Improving Product Search with Session Re-Rank

a Walmart data mining project

Charles Celerier (cceleri), Bill Chickering (bchick), and Jamie Irvine (jirvine)

June 9, 2013

Walmart.com maintains an online catalog of over 2M products. Consequently, enabling users to quickly find products that conform to their specific needs and tastes is especially challenging. Given the difficulty of its task, Walmart.com's product search engine does an impressive job of interpretting the user-provided query and rapidly returning relevant results. Yet, there remains highly significant information that is not fully leveraged. The details of a user's online shopping session are indicative of a user's intent and compliment—indeed, provide context for—the user-provided query. In this report we describe and analyze a ranking scheme we call *Session Re-Rank* (*SRR*) that can potentially induce a large increase in both click-through-rates and conversions on the first page of query results.

1 The Technique

SRR works by comparing previously clicked items with the top N items returned by the search engine in reponse to a query. Items in the top N that are sufficiently similar to previously clicked items are promoted. The extent (i.e. number of positions) of the promotion for a particular item is a function of its similarity to previously clicked items, its original position, and the promotions of other items.

The similarity between an item to be shown and a previously clicked item is determined within five distinct vector spaces: *click-space*, *cart-space*, *query-space*, *title-space*, *item-space*. The non-unique representation of an item within each of these spaces may be thought of as a binary vector or a set of objects. (MapReduce jobs process historic query data to construct indexes whose keys are itemids and values are lists of the appropriate objects. Great care went into ensuring that index entries can be accessed in $\mathcal{O}(1)$ and that two entries can be merged to compute their intersection or union in linear time.) The similarity $J_s(A, B)$ of two items, A and B, within a particular space s is determined using Jaccard similarity. Similarities within particular spaces are then weighted and summed to determine the composite similarity

$$S(A, B) = \sum_s C_s (J_s(A, B))^{\alpha_s}, \tag{1}$$

where C_s and α_s are tuning parameters. The score σ attributed to an item to be shown is then the summation of composite similarities between itself and all previously clicked items plus the click-through-rate (CTR) Γ_i of the item's original position i

$$\sigma(A) = \sum_{B \in P} S(A, B) + \Gamma_i,$$

where P is the set of previously clicked items.

Another important parameter of *SRR* is the insert position I_0, which indicates that the top I_0 positions of the original ordering are to remain fixed. For the results discussed in this report, we use $I_0 = 2$, meaning that we never reorder the first two items of query results. We found this configuration maximized our metrics, although, we suspect this may largely be due to users' bias toward clicking on the first one or two positions indepedent of what is shown there. That is, $I_0 = 0$ might prove optimal for an online implementation.

2 Similarity Spaces

The premise behind *click-space* is that two items are similar if they are both clicked within the same online shopping session. The dimensions, or objects, of this space are therefore past user-sessions. The *clicks-index*

for the data presented in this report was contructed using approximately half of the provided data, or about 60M queries (about 120M page views).

Cart-space is based on the notion that two items are similar if they ever appear in a shopping cart together. The objects of this space are therefore shopping carts. The *clicks-index* for the data presented in this report was contructed using approximately half of the provided data.

Items are also considered similar if they appear in a query together. The objects of *query-space* are therefore queries. We make a distinction, however, between *user-queries* and *unique-queries*. The former are the well-defined entities within the raw Walmart data. The latter is an abstraction based on the notion that multiple *user-queries* can correspond to a single *unique-query*. To derive *unique-queries* from our data, we cluster *user-queries* as follows: two *user-queries* with the same search attributes (e.g. category or price filters) are considered the same *unique-query* if the strings constructed by concatenating the space-seperated, stemmed (we use the Python stemming.porter2 module), forced to lower-case terms from each of their rawqueries are equal. We point out that while we achieved better results with this policy compared to simply using *user-queries*, we have no reason to believe that this is the ideal way to cluster queries for use within *SRR*. Indeed, we believe one way to improve *SRR* is to optimize the query clustering policy.

Title-space is straightfoward: each item is associated with a set of terms from its title. We ignore case, but at present do not stem, discard stop words, or weight terms in any way.

