Beyond Relevance in Marketplace Search

Nish Parikh Neel Sundaresan
eBay Research Labs
2065 Hamilton Ave,
San Jose, CA, 95125, USA.
{nparikh, nsundaresan}@ebay.com

ABSTRACT
In this paper we study diversity and its relations to search relevance in the context of an online marketplace. We conduct a large-scale log-based study using click-stream data from a leading eCommerce site. We introduce 3 main metrics – selection (diversity), trust, and value. In our analysis we also show how these interact with relevance in different ways. We study the benefits of diversity and also show why guaranteeing diversity is important.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*.

General Terms
Economics, Experimentation, Measurement.

Keywords
eCommerce, search, diversity, relevance, trust, value, selection

1. MOTIVATION AND BACKGROUND
Relevance in ranking for keyword search has been extensively studied. Since search queries are likely to have different interpretations for different users, showing diverse set of results is necessary. For instance, for a generic query like "mp3 players" different kinds of mp3 players should be returned. Ideally, the result set should properly account for the interest of the overall user population.

Diversity in information retrieval and web search has also been well studied [2]. [4] describes Maximal Marginal Relevance (MMR) criterion to reduce redundancy while maintaining query relevance in re-ranking retrieved documents. Each document in the ranked list is selected according to a combined criterion of query relevance and novelty of information.

[1] shows efficient query processing techniques that guarantee diversity. Although some attributes of a narrow section of inventory (automobiles) that impact diversity are discussed in, it does not provide enough insight into what dimensions impact user satisfaction and diversity in eCommerce search. Our work focuses on bridging this gap. As little work has been done around diversity in eCommerce search, we study and describe the role of relevance and diversity in eCommerce search by using large-scale retrospective datasets from a leading eCommerce site which has at any point in time millions of items for sale in different categories by different sellers.

In section 2 we describe various dimensions that affect search results. In section 3 we describe experiments to discover the interactions between these dimensions and also show why guaranteeing diversity in product search is essential. Finally, we note our conclusions and plan for future work in section 4.

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'11, October 24 - 28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10...$10.00.

2. ECOMMERCE SEARCH DIMENSIONS
For online marketplaces supporting C2C and B2C transactions sellers pay to list their inventory, however, the engine has to sort items based upon what the buyer is likely to buy and also provide a good buying experience. Various factors which influence buyer satisfaction, marketplace reputation and sell-through of items for sale come into play. Some of these factors which need to be considered in building the item result set are discussed below.

2.1 Trust
For online retailers like Wal-Mart and Target the trust or reputation associated with the products is that of the online retailer itself as they control and manage the entire inventory sold on their site. However, in a multi-buyer to multi-seller marketplace, as there are many sellers selling different items, the trust factor associated with every item is different. If the same product is being sold for the same price by two different sellers with different reputations, buyers would surely want to buy it from the seller with a better reputation. The trust associated with an item is based on the reputation of the seller who is selling the item. The reputation score for a seller could again be influenced by various factors like feedback from former buyers, return policy for items sold by the seller, shipping options offered by the seller and other factors. Total trust associated with a result set which we refer to as τ would be dependent on the collection of trust values of every individual item in the result set. For our experiments we use seller data and feedback from former buyers to compute the trust value of every item. τ determines the long term health of the marketplace.

2.2 Value
In today's online world, buyers have a lot of choice. A compelling value proposition is something that many buyers look for, before completing their purchase. Given two exactly same items with the same trust score, buyers are more likely to buy the

one with a lower price. Thus, value proposition provided by an item to the buyer is a function of the price of the item. If the price of an item is less than the typical selling prices of similar items then the value proposition for the item is high and vice versa. The total value aspect associated with a result set would be dependent on the average value provided by all the items that the result set comprises of and we refer to it as v .

2.3 Selection

Diversity of an item set is an indicator of the variety of items found in the result set. On eCommerce sites, the different factors determining diversity of result set could be types of sellers, auctions vs. fixed price items, shipping options, payment options and also factors like what product types or categories the surfaced inventory belongs to (For e.g., if a user issues a vague or ambiguous query like "peru" the items in the corresponding result set could belong to different categories like "Stamps", "Coins and Paper Money", "Postcards", "Art" and many others).

If relevance is not impacted, then providing a wider selection to buyers is good in general. We calculate diversity of a result set based on 3 different factors; the diversity of sellers whose items are present in the set, the diversity of the items based on format of the item and the diversity measured in terms of the amount of people whose interests would be satisfied by the result set.

