Challenges and Research Opportunities in eCommerce Search and Recommendations

Manos Tsagkias
904Labs, The Netherlands
manos@904labs.com

Tracy Holloway King
Adobe, USA
tking@adobe.com

Surya Kallumadi
The Home Depot, USA
surya@ksu.edu

Vanessa Murdock
Amazon, USA
vmurdock@amazon.com

Maarten de Rijke
University of Amsterdam & Ahold Delhaize,
The Netherlands
derijke@uva.nl

Abstract

With the rapid adoption of online shopping, academic research in the eCommerce domain has gained traction. However, significant research challenges remain, spanning from classic eCommerce search problems such as matching textual queries to multi-modal documents and ranking optimization for two-sided marketplaces to human-computer interaction and recommender systems for discovery and browsing. These research areas are important for understanding customer behavior, driving engagement, and improving product discoverability and conversion. In this article we identify the challenges and highlight research opportunities to improve the eCommerce customer experience.

1 Introduction

As consumer purchase habits shift from shopping at traditional brick-and-mortar stores to shopping online, online product discovery has gained significance. Originally, web search was the primary source for product discovery, but recent surveys show the growing clout of eCommerce search engines as they become the option of first resort for product search [Degenhardt et al., 2019, U.S. Department of Commerce, 2017].

Product discovery in the eCommerce domain is primarily performed using search and recommendations. While one of the primary objectives of web search is information discovery, with eCommerce search and recommendations, the goal is to replicate customer in-store interactions and experiences that lead to purchases. An online shopper's intents can range from building knowledge of a domain to knowing exactly what they want [Su et al., 2018]. eCommerce sites aim

to improve customer engagement by satisfying these diverse information needs and reducing the friction associated with product search, discovery, and purchases.

In this paper, we categorize the aspects of eCommerce search and discovery into (1) customer goals, (2) business goals, and (3) data logistics. After discussing these in Section 2, we take deep dives into three eCommerce research areas: matching and ranking (Section 3.1), conversational search (Section 3.2), and fairness, confidentiality and transparency (Section 3.3). We outline experimental challenges around datasets in eCommerce research (Section 4) and conclude in Section 5.

2 What makes product search unique

Product search has two main stakeholders whose interests cooperate but also compete: customers and business owners. Customers need what businesses offer and businesses need customer purchases to survive: this forms the ground for cooperation. However, their interests also compete: customers want to find the best quality at the cheapest price and businesses want to maximize profit, which translates into higher prices for customers or lower costs for businesses. Data logistics adds another layer of dynamics due to a rapidly evolving inventory and privacy requirements. Below we discuss the needs of each stakeholder and how they link to design decisions for eCommerce search and recommendation.

2.1 Customers

Customers visit eCommerce sites to accomplish a goal. The goal can be simple (for example buying a coffee machine) or complex (such as fixing a hole in the wall). Complex goals typically involve a combination of several simple goals. Goals manifest themselves through user behavior, such as search queries, clicks, page views, and purchases. By analyzing user behavior, we can infer common strategies that users follow to achieve a goal. At the most granular level, user behavior manifests itself as search queries and interactions. When these are grouped together, we can infer semantic labels which map to customers' intents. Following intent over time, we find visitors following a customer journey, which can consist of one or multiple intents. Customer journeys are not limited to a single session and site but span multiple sites over long periods of time. Zooming out further, at the macroscopic level, we find customer journeys that extend from online to offline and vice versa and from simple to complex goals.

Query intent. Why do people visit eCommerce sites? Looking at research on understanding customer intent from different eCommerce companies [Sondhi et al., 2018, Rowley, 2000, Su et al., 2018] we find that, viewed abstractly, the taxonomy of intent of eCommerce queries is similar to that of web search queries: navigational, informational, transactional [Broder, 2002]. However, each intent takes a different form in eCommerce. Here, navigational queries are product serial numbers or names and help queries geared towards contacting customer support, tracking orders, or returning items. Informational queries are broader and typically map to a leaf in the inventory taxonomy or seek information about specific product attributes (e.g., *are batteries included with this radio*). Transactional queries are a mix of navigational and informational queries followed by

a purchase. Su et al. [2018] suggest three query intents for eCommerce: target finding, decision making, and exploration. Sondhi et al. [2018] from Walmart propose a larger set of intents: shallow exploration, targeted purchase, major-item shopping, minor-item shopping, and hard-choice shopping. Home improvement eCommerce sites include a project query intent, which does not exist for all eCommerce domains. This suggests that there is no single taxonomy for eCommerce customer intent because the taxonomy relies on the domain of the business, their inventory and their services. We propose a high-level, eCommerce-specific customer intent taxonomy of:

1. product intent (the majority of queries)
2. help intent (e.g., *contact customer support*, *track my order*)
3. off-site web-search intent (e.g., *facebook login*)

with the work from [Su et al., 2018, Sondhi et al., 2018] a specialization within the product intent leaf. Identifying the user's query intent allows an eCommerce site to provide a more targeted search and recommendation experience.

On-site customer journey. Is there one or multiple intents per customer visit? Su et al. [2018] suggest that customers can have a number of goals depending on where they are in their customer journey. During their customer journey, the customer is contemplating multiple products and interleaving tasks, discovering related products and gaining the confidence to purchase a product. Consider the following: Customers start their journey via a funnel [Blake et al., 2016]. First they issue broad queries (such as, *jeans*), followed by refinements (either by explicit "left rail" refinements or by reformulation), often examining multiple products before making a decision. Since a large portion of eCommerce customer journeys are initially exploratory, recommendations are valuable for inspiration, serendipitous discovery, and basket building. Search becomes more important once the customer has shaped their view of what they want. To provide a seamless customer experience, search, recommendations and the user interface should all be aware of the state of each customer's journey, and accomplishing this involves understanding product attributes, customer behavior, and the query context.

