Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models

Anjan Goswami[**]
agoswami@ucdavis.edu
University of California Davis

Prasant Mohapatra
pmohapatra@ucdavis.edu
University of California Davis

Chengxiang Zhai[†]
czhai@illinois.edu
University of Illinois
Urbana-Champaign

ABSTRACT

The demand generation and assortment planning are two critical components of running a retail business. Traditionally, retail companies use the historical sales data for modeling and optimization of assortment selection, and they use a marketing strategy for demand generation. However, today, most retail businesses have e-commerce sites with rapidly growing online sales. An e-commerce site typically has to maintain a large amount of digitized product data, and it also keeps a vast amount of historical customer interaction data that includes search, browse, click, purchase and many other different interactions. In this paper, we show how this digitized product data and the historical search logs can be used in understanding and quantifying the gap between the supply and demand side of a retail market. This gap helps in making an effective strategy for both demand generation and assortment selection. We construct topic models of the historical search queries and the digitized product data from the catalog. We use the former to model the customer demand and the later to model the supply side of the retail business. We then create a tool to visualize the topic models to understand the differences between the supply and demand side. We also quantify the supply and demand gap by defining a metric based on Kullback-Leibler (KL) divergence of topic distributions of queries and the products. The quantification helps us identifying the topics related to excess or less demand and thereby in designing effective strategies for demand generation and assortment selection. Application of this work by e-Commerce retailers can result in the development of product innovations that can be utilized to achieve economic equilibrium. We can identify the excess demand and can provide insight to the teams responsible for improving assortment and catalog quality. Similarly, we can also identify excess supply and can provide that intelligence to the teams responsible for demand generation. Tools of this nature can be developed to systematically drive efficiency in achieving better economic gains for the entire e-commerce engine. We conduct several experiments collecting data from Walmart.com to validate the effectiveness of our approach.

CCS CONCEPTS

• **Information systems** → *Business intelligence.*

KEYWORDS

E-commerce search; Information retrieval; Business Analytics; Topic Models; Marketplace economics

ACM Reference Format:
Anjan Goswami, Prasant Mohapatra, and Chengxiang Zhai. 2019. Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316605

1 INTRODUCTION

Demand generation and assortment selection are two critical components of a retail business. Demand generation is essential for growing the business by acquiring customers interested in purchasing from the retail shop while optimal assortment selection contributes to the maximization of sales and revenue. These two areas always attracted substantial investment from retail businesses. The demand generation is typically achieved by employing effective marketing strategies. Often, this involves making statistical models of the customer segments who may be interested in some of the products of a retailer and reaching out to those people. The assortment selection algorithms use actual sales and product similarity data for optimizing the selection of assortments for maximizing the sales or revenue. However, today, most retailers also posses an e-commerce business, and the market share of sales from the e-commerce business has been rapidly increasing compared to the brick and mortar stores throughout the last decade. The e-commerce sites typically maintain a vast amount of digitized product data that includes product title, description, it's pictures, brand, price and many other attributes. E-commerce websites log user information and actions, require identification for purchase, and provide mechanisms by which users can explicitly state their demand. This is most notable in e-Commerce search where a user can query for the products of interest. Hence, by understanding the user queries, e-commerce companies can gain insight into the demand. Moreover, this demand data can then be coupled with product supply data to further understand the dynamics of supply and demand on e-commerce sites.

The preponderance of such data allows us to utilize statistical methods in understanding the real demand and supply for the products and develop plans to achieve an economic equilibrium as well as guide investment decisions on search engine optimization

[*] The author worked for Walmart Lab. when he conducted this research.

[†] The author was an advisor for R&D teams at Walmart Lab when he worked on this problem.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316605

and marketing (demand generation techniques), assortment choices, and technology.