Finally, the structure of *item-space* is unique because it involves a level of indirection. The premise here is that if items A and B are clicked in a single user-session and items A and C are clicked in another user-session, that items B and C are similar because they have item A in common. In this way, a large number of relationships between items is created. *Item-space* resembles *click-space* in that if two items are clicked during a single session, they will have nonzero similarity. It differs from *click-space* in two key respects, however. First, items that have historically never been clicked in the same session can have nonzero similarity if they were each clicked with a common third item. Second, if items are clicked together in many sessions this will increase their Jaccard similarity in *click-space* but not in *item-space*.

3 An Example

To illustrate the efficacy of our technique, we present a real query example. The only fictitious part of the example will be our shopper's name, David. David is interested in the "Primo Ceramic Crock Water Cooler with Stand" and clicks on this item during his session. Sometime later he navigates to the "Grocery →Beverages →Water" category and searches for "water". He is presented with 300+ results and clicks the 90th item, a 3 liter jug of water.

The top six original results are compared to the *SRR* results in Table 1 where the first and third column represent the original ranking presented to David. Note that we rerank the 90th item from the original results to be the 3rd item in the *SRR* results. Tables 2 and 3 show the similarity scores from each index for each

Original Ordering		***SRR* Ordering**	
1	Great Value Purified Water, 24ct	1	Great Value Purified Water, 24ct
2	Nestle Waters Bottled Spring Water, 24ct	2	Nestle Waters Bottled Spring Water, 24ct
3	Voss Water, 16.9 oz (Pack of 24)	90	Arrowhead Mountain Spring Water, 3 l
4	Clear American Cherry Sparkling Water, 1 l, 12pk	63	Great Value: Distilled Water, 1 Gal
5	Clear American Water, 1 l, 12ct	38	Arrowhead Mountain Spring Water, 2.5gal
6	Clear American Peach Sparkling Water, 1 l, 12ct	8	Clear American Mandarin Orange Sparkling Water, 1 l, 12pk

Table 1: Original Ordering vs. *SRR* Ordering for "water" query

item in the two orderings compared to David's lone previously clicked item. In each of these tables, the first column correponds to Table 1 and the second column is the similarity metric from Equation (1) using our optimized tuning parameters.

	σ	CTR	Clicks	Items	Carts	Queries	Titles
1.	0.943 75	0.075 40	0.525 00	0.160 48	0.000 00	0.153 49	0.029 37
2.	1.202 11	0.039 00	0.686 16	0.210 10	0.000 00	0.239 45	0.027 40
3.	0.357 06	0.025 40	0.000 00	0.194 19	0.000 00	0.111 77	0.025 70
4.	0.752 68	0.019 50	0.451 34	0.198 29	0.000 00	0.060 63	0.022 92
5.	0.736 73	0.015 30	0.457 94	0.146 90	0.000 00	0.093 68	0.022 92
6.	0.174 83	0.012 90	0.000 00	0.096 78	0.000 00	0.042 24	0.022 92

Table 2: Index similarity scores of the top six original results to "Primo Ceramic Crock Water Cooler with Stand"

	σ	CTR	Clicks	Items	Carts	Queries	Titles
1.	0.943 75	0.075 40	0.525 00	0.160 48	0.000 00	0.153 49	0.029 37
2.	1.202 11	0.039 00	0.686 16	0.210 10	0.000 00	0.239 45	0.027 40
3.	0.953 10	0.000 23	0.657 38	0.222 59	0.000 00	0.045 52	0.027 40
4.	0.949 54	0.000 49	0.562 07	0.266 41	0.000 00	0.093 17	0.027 40
5.	0.937 69	0.000 91	0.645 39	0.262 01	0.000 00	0.000 00	0.029 37
6.	0.896 71	0.010 00	0.655 62	0.208 17	0.000 00	0.000 00	0.022 92

Table 3: Index similarity scores of the top six *SRR* results to "Primo Ceramic Crock Water Cooler with Stand"

The similarity scores from *title-space* are easily calculated by hand. We will show the calculation for the *item-space* similarity for the 90th item in the original ranking. By referring to the corresponding index for *item-space*, we can recall that the 90th item was clicked in a same session with 39 different items and the previously clicked item was clicked in a same session with 455 different items. The titles of the 13 items found in common are:

Great Value: Distilled Water, 1 Gal

Nestle Waters Bottled Spring Water, 24ct

Primo Mineral Water, 5 gal

Deer Park Sumo Bottle Natural Spring Water, 3l

Arrowhead Mountain Spring Water, 3l

PUR Advanced Faucet Water Filter Vertical - Chrome

Ozarka Natural Spring Water

Formula 409 All Purpose Lemon Scented Cleaner, 32 fl oz

Great Value Spring Water, 1 gal

Nestle Pure Life Purified Water, .5l, 35pk

Arrowhead Mountain Spring Water, 2.5gal

Primo Ceramic Crock Water Cooler with Stand

Augason Farms Emergency Water Storage Kit

We then calculate the similarity of the 90th item and the previously clicked item in *item-space* to be $13/(455+39-13) = 0.0270$.

We believe this example illustrates how our technique can form subtle relationships between items based on historical user session data. In this case, we have a shopper who had an interest in a water cooler and subsequently made a query for "water". If the shopper had been in a brick-and-mortar store, he likely would have been directed to the section of the store selling water jugs to use with the cooler he had picked up. In this case, *SRR* recognized David's previous click on a water cooler, related that water cooler to large water jugs, and showed David the water jug he was looking for all along.

4 The Data

Walmart.com has generously supplied us with a large dataset consisting of about 250M pageviews comprising about 120M query results which occurred over about 30 days. The data includes the user-provided rawqueries together with search attributes, visitorIds and sessionIds, shown items, clicked items, which items were placed in a shopping cart, and which items were ultimately purchased. In addition, they have provided detailed item information including title, description, category, and other details. The query data was randomized with respect to search time and then segregated into three disjoint sets. The first set, which consists of about half of the data, was re-structured into indexes that form four of the similarity spaces (*click-space*, *cart-space*, *query-space*, and *item-space*) we use to identify relationships between items in realtime (the remaining similarity space, *title-space*, was compiled separately using the provided item data). The second set, which consists of less than 5% of the data, was used for testing and optimization, allowing us to refine our technique and tune its parameters. And the third set, which includes about 10% of the data, was used in the experiments described and analyzed in this report.

5 The Technique vs The Experiment

An important distinction should be made between the *SRR* technique and the experiment described in this report. Both the technique and the experiment leverage the provided data—however, the experiment is a simulation and a limited one at that. A key limitation is that the provided query data is confined to what the user was actually shown. That is, the search engine may have identified several pages worth of results in response to a *user-query*, but our dataset consists only of those pages actually seen by the user. Meanwhile, the concept behind the *SRR* technique calls for a search engine to deliver to the algorithm the top N items in response to a *user-query independent of the number of items ultimately shown to the user*. As a consequence, it is difficult, if not impossible, to simulate our technique using shown query results that are truncated because a user only viewed one or two pages. Even more generally, the use of historic data to demonstrate the consequences of a online ranking algorithm is intrinsically limited by the fact that one cannot be certain how users would have behaved if presented with different results. Nonetheless, we have done our best to conduct the most fair and informative experiment and analysis.

6 The Experiment

A key feature of *SRR* is that it can only re-rank the results of a query if a user has previously clicked on an item during an online session. Consequently, *SRR* can only affect the subset of queries that occur in a session with previous clicks. We denote this subset of queries as ζ and limit our experiment and analysis to this set. It turns out that 25% of all queries are in ζ. Moreover, 28% of all clicks and 33% of all purchases occur within the query resultsets of ζ, since there is a correlation between previous clicks and clicks/purchases in a query. Thus an impact to this subset can have a significant impact overall.

The goal for the experiment is to simulate *SRR* using the provided historical query data, which is limited to what users were actually shown. An online implementation of our technique would receive the top N items from the search engine and re-rank them prior to showing any results to a user. Because of this, the final ranking would be independent of the total number of items actually seen by a user (determined by the number of pages a user clicked through). Our test set χ therefore consists solely of queries within ζ where either all items in the query resultset or at least $N = 100$ items were shown to the user. For example, if a user stops searching after viewing only the 16 items on the first page (the default number of items on the first page), this query is not included in χ, since it is unknown which other items would have been considered for re-ranking. On the other hand, if the search engine found only 13 items in response to a query, we have the complete query resultset and can therefore determine how *SRR* would have reordered the shown items. Similarly, if more than $N = 100$ were shown to the user, we can determine the reordering regardless of whether the query resultset is truncated since *SRR* only considers and re-ranks the first $N = 100$ items. χ makes up 25% of ζ, accounting for a total of 6% of all queries.