Global customer journey. What is the span of a customer journey? The customer journey can span multiple sites and offline interactions. Customers search for competing brands and products within and across eCommerce businesses, both online and at physical stores. This is partly because a given eCommerce site has a limited set of products and brands, but also because of the customer's interest in finding the best quality at the best price. This behavior is known as comparison shopping. For example, Lowes carries Craftsman brand products (a line of tools, equipment, and work wear) while Home Depot does not. If a Home Depot customer searches for *craftsman drill*, the Home Depot search engine can identify drills from comparable brands and surface them to the customer, ideally with an indication of why they are a suitable substitute. This type of search page blurs the boundaries between search and recommendations [Rosset et al., 2020], while raising the need for explainability (Section 3.3). Realization of such systems requires two key ingredients: (1) access to knowledge outside of what is available in the catalog and (2) access to the global state of customer journey, namely, where the customer has been before and their

interactions. With this information a system can understand products and pricing of competitors and trigger changes in ranking, recommendations, and pricing based on the customer's journey. Although ingesting knowledge outside of the catalog is relatively easy, tracking customer visits and behavior across sites is technically difficult and must comply with privacy guidelines (Section 3.3).

Simple and complex goals. A simple goal is when a customer is interested in one item, for example, to buy an espresso machine. A complex goal is more abstract and involves completing a set of simple goals. For example, the complex goal *child's birthday party* involves completing a series of simple goals such as buying party favors, decorations like balloons and streamers, and themed tableware. Like simple goals, complex goals have a customer journey, which can be seen as the sum of customer journeys from each of the simple goals. As such, individual journeys can be intertwined and affect the final outcome of the individual and combined journeys. Most current eCommerce systems are not designed for fulfilling complex task-based shopping needs, instead requiring customers to search separately for each item they need. Systems that aim at addressing customer journeys with complex goals will need to identify the individual simple goals and familiarize the customer with the domain of each simple goal. Conversational search is taking steps towards this direction (Section 3.2).

Customers set out to realize a goal, either simple or complex. They go into a journey that spans online and offline stores and can take multiple days to complete. Most research focuses on on-site customer journeys, and in particular, on understanding and modeling how customers complete single goals within a customer journey. Understanding the customer's informational needs as their journey unfolds is important for search and recommendation relevance (Section 3.1). Research on conversational search (Section 3.2) widens this window of research from simple goals to on-site customer journeys.

2.2 Business

Customer satisfaction is important for business, but is only one of the many criteria that a business needs to track towards the goal of optimizing profit. Below we discuss strategies for achieving this goal as they relate to search and recommendations and we describe constraints that affect the design of such systems.

Sales strategies and short- and long-term effects. Profit optimization comes from several strategies which have short-term and long-term effects. Typical strategies include cross-selling (enticing customers to buy additional products), up-selling (tempting customers to buy a more profitable version of a product) and down-selling (encouraging customers to buy by matching their budget). How each strategy is applied balances short- and long-term effects on profit. For example, a business may use an aggressive down-selling strategy to sell off inexpensive inventory that it no longer wants. However, exposure to this cheaper, lower-quality inventory may make customers not to return in the future, which has negative long-term effects on profit. One way to cross-sell is to backfill search results with recommendations that are related to the search results. This results in closely related items that can be labeled as being recommendations, providing

transparency to the customer. Backfilling is especially useful when there are only a few search results for the customer's query.

Brand image and inventory. A business's brand image influences its inventory. Some businesses opt for higher end products while others offer less expensive ones. Showing customers products that align with the business' brand image is important for consistency. To this end, search and recommendation engines become marketing tools which frame the catalog according to the chosen brand image. Sorokina and Cantu-Paz [2016] give an example for Amazon, where for the popular query *diamond ring* the business wants to rank genuine diamond rings first; however, without careful sampling of training data, the ranking system can show cheap zirconium rings instead and make the Fashion Department look like a flea market. The design and implementation of search and recommendation systems should work together with the business's marketing department to align on brand image concerns.

Online marketing and ranking. eCommerce search engines include business logic that reflects marketing decisions. This business logic, for example, boosts all red products for Valentine's Day, shows chocolate bunnies for Easter regardless of the query, and ranks advertisements above organic search results. Such business decisions may come and go abruptly and clash with the working assumptions of search and recommendation systems that want to be in full control of the results, especially those based on Learning to Rank (Section 3.1.2). Designing systems and user interfaces while taking into consideration these scenarios can help the overall brand image and amortize risks that online marketing may take. Continuing the Valentine's day example, for the query *yellow shoes*, showing red shoes as the first result following the directive from marketing but also showing yellow shoes high in the ranking allows the customer to select what they want.

Offline marketing and ranking. eCommerce businesses having both online and physical presences creates a unique blend of organizational and infrastructure challenges. The search and marketing teams should work together so that the customer experience transitions smoothly from offline to online and vice versa. Products advertised on billboards should appear prominently in search results, in recommendations, and on the initial landing page. Without considering both worlds in tandem, it is difficult to have a complete picture of how well the systems perform.

Regulatory and business restrictions. Regulatory and business constraints govern which products can be shown to which customers. Some of these depend on location: for example, hunting knives cannot be sold to customers in the UK and alcohol cannot be sold online in certain locations in the US. Others are at the customer level, for example, only adults can view or buy certain products. Although most eCommerce sites have business logic at the time of checkout to determine whether a product can be purchased and shipped to a given customer, showing products in the search results that cannot be purchased by the customer provides a frustrating experience and in some cases, even viewing such products is prohibited. At the business level, competing brands can have agreements with an online retailer to restrict showing their products with those of their competitors, such as requiring them to be two result pages apart. It is important to understand what customers are searching for and what they are not finding. Understanding these inventory gaps and replacing them with alternatives or notifying the customer informs the

customer's journey. The integration of commerce-specific knowledge graphs with search is one way to tackle this issue. Another, complementary way is to incorporate these constraints in the optimization problem that the search and recommendation engines aim to solve (Section 3.1.2).

eCommerce search uses several strategies to optimize profit while taking into consideration a wide spectrum of constraints. Systems need to find the best strategy at any point in the customer journey to decide how to balance between short- and long-term effects. Marketing and search need to work together to ensure that brand image is consistent over domains, queries, and products in offline and online settings.

2.3 Data logistics

Data plays a key role in product search and recommendations. Services where the eCommerce website has multiple vendors bring in dynamics with regards to quality and consistency of the content, fraud detection, and pricing. Inventory updates often have service-level agreements that ensure customer-facing changes within strict time frames. Finally, the products themselves are multi-modal documents and the success of a search depends on matching customer intent to all aspects of the document.

