Automating Relevance Banding in e-Commerce Search using Click Model

Deependra Singh*, Vinay Deolalikar†

Search and Data Mining
Groupon
Palo Alto, CA 94306
E-mail: *deeps@seas.upenn.edu, †deolalikar.academic@gmail.com

***Abstract*—e-Commerce is burgeoning: in the past five years, both in the USA and the UK, e-commerce retail sales having overtaken brick-and-mortar stores for the first time. Search is a primary means for users engaging in e-commerce. e-commerce companies often perform laborious human-intensive mappings of queries to various categories in their product taxonomies—a process called "relevance banding." This is done in order to improve recall and precision of their search engines. In this paper, we propose fully automated alternatives 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 two algorithms—probability banding, and entropy banding—that perform banding in a fully automated manner. In large-scale A/B testing, our algorithms demonstrate considerable revenue and orders increase over the manual banding baseline. Our algorithms are now deployed at scale at CorpX—a multibillion dollar e-commerce major.**

***Index Terms*—e-commerce search, relevance banding, click model**

I. Introduction

When a user submits a query to an e-commerce search engine, the search engine tries to understand the intent behind the query, and subsequently selects from its inventory those items that match the query intent. In order to specify the query intent, 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. If present, such a taxonomy provides a human-curated layer of abstraction and classification over the actual items. Since the sets of categories mapped to a query are also known as the "relevance bands" of the query, this process of mapping queries to sets of relevant taxonomy nodes is called "relevance banding". The rest of the paper focuses on e-commerce search engines having a product taxonomy.

At most e-commerce companies, including ours[1], it is a prevalent practice to *manually* perform the 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:

> Can we automate relevance banding using statistical methods?

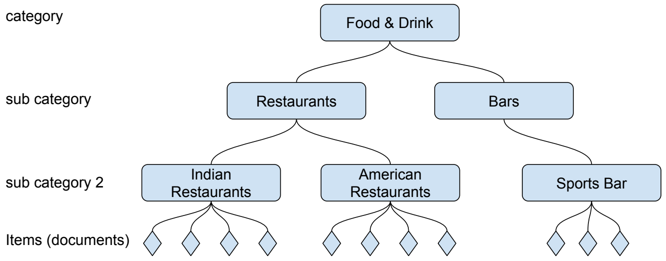

Fig. 1. A part of DAG of the taxonomy used at CorpX. Since our work pertains to e-commerce, we compute the click model on sub-category-2 using items, as opposed to documents for web retrieval. Details in text.

We approach this problem by using the statistical properties of "click model", a well researched technique in the domain of document retrieval. Click models mine the query-click logs to find the query-document associations.

In this paper, we propose two algorithms based on click model, to automate relevance banding: *probability banding*, and *entropy banding*. Not only do these algorithms fully automate this process of relevance banding, they also outperform manual banding in search efficacy. Our algorithms demonstrate considerable revenue and order increases over the manual banding baseline in large-scale A/B testing.

II. Related Work

In web search domain, the problem of associating queries to a set of categories has been studied in detail. Queries can be classified along multiple dimensions, including search goals [4], [10], location-sensitivity [15], and semantic topics [14], [11]. Of these, the most closely related works are on topical query classification. Authors in [14] train a bridging classifier on an intermediate web taxonomy, and tackle the sparseness of query features by augmenting queries with search engine

* † Work done while authors were at Groupon Technologies, Inc.

[1] A multi-billion dollar Fortune-100 e-commerce major that we will call CorpX.

results. In [11], instead of improving feature representation, authors chose to increase the amounts of training data by semi-supervised learning with click-graphs. However, there is scant literature on classifying queries into the product taxonomy nodes in e-commerce domain, and its subsequent application to search relevance. One key difference between web documents and e-commerce items is that latter are pre-mapped into the target categories as opposed to former.

Similarly, there is extensive literature on click models and their applications to web search. We refer the reader to an excellent survey [6] on the increasingly rich landscape of click models. In recent years, many click models have been proposed and used in various ways, including towards improving document ranking in web-search [3] and towards evaluating web search results [7]. In the context of e-commerce, authors in [12] 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 e-commerce for search relevance. The presence of product taxonomies in e-commerce search engines 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 e-commerce.