To construct χ we must discard all queries with a number of shown results less than $N = 100$ that are also divisible by 16. The reason for this is that Walmart.com provides two options for the number of items shown per page: 16 or 32. Thus, by performing the experiment on this subset of the data we precluded queries where the top N items are not available to our algorithm. The choice of $N = 100$, meanwhile, is somewhat arbitrary and was made by balancing our desire for a large test set with our desire to use a value appropriately large for an online implementation. It is therefore quite possible that a larger value of N (e.g. 1000) would achieve better results in the actual online scenario.

While we must constrain χ in this way due to the nature of the available data, we stress that this subset is certainly biased with respect to queries in general. For starters, queries with short resultsets are more likely to have all of their resultset seen by a user, and therefore, are more likely to be included in χ. It is not clear, however, if this particular bias tends to under- or overestimate the effectiveness of *SRR* since, as we will show, *SRR* is more effective on longer query results. Similarly, highly qualified queries—e.g. through the use of category or price filters—tend to have shorter resultsets, and hence, are more likely to be included in χ.

Just as interesting are the ways in which the queries and resultsets of χ are not biased. In Fig. 1 we show click-through-rate (CTR) as a function of position of the original data (i.e. not re-ranked) for both χ and query results in general. The two curves shown in the figure are quite similar, indicating that the quantity and distribution of clicks within χ are essentially representative of those in general. A few other features of this figure warrant brief comment. First, we see that χ has a larger CTR for the top two positions. This is likely due to the fact that highly qualified queries, which have shorter results and higher CTRs, make up a higher proportion of χ than queries in general. While this difference does results in a slightly higher frequency

of clicks in χ, we have no reason to believe that this alone results in a significant bias. Next, we see in the red curve a discontinuity appears at position 17 as a result of the pagebreak. In the entire dataset, the majoriy of queries are truncated at the first page, leading to significantly fewer views of items on other pages and thus fewer clicks. This discontinuity is absent from the blue curve since χ has a less severe dropoff in viewership from one page to the next. The bump at position 17 can be ascribe to users' tendency to disproportionately click on the topmost item shown on a page. Indeed, if we normalize for number of pages viewed by a user, we see this bump at 17 in the overal dataset as well. Consistent with this analysis, we find another smaller drop in the red curve and smaller bump in the blue at possition 33.