Third party content. Some eCommerce sites such as Amazon, Taobao, and eBay serve as a place for other companies to sell products. The products from these companies are referred to as third party content. There may be specific infrastructure requirements for this type of content. For example, there may be contractual agreements as to how quickly the products become searchable and how quickly changes to pricing and availability are reflected. The data for the third party products may need to be reformatted or supplemented before indexing. For example, if the brand of the product is not provided as structured data by the vendor, it may be possible to extract it from the product title. Fraud detection is another aspect of such services. Vendors might engage in illegitimate behavior to abuse the customers, and the service has to rapidly detect and prevent this behavior. Detecting fraudulent cases early on is key for ensuring long-term customer retention.

Volatile inventory. One of the biggest challenges of eCommerce search and recommendation is that the inventory is constantly changing. Prime examples are eBay and other similar businesses with large but ephemeral or highly evolving inventory [Trotman et al., 2017]. New items need to be added quickly, especially for in-demand products like new music, book and film releases. Products change in price and availability very frequently and the availability issue is exacerbated for eCommerce sites that provide information about local in-store availability, where the inventory systems of the stores and search must be synced in real-time. Query suggestions are also affected by volatile inventory as they may suggest queries that no longer return results, creating a frustrating user experience.

Multi-modal documents. In eCommerce search, the indexed documents, i.e., the products customers are looking for, are combinations of images, unstructured text such as titles, descriptions, and reviews, and structured data such as price, brand, ratings, and seller location [Trotman

et al., 2017]. This rich combination of data raises interesting research opportunities, including improving document data extraction by using signals from different types of data such as providing more detailed color information for clothing and using image similarity for recommendations and as a way for customers to query the search engine.

Data is key for providing a high quality search and recommendation experience that meets customer and business needs. eCommerce data is generally multi-modal and provides rich signals. However, its quality depends on the merchant and aspects of the data change frequently, especially for businesses with both online and physical stores.

3 eCommerce research area deep dives

Above we introduced the three key ingredients of eCommerce search and recommendations: customers, business, and data. Each of these areas offers a rich collection of research areas. In this section, we consider three areas in depth. The first is the design of matching and ranking for eCommerce search. Matching and ranking are at the heart of search systems and information retrieval (IR) research, and eCommerce provides some unique challenges and opportunities. Second, the customer journey underlying eCommerce is complex and only beginning to be understood. We provide a deep dive into conversational eCommerce that has promise to enable the smooth shopping experience provided by expert shop assistants. Third, we discuss issues of fairness, confidentiality and transparency which are at the heart of maintaining customer trust while providing personalized eCommerce experiences.

3.1 Relevance: Matching and ranking

To better understand ranking in eCommerce, we need to look at how it has evolved. eCommerce appeared with web search in the early 2000s. Initially, there was not much search in eCommerce. Product catalogs were organized by a taxonomy which customers navigated by clicking. At the taxonomy leaves, products from that leaf were shown. Sometimes there were too many products and so sorting was introduced to quickly jump to the top and bottom of the list based on product attributes such as title or price. These sorts were very similar to those in databases. Sorts evolved, becoming more complex and more useful to the customer. Some sorts, in particular popularity or trending, have become the norm. As catalogs grew larger, a natural next step to exploring the products within a category was to allow filtering, including limiting the products in a category given a search term. This resulted in the search box appearing on eCommerce sites. Search, however, was merely about finding exact matches of the query term against the product title. In the 2010s, eCommerce sites began adopting search the way we define it in IR. The use of open source search engines in popular eCommerce suites such as Magento and the ability of third-parties to add functionality to these suites propelled the adoption of search and recommendation engines in today's websites. This history makes it easier to understand why eCommerce search is not as thoroughly researched as web search. Here we touch on matching and ranking topics that we believe will be interesting research topics in the coming years.

3.1.1 Matching

Different types of customer intent yield different types of queries. These query types range from short navigational ones (a serial number) to long informational ones (*are batteries included with this watch*). Each query type poses idiosyncrasies for matching. Navigational queries need exact matches to product serial numbers, product titles or category names, while informational queries need semantic parsing and more elaborate indexing before they can be answered. Some queries may require a different user interface; for example a tabular layout is better for answering comparison queries (*differences between iPhone 11 and iPhone SE*). Relevance is inextricably linked to retrieval as it dictates what should be retrieved. First we discuss the peculiarities of relevance in eCommerce, then query and document modeling, and finally query understanding, also referred to as query intent engines.

Relevance. Relevance has been a long-standing puzzle for IR researchers [Saracevic, 1975]. At first, relevance was considered a universal, dimensionless quantity that described a document. Later, relevance was thought to be universal but query dependent. TREC benchmarks are based on this assumption. Then, relevance moved from binary grades to richer scales. Finally, relevance was determined not to be universal but instead user dependent: what is relevant for one person is not necessarily relevant for another. Customer journeys (Section 2.1) add a temporal dimension to relevance [Radinsky et al., 2013]. A student looking for a bed is likely to buy a single bed but a few years later she may opt for a double bed. Seasonality also reflects relevance's dependence on time. The query *jacket* has a different set of relevant documents in winter than summer. Context in eCommerce is influenced by what category the customer is in. In some categories new items should rank higher, while in others they should not. In sum, eCommerce relevance is context-dependent and has four dimensions: (1) user, (2) time, (3) query, (4) context (e.g., category). Understanding these dimensions can help us better understand customer behavior and design systems that can take these dimensions into account.

Matching queries and products. eCommerce search is as much about exploration as it is about finding the best exact match [Sondhi et al., 2018, Carmel et al., 2020a], and different types of query require different matching strategies. Finding exact matches needs strict matching, while broad queries that correspond to categories (e.g., *shoes*, *kitchen tables*; see [Rowley, 2000]) may need careful crafting of synonyms to match a customer's vocabulary to that of the business. As with all types of search, tokenization, including word breaking, decompounding, and punctuation handling, lemmatization or stemming, and stopword identification are important for identifying relevant products, especially when handling multi- or cross-lingual search, which occurs frequently in eCommerce. Attempts to encode all query types into a single matching function have met with limited success. In research, although exact matches are considered solved, bridging the vocabulary gap between queries and documents and query understanding more generally is a long-standing and active research topic [Van Gysel, 2017].