We calculate the seller diversity using this Simpson's index metric.

$S_d = 1 - \frac{\sum_n n(n-1)}{N(N-1)}$, where S_d is the seller diversity, N is the total number of items in the result set, and n is the total number of items in the result set from one particular seller.

Format diversity F_d is calculated in a similar way. For our work, we have considered two formats: fixed price items and auction items.

We use data from a leading eCommerce site which has individual items for sale, and where many of the long-tail items are not necessarily grouped by product ids, or UPC/EAN/ISBN numbers.

As most users don't look beyond the first couple of pages of the result, the goal is to be able to show all kinds of desirable items for sale on the first page of results. Although two individuals might issue the same query, their intents could be different. We look at past history of user activity to associate desirable items with a given query. As described in [3] we mine the user click-through, buying and bidding behavior to find out a mapping from queries to features. For example, if users in the past issued a search for "roger federer" and then ended up buying "t-shirts" , "collectibles", and "rackets", then we will have a mapping from the query "roger federer" to the features ("t-shirts", "collectibles", "rackets") with some weights (w_1, w_2, w_3). This query to feature mapping is based on a desirability metric. So if lot of "roger federer posters" were available on sale but none of them got bought on site after querying for "roger federer", then the vector for the query "roger federer" would have the feature "poster" with a negative weight. Given a result set R for a query Q, we look at the feature vector V for query Q and try to see how many of the positive features for Q are found in R, which gives a measure of desired item diversity I_d . The feature set of R is the collection of terms found in the titles of any of the items found in R. So if for the query "roger federer" t-shirts, collectibles and rackets were found in the result set, the item diversity score I_d would be 1.0, else it would be some number between 0 and 1.0 depending on which features for the query "roger federer" were found in the result set R and the weights for those features.

We combine these individual diversities to get a composite selection score, $S = \phi(S_d, F_d, I_d)$. Our approach to learning the function ϕ is similar to the approach used in computing Maximal Marginal Relevance (MMR) [4]. We use a linear combination function to express total selection score S as a linear combination of individual diversities S_d , F_d and I_d .

$S = \alpha S_d + \beta F_d + (1 - (\alpha + \beta)) I_d$, where

$0 \le \alpha \le 1 \; and \; 0 \le \beta \le 1 \; and \; 0 \le (1 - (\alpha + \beta)) \le 1$.

We learn the weights α and β through human judgments and use a function ϕ which linearly combines S_d , F_d and I_d in proportions 0.2, 0.4 and 0.4 respectively.

2.4 Relevance

As a simple measure of relevance, we see how every individual item in the set is relevant to the query. The total relevance of the item set for the query is then an average of the relevance of individual items for the query.

The relevance of an item I for a query Q is again calculated using the feature vector V for the query Q which was described in section 2.3. The vector V for query Q is comprised of different features with positive and negative weights. An item I can be represented as a vector $V_i = (t_1, t_2, \ldots, t_n)$ where t_i are the terms in the title of the item. The relevance value of item I for query Q is, $r_i = \left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} |w_i|} \right)$, Where w_i is the weight for the term t_i in the feature vector V for query Q. If a term t_i is not present in the feature vector for query Q then $w_i = 0$.

Thus $-1 \le r_i \le 1$. The average of the relevance values r_i for all the items in the result set gives the relevance ρ ($-1 \le \rho \le 1$) for the result set.

Most eCommerce sites provide an option for users to sort the retrieved item set using deterministic sort factors like "Price: Low to High" or "Best-selling" or "Time: Newly Listed" or "Distance: nearest first" etc. If the search engine is Boolean and ranks the items based on some factors like price of item or auction ending time, there is a higher probability that irrelevant items will show up in the result set. This is true in case of queries like "ipod nano" where the intent of most buyers is to

buy an mp3 player but a Boolean search with *TimeRank* sort might show only "skins", "chargers" and "batteries" depending on the available inventory for each of these kinds of items.

2.5 Perturbation

As discussed in sections 2.3 and 2.4, we use historical click-through and user activity information to map a query to features, and this information is used to calculate selection (S) and relevance (ρ). Thus, it becomes important that in construction of the ideal result set, room be left for a perturbation factor P , so that new items (not indicated in the query to feature mapping), also have some non-zero probability of making it to the ideal result set. Some queries to feature mappings remain stable over time while others do not. For example, we extracted some historical data to see how the top feature related to the query "poster" changes with time. It changes rapidly and dynamically as new movies or celebrities get popular. Similar changes in features for the query "harry potter" were observed as new books and movies were released. Identical effects are observed when newer models of electronic products get released as well.