A tangential area from which relevant ideas can be drawn in future is recommender systems, which seeks to produce individualized recommendations. We refer the reader to [1] for a comprehensive treatment of this mature research field. For a more recent survey, refer to [13], [2].

III. Background

A. Click Model

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. Over a certain time period, we aggregate these user clicks for each query. These aggregated clicks, when normalized, form a probabilistic distribution over categories. These resulting distribution of clicks are formalized into the *click model* [6].

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ be the set of all queries, and $\mathcal{T} = \{c_1, c_2, \ldots, c_m\}$ be the taxonomy. Let w_{q_i,c_j} denote the number of user clicks for query $q_i \in \mathcal{Q}$ pertaining to the items in result-set belonging to category[2] c_j.

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 $P_{q_i}(\mathcal{T})$, and is defined as

$$P_{q_i}(\mathcal{T} = c_j) = \frac{w_{q_i,c_j}}{\sum\limits_{j=1}^{m} w_{q_i,c_j}}$$

We also use $p(c_j)$ to denote $P_{q_i}(\mathcal{T} = c_j)$ when the query q_i is clear from the context. 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 practice, for a given query, this proportion turns out to be significant only for a few categories.

Although clicks provide an important source of implicit feedback, and have been shown to be an effective method to extract relevance information, research shows that clicks suffer from a number of biases. In the following paragraphs, we explain these biases and the corrective measures we have employed.

1) Appearance Bias Correction: When presented with the search results, a user tends to click on a result based on the accompanying meta-data or abstract. Therefore, a result may spuriously trick the user into clicking the result. We use the information whether the clicked item was purchased as a proxy to determine the importance of a click.

Let p_{q_i,c_j} denote the number of aggregated purchases on query q_i pertaining to the items belonging to category c_j. Consequently, we modify

$$P_{q_i}(\mathcal{T} = c_j) = \frac{w_{q_i,c_j} + \alpha \cdot p_{q_i,c_j}}{\sum\limits_{j=1}^{m} (w_{q_i,c_j} + \alpha \cdot p_{q_i,c_j})}$$

The constant α measures the relative weights of purchases versus clicks. We use an empirically determined value of 30 for α.

2) Position Bias Correction: Studies show that a user is less likely to click on a result at a lower position in the result set, albeit being relevant [9]. To correct for this bias, we apply a position correction factor β to each click/purchase activity. The basic idea is to give more weight to an activity at a lower position. Such a function is application specific; we use the following function in our system, starting at position $l = 1$:

$$\beta(l) = 1 + \frac{\log(\min(l, C))}{\log C}$$

Where C is the position after which β becomes constant ($\beta(l) = 2$ for $l \geq C$). We empirically tune the value of C to be 30 for our system.

Therefore, if A is the actual activity performed by a user at position l, then position corrected activity A_l is $A_l = \beta(l) \cdot A$. (Here, A could either be a click or a purchase.)

3) Temporal Bias Correction: User interests and behaviors in e-commerce may change over time [5]. Accordingly, we wish to stress recent activity (clicks/purchases) while mining query to category associations, and de-stress past activity. Therefore, each activity is decayed with a daily time decay factor $\theta \in (0, 1]$. This also enables the click model to adapt to changes in the taxonomy.

If $A(t)$ is the actual activity performed by a user at time t, and t_f is the time when click model is constructed, then the time decayed count of activity $A(t_f)$ is

$$A(t_f) = A(t) \times \theta^{(t_f - t)}$$

We apply these corrective measures to each activity, and calculate the probabilistic distribution from the resulting data.

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

B. Information Theory

The (information theoretic) entropy $H(X)$ of a discrete random variable X is defined by:

$$H(X) = -\sum_{x} P(X = x) \cdot \log P(X = x)$$

IV. Algorithms

We now gather all the components defined in the previous section to compute the relevance bands for a given query q_i . We propose two algorithms based on the click model query-category probability distribution: probability banding, and entropy banding.