Query understanding methods include pseudo-relevance feedback [Lee and Chau, 2011], query-click graphs [Cao et al., 2008, Santos et al., 2010, Adhikari et al., 2018], word embeddings [Kuzi et al., 2016], and multi-modal methods that combine text and visual cues [van den Akker et al., 2019, Lin et al., 2019]. Query intent engines are an increasingly used component in pre-matching [Li

et al., 2008, Ahmadvand et al., 2020]. Query intent engines parse the query to extract catalog specific attributes for driving matching and ranking. Their goal is to return a set of boolean, fielded queries given a free text query, mapping from unstructured to structured text. For example, for the query *red sneakers*, a query intent engine will return fielded queries such as COLOR:RED and SHOE_TYPE:RUNNING and CATEGORY:SHOES. The spectrum of capabilities of query understanding engines ranges from simple matching of query terms to a predefined set of product attributes to more elaborate semantic methods [Wang et al., 2015, Baeza-Yates, 2017, Dai and Callan, 2019]. The ultimate goal of a query intent engine is to return structured, personalized queries for all customer queries.

In eCommerce, query understanding has to adapt to the behavior of a business's customers, products and services. The level of difficulty in addressing this depends on dimensions such as customer demographics, niche or not products and services, and the volume of available data. Mappings between generic and customer-specific vocabulary and what the business offers must be based on business-specific data. Small businesses targeting niche areas are likely to have access to less data and therefore have less effective query understanding methods, despite the fact that they need them the most for customers who are unfamiliar with their domain-specific vocabulary. More generic stores attract a more diverse audience that potentially uses a much wider variety in vocabulary, increasing the need for more effective query understanding methods.

3.1.2 Ranking

How to rank the results shown to customers is one of the most complex issues in eCommerce. Ranking in eCommerce started from sorting on title or price, and it was later extended with basic boolean retrieval. Practitioners have put effort into deriving a single ranking function that mixes boolean or tf.idf-based ranking algorithms with other signals, such as recency or popularity. This has had mixed results because some signals work for some queries but not for others. For example, due to collection-specific reasons, the query *striped t-shirts* may rank highly striped products other than t-shirts because "striped" has a substantially higher idf than "t-shirts." As the number of signals increases, it becomes impossible to derive a ranking function that returns optimal results over the entire spectrum of queries, customer intents, and business goals. In recent years, Learning to Rank has been increasingly used to solve such complex optimization problems, either in a pointwise fashion (by minimizing the loss with respect to gold standard scores), in a pairwise fashion (by minimizing the loss incurred by having out-of-order pairs of items), or in a listwise fashion (by minimizing a listwise loss function defined on the predicted list and the ground truth list). Below we discuss ranking signals, designing multi-objective functions, feedback loops, and practical limitations of Learning to Rank.

Extending the product representation. Queries and documents can be represented over the terms in the vocabulary or over a latent embedding space. These representations must be shared between queries and documents since they are the basis of matching. Documents have many features beyond how closely they match the query terms, such as how many times they have been purchased, how many times they have been clicked, and the ratio of clicks versus purchases. Systems can extend the query-document-dependent representation with a document-dependent representation that includes each of these document-specific features. The representation space

can further be expanded to capture higher-order relations, for example, the rank of a document within a category combined with the similarity score of a query and the document given one or multiple similarity functions. Creating a wide and rich representation space for documents allows Learning to Rank systems to learn more complex definitions of objective functions.

Ranking signals and optimization criteria. eCommerce search and recommendation systems must optimize for multiple criteria (also known as objective functions) simultaneously [Carmel et al., 2020b]: one encoding customer preferences and one encoding business preferences. These objective functions compete against each other, raising research questions as to the best learning strategy for the short- and long-term. Before designing optimization functions that encode both customer satisfaction and business success, we look into signals that capture the goals of each stakeholder individually.

Customer satisfaction is measured over multiple signals [Hong and Lalmas, 2019] including click-through rate, hover and dwell time, satisfied clicks, query reformulations, session length, number of queries before checkout, add-to-baskets, purchases, time-to-next-visit, product returns, and calls to customer service. Business success is measured over several key performance indicators (KPI) including inventory-oriented measures (such as average turn-around time, number of unique products sold per time period), revenue-oriented measures, profit-oriented measures, visitor-oriented measures (such as total and unique number of visitors, number of new and returning visitors), and basket-oriented measures (such as average number of items per basket, basket average value). Each signal can translate to an objective function and a linear interpolation of them can define a global objective function. Given their short- and long-term goals and marketing decisions, businesses may want to weigh individual objective functions or to learn a meta objective function that optimizes the combination of objective functions.

Not all signals are equal. Objective functions over multiple signals can bias towards more abundant signals. By nature, some signals are more abundant than others, and some are stronger indicators of preference than others. For example, a purchase is a more explicit preference indicator than a click but it is much less frequent. A purchase that was not returned is a stronger signal than a purchase but again is less frequent. The stronger the preference indicator, the lower the volume. Objective functions should take into account this difference in signal strength versus signal abundance: otherwise stronger, but less frequent signals may be lost in the sheer volume of weaker signals. This is not an issue when signals correlate positively but is when they do not correlate or correlate negatively, e.g., people click on expensive products but purchase affordable ones. Normalizing by the volume of each signal is one way [Tsagkias et al., 2010] to counter this; another is to consider individual objective functions per signal and then ensure they are represented adequately at interpolation time.

Positive, negative, and delayed feedback loops. Measuring the success of the system and informing the system about these measurements creates a feedback loop. This is the reinforcement learning paradigm, in which the system takes an action and receives feedback to update its internal state towards maximizing (or minimizing) an objective function [Hu et al., 2018]. Some criteria can be measured immediately after the search results are displayed while others take longer. Clicks,

reformulations, or abandonments define a rapid feedback loop that spans from a few seconds to less than an hour [Su et al., 2018].

In eCommerce there are longer feedback loops where the feedback occurs well after the system has shown results to the user. Examples include adding products to the basket and checking it out at a later date, adding products to a wish list and checking them out at specific events (e.g., Christmas, birthdays), or buying several sizes of an item to return all but the one that fits. Removing items from the basket or a wish list and product returns are interesting types of delayed feedback because the delay can be weeks. These delays make tracking and attaching feedback to a specific ranking challenging because it requires persisting records of the rankings for long periods of time. It is an open question how to engineer systems that capture delayed feedback loops in a technically feasible way. Finally, good priors for feature values and business logic are needed to overcome issues with the lack of behavioral data for new items, referred to as cold start.

Feedback of the types listed above can be used to Learning to Match (LtM) [Van Gysel et al., 2016] to optimize individual rankers or Learning to Rank (LtR) to optimize combinations of individual rankers [Karmaker Santu et al., 2017]. Increasingly, the community is considering the use of historical feedback data for this purpose to avoid interventions that may negatively impact the customer experience [Jagerman et al., 2019]. How to make effective use of the broad diversity of feedback signals with widely differing timescales for counterfactual LtR remains a matter of research.