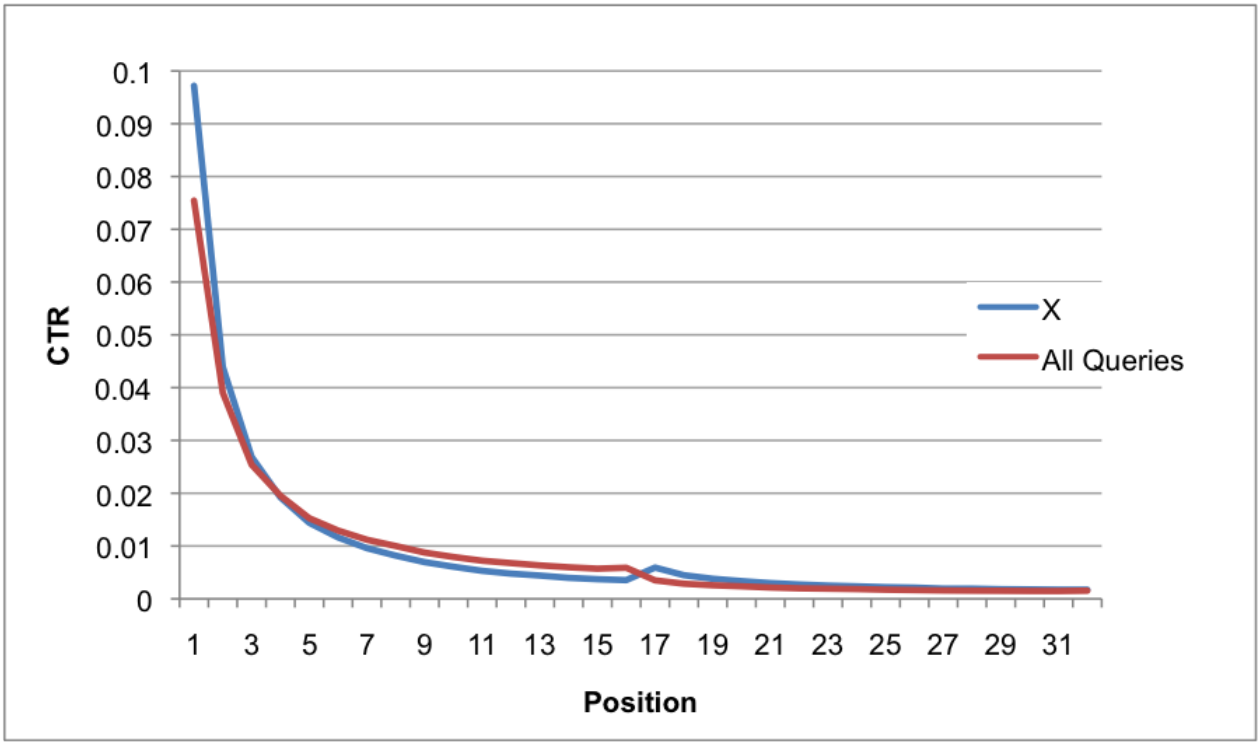

Figure 1: CTR as a function of position from the original data (i.e. not re-ranked) for the test set χ (blue) and all queries (red).

7 Metrics

A true test of the effectiveness of *SRR* would require online A/B testing. In the meantime, we can simulate the effect of *SSR* by running it on historical data and examining the new positions of clicked and purchased items. Assuming the user would have clicked or purchased the same items in this new ordering, we can compare the distribution of clicks in the original ranking to that yielded by *SRR*.

To compare two rankings of shown items, we focus on three key metrics. Our primary metric is the first-page CTR $\mathcal{C}$, defined as the likelihood that an item presented on the first page receives a click. As noted above, 75% of all queries are only one page long. This shows the importance of bringing desirable items to the first page and it motivates our focus on $\mathcal{C}$. We calculate this metric by counting the number of items in first-page positions that were clicked and dividing by the number of total items in first-page positions. More formally

$$\mathcal{C} = \frac{\sum_{q \in Q} \sum_{i=1}^{L_q} \mathbb{1}\{click\ @\ i \wedge i \leq 16\}}{\sum_{q \in Q} \sum_{i=1}^{L_q} \mathbb{1}\{i \leq 16\}},$$

where Q is a set of query results, L_q is the number of results for query q, and the number 16 is due to the fact that 16 items are shown on a page.

Our second metric is the purchasing rate of items on the first page $\mathcal{P}$. This is similar to $\mathcal{C}$, except here we consider purchases per first-page item instead of clicks. It is calculated as the number of items in first-page positions that were purchased divided by the total number of items in first-page positions. The importance

of position for purchases is even stronger than that for clicks. While 70% of all clicks were presented on the first page, that number is 85% for purchases. Formally, we have

$$\mathcal{P} = \frac{\sum_{q \in Q} \sum_{i=1}^{L_q} \mathbb{1}\left\{purchase \ @ \ i \wedge i \leq 16\right\}}{\sum_{q \in Q} \sum_{i=1}^{L_q} \mathbb{1}\left\{i \leq 16\right\}}.$$

The first two metrics focus on whether or not a desirable item was presented on the first page. To obtain a more granular picture of where desirable items are positioned, we also compute a third metric which we call *click-position score* $\mathcal{S}$. Somewhat similar to normalized discounted cumulative gain (NDCG), which is a common metric of search engine results, this score weighs the value of a clicked item by its position, giving higher weights to items closer to the top. For $\mathcal{S}$, the weight given to a click in position i is the CTR at position i Γ_i. In this way, we equate how often users click on a certain position to how valuable it is to put a desirable item there. Formally, we define *click-position score* as

$$\mathcal{S} = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i=1}^{L_q} \mathbb{1}\left\{click \ @ \ i\right\} \Gamma_i.$$

These metrics give a sense of how successful a ranking scheme is. Each one looks at a slightly different aspect of the ordering. Indeed, optimizing for one metric does not necessarily optimize for the others. We focus our optimizations and primary analysis on $\mathcal{C}$ as we believe it is the simplest and has the clearest impact to overall CTRs.