In the case of large-scale e-commerce businesses that operate highly trafficked sites, the volume and variety of queries are well beyond meaningful human interpretation. To utilize this data and to make supply and demand decisions, e-commerce retailers require a scalable way to understand the broader themes on both sides and to be able to connect them. Topic models are unsupervised machine learning algorithms for mining data in large corpora that can uncover the underlying semantic structure of large data sets. They have been successfully applied to various types of data such as text, images, and biological data amongst others. In this paper, we present a novel application of topic models to systematically identify the supply and demand gap of an e-commerce engine from the textual contents of the query and the product data. We then correlate the gap with the revenue to quantify this. Identifying and quantifying the supply and demand gap is a practical problem for an e-commerce business. Using this, we can determine the cluster of queries that represent unmet demand on the site and the items that are rarely queried for and may need to be replaced. In this paper, we show that computing a distance between the topic models of a representative sample of queries on the topic space of the items for the e-commerce site can be a systematic way to obtain insights for both demand generation and the assortment planning. We evaluate our algorithm with a simulation study conducted using an extensive data set consisting of the historical search logs and the product catalog obtained from Walmart.com site.

Our paper makes the following key contributions:

- We address the problem of gaining insight about the supply and demand side gap for an e-commerce site from its search queries data and the product catalog using textual information.
- We propose an algorithmic framework that constructs topic models from query and items data and quantifies the gap between supply and demand.
- We provide an understanding of the economic dynamics of an e-commerce site so that areas most pertinent to a disequilibrium in supply and demand side can be identified using visualization and quantification.

2 BACKGROUND

In this paper, we extend that idea to show an emerging application of using topic models in providing insights about the potential directions of demand generation and assortment planning.

2.1 Topic models

Topic models are probabilistic generative models of the documents in a corpus. It assumes that each document consists of a mixture of topics that are shared across the corpus and each topic is a distribution over words. Once a model is learned, a document can be represented by a vector of topic probabilities. The strength of topic modeling is that it is an unsupervised algorithm and it requires only the corpora as input to compute the semantic relationship of the documents in the form of topics. One of the popular approaches of topic modeling is latent Dirichlet allocation (LDA) [7] that uses a Bayesian generative model. LDA algorithm has been proved to be successful in various applications beyond information retrieval and text mining including collaborative filtering, computer vision, and bioinformatics [4, 7, 16, 17]. The LDA algorithm's parameters are estimated using either sampling-based algorithms which attempt to approximate the posterior distribution from an empirical distribution or variation-based algorithms which try to find a distribution of best fit to the posterior by using a parameterized family of distributions over it [5]. Typical applications of LDA primarily use sampling-based algorithms, the most popular of which is Gibbs sampling which has been initially demonstrated by Griffiths and Steyver [12]. In this paper, we use the LDA algorithm to construct topic models for the set of queries in a fixed period and the set of documents from the product catalog that are in the inventory during that time. The queries and product titles are short text, and a direct application of LDA will not work. Hence, we first construct clusters of similar queries and items. Researchers [19] used similar clustering techniques for building topic models for short text.

2.2 Retail studies

Demand generation in retail is related to investing in marketing and advertising to attract potential customers to its product offerings. This involves traditional marketing [2], communicating promotions and deals to a potential audience [1], advertising the products in search engines or social networks [13, 21] etc. The assortment planning also has a rich literature in retailing [15]. It depends on many factors and often the process can be very complicated because several aspects of bringing a new product lines need to be considered [18]. It has been understood in retailing literature that it makes sense to get the insight about assortment planning from the consumer's search of products [8]. In this paper, we aim to obtain insight from actual e-commerce search data. Note that our algorithm intends to help in identifying some areas which are worth investigating for demand generation and assortment planning and hence it can be viewed as a complementary tool for the existing research in retailing in this area.

2.3 Distance between distributions

This is a classic area in statistics [3]. We use Kullback-Leibler divergence (KL-divergence) which is a popular such distance function in information retrieval [22] and has been in the past used to measure the quality of topic models [14].

3 TOPIC MODELS FOR E-COMMERCE SUPPLY AND DEMAND GAP

In this paper, we are interested in measuring the supply and demand gap of an e-commerce engine to understand (a) marketing needs to generate demand (b) assortment selection for unmet demand. Our basic idea is straightforward. We construct topic models from the textual content of the queries and the items. The topic models capture the underlying topic structure of the implicit economics of the e-commerce engine and define a distribution of topics in each corpus. We quantify the difference between these two topic distributions with a statistical distance function which measures the extent of the gap or disequilibrium between the supply and demand of an e-commerce engine. We then also use the existing click data to understand the business implication of the gap in terms of a click based engagement metric. Note that, we use only the click data but it is simple to extend this using sales or revenue data and we can then also obtain the expected impact on sales and revenue based business metrics. We visualize the differential word clouds from the top words of topic models from the supply and the demand side to obtain a visual insight about the supply and demand gap.

