Automating Relevance Banding in eCommerce Search using Click Model

Deependra Singh, Vinay Deolalikar
Search, Data Mining, Machine Learning & Personalization
Groupon
Palo Alto, CA 94306

Abstract. eCommerce is burgeoning: in the past five years, both in the USA and the UK, eCommerce retail sales have overtaken brick-and-mortar stores for the first time. Search is a primary means for users engaging in eCommerce. eCommerce companies often perform laborious human-intensive mappings of queries to various categories in their product taxonomies—a process called "banding." This is done in order to improve recall and precision of their search engines. In this paper, we propose a fully automated alternative to this manual process. We use statistical properties of the click-model that is constructed using query-click logs in order to automate this process. We propose an algorithm—probability banding—that performs banding in a fully automated manner. In large-scale A/B testing, our algorithm demonstrates considerable revenue and orders increase over the manual banding baseline. Our algorithms are now deployed at scale at CorpX—a multi-billion dollar eCommerce major.

1 Introduction

When a user submits a query to an eCommerce search engine, the search engine tries to understand the intent behind the query. In many eCommerce search engines, the most important queries (by frequency of occurrence, and monetization) are *mapped* to sets of nodes of the product taxonomy. A taxonomy is a DAG of many nodes (also called *categories*): a main-category has children categories, each of whom further has children sub-categories, as shown in Fig. 1. Each product in the inventory is placed into one or more categories in the taxonomy. Since the set of categories mapped to a query is also known as the *relevance band* of the query, this process of mapping queries to sets of relevant taxonomy nodes is called *relevance banding*.

At most eCommerce companies, including ours[1], it is a prevalent practice to *manually* perform relevance banding. This manual banding is typically performed, and maintained by expert search analysts, over a period of time. Clearly this is a time-consuming process, prone to human error, and subject to an intensive amount of human bookkeeping and tracking. This leads to our problem statement:

[1] A multi-billion dollar Fortune-100 company that we will call CorpX

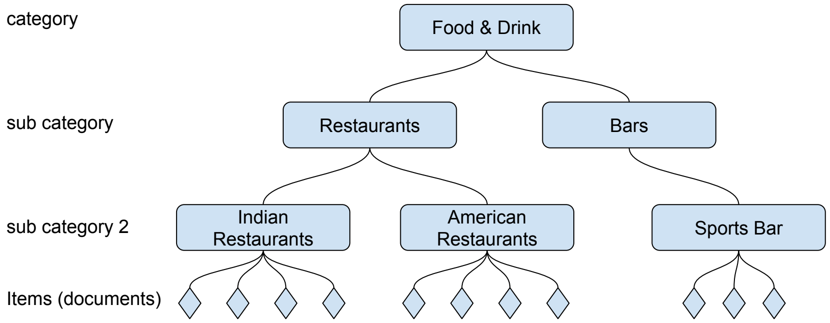

Fig. 1. A part of DAG of the taxonomy used at CorpX. Since our work pertains to eCommerce, we compute the click model on sub-category-2 using items, as opposed to documents for web retrieval. Details in text.

Can we automate relevance banding using statistical methods?

A well researched technique in the domain of *document* retrieval is *click model*. Click Models mine the query-click logs to find the query-document associations. We refer the reader to an excellent survey [1] on the increasingly rich landscape of click models. In recent years, many click models have been proposed and used in various ways: towards *improving document ranking in web-search* [2–4], towards *evaluating web search results* [5–7], and even towards *displaying advertisements* [8, 9]. In the context of eCommerce, authors in [10] discuss an application of click model for *atypical query identification*. However, to the best of our knowledge, there is no study on applying click models to eCommerce for search relevance. The presence of product taxonomies and their central role in search engines in eCommerce means that standard click model techniques from document retrieval do not carry over to this domain. In this paper, we address this gap in literature, applicable to a domain of high commercial importance, by applying the click model to automate relevance banding for eCommerce.