8 Results

Figure 2 shows the clicks-position score $\mathcal{S}$ as a function of each of the coefficients C_s discussed in sections 1 and 2. Here, we vary a single coefficient, corresponding to one of the five similarity spaces—*click-space*, *cart-space*, *query-space*, *title-space*, or *item-space*—while setting the others to zero. The ranking score for each of the top $N = 100$ items is then determined by only two terms, as per Eqs. 1 and 1, as

$$\sigma(A) = \sum_{B \in P} C_s (J_s(A, B))^{\alpha_s} + \Gamma_i,$$

where, as before, P is the set of previously clicked items and Γ_i is the CTR of the original position of item A. (Note that the exponents α_s are held fixed during these measurements.) Thus, since Γ_i is a monotonically decreasing function of position i, when $C_s = 0$, *SRR* returns the original ordering.

In each case, as C_s is increased from zero, the degree of reordering is enhanced. And for each coefficient, this reordering is accompanied by an increase in $\mathcal{S}$ indicating a greater concentration of clicked items, on average, in the top positions of query results as compared to the original ordering. Moreover, with the exception of *title-space*, the score increases monotonically with each coefficient. In the case of *title-space*, a broad maximum can be seen around $C_s = 0.05$. This indicates that an optimal combination between the contributions from *title-space* and CTRs exists, which maximizes $\mathcal{S}$. For all other coefficients, however, a maximum cannot be found. Rather, the value of $\mathcal{S}$ asymptotically increases with value of C_s, which demonstrates that the optimal average ordering, according to the metric $\mathcal{S}$, is determined solely by the similarity space irrespective of the original ordering.

With a locally optimized combination of coefficients we find that *SRR* significantly outperforms Walmarts original ordering on queries in χ. As shown in Figure 3, the re-ranking achieves a 16% increase in $\mathcal{C}$ compared to Walmarts original ordering. The implication is that items on the first page of the re-ranking are 16% more likely to be clicked than those in the original ranking. With 1M queries in our test set χ, these results are statistically significant, producing a 95% confidence interval of $[15.8\%, 16.2\%]$ for $\mathcal{C}$.

Notably, we see these results despite the benefit the original ranking receives from position bias, the well-established phenomenon that users tend to click on items presented at the top of a list. The average CTR of this data broken down by position (Fig. 1) suggests that Walmarts search is no exception to this phenomenon. Antoher obstacle overcome by *SRR* is that there is little room for improvement in first-page CTR, since the vast majority of clicks in χ (80%) are already on the first page.

To ensure that these results are not due to an underlying structure of the data, we constructed a Random Re-Ranker *RRR*. *RRR* operates exactly as *SRR* does, except instead of calculating various similarity scores, it generates a random number. This randomized algorithm significantly hurts the results in all three metrics, suggesting that *SRR* achieves its success by intelligently deciding which items are more likely to be clicked.

