an existing taxonomy of the data. Diversified retrieval based on an existing taxonomy has also been studied in different settings. Zhai et al. [26, 27] formalized the problem of diversity as subtopic retrieval where the goal is to find documents that cover many different subtopics of a query. Yue and Joachims [25] assumed the topic coverage of a document is unknown and proposed a structural SVM framework to learn a mapping from words in a document to the topic coverage. This approach penalizes low diversity in the loss function of the structural SVM. Li et al. [13] extended the structural SVM framework by explicitly adding constraints for diversity, coverage and balance respectively and learning a function to select a subset of sentences for document summarization. Ziegler et al. [31] applied topic diversification in book recommendation where they measured the similarity between two product sets based on the taxonomy and demonstrated that the users prefer diversified recommendations. Agrawal et al. [1] studied the problem of answering ambiguous web queries and proposed a greedy approach *IA-Select* to minimize the risk of dissatisfaction of the average user. Different from previous work, it takes into account both the importance of a category and the relevance of a document for a category. Welch et al. [24] proposed a search diversification algorithm where the users require multiple relevant documents to satisfy their needs.

Only a few approaches have addressed the diversity in online search engines. Gollapudi and Sharma [9] pointed out it is key to understand user intents in designing an effective ranking system in the search engine. They proposed a set of axioms for result diversification, provided three diversification objectives based on these axioms and demonstrated its effectiveness on semantic and product disambiguation. Radlinski and Dumais [17] used related search to address the diversity in personalized web search. Their approach first finds the related queries of the search query and then outputs a combination of the results from the related queries. Vee et al. [22] studied the problem of diversity in online shopping applications. They assumed each item is in the form of attribute-value pairs and the goal is to find an ordering of attributes to maximize diversity. Radlinski et al. [18] learned a diverse ranking of documents using multi-armed bandits. Their approach works in an online fashion where it requires user feedback in the form of clicks. Brandt et al. [9] proposed a dynamic ranking framework where the search engine dynamically generates the rankings based on user feedback.

7. CONCLUSION

We develop a Latent Dirichlet Allocation based diversified retrieval approach to address the unique challenges of product search on e-commerce sites. We also propose a Multivariate Bernoulli LDA model to target the short text without

duplicates in the e-commerce data. To evaluate the quality of the search results, an offline metric *average satisfaction* is proposed to measure user satisfaction. Our empirical study on the eBay data shows that the LDA model can discover meaningful user intents and the LDA-based approach outperforms the eBay production ranker and three other diversified retrieval approaches. In particular, our approach can improve the eBay production ranker on average satisfaction by as much as 6% on average and more than 10% on the ambiguous queries. Finally, we shed light on when a diversified retrieval approach should be applied and which diversified retrieval approach we should use so that users can benefit the most from introducing the diversity in the search results on e-commerce sites.

For future work, we would like to extend the LDA model to include other item-level metadata and to train the LDA model on other types of user behavior data to further improve user satisfaction. In addition, we would like to evaluate user feedback on including the "See more items like this" option in the search result page.

Acknowledgments

We would like to thank Daniel Miranda and Nadia (Ghamrawi) Vase for their helps on the eBay title relevance model.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] A. Anagnostopoulos, A. Z. Broder, and D. Carmel. Sampling search engine results. In *WWW*, pages 245–256, 2005.

[3] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] A. Bookstein. Information retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34(5):331–342, 1983.

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and reproducing summaries. In *SIGIR*, 1998.

[6] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pages 429–436, 2006.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[8] A. Dubey, S. Chakrabarti, and C. Bhattacharyya. Diversity in ranking via resistive graph centers. In *KDD*, pages 78–86, 2011.

[9] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW*, pages 381–390, 2009.

[10] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

[11] S. Guo and S. Sanner. Probabilistic latent maximal marginal relevance. In *SIGIR*, pages 833–834, 2010.

[12] M. Hoffman, D. M. Blei, and F. Bach. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[13] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *WWW*, pages 71–80, 2009.

[14] M. R. McLaughlin and J. L. Herlocker. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *SIGIR*, pages 329–336, 2004.

[15] Q. Mei, J. Guo, and D. Radev. Divrank: the interplay of prestige and diversity in information networks. In *KDD*, pages 1009–1018, 2010.

[16] D. P. Putthividhya. ebay's internal report. 2011.

[17] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR*, pages 691–692, 2006.

[18] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, pages 784–791, 2008.

[19] S. E. Robertson. Readings in information retrieval. chapter The probability ranking principle in IR, pages 281–286. Morgan Kaufmann Publishers Inc., 1997.

[20] J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *SIGIR*, pages 757–758, 2007.

[21] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR*, pages 11–18, 2006.

[22] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *ICDE*, pages 228–236, 2008.

[23] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *AAIM*, pages 301–314, 2009.

[24] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *WWW*, pages 237–246, 2011.

[25] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *Proceedings of the 25th international conference on Machine learning*, pages 1224–1231, 2008.

[26] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17, 2003.

[27] C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55, 2006.

[28] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR*, pages 504–511, 2005.

[29] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR*, pages 81–88, 2002.

[30] X. Zhu, A. B. Goldberg, J. V. Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, pages 97–104, 2007.

[31] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW*, pages 22–32, 2005.

