SurfShop: combing a product ontology with topic model results for online window-shopping.

Zofia Stankiewicz and **Satoshi Sekine**
Rakuten Institute of Technology, New York
215 Park Avenue South
New York, NY 10003, USA
{zofia.stankiewicz,satoshi.b.sekine}@mail.rakuten.com

Abstract

At present, online shopping is typically a search-oriented activity where a user gains access to products which best match their query. Instead, we propose a *surf-oriented* online shopping paradigm, which links associated products allowing users to "wander around" the online store and enjoy browsing a variety of items. As an initial step in creating this experience, we constructed a prototype of an online shopping interface which combines product ontology information with topic model results to allow users to explore items from the food and kitchen domain. As a novel task for topic model application, we also discuss possible approaches to the task of selecting the best product categories to illustrate the hidden topics discovered for our product domain.

1 Introduction

Query based search remains the primary method of access to large collections of data. However, new interfacing options offered by mobile and touchscreen applications lead to decreased reliance on typed search queries. This trend further fuels the need for technologies which allow users to browse and explore large amounts of data from a variety of viewpoints. Online store product databases are a representative example of such a data source. At present, online shopping is typically a search-oriented activity. Aside from suggestions of closely matching products from a recommender system, internet shoppers have little opportunity to *look around* an online store and explore a variety of related items. This observation led us to define a novel task of creating a *surf-oriented* online shopping interface which facilitates browsing and access to multiple types of products. We created the prototype *SurfShop* application in order to test whether we can combine knowledge from a product ontology with topic modeling for a better browsing experience.

Our aim is to design an application which offers access to a variety of products, while providing a coherent and interesting presentation. While the product ontology provides information on product types which are semantically close (for example *spaghetti* and *penne*), it does not provide information about associations such as *pasta* and *tomato sauce*, which may be mentioned implicitly in product descriptions. In order to obtain semantically varied product groupings from the data we integrated topic model results into the application to display products which are related through hidden topics.

The data used for this project consists of a snapshot from the product database of a Japanese Internet shopping mall Rakuten Ichiba obtained in April 2011[1]. We limited our prototype application to the food and kitchen domain consisting of approximately 4 million products. The textual information available for each product includes a title and a short description. Furthermore, each product is assigned to a leaf category in the product hierarchy tree.

We use standard LDA (Blei et al., 2003) as the topic model and our prototype can be treated as an example of applied topic modeling. Although there exist browsers of document collections based

[1]For a version of Rakuten product data made available for research purposes see http://rit.rakuten.co.jp/rdr/index_en.html.

Proceedings of the NAACL-HLT 2012: Demonstration Session, pages 13–16,
Montréal, Canada, June 3-8, 2012.

on topic modeling [2], they have been constructed as direct model result visualizations. In contrast, we incorporate the LDA results into the output by combining them with product category information and search to produce a full blown application with a topic model serving as one of its components. We provide a more detailed overview of the entire system in section 2.

In LDA literature, the topics discovered by the model are typically represented by top n most probable words for a given topic. In integrating topic model results into our application we faced a challenge of creating theme pages which correspond to hidden topics, and selecting product categories which best illustrate a given topic. In section 3 we discuss a preliminary evaluation of the application's theme pages which suggests that combining topic knowledge with ontology structure can lead to more coherent product category groupings, and that topic interpretation and labeling based solely on top n words may not be sufficient for some applied tasks. We conclude by summarizing plans for further development of our prototype.

2 System overview

The initial input to the *SurfShop* system consists of a product database and a product ontology with node labels. All products were indexed for fast retrieval by the application[3]. A chart of application components is presented in Figure1.