A. Probability Banding

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

Algorithm 1 ProbabilityBanding (max_cats, min_prob)

```
Require: $sorted\_cats$ : list of categories sorted in decreasing
    order of probability given by $P_{q_i}(\mathcal{T})$
 1: $band \leftarrow \{ \}$;                    ▷ Initializing the band
 2: for Each category $c_j$ in $sorted\_cats$ do
 3:     if ($band$.size() $\geq max\_cats$) then
 4:         break;
 5:     end if
 6:     if ($p(c_j) \leq min\_prob$) then
 7:         break;
 8:     end if
 9:     $band$.add($c_j$);
10: end for
11: return $band$;
```

We walk the reader through the probability banding algorithm. First, it obtains the list of categories sorted in decreasing order of the probability given by $P_{q_i}(\mathcal{T})$. Then, 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 3-5)
- hits a category with probability less than the minimum threshold (lines 6-8)

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

b) Determination of parameters:

- max_cats: 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 allowed in the band to 4.
- min_prob: We chose a value of 0.02 since a lower probability than that cannot be justified to be statistically significant, and is deemed to be noise.

B. Entropy Banding

Probability banding considers a category with a high probability and a category with relatively lower probability equally important, by placing them in the same band. For example, in a scenario where a query, say, “pizza” is associated with the category “pizza” with a probability of 0.70, and with the category “italian-restaurants” with a probability of 0.25, probability banding will place both the categories in one band even though the given distribution implies that the category “pizza” is significantly more associated with this query than the category “italian-restaurants”. This leads us to the conclusions that the probability distribution should be interpreted in a more sophisticated manner to generate a *list of bands* than merely to filter the most important categories.

There are, therefore, two questions we must answer: how many bands should be created, and what should be the “cutoffs” for each band? We build an algorithm around information theoretic entropy to answer both.

Intuitively, we agree that categories with similar probability should reside in one band. Let us make precise the notion of similarity as follows.

> If the probability of a new category deviates more than a certain proportion of the average probability of categories in the band, that new category will be considered dissimilar to the ones in the band.

It remains to define the proportion used in the definition above. We denote this proportion by λ, and use information theoretic entropy in order to define it. We explain this in the determination of parameters later in this section.

With this idea, we describe our algorithm for entropy banding for a given query q_i and its probabilistic distribution over categories as $P_{q_i}(\mathcal{T})$ in Algorithm 2.

Algorithm 2 EntropyBanding(λ)

```
Require: $sorted\_cats$ : list of categories sorted in decreasing
    order of probability given by $P_{q_i}(\mathcal{T})$
$band\_prob\_avg \leftarrow 0.0$;
$prev\_prob \leftarrow 0.0$;
$band\_list \leftarrow [\ ]$;
$current\_band \leftarrow \{ \}$;
for Each category $c_j$ in $sorted\_cats$ do
    if $(prev\_prob - p(c_j)) > \lambda \cdot band\_prob\_avg$ then
        $band\_list$.add($current\_band$);
        $current\_band = \{ \}$;
        $previous\_prob = 0.0$;
    end if
    $current\_band$.add($c_j$);
    $prev\_prob = p(c_j)$;
    update $band\_prob\_avg$;
end for
return $band\_list$;
```

We walk the reader through the entropy banding algorithm. First, it obtains the list of categories sorted in decreasing order of the probability given by $P_{q_i}(\mathcal{T})$. It then iterates over the sorted list of categories. If the probability of the next category significantly changes w.r.t. to the current band, a new band is created (lines 6-10).

a) Determination of parameter λ: The deviation factor λ represents the proportion of average probability of the band, by which the next category is allowed to deviate, in order to still be considered similar to the existing categories in the band. Intuitively, a large value of λ implies that a broader range of probabilities can be part of a single band.

Turning this reasoning around, we see that for those queries whose probability distribution is concentrated, λ should be large; conversely, for those queries whose distribution is spread more evenly over many categories, λ should be small.

We know that entropy of a random variable quantifies exactly the concentration of its probability distribution. Therefore, to model the above explained behavior, we define λ for a query q_i as:

$$\lambda_{q_i} = 2^{-H(P_{q_i}(\mathcal{T}))}$$

V. Empirical Validation

We performed two experiments to assess the efficacy of our algorithms. In the first experiment, we compare the probability banding algorithm against the baseline manual banding. In the second experiment, we compare the entropy banding algorithm against the probability banding algorithm.

