Getting Recommender Systems to Think Outside the Box

Zeinab Abbassi[‡], Sihem Amer-Yahia[†], Laks Lakshmanan[‡],
Sergei Vassilvitskii[†], Cong Yu[†]

[†]Yahoo! Research, {sihem,sergei,congyu}@yahoo-inc.com,
[‡]University of British Columbia, {z.abbassi@gmail.com, laks@cs.ubc.ca}

ABSTRACT

We examine the case of *over-specialization* in recommender systems, which results from returning items that are too similar to those previously rated by the user. We propose Outside-The-Box ($\mathcal{OTB}$) recommendation, which *takes some risk* to help users make *fresh discoveries*, while *maintaining high relevance*. The proposed formalization relies on *item regions* and attempts to identify regions that are under-exposed to the user. We develop a recommendation algorithm which achieves a compromise between relevance and risk to find $\mathcal{OTB}$ items. We evaluate this approach on the MovieLens data set and compare our $\mathcal{OTB}$ recommendations against conventional recommendation strategies.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

otb, outside the box, diversity, serendipity, recommendation

1. INTRODUCTION

Recommender systems have recently witnessed impressive gains in their research methodology and practical success (c.f., Amazon and Netflix). Lately, recommendations are appearing in social content sites such as del.icio.us and Yahoo! Travel [3]. A typical problem with recommenders is *over-specialization*: users frequently see items that are very similar to what they liked in the past. While this approach produces relevant items, anecdotal evidence suggests that they may not be the most *useful* recommendations, due to their lack of novelty. The push for relevancy leads recommender systems to produce these *safe* items, and reduces the chances that a user will be exposed to items that she may actually like, had she known about them (serendipitous recommendation) [9, 10]. In this paper, we formalize *outside-the-box, $\mathcal{OTB}$, recommendations* which aim at helping users find *surprisingly good* items.

Seeking $\mathcal{OTB}$ recommendations and finding diverse recommendations as in [10, 11, 12], are two different and complementary tasks. A user who is recommended diverse items, e.g., a comedy, a drama and an adventure movie, may still consider them too similar to previously rated ones if the user rated many such movies in the past. That same user, if recommended a Sci-Fi and two thrillers, will find them more *interesting* despite being less diverse. Moreover, a user coming to an $\mathcal{OTB}$ *system* expects a different experience from a conventional recommender, which aims to maximize relevance only. Hence, *$\mathcal{OTB}$-ness* should be viewed as complementary to diversification and to conventional recommendations.

We introduce a notion of *regions* over items, and define it based on item attributes (e.g., the genre of a movie) or user behavior (e.g., movies liked by the same group of users). Over-specialization occurs when the recommended items to a user overwhelmingly fall into regions the user is familiar with. For example, recommending the movie "X-Men Origins: Wolverine" to a user who has watched and liked a lot of Sci-Fi action/thriller movies is, although highly relevant, less *useful*, since the user is likely to know about this movie already. Intuitively, users would experience serendipity as they are recommended items from less familiar regions.

To this end, we introduce *stickiness* which measures a user's familiarity with items in a region. Intuitively, a user is familiar with a region if her stickiness for the region exceeds the global stickiness for it. Observed unfamiliarity with a region can arise either when a user truly does not like items in the region and chooses not to rate them or when the user has not been *exposed* enough to the region. It is in the latter case that a region is likely to contain surprising yet relevant items. Therefore, in *Collaborative Filtering (CF)*, where recommended items are those popular among users who are similar to the given user (the user's network)[1], we detect *under-exposure* by comparing a user's stickiness to a region with her network's. In Section 2.1, we formalize this intuition into the notion of region *$\mathcal{OTB}$-ness*.