Practical limitations of Learning to Rank. Most eCommerce search engines based on LtR work in two steps: a recall-oriented step and a precision-oriented step. In the recall-oriented step, the system issues queries to several rankers (generally simpler IR systems that return a small set of results), pools the results and extracts features for each document. In the precision-oriented step, a re-ranker applies the currently learned weights to the extracted features, scores each document, and ranks them. The re-ranked list is shown to the user; if there are more results, they are ranked by a simpler ranker and shown after the LtR results. This implementation of LtR has proven to be effective in terms of IR and business metrics and the principle of re-ranking remains effective across tasks regardless of the algorithm for learning the weights, the feedback loop, or the features [Carmel et al., 2017, Dai et al., 2011, Karmaker Santu et al., 2017].

LtR depends on the effectiveness of the individual rankers. Designing effective rankers is challenging in eCommerce because broad exploratory queries, like *shoes*, can have multiple aspects (e.g., type, brand, color, price). An ideal exploratory ranking should include representative samples for each aspect. To get documents for each aspect we either need one ranker per query aspect or rankers that determine query aspects and account for this diversification internally. In practice, systems instruct a ranker to boost a particular product attribute according to predictions derived from a query intent engine that infers the importance of a product attribute for a query.

Next, consider LtR's issue with the discontinuity in usefulness of results. For efficiency reasons each ranker returns a limited set of results but a broad query can yield thousands of results, leaving the majority out of the candidate documents. This remainder appears after the re-ranked results. Consider a product inventory with thousands of products, hundreds of TV accessories, and 100 TVs. This inventory is accessed with two rankers that return 50 results each. A customer issues the query *tv*. The two rankers return a largely overlapping set of TVs, yielding 75 unique results, 50 of which are TVs. LtR ranks the products so that the 50 TVs show at the top of the 75

candidates. The customer browses the page full of 50 TVs and then sees 25 results which are not TVs followed by results from the naive ranker. If the naive ranker ranks TV accessories first, then the customer may stop examining the results, although there are 50 TVs further down. This is not surprising. Customers think that when there is a shift from useful to non-useful results it signals the end of useful results and so stop examining the list [Robertson, 1977]. However, because we have two results lists, two shifts can exist, but the customer sees only one result list and expects one shift. One way to tackle this issue is to increase the number of documents returned by each ranker which increases response time. It is an open question how to operationalize re-ranking for larger parts of the catalog. Recent research in neural IR [Mitra and Craswell, 2018] addresses such scenarios where the re-ranker is applied to all documents in the collection but has a more opaque model that is harder to understand and optimize.

Although much studied within industry, relevance for eCommerce, both matching and ranking, is a rapidly evolving field with many open research areas. The next section discusses conversational eCommerce which uses multi-turn interactions with the customer to hone their search results and help them through the customer journey.

3.2 Conversational eCommerce

Conversational eCommerce engages the customer in a dialog, by voice or by chat, to elicit the information needed to find or recommend a product (or a set of products) or to answer questions about a product or an order. Regardless of whether the interaction is by voice or by chat, the dialog must be succinct, natural and informative. It must be able to determine when a customer is being facetious or becoming frustrated. Conversational system applications include product search and recommendation, product question answering, customer care, and notifications (such as order tracking and delivery notifications, alerts for product availability and discounts). Conversational eCommerce is a relatively new area, driven by the development of voice assistants. In this section we focus on challenges presented by conversational systems.

The benefit of having a conversational system is that it engages the customer in a "natural" interaction, which is intended to build trust and to elicit additional information to satisfy the customer need. While customers clearly value systems that produce reliable results, if the conversation is not natural (that is, if the system misunderstands the customer or is too verbose) the customer will not engage. Beyond being natural, people prefer systems that are "attractive", trust their results more, and find them subjectively more accurate [Yuksel et al., 2017]. This suggests that the manner in which the information is delivered by the conversational agent is at least as important as the accuracy of the results.

eCommerce systems have additional challenges in understanding customer intent. The system must disambiguate the overall intent of the request (such as, a shopping request, a customer-care request, or a question about a product or an order). If it is a shopping request, the product may be ambiguous. For example, if a customer asks *where can I get coffee* are they trying to find whole bean or ground coffee? Or perhaps they want to find a cup of coffee. Depending on the context of the customer, they may be asking about the availability of a product (*where can I get toilet paper*) and would be happy with an online purchase, or they may intend to go to a physical store (*where can I get an omelet*).

In addition to understanding and disambiguating the customer request, the system should present the results in a device-appropriate way, recognizing that many customers use mobile phones or bespoke voice assistants for eCommerce. If the results are presented by voice, the system should emphasize precision because of the difficulty of browsing a results list. Beyond getting the right result at rank one, each result presented should be summarized in a voice-friendly way and not be overly verbose.

Conversing with the customer. As stated above, one purpose of a conversational system is to elicit information from the customer to improve the results, in a sense to mimic the patter of a human shopping assistant. A key difference between a human assistant and a digital assistant is that human interactions are less efficient because of our need to be pro-social. A digital assistant asking irrelevant questions while helping a customer would be seen as incompetent. For this reason, any follow-on questions asked by the system must be clearly related to satisfying the customer need.

An early conversational question-answering system for customers ("How May I Help You") was developed by AT&T in the late 1990s [Goran et al., 1997, 2002]. It had a broad vocabulary to handle the most common customer needs and was able to discern when a customer was becoming frustrated [Walker et al., 2000]. It classified customer questions into a small number of intents, and had a pre-determined template-based dialog subsystem for each intent. In modern systems, follow-on question templates are generated automatically. Zamani et al. analyze query reformulations to generate question templates, and then use a slot-filling algorithm to suggest facets to refine broad queries [Zamani et al., 2020, Hashemi et al., 2020]. In their system, the facets suggested to the users were clickable to present results for the corresponding search reformulation. Christakopoulou et al. [2016] present a bandit-based approach for modeling recommendation dialogs that mimic human recommender dialogs in that each turn seeks to elicit information that informs the next turn.