3.1 LDA on Search Query and Items data

The demand side of an e-Commerce platform can be represented by the set of all queries that are submitted to the search engine in a fixed period of time. The supply side can similarly be represented by the set of all items in the inventory during the same period of time. Each item can be associated with some textual data that typically includes the title, and the description. This textual data can be used for topic modeling. The title is expected to convey a good summary of any items. The description can have a lot of information and can be somewhat noisy. In our experiments, we combine the title and the description data that is generally available for most of the items in a catalog for a retail company. We henceforth mention the combination of this textual data from the items as documents. We compute two topic models, the first one is from the set of queries that represents the topics on the demand side and the second one is from the items that represent the supply side.

We intend to make these topic models per category. We thus first start preprocessing the corpus removing the stop words, punctuation, and other semantic elements that do not have meaning in the context of topic models. The documents are tokenized and tokens with low term frequency (TF) and low inverse document frequency (IDF) [20] are removed from the set. Then, we construct a fixed vocabulary of tokens from a category that we use for topic modeling. Let's call this vocabulary V. One of the challenges of applying topic models for the query data is that queries are very short snippets of text and we cannot build topic models only from such short texts. We thus construct a mechanism to cluster similar queries by grouping them based on a common item that is clicked when it is shown in search results for these queries. We filter out item-query document pairs below certain traffic and click threshold. This gives us sets of small clusters of similar queries. We do not have a similar challenge to apply topic modeling for the items data since typically items come with a document with product description and attributes along with the product titles. We use the topic modeling based on latent Dirichlet allocation (LDA) for our problem. It works first by assuming a fixed number of topics K and then by drawing the topics from a distribution over the fixed vocabulary V. This distribution is assumed to be Dirichlet. After this, we generate a proportion of topics per document using another K dimensional Dirichlet distribution. We then further generate a topic assignment per word per document using a multinomial distribution and also generate the words for the topics selected using yet another multinomial distribution. The central computational problem for LDA is the estimation of the joint posterior probability of hidden topics given the observation of the documents. We use a classic variant of LDA algorithm described in a paper by Blei et al [6] where this problem is solved using Gibb's sampling and variational inference. Readers can find the details of the LDA algorithm in the paper by Blei et al [6].

We discuss the computation of the supply-demand gap in section 3.2.

3.2 Quantifying Supply-Demand Gap

In this section, we provide an algorithm to quantify the supply-demand gap. We define two distributions p_d and p_s as the distribution over words for demand and supply respectively. The distance function is defined as $dist(p_s, p_d)$. We use Kullback-Leibler divergence to measure the similarity between the two probability density distributions:

$$D_F(p,q) = \sum p(i) \log \frac{p(i)}{q(i)}$$

KL divergence is non-negative and is 0 only when the two distributions are themselves equal. KL divergence is, however, asymmetrical. In order to make it symmetric, we define our distance function $dist(p_s, p_d)$ as:

$$dist(p_s, p_d) = \frac{D_F(p_s, p_d) + D_F(p_s, p_d)}{2}$$

4 EXPERIMENTAL EVALUATION

We use the data from Walmart's e-commerce site which is a high traffic website and one of the largest e-commerce site in the United States. We evaluate our algorithm using six month's of search query data and the product catalog data available during that time from Walmart's e-commerce platform. We create two data sets. In one data set, we have the query, items that are shown for the query, and the action taken by the customers such as clicks, sales. We also keep the price data to compute revenue. On the other hand, the product data consists of title, description and set of attributes of the product.

We use the R package "lda" [9] that uses a Gibbs sampling-based method for inference. We generated a 100-topic LDA model fit to the search query and the item data.

We show the results of four experiments.