A. 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 $16,727,410$ purchases distributed over the categories in the product 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 with probability more than 0.02.

B. Experiment 1: Probability Banding vs. Manual Banding

In this experiment, we wish to compare the probability banding algorithm against the manual banding.

To compare the two methods, we use an A/B testing framework. Our baseline comprises of manual relevance bands developed by a team of expert search analysts at *CorpX* over a period of 6 months, 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 ($5,861,895$ sessions with manual banding and $5,857,063$ sessions with probability banding), and the inventory comprised of $1,089,833$ items. We observed a statistically significant increase in both net revenues and in orders.

C. Experiment 2: Entropy Banding vs. Probability Banding

In this experiment, we wish to compare the entropy banding algorithm against the probability banding algorithm.

To compare the two methods, we use an A/B testing framework. Our baseline comprises of relevance bands computed through probability banding algorithm, and our variant has relevance bands computed through entropy banding algorithm. Again, we diverted equal amount of search requests to entropy banding and probability banding during the period of September 21, 2016 to October 9, 2016. During this time period, set of $148,737$ queries were issued through $6,654,113$ sessions ($3,321,295$ sessions with probability banding and $3,332,818$ sessions with entropy banding). Once again, we observed a statistically significant increase in both net revenues and orders. Consequently, a variant of our algorithm was employed in our product.

D. Discussion

Our broad hypothesis was that automated methods, based on click model, are not only more efficient operationally, but can outperform manual banding as far as search efficacy. This is proved by the significantly positive results of both our experiments.

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.

Let us now compare the two experiments. Note that when the entropy of a query's click model distribution is low, the first band will likely have all the significant categories. In that case, there is not much difference between the two algorithms. On the other hand, when the entropy of a query is high, we would expect entropy banding to make a difference. As it stands, queries with high entropy (≥ 2) are a significant minority (8% of total $148,737$ queries). It is on these smaller set of queries that entropy banding will generate an increase in revenue. This is borne out by the results.

VI. Conclusion

We began our work with the hypothesis that the intensive and error-prone process of manual banding can be improved using automated statistical methods. We proposed two algorithms: probability banding, and entropy banding, in order to accomplish this automation. Both of them rely on the query-click model.

Probability banding is conceptually simple, and generates a single band of highly associated categories. Entropy banding aims to outperform probability banding on high-entropy queries. It generates (potentially) multiple bands, using information theoretic criteria to determine the band boundaries. Our A/B tests show significant increases in revenue and order generation over manual banding as a result of both these algorithms.

Our system is now deployed at scale at CorpX—a multi-billion dollar e-commerce major—, and further innovations based upon it are being experimented with. Although we are not at liberty to disclose precise revenue increase figures, they are in the millions of dollars (per financial quarter) range.

Future work points to algorithms that segment queries based on other statistical properties (including ambiguity), and tailoring of algorithms towards each segment. We are currently experimenting with a combination of entropy, frequency, and monetization. Another line of future work is to use these algorithms on query categorization using knowledge resources [14], [8].

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 transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[3] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A neural click model for web search. In *Proceedings of the 25th International Conference on World Wide Web*, pages 531–541. International World Wide Web Conferences Steering Committee, 2016.

[4] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[5] M.-C. Chen, A.-L. Chiu, and H.-H. Chang. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4):773–781, 2005.

[6] A. Chuklin, I. Markov, and M. d. Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

[7] A. Chuklin, P. Serdyukov, and M. De Rijke. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 493–502. ACM, 2013.

[8] L. Hollink, P. Mika, and R. Blanco. Web usage mining with semantic analysis. In *Proceedings of the 22nd international conference on World Wide Web*, pages 561–570. ACM, 2013.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7, 2007.

[10] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400. ACM, 2005.

[11] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008.

[12] N. Pradhan, V. Deolalikar, and K. Li. Atypical queries in ecommerce. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1767–1770. ACM, 2015.

[13] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[14] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 131–138. ACM, 2006.

[15] X. Yi, H. Raghavan, and C. Leggetter. Discovering users' specific geo intention in web search. In *Proceedings of the 18th international conference on World wide web*, pages 481–490. ACM, 2009.