Specifically, $\mathcal{OTB}$ is a recommendation strategy that takes into consideration which regions items come from, and favors items that come from regions with high $\mathcal{OTB}$-ness. The process of recommending $\mathcal{OTB}$ items involves *combining region $\mathcal{OTB}$-ness with item relevance*. Computing item relevance is a substantial challenge in itself since neither the user nor her network know much about those regions. Hence, we propose the notion of *expanded region network* which relies on *region-region correlations*. For example, for a user John who has high $\mathcal{OTB}$-ness for French restaurants (i.e., neither he nor his network knows much about French restaurants), his network becomes useless in estimating the expected ratings

RecSys'09, October 23–25, 2009, New York, New York, USA.

of French restaurants. However, if it is globally known that French restaurants are correlated with Mediterranean ones, the expanded network built from users who like Mediterranean cuisine can be helpful in computing the relevance of French restaurants for John. This is further detailed in Sections 2.2 and Section 2.3, which describes our algorithm.

Finally, in Section 3, we perform comprehensive experiments over the 10M ratings MovieLens data set. We show that our strategy not only finds items which are different from conventional recommenders but also preserves item quality as measured by nDCG.

2. GOING OUTSIDE THE BOX

We start by defining a *region* (i.e., the "box") as a group of similar items. $R^{\mathcal{I}}$ denotes a (potentially overlapping) assignment of items $\mathcal{I}$ into regions. Regions in $R^{\mathcal{I}}$ are produced using similarity distances between items. We explore two such similarities: *attribute-based* and *activity-based*.

DEFINITION 1. **[Attribute-Based Similarity]** *Let A be a set of item attributes, called* region dimensions. *Then we can define an attribute based similarity function d_A. For any two items i and j the distance $d_A(i,j) = 0$ iff $\forall a \in A$ we have $i.a. = j.a$, and $d_A(i,j) = \infty$ otherwise.* ■

For movies, region dimensions may include movie attributes like genre and directors. A region instance is often identified by its dimensions and their values (e.g., {(genre=comedy), (producer=Disney)}). For attributes with a large number of distinct values (e.g., actors), a taxonomy can be introduced to reduce the total number of regions. For example, Jim Carrey and Adam Sandler are similar under "comedians." Our framework also supports sophisticated distance functions, e.g., a weighted combination of distances on item attributes.

DEFINITION 2. **[Activity-Based Similarity]** *For any two items i and j, and action a let $a(i)$ and $a(j)$ define the respective sets of users that performed action a on the item. Then let $d(i,j)$ be the Jaccard dissimilarity between $a(i)$ and $a(j)$:*

$$d(i,j) = 1 - \frac{|a(i) \cap a(j)|}{|a(i) \cup a(j)|}$$

■

Here, regions identify items rated by a large enough number of similar users.

To produce an assignment of items into regions given a similarity function, we turn to the k-means^{++} algorithm [4]. The algorithm is known to converge quickly, and for clustering n items requires only $O(nkr)$ distance computations, where k is the number of clusters and r is the number of rounds performed by k-means.

2.1 Region $\mathcal{OTB}$-ness

We let $\texttt{items}(u)$ denote the set of items that are rated by u, $\texttt{items}(r)$ the set of items belong to region r, $\texttt{items}(u,r)$ the set of items belong to region r that are rated by u, i.e., $\texttt{items}(u,r) = \texttt{items}(u) \cap \texttt{items}(r)$ and, $\texttt{rating}(u,i)$ the known rating of user u for item i;

DEFINITION 3. **[User Region Stickiness]**
The stickiness of a user u to a region r, $\texttt{stick}(u,r)$ is the fraction between the number of items rated by u which belong to r over the total number of items rated by u. That is, $\texttt{stick}(u,r) = \frac{|\texttt{items}(u,r)|}{|\texttt{items}(u)|}$. ■

A user who rated 500 movies, 50 of which are Drama, would have a stickiness of 10% for the region {(genre=Drama)}. Intuitively, stickiness measures the degree of familiarity of a user toward a given region: the higher the stickiness, the more likely the user already knows about items within the region. Note that if the given region is the entire set of items ($\mathcal{I}$), then $\texttt{stick}(u,\mathcal{I}) = 1$ for any user u. Similarly, we can measure the stickiness of a group of users to a region:

DEFINITION 4. **[Network Region Stickiness]**
The stickiness of a group of users (i.e., a network) N to a region r, $\texttt{stick}(N,r)$ is the average of each individual member's stickiness. Hence, $\texttt{stick}(N,r) = \frac{1}{|N|}\Sigma_{u \in N}(\texttt{stick}(u,r))$. Furthermore, we have the deviation of stickiness:
$\texttt{stickDev}(N,r) = \sqrt{\frac{1}{|N|}\Sigma_{u \in N}(\texttt{stick}(u,r) - \texttt{stick}(N,r))^2}$.
■

The network stickiness measures the familiarity toward the given region by a group of users collectively. The deviation of stickiness measures how consistent each member's stickiness is with the others. The lower the deviation, the more likely every member in the group is familiar (or unfamiliar) with items in the given region. When N is the entire group of users ($\mathcal{U}$), we have the global stickiness, $\texttt{stick}(\mathcal{U},r)$, and deviation, $\texttt{stickDev}(\mathcal{U},r)$, for the region.

There are two main factors in measuring a region's $\mathcal{OTB}$-ness for a given user: the level of unfamiliarity and the (under-)exposure potential. We combine those two factor in the definition of $\mathcal{OTB}$-ness:

DEFINITION 5. **[User Region $\mathcal{OTB}$-ness]**
The $\mathcal{OTB}$-ness for a region r by a given user u is defined as $\texttt{otb}(u,r) = \texttt{otbBase}(u,r) \times \texttt{otbFactor}(u,r)$. *where the* level of unfamiliarity *for r by u is defined as:*
$\texttt{otbBase}(u,r) = \frac{\texttt{stick}(\mathcal{U},r) - \texttt{stick}(u,r)}{\texttt{stickDev}(\mathcal{U},r)}$,
if $\texttt{stick}(\mathcal{U},r) > \texttt{stick}(u,r)$, 0 *otherwise. And the* exposure factor *for r by u is defined as:*
$\texttt{otbFactor}(u,r) = \frac{\texttt{stick}(\mathcal{U},r) - \texttt{stick}(N,r)}{\texttt{stickDev}(\mathcal{U},r) + \texttt{stickDev}(N,r)} \times 2$,
if $\texttt{stick}(\mathcal{U},r) > \texttt{stick}(N,r)$, 0 *otherwise, N is u's network.*
■

Here, normalization by the global deviation in `otbBase` is done to identify the regions whose unfamiliarity are the most statistically significant. And, a region has a high `otbFactor` if the user's network is unfamiliar with items in the region.

2.2 Region-Based Relevance

Identifying good items within $\mathcal{OTB}$ regions now becomes a challenge since neither the user nor the user's network knows much about items in those regions with a high $\mathcal{OTB}$-ness for that user (according to the definitions in Section 2.1). As a result, computing the user's expected rating, i.e., relevance, for items within those regions requires special attention. To address this question, we propose the notion of *region-region correlation* to identify the set of regions that *implies* each $\mathcal{OTB}$ region. We then construct an *expanded region network*, which consists of users who are similar to the target user based on items in those correlated regions.

We use association rules [2] to identify region-region correlations of the form $r \Rightarrow r'$ where r and r' are different regions in $R^{\mathcal{I}}$.