Raw product descriptions from our data would constitute a large corpus including meta-data such as shipping or manufacturer information, which are not relevant to our task. Thus, fitting a topic model over this corpus is not guaranteed to provide useful information about related product types. Therefore, we decided to aggregate the product information into a collection of product category documents, where each document corresponds to a node in the product ontology tree (1088 nodes total). Each document consists of sentences extracted from product descriptions which potentially describe its relationship to other product categories (based on the occurrence of category name labels). We can then use

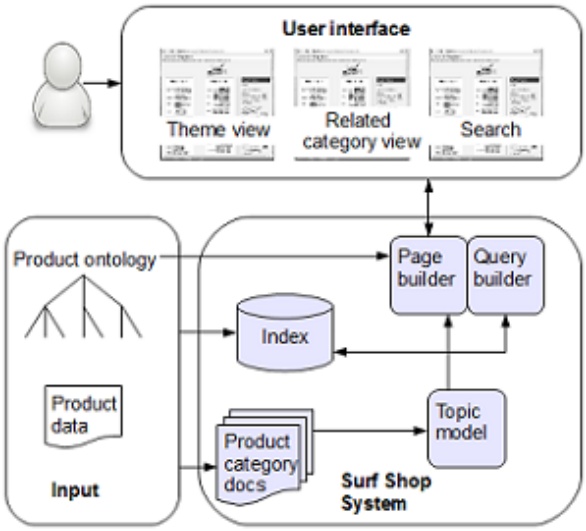

Figure 1: System overview

this artificially constructed corpus as input to LDA to discover hidden topics in the collection[4].

The topic model results, as well as product ontology information are combined with product search in order to build pages for our *SurfShop* application. In the prototype the user can move between search, browsing related categories, as well as browsing thematic product groupings. In the search mode, we use the query and the top n search results to infer which product category is most relevant to the query. This allows us to display links to related category groups next to the search results.

Given a product category, the users can also explore a related category map, such as the one shown in Figure 2 for *cheese*. They can browse example products in each related category by clicking on the category to load product information into the right column on the page. To provide example products, a query is issued under the relevant ontology node using the product category label and topic keywords, to ensure that we display items relevant to the current page. The product browsing functionality is similar for theme pages which are discussed in the next section.

[2]For an example see http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-list.html.

[3]We used an Apache Solr index and a JavaScript Ajax-Solr library from https://github.com/evolvingweb/ajax-solr.

[4]For LDA we used the Lingpipe package (http://alias-i.com/lingpipe/).

Figure 2: Related category page example. Category and theme labels have been translated into English.

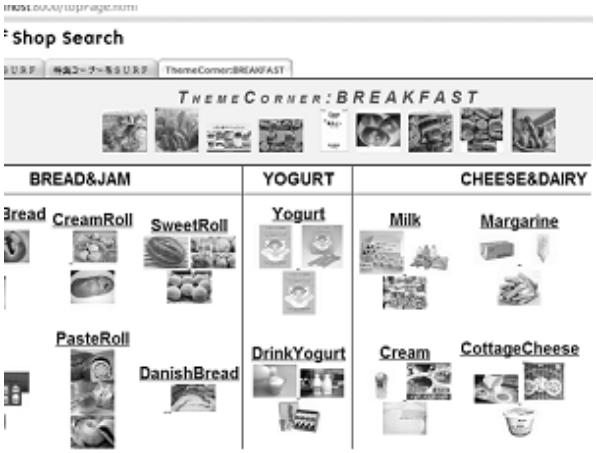

Figure 3: Theme page fragment. Category and theme labels have been translated into English.

3 Theme pages

An example of a *breakfast* theme page view is shown in Figure 3. It includes clusters of product categories which exemplify the page theme, such as *bread and jam* or *cheese and dairy*. Each theme page corresponds to a hidden topic discovered by the LDA model[5]. Human interpretation of topic models has been a focus of some recent work (Chang et al., 2009; Newman et al., 2010; Mimno et al., 2010). However, previous approaches concentrate on representing a topic by its top n most probable words. In contrast, our goal is to illustrate a topic by choosing the most representative documents from the collection, which also correspond to product categories associated with the topic. Since this is a novel task, we decided to concentrate on the issue of building and evaluating theme pages before conducting broader user studies of the prototype.