3. EXPERIMENTS

Our experiments consisted of running queries against the search engine, obtaining the item set from the first page of search results, and calculating various measures for those item sets. For our experiments we used 3 different sets of queries. The first set consists of 50 most frequent queries on the site which cover close to 2% of total volume of traffic because of the power-law nature of search traffic and we refer to it as *TopQueriesSet*. We also looked at top 1000 queries and randomly picked 100 queries out of those 1000. This forms our second set which we refer to as the *RandomQueriesSet*. We mined user sessions to find the sessions in which users issue a query and then exit the site i.e. they don't do any other viewing, bidding or buying activity after issuing the search. We collected such queries and these queries comprise our third set which we refer to as *ExitQueriesSet*.

For our experiments we have used two different ranking options. The first option is ranking by time (*TimeRank)*, where the items that are about to expire show up first. The second option is ranking by RelevanceMatch (*RelevanceMatchRank*) which is provided by the site to sort items by some multi-dimensional relevance function. We calculated the trust, value, selection and relevance scores for item sets returned by queries from *TopQueriesSet* and *RandomQueriesSet*. We performed the experiments using both *TimeRank* and *RelevanceMatchRank* and analyzed various dimensions.

The trust scores are better for *TopQueriesSet* as compared to *RandomQueriesSet* irrespective of the ranking used. This shows that popular and in-demand items (head inventory) are usually available to be bought from reputable sellers. However, the not so popular items in the long tail may not always be available from reputed sellers. We also observed that with either set, the values of τ (trust) and ρ (relevance) are higher with *RelevanceMatchRank* as compared to *TimeRank* as the conditional ranking algorithm in *RelevanceMatchRank* uses some of the features that we use to measure τ and ρ of the retrieved item set. For *TimeRank,* values of ρ are low for many queries irrespective of the query set being *TopQueriesSet* or *RandomQueriesSet.* Selection S is not dependent on the query set (*TopQueriesSet* or *RandomQueriesSet*), neither is it dependent on the ranking. S is different for different queries (varied from 0.35 to 0.78) and it is observed that S is a function of how vague, ambiguous or generalized the query is. ν is high and stable irrespective of the ranking algorithm applied or the query set used, implying that generally the result item sets do offer a high value proposition.

In order to understand and observe the importance of diverse result sets, we experiment to observe the correlation between the diversity of the result set and the exit rate of users from site.

There are many users who search for items and then exit without performing any other on-site activity. These search terms are available in *ExitQueriesSet*. We see that some users exit because no items are returned for their query or few items are returned for their query and those items don't pique their interest.

However, there is another section of users, who issue queries, and who also see a full page of item results, but they still quit the site, without even clicking on the items. This could happen if the items were irrelevant, so relevance does impact exit rates. However, there are user sessions where exits happened after search for which items were shown with a reasonably high value of the relevance factor. We try to observe if this has any correlation with diversity. An item set with reduced diversity although relevant, may not be able to account for the interests of the overall user population.

Some queries for which S was low but ρ was high and led to users exiting from site are shown below in Table 1.

Table 1 Values of S for some queries from *ExitQueriesSet* which had a high value of ρ . Note that only items that are available for sale, tagged as desirable based on historical query mappings but still were not found in the result set are discussed. Only a couple of representative items are discussed for every query.

Query (S)	**Some observations from Result Set**
Oeuf (0.20)	Did not find armbands and measuring cups. Did find some form of Tupperware.
Johnsons (0.30)	Found vintage items but did not find matchboxes and fishing lures.
Alfa 75 (0.37)	Found lights for alfa 75 but did not find alternator.
Albert (0.38)	Found jerseys and vintage items but did not find diamond pendants and green wood tree salad plates.
Game board (0.40)	Found backgammon, wooden and vintage game boards but did not find game boards for Parcheesi and checkers.

It can be seen that many queries in Table 1 are generic. It is difficult to know which game board a user is looking for when a user searches for "game board". However, for users who are specifically looking for something but using generic queries; not finding what they want is frustrating and those users might just walk away. As seen from column 2 of Table 1, items shown for

these queries were relevant, but as there were lot of items that could be reasonably relevant (based on historical click-through data) to these vague queries, probably items that the user was looking for were missing from the result set and the user just exited the site after the search.

