ABSTRACT
Most of current recommendation systems use numerical ratings to suggest a content (e.g., movie, restaurant) to a user. Instead, we apply probabilistic topic models to text reviews. We profile contents in a latent space where we compute distances that can be used for cold start recommendation.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: information filtering, retrieval models

Keywords
recommendation systems, content profiling, topic models

1. INTRODUCTION
Recommendation systems exploiting users’ ratings, they mostly recommend content one would like, ignoring content similarities. Ideally, a recommender should build a user profile and make recommendations that match this profile, independently of whether the user will like it or not.

We propose to address this issue with a probabilistic topic model applied to full text reviews instead of scores. Our approach is to (1) filter meaningful words in a database of text reviews, (2) extract topics using LDA and (3) measure similarity between contents given their topic distribution and a distance metric. We show encouraging results and identify future directions for this work.

2. RELATED WORK
Most commercial recommender systems exploit user ratings. According to [1], rating patterns can help diversify recommendations. Our objective is to show that text reviews can provide additional diversity.

In [3], content and user profiling is performed adding social tags to usual ratings. We adapt these methods to raw text reviews input as they include richer assessment of the user’s opinion about the content. We use Latent Dirichlet allocation (LDA) [2], a probabilistic model that infers hidden topics given a text corpus where each document of the corpus can be then represented as topic probabilities. LDA has been widely used in recent years, either for collaborative filtering purpose [5] or tag recommendation [4].

3. METHODOLOGY
In the following, a corpus is a list of documents. Each document is the concatenation of the reviews of one content. In this work, content can be a movie or a restaurant.

3.1 Pre-processing the reviews
Many words in reviews are not relevant to recommend and profile contents as they convey neither qualitative nor descriptive information (e.g., stop words), or they appear too frequently in reviews (e.g., movie, film, scene in movie reviews). These words create noise in LDA topics, although extensions of the LDA model may deal with these words. We made the decision to eliminate these words by hand, using a dictionary for movies and restaurants that we extracted from our two databases.

Remaining words are filtered by frequency using the Term Frequency - Inverse Document Frequency score (TF-IDF). TF-IDF measures the importance of a word in a corpus. It increases with the number of occurrences in the document and decreases with the frequency in the corpus. We compute TF-IDF for each word of each review in the corpus and select the 10,000 words with the highest score.

3.2 Topic Extraction
LDA extracts K latent topics from a corpus of text documents. K is an input to LDA. Each topic corresponds to a distribution on the 10,000 extracted words. For each document LDA infers a distribution on the K topics. In the corpus, as each document corresponds to a single content, we use the distribution on the K topics as a profile of a given content. In practice, the inference is done using a variational EM, described in [2].

3.3 Similarity
Content similarity is an important factor in recommendations. We cluster movies and restaurants based on their topic distribution and a distance metric. We show encouraging results and identify future directions for this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
contents \( d \neq c \) and sort these distances from the closest to the furthest content to \( c \).

4. DATASETS

We apply our similarity clustering methodology to two reviews datasets. The restaurant dataset contains 68,000 words spread over 58,432 reviews of 1,992 restaurants in a big US city. The movie dataset has 122,000 words, 53,166 reviews and 1,886 movies and is also collected in the US. We are not authorized by the data owners to share more information about the dataset but we are happy to help other researcher access our data.

5. RESULTS

For each movie \( c \) given its profile \( \theta^c \), the number of representative topics of \( c \) (i.e. the number of high components in \( \theta^c \)) is low - generally less than 5. The distance between movies increases as the number of shared topics decreases.

We can observe that topics do not exhibit consistency with qualitative argument developed by reviewers (probably because they all use different words and LDA fails, as with users). However, topics are quite consistent around a genre, an actor, a director or a sequel. As a consequence, the consistency of our similarity clusters is quite high. The least reviewed genres - e.g. animation, western - spread on one or two topics. The most reviewed genres - e.g. action, horror - spread on almost ten topics. For action movies, our method extracts topics associated to action sub-genres such as crime or war and topics associated to famous action actors or directors such as Bruce Willis or Quentin Tarantino.