DEFINITION 6. **[Source Region]** A region s is a source region of a region r if and only if at least x% of the users who rate items in s also rate items in r, where x is a custom defined threshold. ■

Source regions indicate general trends such as *people who rate Woody Allen movies also rate David Lynch movies*. We use $\texttt{sources}(r)$, to denote the set of all source regions of a region r. We show an example of source regions found in our data set in Section 3. Based on this, $\texttt{exSim}(u, u', r)$, the expanded similarity of two users given a region can be defined as:

$\texttt{exSim}(u, u', r) = max_{r' \in \texttt{sources}(r)} \texttt{userSim}(u, u', r')$

where $\texttt{userSim}(u, u', r)$ is a similarity between two users restricted to region r, formally defined as:

$$\texttt{userSim}(u, u', r) = \frac{|\{i | i \in \texttt{items}(u,r) \wedge i \in \texttt{items}(u',r) \wedge |\texttt{rating}(u,i) - \texttt{rating}(u',i)| \leq 2\}|}{|\{i | i \in \texttt{items}(u,r) \vee i \in \texttt{items}(u',r)\}|}$$

where two ratings with a difference less than 2 (on a scale of 0 to 5) is considered to be similar. This is number is user customizable and chosen based on our experience.

DEFINITION 7. **[Expanded Region Network]** The expanded region network, $\texttt{exNet}(u, r)$, for a user u and a region r is the set of users $u' \in \mathcal{U}$ such that $\texttt{exSim}(u, u', r) \geq \theta$ where θ is an application-dependent threshold. ■

Intuitively, $\texttt{exNet}(u, r)$ is the set formed by users who share similar interests with u over source regions of r. We can now have:

$\texttt{relevance}(u, r, i) = \Sigma_{u' \in \texttt{exNet}(u,r)} \texttt{exSim}(u, u', r) \times \texttt{rating}(u', i)$

2.3 Consolidation

Finally, we define the *overall score* of an item i as follows:

$\texttt{overall}(u, i) := \Sigma_{r \in \texttt{regions}(i)} \texttt{otb}(u, r) \times \texttt{relevance}(u, r, i)$

where $\texttt{regions}(i)$ is the set of all regions an item belongs to, $\texttt{otb}(u, r)$ denotes the $\mathcal{OTB}$ score of region r for user u and $\texttt{relevance}(u, r, i)$ is the region-specific relevance score of item i for user u.

Algorithm 1 summarizes our recommendation strategy. In order to efficiently generate the top-k items, we need to determine those regions r with $\texttt{otb}(u, r) > 0$. Furthermore, for each such region r, we need to create a list of items from r sorted in decreasing order of relevance (line 1). Given this information, we can generate the top-k items by a simple adaptation to a standard algorithm such as NRA or TA [6]. The two required changes are: the score of an item from a list $\mathcal{IL}_u^r$ should be weighted by the $\mathcal{OTB}$-ness of r and, an item's score needs to be aggregated across all its regions. The algorithm maintains a heap of current candidate items to recommend and stops when the expected overall score can not exceed the current k^{th} score in the heap (line 6).

Algorithm 1 $\mathcal{OTB}$ Recommendation Strategy

Require: A user u and all regions $r \in R^{\mathcal{I}}$ s.t. $\texttt{otb}(u, r) > 0$;
1: Retrieve relevance lists $\mathcal{IL}_u^r$ for each $r \in R^{\mathcal{I}}$
2: Cursor $cur = \texttt{getNext}()$ round-robin across each $\mathcal{IL}_u^r$;
3: **while** (cur <> NULL) **do**
4: Get item i at cur;
5: **if** not($\texttt{inHeap}(\texttt{topKHeap}, i)$) **then**
6: **if** ($\texttt{computeMaxScore}(i) \geq \texttt{topKHeap}.k_{th}\texttt{overall}$) **then**
7: Probe $\mathcal{IL}_u^r$ to compute overall score $\texttt{overall}(u, i)$;
8: $\texttt{topKHeap.addToHeap}(i, \texttt{overall}(u, i))$;
9: **end if**
10: **end if**
11: $cur = \texttt{getNext}()$;
12: **end while**
13: **return** topKList(topKHeap);

# users	# movies	# ratings
71,567	10,681	10,000,054

Table 1: MovieLens 10M Data Set Statistics.