As the queries get more specific usually the desirable diversity found in result set also goes up because more specific the query, less the number of different desired items matching the query and higher the chance that most of those themes get captured by the item set. For example, the query "game board" has a low S of 0.40 but the query "monopoly game board" has a selection score S of 0.65.

3.1 Correlation between S and ρ

In order to see if diversity S and relevance ρ are correlated we calculated those scores for different queries. A high value of selection S did not necessarily mean that the result set had a good relevance ρ and vice versa.

The queries with high values of S as well as ρ lead to a good buyer experience. The queries with low values of both S and ρ lead to a poor buyer experience. However, both values would be low with a *RelevanceMatchRank* sort only if appropriate inventory is not available on the site, and is more of an inventory availability problem than a search or ranking problem.

Queries leading to result sets with high S and low ρ usually have good variety of items but are missing variety within popular items. To please most users with optimal satisfaction a good balance of S and ρ is needed.

Queries leading to result sets with high ρ and low S are of interest to our study. As described in section 3 , such result sets cause some users to walk away as they don't see any items matching their interest. The average selection score for queries that led to exit is 0.61. As a comparison the average selection score of top queries was 0.72. Table 2 shows values of S for some popular ipod related queries leading to high user satisfaction. The higher satisfaction was measured in terms of number of sales in user sessions in which those queries were issued. It can be seen that queries that lead to higher satisfaction have higher selection (S) scores for corresponding result sets.

Table 2 Table indicating Selection (S) values for popular ipod related queries which led to a satisfying buyer experience. The S values for these queries are much higher than those seen for *ExitQueriesSet*

Query	(S)	Query	(S)
ipod touch 8gb	0.85	iphone 3g	0.84
ipod touch	0.87	unlocked cell phones	0.80
apple ipod touch 32 gb	0.82	ipod nano 4th generation 8gb	0.82

3.2 Do S, ρ relate to User Satisfaction?

In order to precisely measure and quantify the impact of diversity on satisfaction of users, we looked at a large sample of click-stream logs (More than 1M queries conducted by users). We looked at all search impressions, for which the relevance ρ of item set shown was above a threshold $t_r = 0.4$ determined empirically. We term these impressions as data points with $\rho = high$. We looked at all impressions which led to user satisfaction and user dissatisfaction. User satisfaction was defined as a click on one of the items shown to the user. User dissatisfaction was defined as exit from the site. For all these search impressions we also measured the selection score S . Selection score $S \leq 0.3$ was referred to as $S = low$ and selection score $S > 0.3$ was referred to as $S = high$. We observed that

$\left(\frac{P(exit / S = high, \rho = high)}{P(exit \mid S = low, \rho = high)} \right) = 0.62$. This ratio is statistically significant based on the large number of samples from the data we used. This shows that on pages with relevant item sets, the probability of a user being dissatisfied is significantly increased if the diversity of the item sets on the page is lowered. Thus, it is essential to guarantee diversity and proper selection to be able to satisfy most users from the population. None of these factors are highly correlated and in order to satisfy majority of buyer population the result set needs to be optimized based on τ (Trust), υ (Value Proposition), S (Selection / Diversity), ρ (Relevance) and P (Perturbation).

4. CONCLUSIONS AND FUTURE WORK

We described factors that need to be optimized in eCommerce search in order to minimize the risk of dissatisfaction of the average user. We conducted a large-scale empirical study to find out the interplay between these factors – trust, value, selection, relevance and perturbation. We also showed through log analysis that lack of diversity is correlated with metrics like exits of users from site and that a lack of diversity could have a detrimental impact on an online marketplace.

An advanced search interface like the one we describe in [5] gives more control to the user but also requires more inputs from the user, due to which we believe its utility is limited to experts. As part of future work we would like to build a ranking function that can optimize search results based on the above mentioned factors to be able to help the non-expert average users.

5. REFERENCES

[1] Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A. Efficient Computation of Diverse Query Results. ICDE (2008), 228-236.

[2] Agarwal, R., Gollapudi S., Halverson A., Ieong S. Diversifying Search Results. WSDM (2009). 5-14.

[3] Parikh N., Sundaresan N. Inferring Semantic Query Relations from Collective User Behavior. CIKM (2008), 349-358.

[4] Carbonell J., Goldstein J. The use of MMR, Diversity-Based Re-ranking for Reordering Documents and Producing Summaries. SIGIR (1998), 335-336.

[5] Parikh N., Sundaresan N. A User-Tunable Approach to Marketplace Search. WWW (Companion Volume) 2011. 245-248.