(1) First we show visualizations of topics from queries to explain the effectiveness of our algorithm for applying LDA to the set of queries.
(2) We then show some top words per category to show the effectiveness of applying LDA to the items.
(3) We show the word clouds from query and the item side. These word clouds can be excellent tool to provide visual insight on the supply and demand gap in an e-commerce site.
(4) We also conduct an experiment to show the effectiveness of our KL-divergence based statistical distance between topic distribution to capture the relationship with engagement. Intuitively, we expect to observe that if the distance is more then there will be less clicks.

4.1 Visualization of Search Query Topics

In table 1, we show five top words from seven topics from the LDA models of the query clusters. This can provide an idea about effectiveness of our application of LDA to our search query (cluster) data.

Topics	word 1	word 2	word 3	word 4	word 5
1	frozen	disney	doll	elsa	queen
18	curtain	decorative	drape	sheer	panel
23	kayak	boat	fishing	inflatable	rod
31	sofa	futon	couch	leather	sectional
53	bathroom	shower	bath	towel	mainstays
64	stroller	carseat	graco	britax	combo
97	bed	full	twin	frame	platform

Table 1: Seven topics from a 100-topic LDA model fit to search log. A simple inspection of the words tells us the following about the topics: topic 1 is about frozen, the movie, topic 18 is about curtains and drapes, topic 23 is about fishing and kayaks, topic 31 is about sofas, topic 53 is about accessories of bathing, topic 64 is about carseats, and topic 97 is about beds.

4.2 Visualization of Items Related Topics:

In figure 1 we show six plots where the X-axis shows the uniqueness of the words which is the inverse document frequency and the Y-axis shows the term frequencies. The terms on the upper right corner are high in uniqueness and also high in frequency. We select some of the top words from some of the topics generated from the product catalog dataset. It is easy to see that each topic represents a specific product theme.

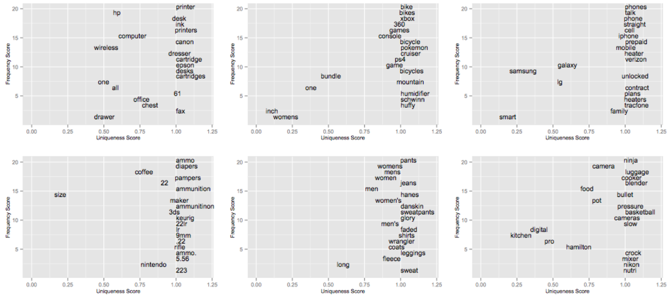

Figure 1: Frequency vs. uniqueness plots of words for select topics. Upper left plot can be considered about Samsung TV, the upper middle plot may have a few different products from furniture, upper right one is about smartphones, lower right one is about carseat, lower middle one is about laptops.

4.3 Visualization of Supply and Demand Gap

Now, we show some visualizations of the supply and demand sides topics using word clouds. This allows us to identify areas where there is excess supply and demand. In figure 2 we show a word cloud of the demand and supply of a given topic side by side. In order to obtain the visualizations, we weigh the demand side words by query traffic obtained for search log data and for the supply side, we weigh the item titles by the item count in the inventory from the product catalog. From visual inspection, we can identify that some large words like “women" and “shoe" are common in both our demand and supply sides. This indicates that customers may be searching for “Women’s’ Shoe” and the inventory also have adequate stock for Shoes.