region	source regions
Action	Adventure, Comedy, Crime, Drama, Mystery, Thriller, War
Adventure	Crime, Drama, Horror, Mystery, Sci-Fi
Animation	Children, Comedy, Sci-fi
Children	Comedy, Documentary, Drama, Fantasy, Romance
Comedy	Crime, Drama, Fantasy, Romance
Crime	Documentary, Drama, Thriller, War
Documentary	Drama
Drama	Action, Comedy,Romance, Sci-Fi, Thriller
Fantasy	Sci-Fi, Thriller
Noir	Drama
Horror	Adventure, Thriller
Imax	Animation
Musical	Romance
Mystery	Thriller
Romance	Drama, Thriller, War
Sci-Fi	Documentary, War
Thriller	War
War	Adventure, Western
Western	Drama, Action

Table 2: Region-Region Correlations.

3. EXPERIMENTAL ANALYSIS

Our experiments demonstrate two main points. First, $\mathcal{OTB}$ recommendations are of high quality, comparable to those from traditional CF strategies, as measured by standard metrics. Second, $\mathcal{OTB}$ recommendations do differ significantly from traditional ones, and this difference is *not* simply in the tail end of the recommendations—we find that items produced by $\mathcal{OTB}$ contribute significantly to the overall quality. This argues that $\mathcal{OTB}$ recommendations are complementary to conventional ones, and that both novelty and relevance are necessary to achieve user satisfaction.

3.1 Data Analysis

We adopt the MovieLens 10M ratings data [7] (statistics in Table 1.) For attribute-based regionization, we cluster movies based on genre which results in 19 regions. Consequently, a movie may belong to more than one region[1]. For activity-based regionization, we use k-means to produce the same number of (non-overlapping) regions, with the distance between two movies defined based on the two sets of users who rated them (Definition 2.) The size of each region varies from 50 to 2000 movies in the attribute-based regionization, and from 160 to 1050 movies in the activity-based one.

We further compute region-region correlations to generate expanded region networks (see Definition 7.) Table 2 reports source regions for each attribute-based region using a threshold of 60%.

3.2 Evaluation Methodology

We studied three strategies. **CF** uses conventional collaborative filtering, while **OTB-ATT** (resp. **OTB-ACT**) pro-

[1]The genres drama and comedy for a movie are ignored if it also belongs to other genres: this ensures that the drama and comedy regions stay within a reasonable size.

	OTB-ATT	OTB-ACT
top-10	2.7	3.1
top-50	25.1	22.9

Table 3: Average overlap with CF recommendations.

duces $\mathcal{OTB}$ recommendations using attribute-based (resp. activity-based) regions.

Movie ratings range from Jan 9, 1995 to Jan 4, 2009. We used ratings by 700 active users, i.e., users who rated movies for more than a year. For each user, we divided the data into a training set which covers the first 80% of each user's rating period and, a test set, the remaining 20%.

To evaluate the quality of the results we adopt the Normalized Discounted Cumulative Gain (nDCG) [8] measure. Each recommendation strategy generates a ranked list. For a given list of size n, we compute the discounted cumulative gain, DCG, as follows:

$$DCG = \sum_{i=1}^{n} \frac{2^{rel_i}}{\log_2(1+i)}$$

where rel_i is the rating of the movie at position i in the test set. To compute nDCG, we reorder the list such that the most relevant items appear first and compute the ideal DCG in the same way:

$$DCG_{ideal} = \sum_{j=1}^{n} \frac{2^{rel_j}}{\log_2(1+j)}$$

The value $\frac{DCG}{DCG_{ideal}}$ is then taken as the nDCG value, falls between 0 and 1, regardless of the test set size.

3.3 Result Analysis

Figure 1 summarizes the average nDCG (and the standard deviation) across all 700 users for **CF**, **OTB-ATT** and **OTB-ACT**. All three produce recommendations with statistically comparable nDCG values. This validates that taking more risk to produce $\mathcal{OTB}$ recommendations *does not hurt* overall quality.