Due to space constraints, we only show two representative clusters (Table 1) from our movie dataset. However, all topics and clusters can be found on our server for \( K = 64 \). 1

Animation movies profiles have a very high component for one common topic and low components for all the other topics. As a result, their \( KL \) distance is very low. The order of the movies in the Shrek similarity cluster highly depends on the frequency of words in the reviews. The word shrek occurs many times in Hoodwinked! reviews, as users tend to compare this movie with Shrek. This explains why Hoodwinked!, Mulan and Pinocchio are closer to Shrek than Shrek the Third.

The movie Inglourious Basterds (in Table 1) has a high component in a topic represented by words nazi, war, german and another high component in a topic represented by words tarantino, kill, bill, pulp, fiction. As a result, movies found in the similarity cluster are a mix of World War II movies (e.g. The Pianist, The Sum of All Fears) and Tarantino movies (e.g. Kill Bill, Pulp Fiction, Django Unchained). Ocean Twelve and Southern Comfort are explained by the action nature of these three movies. As Bruce Willis acts in several Tarantino’s action movies, Inglourious Basterds profile has a Bruce Willis component, shared with Unbreakable. Knife in the Water and Will Penny are outliers. Knife in the Water’s profile shares several low importance components with Inglourious Basterds. Will Penny is associated to the Bruce Willis topic as the word bruce occurs several times in its reviews (Bruce Dern is one of the actors). We conjecture that as this movie has not been reviewed a lot (~ 50 reviews), the frequent occurrence of the word bruce has a strong impact on the profile.

<table>
<thead>
<tr>
<th>Movie</th>
<th>KL</th>
<th>Inglourious Basterds KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shrek</td>
<td>0.06</td>
<td>Django Unchained 1.39</td>
</tr>
<tr>
<td>Shrek Forever After</td>
<td>0.30</td>
<td>The Pianist 2.84</td>
</tr>
<tr>
<td>Hoodwinked!</td>
<td>0.38</td>
<td>Kill Bill: Vol. 1 2.93</td>
</tr>
<tr>
<td>Mulan</td>
<td>0.43</td>
<td>Ocean’s Twelve 3.26</td>
</tr>
<tr>
<td>Pinocchio</td>
<td>0.46</td>
<td>The Sum of All Fears 3.60</td>
</tr>
<tr>
<td>Shrek the Third</td>
<td>0.56</td>
<td>Knife in the Water 3.82</td>
</tr>
<tr>
<td>The Lion King</td>
<td>0.76</td>
<td>Pulp Fiction 3.83</td>
</tr>
<tr>
<td>Tangled</td>
<td>0.86</td>
<td>Southern Comfort 3.89</td>
</tr>
<tr>
<td>Beauty and the Beast</td>
<td>0.95</td>
<td>Will Penny 3.96</td>
</tr>
<tr>
<td>Song of the South</td>
<td>0.98</td>
<td>Unbreakable 3.97</td>
</tr>
</tbody>
</table>

Table 1: Similar contents, \( K = 64 \) topics

The impact of \( K \) on our similarity clustering is quite interesting. If \( K \) is low, LDA topics mix close genres. If \( K \) is high, LDA tends to affect one topic to each movie. In other words, extreme values of \( K \) reduce profiling precision as two close movies in \( KL \) distance do not necessarily share common characteristics.

We also tried to apply our methodology to users. A document is now defined as all reviews written by a given user. Interestingly, LDA can not extract discriminative information because users that have reviewed movies from different genres use a diverse vocabulary where it is difficult to find meaningful topics. Unlike contents, similarity between users can not be manually checked.

6. FUTURE WORK

Given that qualitative aspects in the reviews do not seem to impact the topic construction and later the similarity clustering, we will integrate ratings to our model next. We also want to investigate user profiling in more depth. Very popular actors (e.g. B. Willis, S. Stallone) influence resulting clusters as they tend to occur frequently in the reviews. The method then extracts consistent topics around them. As a result, if a given movie \( c \) casts a popular actor, the closest movies might also cast the same actor. Introducing users’ opinion would help diversify resulting clusters.

7. REFERENCES


1 topics are available on www.di.ens.fr/~cdupuy/.