APPENDIX

A. THE COLLAPSED GIBBS SAMPLER FOR MULTIVARIATE BERNOULLI LDA

Assume we have D documents in the corpus, a vocabulary of size V and a predefined number of topics K. Associated with each term v in document d, there exists a topic assignment $Z_{d,v}$ and a term occurrence $W_{d,v}$ indicating whether term v appears in document d. The topic proportion θ_d specifies a multinomial distribution of topics associated with document d and the probability of occurrence $\beta_{k,v}$ specifies a Bernoulli distribution of term v in topic k. α and η are the hyper parameters of the prior distributions where θ and β are generated from. The joint probability of the MB-LDA model is written as follows.

$$P(W, Z, \theta, \beta; \alpha, \eta)$$
$$= \prod_{k=1}^{K}\prod_{v=1}^{V} P(\beta_{k,v}; \eta) \prod_{d=1}^{D} P(\theta_d; \alpha) \prod_{v=1}^{V} P(Z_{d,v}|\theta_d) P(W_{d,v}|\beta_{Z_{d,v},v})$$

The collapsed Gibbs sampler integrates out the hidden variables θ and β so that the Gibbs sampling converges faster. After integrating out θ and β, we get the following marginal probability.

$$P(W, Z; \alpha, \eta) = \prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(n^{\cdot}_{d,k,\cdot} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} n^{\cdot}_{d,k,\cdot} + \alpha_k)} \prod_{k=1}^{K}\prod_{v=1}^{V} \frac{\Gamma(\eta_1 + \eta_0)}{\Gamma(\eta_1)\Gamma(\eta_0)} \frac{\Gamma(n^1_{\cdot,k,v} + \eta_1)\Gamma(n^0_{\cdot,k,v} + \eta_0)}{\Gamma(n^1_{\cdot,k,v} + \eta_1 + n^0_{\cdot,k,v} + \eta_0)} \tag{6}$$

where $n^1_{d,k,v}$ is 1 when term v occurs in document d and has topic assignment k and is 0 otherwise. 1 on the superscript specify the occurrence of term v in document d. We use $\cdot$ to indicate that we sum over all configurations of that variable. For example, $n^0_{\cdot,k,v}$ is the number of documents where term v does not occur and has topic assignment k.

The collapsed Gibbs sampler approximates the conditional distribution $P(Z|W; \alpha, \eta)$ by sampling the topic assignment iteratively. More specifically, it samples the topic assignment $Z_{a,b}$ for term b in document a from the conditional distribution $P(Z_{a,b}|Z_{\neg(a,b)}, W; \alpha, \eta)$ while keeping the other topic assignments $Z_{\neg(a,b)}$ fixed. For topic k', the conditional probability $P(Z_{a,b} = k'|Z_{\neg(a,b)}, W; \alpha, \eta)$ is proportional to the marginal distribution in Equation 6. Finally, the collapsed Gibbs sampler for the MB-LDA model can be written in the following simple form.

$$P(Z_{a,b} = k'|Z_{\neg(a,b)}, W; \alpha, \eta)$$
$$\propto (n^{\cdot,\neg(a,b)}_{a,k',\cdot} + \alpha_{k'}) \frac{n^{W_{a,b},\neg(a,b)}_{\cdot,k',b} + \eta_{W_{a,b}}}{n^{\cdot,\neg(a,b)}_{\cdot,k',b} + \eta_1 + \eta_0}$$

Once the collapsed Gibbs sampler converges, we can estimate the parameters θ and β using the hidden topic assignments.

$$\theta_{d,k} = \frac{n^{\cdot}_{d,k,\cdot} + \alpha_k}{\sum_{k'=1}^{K} n^{\cdot}_{d,k',\cdot} + \alpha_{k'}}$$

$$\beta_{k,v} = \frac{n^1_{\cdot,k,v} + \eta_1}{n^1_{\cdot,k,v} + \eta_1 + n^0_{\cdot,k,v} + \eta_0}$$

B. USER INTENTS OF MORE QUERIES

Table 5: The user intents of query *keyboard*, *hello kitty*, *mirror*, and *timberland* using the MB-LDA model with 4 topics.

keyboard

computer keyboard	usb "cat-Keyboards, Mice & Pointing" black new dell hp gaming
silicone keyboard	"cat-Keyboards, Mice & Pointing" silicone usb roll foldable computer
musical keyboard	"cat-Electronic Instruments" electronic electric piano key 61 casio music yamaha
keyboard combo	"cat-Keyboards, Mice & Pointing" mouse combo wireless bluetooth mini logitech

hello kitty

hello kitty watch	watch "cat-Watches" girl quartz pcs wrist gift new lady
hello kitty case	case for iPhone cover "cat-Cell Phone Accessories" 4th 3rd
hello kitty purse and bag	bad pure "cat-Women's Handbags & Bags" tote handbag shoulder
hello kitty necklace	"cat-Animation Characters" necklace crystal girl gift fashion bow

mirror

wall mirror	wall "cat-Home Decor" decorative decor new vintage large
antique mirror	antique "cat-Decorative Arts" vintage wood with gold
motorcycle mirror	"cat-Motorcycle Parts" harley motorcycle chrome for honda black
car mirror	side car "cat-Car Exterior" view power door left rear

timberland

timberland shirt	men shirt size "cat-Men's Clothing" jacket new nwt
timberland men boots	boots "cat-Men's Shoes" men wheat work waterproof
timberland women boots	boots "cat-Women's Shoes" women leather new size
timberland shoes	"cat-Men's Shoes" shoes classic boat men brown