1.1 Click Model

When trying to understand query intent, the log of clicks made by the users issuing the query is a often used proxy for ground truth. Namely, items clicked are, on average, relevant to the query intent. Besides, query click data is cheap and is readily available, making it a natural choice for improving the search system.

In the context of a taxonomy, when a user makes a query, and then clicks on items in the result set, these clicks are recorded against the categories that the items are in. For each query, we aggregate these user clicks over a certain time period. These aggregated clicks, when normalized, form a probabilistic distribution over categories. This distribution of clicks that results from a query are formalized into the *click model* [1].

In the following formalization, we have used the generic term "category" to mean a category, a subcategory-1, or a sub-category-2 in the taxonomy. In practise, our click model is computed for sub-category-2 level nodes.

Definition 1 (Click Model). *The probabilistic distribution of association of a query $Q_i \in \mathcal{Q}$ to the taxonomy $\mathcal{T}$ is denoted by $\Pr_{Q_i}(\mathcal{T})$, and is defined as*

$$\Pr_{Q_i}(\mathcal{T} = C_j) = \frac{clicks_{Q_i,C_j}}{\sum_{j=1}^{j=m} clicks_{Q_i,C_j}}$$

Where $clicks_{Q_i,C_j}$ denotes the number of user clicks for query Q_i pertaining to the items in result-set belonging to category C_j.

Click model probability of a category being associated with a query can be understood to be the proportion of clicks that category accounts for in the query result-set over time. In practise, for a given query, this proportion turns out to be significant only for a few categories.

2 The Probability Banding Algorithm

We now gather all the components defined in the previous section to compute the relevance band for a given query Q_i . We propose an algorithm based on the click model query-category probability distribution: probability banding.

Idea and Algorithm. The probability banding algorithm uses the click model probability distribution $\Pr_{Q_i}(\mathcal{T})$ for Q_i to place the most important categories from the taxonomy $\mathcal{T}$associated to Q_i into the relevance band.

We walk the reader through the probability banding algorithm. First, it obtains the list of categories sorted in decreasing order according to probability given by $\Pr_{Q_i}(\mathcal{T})$. It iterates over the sorted list of categories, adding the categories into the band until any of the following happens:

- the number of categories added reach the maximum number of categories allowed for the band (lines 4-6)
- hits a category with probability less than the minimum threshold (lines 7-9)
- cumulative sum of the probabilities reaches the cumulative sum threshold(lines 10-12).

Finally, it returns the single relevance band it has constructed, comprising the most important categories for Q_i .

Determination of parameters.

- *max_categories*: Empirically, we determine that on an average, the probability of top four categories sums up to 0.9 for the queries in our database. Accordingly, we set the maximum number of categories at 4.
- *min_prob*: A value of 0.02 was chosen since a lower probability than that cannot be justified to be statistically significant, and is deemed to be noise.
- *max_cum_prob* is set to 0.90 as a converse to the previous setting; namely, categories left after reaching 0.90 cumulative probability are typically noise.

Algorithm 1 Probability Banding (max_categories, min_prob, max_cum_prob)

```
Require: sorted_category_list : list of categories sorted in decreasing order of
    probability given by $\Pr_{Q_i}(\mathcal{T})$
band ← { };                                          ▷ Initializing the band
cum_prob ← 0.0;
for Each category $C_j$ in sorted_category_list do
    if (current_band.size() ≥ max_categories) then
        break;
    end if
    if ($\Pr_{Q_i}\{\mathcal{T} = C_j\} \leq$ min_prob) then
        break;
    end if
    if (cum_prob ≥ max_cum_prob) then
        break;
    end if
current_band.add($C_j$);
cum_prob += $\Pr_{Q_i}\{\mathcal{T} = C_j\}$;
end for

return band;
```

3 Empirical Validation

3.1 Dataset

We use real time data from *CorpX* search engine to train the click model, and to assess the performance of our algorithm. We first describe the details of the historical data used to train the click model, afterwards we give the details of experiment datasets.