There are a few possible ways to select documents which best represent a topic. The simplest would be to consider the rank of this topic in the document. Alternatively, since the model provides an estimate of topic probability given a document, the probability that a product category document belongs to a topic could be calculated straightforwardly using the Bayes rule[6]. Yet another option for finding categories related to a given topic would be to assign a score based on KL divergence between the topic word multinomial and a product category multinomial, with the probability of each word w in the vocabulary defined as follows for a given category:

$$P(w) = \sum_t (P(w|t_i) * P(t_i|c_j)) \qquad (1)$$

Finally, we hypothesized that product ontology structure may be helpful in creating the theme pages, since if one product category is representative of the topic, its sibling categories are also likely to be. Conversely, if a category is the only candidate for a given topic among its neighbors in the tree, it is less likely to be relevant. Therefore, we clustered the topic category candidates based on their distance in the ontology, and retained only the clusters with the highest average scores.

To evaluate which of the above methods is more effective, we gave the following task to a group of three Japanese annotators. For each topic we created a list of category candidates which included product categories where the topic ranked 1-3 (methods 1-3 in Table 1), top 25 Bayes score and KL divergence score categories (methods 4 and 5), as well as the categories based on ontology distance clusters combined with the Bayes score averages for cluster reliability (method 6). Each annotator was given a list of top ten keywords for each of the topics and asked to choose a suitable label based on the keywords. Subsequently, they were asked to select product cat-

[5]We empirically set the number of topics to 100. We removed top 10% most general topics, as defined by the number of documents which include the topic in its top 10.

[6]We made an additional simplifying assumption that all documents are equiprobable.

Scoring method	Precision	Recall	F-score
1.Rank1	73.83%	43.21%	54.16%
2.Rank1+2	50.91%	59.56%	54.54%
3.Rank1+2+3	41.71%	73.08%	52.77%
4.Top25 KL	53.54%	70.44%	60.45%
5.Top25 Bayes	53.56%	71.25%	60.76%
6.Bayes+Ont	66.71%	69.17%	67.48%

Table 1: Result average for three annotators on Task 1.

egories from the candidate list which fit the topic label they decided on.

In this manner, each annotator created their own "golden standard" of best categories which allowed us to compare the performance of different approaches to category selection. The amount of accepted categories varied, however a performance comparison of candidate sets showed consistent trends across annotators, which allows us to present averages over annotator scores in Table 1. Rank based selection increases in recall as lower ranks are included but the precision of the results decreases. KL divergence and Bayes rule based scores are comparable. Finally, combining the ontology information with Bayes scoring improves the precision, while retaining the recall similar to that of the top 25 Bayes score approach. We chose this last method to create theme pages.

We also wanted to verify how the presence of top topic words affects topic interpretation. In another task, shown in Table 2, the same group of annotators was presented only with product category lists which combined method 5 and method 6 candidates from the previous task. They were asked to assign a topic label which summarized the majority of those categories, as well as mark the categories which did not fit the topic. Even though the annotators had previously seen the same data, they tended to assign broader labels than those based on the top topic words, and included more categories as suitable for a given topic. For example, for the *breakfast* theme shown in Figure 3, one annotator labeled the topic *dairy products* based on topic words, and *bread and dairy products* based on the product category examples. The results of Task 2 led us to use manually assigned theme page labels based on the product category groupings rather than the topic keywords.

Scoring method	Precision	Recall	F-score
5.Top25 Bayes	71.28%	81.83%	76.03%
6.Bayes+Ont	84.11%	75.46%	79.38%

Table 2: Result average for three annotators on Task 2.

The differences in results between Task 1 and Task 2 indicate that, while top topic keywords aid interpretation, they may suggest a narrower theme than the documents selected to represent the topic and thus may not be optimal for some applications. This underscores the need for further research on human evaluation methods for topic models.

4 Future work

We demonstrated a prototype *SurfShop* system which employs product ontology structure and LDA model results to link associated product types and provide an entertaining browsing experience.

In the future we plan to replace the LDA component with a model which can directly account for the links found through the product ontology tree, such a version of the relational topic model (Chang and Blei, 2009). In addition, we hope that further exploration of theme page construction can contribute to the development of topic visualization and evaluation methods.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn.* Res. 3, 993-1022.

Jonathan Chang and David Blei. 2009. Relational topic models for document networks. *Proc. of Conf. on AI and Statistics*, 81-88.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. *NIPS*, 1-9.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum 2011. Optimizing semantic coherence in topic models. *EMNLP, 2011.*

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100-108.