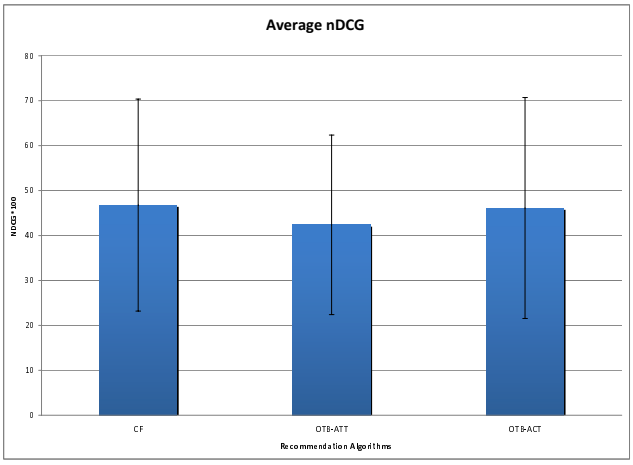

Figure 1: Average nDCG for all three recommendation strategies.

Table 3 shows the average overlap between **OTB-ATT**, **OTB-ACT** and **CF**. The overlap at top-10 is quite small (about 30%.) Even for top-50 lists, the overlap is still reasonably small at about 50%. This suggests that a significant portion of $\mathcal{OTB}$ recommendations are different from the ones found by **CF**, making them novel and complementary.

We further investigate nDCG values achieved by each strategy. As shown in Figure 2, recommendations generated only by our $\mathcal{OTB}$ strategies contribute to a significant portion (over 80%) of the overall nDCG, indicating most of those recommendations rank higher up in the returned list which makes them *different* and *novel*, hence, potentially *useful*.

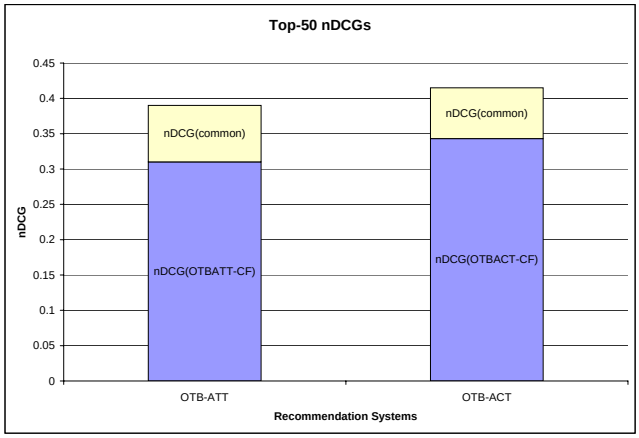

Figure 2: Contribution to nDCG by overlapping and non-overlapping recommendations.

4. RELATED WORK

Despite being a recognized problem, over-specialization is addressed in an ad-hoc manner. Some content-based recommender systems, such as DailyLearner [5], filter out highly relevant items which are too similar to items the user has rated in the past. Most previous work focuses on increasing diversity, such as [11] and [12]. Finally, there are also proposals that incorporate randomness in recommendations [1].

5. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6), 2005.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, 1994.

[3] S. Amer-Yahia, L. Lakshmanan, and C. Yu. Socialscope: Enabling information discovery on social content sites. In *CIDR*, 2009.

[4] D. Arthur and S. Vassilvitskii. K-Means++: the advantages of careful seeding. In *SODA*, 2007.

[5] D. Billsus and M. Pazzani. User modeling for adaptive news access. In *User Modeling and User-Adapted Interaction. 10:2/3. 147-180*, 2000.

[6] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *PODS*, 2001.

[7] GroupLens at University of Minnesota. http://www.grouplens.org/node/73.

[8] K. Jarvelin and K. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*, 20(4), 2002.

[9] J. A. Konstan. Introduction to recommender systems. In *SIGIR*, 2007.

[10] S. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI*, 2006.

[11] S. A. Munson, D. X. Zhou, and P. Resnick. Sidelines: An algorithm for increasing diversity in news and opinion aggregators. In *AAAI*, 2009.

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