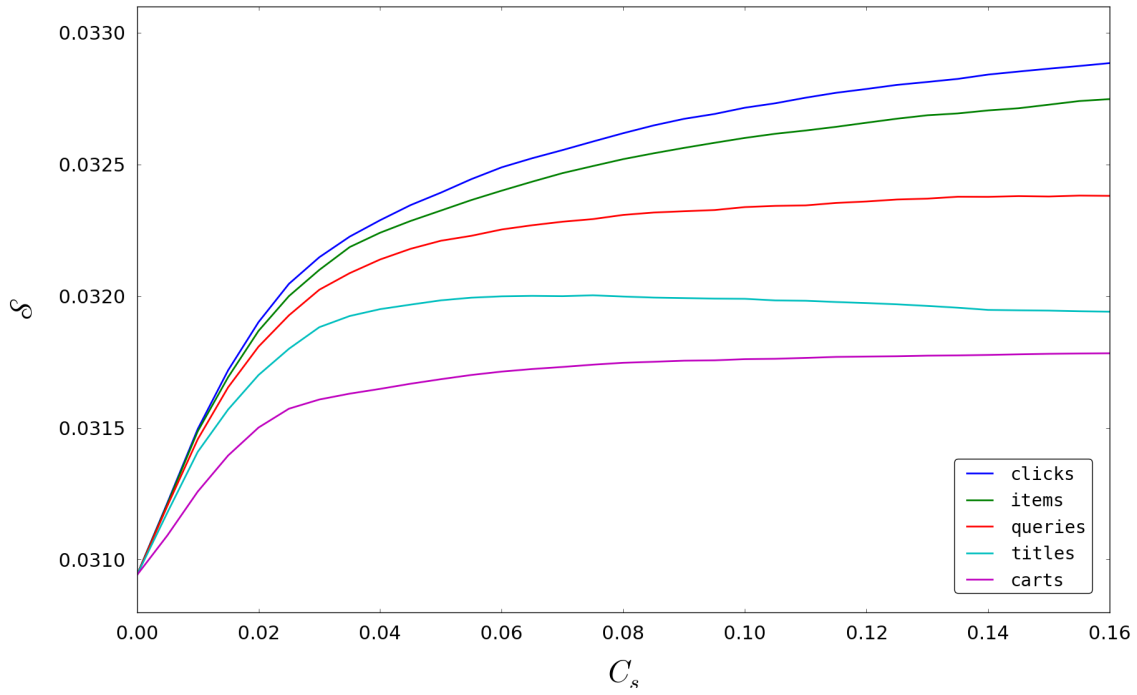

Figure 2: Clicks-position score $\mathcal{S}$ as a function of individual similarity space coefficients C_s. For each curve, a single coefficient is varied with all other set to zero.

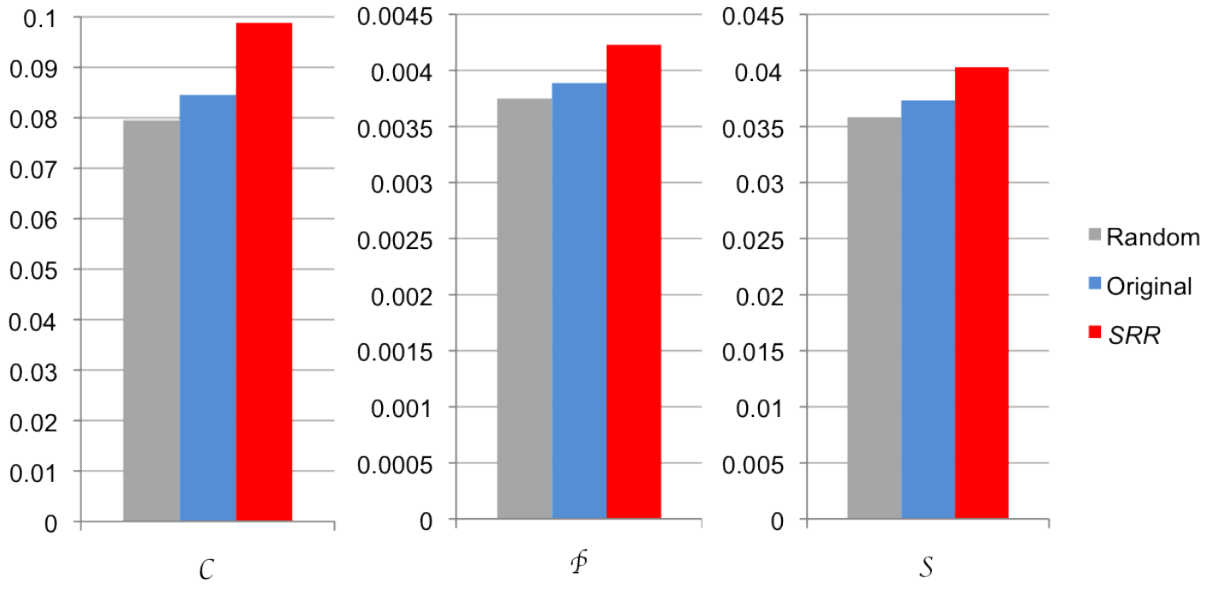

Figure 3: Comparison of the original ranking (blue) to that of randomly re-ranking (grey) and *SRR* (blue) using all three metrics: front-page CTR $\mathcal{C}$, front-page purchase rate $\mathcal{P}$, and click-position score $\mathcal{S}$.