In a product search setting, the follow-on question formulation and the facets would be informed by product attributes and user-generated content such as reviews and product question-answers. If the system is voice-driven, the response to the follow-on question must be parsed for additional information. Natural language answers may also provide more context than a query reformulation facet, and it is an open question how to leverage this context to improve the interaction. For example, if the system offers the customer a disambiguation *Do you want coffee beans or ground coffee?* the customer may respond with *ground, but it's for a Chemex*, and the system should be expected to know what *for a Chemex* entails.

Much of the research on dialog systems over the past 25 years measures how long interactions last, in terms of the number of turns. Often, the conversations are occurring between people rather than between a person and a machine. These studies focus on various information seeking tasks not related to eCommerce [Trippas et al., 2017, Raux et al., 2005] and seek to characterize the nature of the interaction [Qu et al., 2019]. The optimal length of a customer's interaction with an eCommerce system for different tasks is yet to be determined.

Learning from data. As with standard eCommerce search, conversational systems are informed by the product catalog information. However, many of the attributes important to customers are not listed in the product catalog [Moraes et al., 2020]. For example, a catalog may list a color or

a material, but will not list details about the product's usability, durability or compatibility with other products. If a customer asks about such details, the conversational agent may have to depend on user-generated content such as customer reviews and product Q&A. Because eCommerce systems rely heavily on user trust, it is crucial to filter offensive content to avoid injecting it into the conversational system's models. Microsoft released a chatbot in 2016 designed to learn from user interactions on Twitter[1] but had to take it down within 16 hours because adversarial users induced the bot to generate hate speech and other offensive content.[2] Beyond being embarrassing, this would threaten the core business of an eCommerce company.

Result presentation and navigation. When presenting a search result or product summary by voice, the result must be transformed to be voice-friendly. Voice systems may present multimodal results, combining a spoken summary with a display of information. For example, in product search, they may present product information for the top three results while speaking a summary of the top product. There are many ways that information might be presented on a web site or in a chat that would be intolerable with voice. For example, a customer might prefer to see a faceted recommendation in the form of a grid on a web page but a list of options in a chat. It would be onerous and confusing to hear all of the options read out loud; so voice systems typically present one option at a time, with a *hint* to prompt the customer to request the next item. Hints are typically generic such as *You can say 'Next Item' to hear the next result.* It would be more helpful to offer hints that indicate the types of items the system can find. This would produce a guided conversational search centered around the customer need, rather than around the system catalog.

Spina et al. [2017] investigate audio-only summaries of podcasts, addressing some key questions about how users perceive audio summaries. In their work, the documents to be summarized are audio podcasts. In eCommerce, the information to be summarized may be heterogeneous and needs to be transformed to be more succinct with fewer parentheticals omitting product details such as model numbers.

Voice in multi-user environments. Voice systems are often shared by multiple users in the same household and sit in common areas. The assistant must recognize sensitive content that should not be spoken out loud. For example in delivery notifications, the content of the delivery should not be revealed if it would cause embarrassment to the customer or ruin a surprise. What is considered embarrassing differs from person to person. For example, people have different feelings about beauty products such as hair dye and health products such as medicines or ointments.

Since multiple people are often using the same voice device, the system may benefit from being able to identify different speakers. This will help addressing "friendly fraud" where someone not authorized to make an order on an account, such as a child, does so, resulting in a product return. Identifying multiple users also would aid in tailoring the request understanding and result presentation to a specific customer. This would save the customer from having to specify details or reformulate requests to get the most appropriate results.

[1] https://twitter.com/TayandYou, visited May 2020.

[2] https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/, visited May 2020.

eCommerce is moving beyond single-turn search and browse to include conversations that engage the customer in a dialog to elicit the information needed to find a product or to answer questions. Enabling conversational eCommerce depends on a broad range of research areas at the intersection of IR, human-computer interaction, and natural language processing.

3.3 Fairness, accountability, confidentiality, and transparency

Search engines and recommender systems have developed into powerful and ubiquitous tools, especially in eCommerce. Unintended negative side effects of search and recommendation technology are gaining attention [Roegiest et al., 2019]. In the wake of Europe's GDPR legislation, the unforeseen consequences of data-driven technologies have also received political attention [Cramer et al., 2018]. While discussions about side effects of data-driven technologies often focus on fairness and bias in North America, in Europe they often focus on issues such as privacy, interpretability and explainability. Creating search and recommendation systems that provide fairness, accountability, confidentiality, and transparency and are perceived as such is particularly important in eCommerce since customers, both consumers and suppliers, will move to competitors' sites if they feel that a particular site is unsafe or unfair.

Fairness and bias. Recent work on fairness and bias has contributed advances in fair rankings [Singh and Joachims, 2018], predictive models for individual fairness [Kowald et al., 2020], and mitigation mechanisms for different types of individual bias [Oosterhuis et al., 2020]. In the context of eCommerce, fairness refers to both the customer side and the supply side. The interplay between supplier fairness and customer satisfaction in a two-sided marketplace has only recently been addressed [Tadelis, 2016, Mehrotra et al., 2018], with a large number of foundational issues, concerning definitions, metrics, and optimization, still to be addressed. Reputation systems could promote fair transactions, helping consumers and suppliers distinguish transaction quality [Resnick and Zeckhauser, 2001]. However, reputation systems may allow for unfair treatment through inaccurate reviews or biased behavior based on the information provided [Fradkin et al., 2015]. Automatically calibrating reputation ratings and assessing review quality remains an open research question.

Accountability. Accountability is the assumption of accepting responsibility for actions and decisions [Lepri et al., 2018]. Consider query autocompletion [Cai and de Rijke, 2016]. While the purpose of autocomplete algorithms is to steer inquiry, examples of socially unwanted steering have surfaced. Robertson et al. [2019] propose a method for auditing autocomplete methods by recursively submitting a root query and its child suggestions to recover a network of algorithmic associations. Interestingly, the associations generated by autocomplete methods in eCommerce are different from those generated in web search. Returning the most frequent completion is a powerful baseline that is used in virtually all autocompletion setups. However, in eCommerce, catalog-based suggestions are often included to ensure more relevant and applicable suggestions, resulting in different association types. Auditing as proposed by Robertson et al. can support accountability, but it does so by dealing with algorithmic decision processes as a black box, whose inputs and outputs are visible, while their inner workings are not.