Click Model Dataset. To train the click model, we use the click and purchase activities on *CorpX* search engine during the period of *Dec 1, 2014 to May 30, 2016*. Our query space consists of $148,737$ queries. These queries were issued $101,379,388$ times by $36,598,660$ users in the given time period. Our inventory consists of $1,927,349$ items placed into $1,517$ categories in the taxonomy. For the aforementioned queries, there are $210,754,810$ clicks and $16,727,410$ orders distributed over the categories in the taxonomy. While training the click model, we discard the query-category pairs with less than 30 clicks so as to reduce noise. After training, we obtain $863,403$ query-category pairs.

3.2 Probability Banding vs. Manual Banding

Objective. In this experiment, we wish to compare the probability banding algorithm against the manual banding. Our manual bandings have been developed by a team of expert search analysts at *CorpX* over a period of *one year*. We compute the relevance bands through running the Algorithm 1 for each query in the dataset.

Methodology. To compare the two methods, we use an A/B testing framework. Our baseline comprises of manual relevance bands and our variant has relevance bands computed through probability banding algorithm. We diverted equal amount of search requests to probability banding and baseline manual banding for a period of *15 days* from *June 24, 2016 to July 10, 2016.* During this time period, set of $148,737$ queries were issued through $11,718,958$ sessions and the inventory comprised of $1,089,833$ items. Distribution of these sessions across the control and variant is given in Table 1.

Table 1. Experiment 1: Distribution of query sessions

Variant - Probability Banding	Control - Manual Banding
$5,861,895$	$5,857,063$

Results. We observed an increase of +3.366% in net revenues and an increase of +2.922% in orders. Our results are statistically significant. Results are summarized in Table 2.

Table 2. A/B test results comparing Probability Banding and Manual Banding

	Net Revenue Increase	Orders Increase
Increase	+3.366%	+2.922%
T-statistic	2.035	2.953

3.3 Discussion and Conclusion

Our hypothesis was that an automated method, based on click model, is not only more efficient operationally, but can outperform manual banding as far as search efficacy. This is proved by the significantly positive results of our A/B testing experiment. Readers relatively unfamiliar to A/B testing in eCommerce should kindly note that increases of 0.5% are considered significant improvements; whereas our revenue increases are 3.3%.

Furthermore, the entire click model banding process is fully automated, requiring no human intervention at any stage. This contrasts with the intensive manual mapping and re-mapping (upon seeing search results) done previously.

Our work shows that the click model based approach is considerably more efficacious than manual banding. Our system is now deployed at scale at CorpX—a multi-billion dollar company. Although we are not at liberty to disclose precise dollar values of the revenue increases, they are in the range of millions of dollars per financial quarter.

We are currently experimenting with further innovations using information theoretic algorithms based on the click model. We hope to report on these soon.

References

1. Chuklin, A., Markov, I., Rijke, M.d.: Click models for web search. Synthesis Lectures on Information Concepts, Retrieval, and Services **7**(3) (2015) 1–115
2. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th international conference on World wide web, ACM (2009) 1–10
3. Dupret, G., Liao, C.: A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 181–190
4. Borisov, A., Markov, I., de Rijke, M., Serdyukov, P.: A neural click model for web search. In: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee (2016) 531–541
5. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 621–630
6. Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S.: Expected browsing utility for web search evaluation. In: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM (2010) 1561–1564
7. Chuklin, A., Serdyukov, P., De Rijke, M.: Click model-based information retrieval metrics. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM (2013) 493–502
8. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: estimating the clickthrough rate for new ads. In: Proceedings of the 16th international conference on World Wide Web, ACM (2007) 521–530
9. Zhu, Z.A., Chen, W., Minka, T., Zhu, C., Chen, Z.: A novel click model and its applications to online advertising. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 321–330
10. Pradhan, N., Deolalikar, V., Li, K.: Atypical queries in ecommerce. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM (2015) 1767–1770