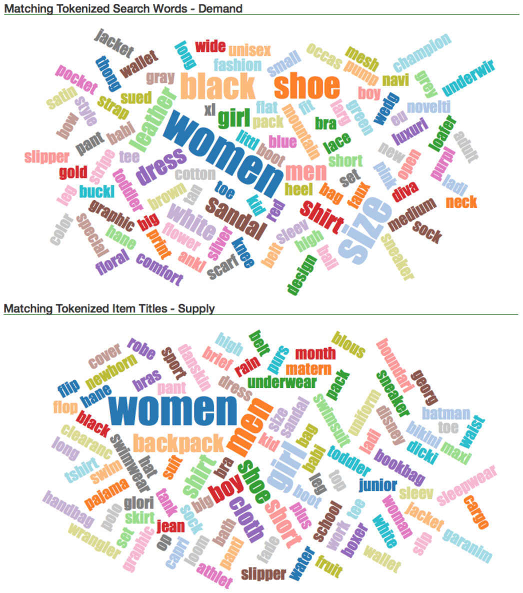

Figure 2: Word clouds on demand and supply for one topic

To identify areas of opportunity, we remove these common words and regenerate the word cloud visualization as shown in Figure 3. The resulting figure is much more informative. We observe now demand for which there is not comparable supply and vice versa. For this topic, we can clearly see that there is excess demand for “jansport" bags for which our supply appears deficient. Our supply side word cloud now consists of tokens of item titles that receive lacking consumer interest such as “textil" (eg. textile) and “fuchsia" for which there is little demand.By generating similar word cloud visualizations of the supply-demand gap for various topics, we can systematically identify areas for improvement in an e-commerce engine.

4.4 Validation of the Supply Demand Gap Metric

In this study, we compute our distance function $dist(p_s, p_d)$ to quantify the supply-demand gap. In order to understand its relationship with an engagement metric, we artificially remove some products or queries from the dataset and generate the topic models and compute the distance. We see, in figure 4, that the clicks are correlated to the gaps between supply and demand. The larger the gap, we see a larger drop in clicks. This informs us that we can use such divergence measures as a metric to understand the degree of supply and demand matching from the text data in an e-commerce site. This will allow us to understand the economics of supply and demand without the historical data. The method also can be fully automated. Hence, this can be a very powerful metric in e-commerce to quickly conduct a first level automated health check of the e-commerce site considering the supply and demand gap.

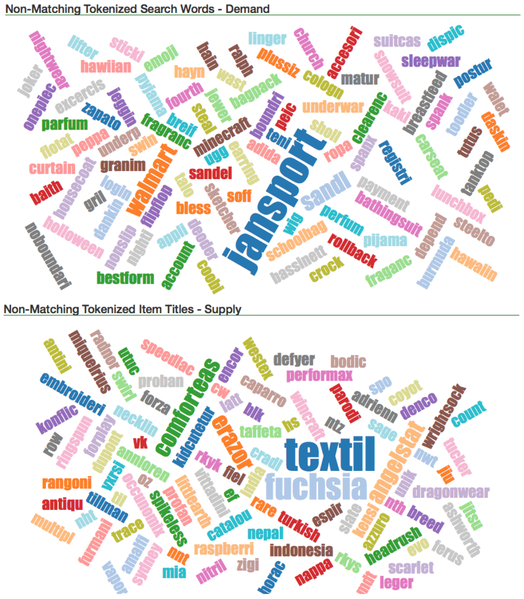
Non-Matching Tokenized Search Words - Demand
jansport
walmart
Non-Matching Tokenized Item Titles - Supply
textil
fuchsia

Figure 3: Word clouds of relative complements

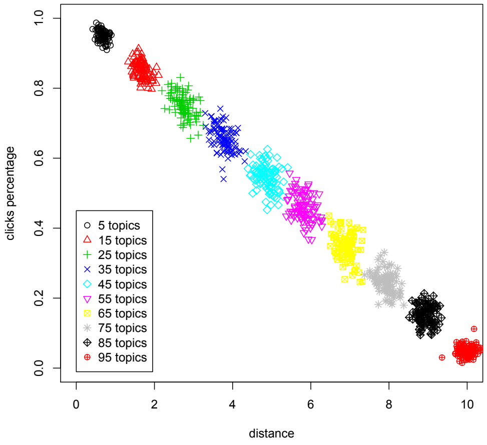

Figure 4: Relationship between supply-demand distance and clicks, we randomly remove 5 to 95 topics from the inventory and draw the correlation plot with clicks. Here, it shows clicks increase monotonically with less statistical distance between the query item topic distributions.