Transparency. Transparency, understood as openness and communication of both the data being analyzed and the mechanisms underlying the models, is a key enabler of accountability. Explanations can increase transparency and support innovation and technological development while keeping the human "in the loop." When a system is too good in understanding users and offering relevant products, it might unnerve people as they find it eerie that a system is that good. Conversely, algorithmic mistakes decrease confidence in the model and hence have a significant impact on customers' perception of the model and the product using it. By being offered an option to receive an explanation of a search result or recommendation, eCommerce customers may gain trust and increase their engagement [Ter Hoeve et al., 2017]. However, explanations of mistakes may trigger "algorithm aversion," where customers prefer to disengage from algorithmic processes, even when they yield more satisfactory results [Lucic et al., 2019]. Algorithm aversion can be overcome by giving users the option to intervene and alter the outcomes. The problem of finding counterfactual explanations, that is the minimal perturbation to an instance such that an algorithmic outcome changes, can be framed as an optimization task.

To obtain satisfactory results in an eCommerce setting, customers may have to reveal information that they may consider private, such as their identity, demographic characteristics, previous purchases or interaction behavior. It is important to investigate how customers can share private information in a secure manner and determine what it is they receive in return, in terms of customization and improved service [Aïmeur et al., 2008].

Research challenges concerning fairness in eCommerce in the broad sense advocated here start with definitions and metrics of fairness, accountability, confidentiality, and transparency for different consumer segments and locales. We then need to establish reliable offline predictors for these dimensions that can inform us without running online experiments that affect customers and suppliers on the very dimensions where we seek to protect them. Finally, we need to develop search and recommendation methods that allow us to jointly optimize for customer satisfaction, supplier success, fairness, accountability, confidentiality, and transparency.

4 Datasets for eCommerce research

Offline and online experimentation allow eCommerce to incrementally measure the impact of changes to models and algorithms on customer engagement. Historically, for eCommerce, as well as other types of search and recommendation research, it has been difficult to access such data, especially in academia. Recently more companies are releasing datasets for collaborating on eCommerce problems across domains ranging from fashion and movies to second hand auctions and paper dolls. Having companies collaborating with external teams on research projects reflects the challenging nature of eCommerce search and recommendations.

Table 1 lists 28 eCommerce IR datasets from 2007 until mid-2020.[3] Encouragingly, the number of datasets released between 2016 and 2020 is more than double the number of datasets published between 2007 and 2014. These datasets reflect the intensifying research activity around eCommerce and the diverse set of user interactions, data types and research opportunities present in this field. Many of these datasets focus on product catalogs and taxonomies. Others involve reviews

[3] An online version of this list is at https://sigir-ecom.github.io/ecomDatasets.html (visited May 2020).

and recommendations. Relatively few provide data at scale for search relevance, either matching or ranking.

Along with data resources, there are also other efforts to attract researchers to eCommerce IR. Yearly eCommerce workshops hosted by SIGIR started in 2017 [Degenhardt et al., 2017] with 60 participants and followed in 2018 [Degenhardt et al., 2018] and 2019 [Degenhardt et al., 2019] with more than double the participants; the 2020 online version[4] is expected to continue this trend. Each year, the workshop partners with an eCommerce company on a data challenge for researchers and practitioners. Additional efforts exist on Kaggle, the data challenges platform, but are more spurious.

With most data challenges and evaluation systems such as Kaggle and TREC, the participants do not evaluate on real users performing complex tasks. To capture and understand the four dimensions of relevance and the complexity of a customer's journey (Section 3.1.1), we need to move beyond static data and to bring our experimentation instrumentation closer to the customer, while keeping the customer and business safe (Section 3.3). A recent initiative, LivingLabs, proposes a new experimentation paradigm allowing live experimentation on real customers. The Living Labs for Information Retrieval Evaluation (LL4IR) held as part of CLEF 2015 presented a unique collaboration between industry and academia [Schuth et al., 2015, Balog et al., 2014]. LL4IR released a benchmarking dataset contributed by REGIO Játék and a platform for evaluating models on a live system with real users. The idea was later also adopted by TREC OpenSearch to facilitate search research in academia [Jagerman et al., 2018]. This also fits naturally with evaluating reinforcement learning approaches, where the models are not trained on a static dataset but are constantly improving from user interactions and feedback. This new experimentation paradigm opens interesting ways for academic researchers to explore aspects of an eCommerce domain with the customer in mind.

[4] https://sigir-ecom.github.io/ecom2020/

Table 1: A list of 28 datasets from 2007 until mid-2020 on eCommerce search and recommendation. The wide array of tasks and data types indicates the diversity and richness of the research challenges in the eCommerce domain.

Year	Dataset name and link	Data type	Task
2020	Rakuten multi modal taxonomy dataset https://sigir-ecom.github.io/data-task.html	Catalog	Taxonomy classification Multimodal retrieval
2020	Semantic web embedded product data https://ir-ischool-uos.github.io/mwpd/index.html	Catalog	Product matching Product classification
2020	Alibaba eCommerce de-biasing dataset https://tianchi.aliyun.com/competition/entrance/231785/introduction	Click logs	Fairness in exposure
2020	Alibaba multimodal recall https://tianchi.aliyun.com/competition/entrance/231786/introduction	Catalog	Multimodal retrieval
2019	Alibaba user behavior dataset http://yongfeng.me/dataset/	Click logs	Recommendations
2019	DeepFashion https://github.com/switchablenorms/DeepFashion2	Images	Image segmentation Multimodal retrieval
2019	eBay high accuracy recall task https://sigir-ecom.github.io/ecom2019/data-task.html	Search relevance	Product recall
2019	eBay ML data https://evalai.cloudcv.org/web/challenges/challenge-page/462	Catalog	Product level equivalence
2018	Alibaba conversational search and recommendation dataset http://yongfeng.me/dataset/	Conversations	Conversational search and recommendations
2018	Olist Kaggle dataset https://www.kaggle.com/olistbr/brazilian-ecommerce	Catalog Transactions	Catalog mining Review mining
2018	Rakuten taxonomy dataset https://sigir-ecom.github.io/ecom2018/data-task.html	Catalog	Taxonomy classification
2017	Amazon baby registry http://yongfeng.me/dataset/	Wishlists	Recommendations
2017	Fashion MNIST – Zalando https://github.com/zalandoresearch/fashion-mnist	Images	Product classification
2017	Flipkart product catalog data https://www.kaggle.com/PromptCloudHQ/flipkart-products	Catalog	Classification
2017	Innerwear dataset https://www.kaggle.com/PromptCloudHQ/innerwear-data-from-victorias-secret-and-others	Catalog	Catalog mining
2017	Lazada product title quality https://competitions.codalab.org/competitions/16652	Catalog	Product title grading
2017	Retail rocket dataset https://www.kaggle.com/retailrocket/ecommerce-dataset	User interactions	Recommendations
2016	Homedepot product search relevance https://www.kaggle.com/c/home-depot-product-search-relevance	Search relevance	Search ranking
2016	Paper doll dataset https://github.com/kyamagu/paperdoll	Images	Image segmentation Vision
2016	Personalized eCommerce search challenge https://competitions.codalab.org/competitions/11161	Search logs Transactions	Search personalization
2015	JD.com reviews and rating dataset http://yongfeng.me/dataset/	Catalog Ratings and reviews	Recommendations
2015	Online retail dataset of transactions http://archive.ics.uci.edu/ml/datasets/Online+Retail	Transactions	Classification, clustering Customer segmentation
2015	Regio - Living Labs dataset https://bitbucket.org/living-labs/ll-api/src/master/	Search relevance	Search ranking
2015	Yoochoose dataset https://2015.recsyschallenge.com	Click logs	Recommendations
2014	Amazon product datasets https://cseweb.ucsd.edu/~jmcauley/datasets.html https://snap.stanford.edu/data/#amazon	Catalog	Catalog mining Review mining Recommendations
2012	Best Buy Kaggle dataset https://www.kaggle.com/c/acm-sf-chapter-hackathon-big/data	Search logs	Click prediction
2012	Online auctions dataset http://www.modelingonlineauctions.com/datasets	Auctions	Bid price prediction
2007	Multi-domain sentiment dataset http://www.cs.jhu.edu/~mdredze/datasets/sentiment/	Reviews	Sentiment mining