Indeed, Table 4 compares the average CTR of items *SRR* moves on or off the first page to those chosen randomly. By accurately determining a users preference for certain items, *SRR* promotes items with a significantly higher CTR than it would by blindly picking items from other pages. Similarly, it demotes first-page items with a lower CTR than an average item on the first page. With these successful decisions, *SRR* manages to outperform the original ordering in the face of a number of disadvantages.

	CTR of demoted items	**CTR of promoted items**
SRR	6.24%	1.79%
RRR	2.42%	3.67%

Table 4: CTR of items promoted to the first page and demoted off the first page. *SRR* siginificantly outperforms random re-ranking in both categories.

Taking a closer look, we see that the performance of *SRR* is correlated to query length L_q. Figure 4 shows the percent increase in $\mathcal{S}$ over L_q. While it on average outperforms the original ordering for queries of any length, it gains greater improvements the longer a query is. This general trend makes sense; with more items to work with, *SRR* has a higher likelihood of finding an items more similar to a previously clicked one. The specifics of the scatter plot are a little less clear. There is a distinct jump in performance after moving to the second page ($L_q > 16$). This could be due to behavioral bias, as all these queries were seen by users who chose to search past the first page. We also see the rate of increase drop off after that jump. We speculate that this is due to diminshing returns for exposure to more items.

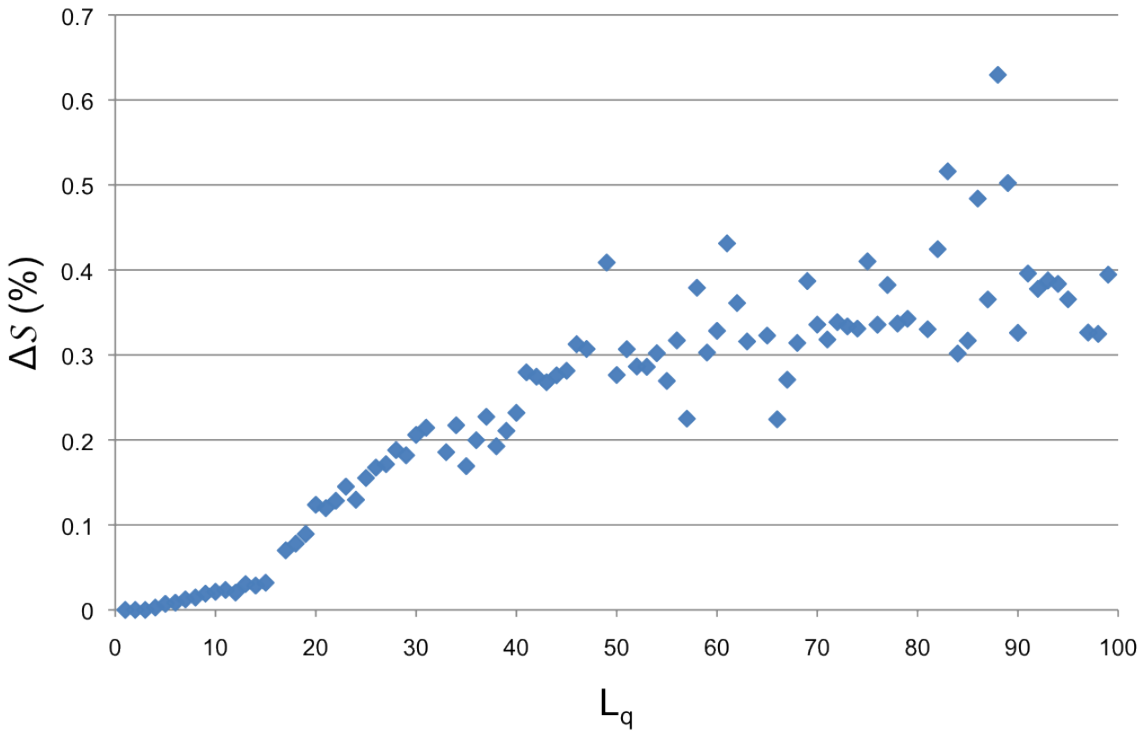

Figure 4: Change in click-position score $\mathcal{S}$ as a function of query length L_q.