5 CONCLUSION

In this paper, we develop a topic models based approach to obtain insight about demand generation and assortment planning of an e-commerce site. We also define a measure of supply and demand gap on an e-commerce site and show that this measure is monotonically decreasing with the click-based engagement metric. Hence, if this measure indicates a large gap, then it is not healthy for an e-commerce site. The technique is based on topic models, is unsupervised, scalable, and can be very powerful in determining the state of supply and demand in a large e-commerce site. The visualizations also can help in obtaining visual insights on the demand and supply side. The method can be used as a powerful unsupervised technique to understand the supply and demand gap in an e-commerce site. Future work can be done to utilize the existing tree structure of categories to analyze supply-demand gaps as they pertain to a retailer's organizational and product hierarchy. Such work can result in a more granular and operationally effective application of topic models to increase economic output for the e-commerce business. Additionally, we can explore other clustering strategies beyond clicks to handle short text documents to improve the topic model such as different metrics like add-to-cart or conversion data and using Bipartite Spectral Graph Partitioning [10]. Additionally, a similar methodology can be used by constructing topic models using any deep learning and word embedding models [11]. We believe that our paper can pave the way for exploring with similar advanced natural language processing based tools to identify demand generation and assortment choices from the vast amount of search log data for e-commerce sites.

ACKNOWLEDGMENTS

We acknowledge Walmart Inc. for providing us the permission to publish the paper. We also acknowledge Angela Jiang and Wei Han for their contribution to this work when they worked at Walmart Lab. An earlier version of this work coauthored with Angela Jiang and Wei Han was presented in 2015 INFORMS data mining and analytics workshop.

REFERENCES

[1] Kusum L Ailawadi, Jonathan P Beauchamp, Naveen Donthu, Dinesh K Gauri, and Venkatesh Shankar. 2009. Communication and promotion decisions in retailing: a review and directions for future research. *Journal of retailing* 85, 1 (2009), 42–55.

[2] Gary Armstrong, Philip Kotler, Michael Harker, and Ross Brennan. 2015. *Marketing: an introduction.* Pearson Education.

[3] Michéle Basseville. 2013. Divergence measures for statistical data processing?An annotated bibliography. *Signal Processing* 93, 4 (2013), 621–633.

[4] Indrajit Bhattacharya and Lise Getoor. 2006. A latent dirichlet model for unsupervised entity resolution. In *Proceedings of the 2006 SIAM International Conference on Data Mining.* SIAM, 47–58.

[5] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

[6] David M Blei, Michael I Jordan, et al. 2006. Variational inference for Dirichlet process mixtures. *Bayesian analysis* 1, 1 (2006), 121–143.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[8] Gérard P Cachon, Christian Terwiesch, and Yi Xu. 2005. Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management* 7, 4 (2005), 330–346.

[9] Jonathan Chang. 2011. R package 'lda'. http://cran.r-project.org/web/packages/lda/

[10] Inderjit S. Dhillon. 2001. Co-clustering documents and words using Bipartite Spectral Graph Partitioning.

[11] Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57 (2016), 345–420.

[12] Thomas L. Griffiths, Mark Steyvers, Thomas L. Griffiths, and Mark Steyvers. 2004. Colloquium Finding scientific topics.

[13] Lisa Harris and Charles Dennis. 2011. Engaging customers on Facebook: Challenges for e-retailers. *Journal of Consumer Behaviour* 10, 6 (2011), 338–346.

[14] Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA'10).* ACM, New York, NY, USA, 80–88. https://doi.org/10.1145/1964858.1964870

[15] A Gürhan Kök, Marshall L Fisher, and Ramnath Vaidyanathan. 2008. Assortment planning: Review of literature and industry practice. In *Retail supply chain management.* Springer, 99–153.

[16] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 61–68.

[17] Marie Lienou, Henri Maitre, and Mihai Datcu. 2010. Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters* 7, 1 (2010), 28–32.

[18] Murali K Mantrala, Michael Levy, Barbara E Kahn, Edward J Fox, Peter Gaidarev, Bill Dankworth, and Denish Shah. 2009. Why is assortment planning so difficult for retailers? A framework and research agenda. *Journal of Retailing* 85, 1 (2009), 71–83.

[19] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and Sparse Text Topic Modeling via Self-Aggregation.. In *IJCAI*. 2270–2276.

[20] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 133–142.

[21] Sha Yang and Anindya Ghose. 2010. Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science* 29, 4 (2010), 602–623.

[22] Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 403–410.