5 Conclusions and call for research

In the future, online shopping will be automated and personalized, from the initial landing page, to product taxonomies, individual product pages and recommendations. Given a small set of business rules, the system will learn from customer behavior to transform the experience around the needs and style of each customer. Although academic and industrial research in eCommerce has gained traction in recent years, significant research challenges remain. The challenges include classic eCommerce search problems such as matching textual queries to multi-modal documents, ranking optimization for two-sided marketplaces, human-computer interaction, and conversational recommender systems for product discovery and browsing, and automated customer care.

While eCommerce shares many of the core IR problems of other domains, the customer and business goals of eCommerce are unique. Search in eCommerce usually entails matching a customer request to a combination of structured (such as product catalog) and unstructured (such as customer review) data. In eCommerce, search may be part of a larger task and may span multiple online sites (commerce and non-commerce) as well as physical stores. eCommerce introduces complexities for merchandizing, inventory management, and customer care. For sites like Amazon, eBay, and Taobao, the system must be fair for third-party sellers, to maintain their trust as well as the trust of the consumers. Emerging information discovery approaches such as conversational systems for eCommerce introduce a new set of unexplored research opportunities spanning information retrieval, recommender systems, natural language processing, and human-computer interaction.

Academic and industrial research both contribute substantially to the progress in eCommerce research. While industry can focus on large scale data and systems and has access to vast amounts of customer interaction data, it is less well equipped to study human-computer interaction or to take risks in developing novel approaches. Because industrial systems must place customer and seller trust above other considerations, they are necessarily more conservative in experiments. Increasing dataset availability and new experimentation paradigms are opening the way for academic researchers to explore more aspects of eCommerce search and recommendations. We hope that this paper will foster closer collaboration between academic and industrial research, and increase the pace of innovation in eCommerce.

6 Acknowledgements

We thank David Carmel, Parikshit Sondhi, Daniel Tunkelang and Wouter Weerkamp for comments and suggestions.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Jon Degenhardt, Surya Kallumadi, Utkarsh Porwal, and Andrew Trotman. Report on the SIGIR 2019 Workshop on eCommerce (ECOM19). *SIGIR Forum*, 53:11–19, 2019.

U.S. Department of Commerce. Quarterly Retail E-Commerce Sales. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf, 2017. Accessed May-2020.

Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *WSDM*, pages 547–555, 2018.

Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. A Taxonomy of Queries for E-Commerce Search. In *SIGIR*, pages 1245–1248. ACM, 2018.

Jennifer Rowley. Product Search in E-shopping: A Review and Research Propositions. *Journal of Consumer Marketing*, 17:20–35, 2000.

Andrei Broder. A Taxonomy of Web Search. *SIGIR Forum*, 36(2):3—-10, September 2002.

Tom Blake, Chris Nosko, and Steve Tadelis. Returns to Consumer Search: Evidence from eBay. In *Electronic Commerce*, pages 531–545. ACM, 2016.

Corby Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. Leading Conversational Search by Suggesting Useful Questions. In *The Web Conference*, pages 1160–1170, 2020.

Daria Sorokina and Erick Cantu-Paz. Amazon Search: The Joy of Ranking Products. In *SIGIR*, pages 459–460. ACM, 2016.

Andrew Trotman, Jon Degenhardt, and Surya Kallumadi. The Architecture of eBay Search. In *SIGIR Workshop on eCommerce*, 2017.

Tefko Saracevic. RELEVANCE: A Review of and a Framework for the Thinking on the Notion in Information Science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

Kira Radinsky, Krysta M. Svore, Susan T. Dumais, Milad Shokouhi, Jaime Teevan, Alex Bocharov, and Eric Horvitz. Behavioral Dynamics on the Web: Learning, Modeling, and Prediction. *Transactions on Information Systems*, 31(3):Article 16, 2013.

David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. Why Do People Buy Seemingly Irrelevant Items in Voice Product Search? On the Relation between Product Relevance and Customer Satisfaction in eCommerce. In *WSDM*, page 79–87. ACM, 2020a.

Christophe Van Gysel. *Remedies against the Vocabulary Gap*. PhD thesis, University of Amsterdam, 2017.

Alice Lee and Michael Chau. The Impact of Query Suggestion in E-Commerce Websites. In *E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life*, pages 248–254, 2011.

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-Aware Query Suggestion by Mining Click-through and Session Data. In *SIGKDD*, pages 875–883. ACM, 2008.

Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Exploiting Query Reformulations for Web Search Result Diversification. In *WWW*, pages 881–890, 2010.

Bijaya Adhikari, Parikshit Sondhi, Wenke Zhang, Mohit Sharma, and B. Aditya Prakash. Mining E-Commerce Query Relations Using Customer Interaction Networks. In *The Web Conference*, pages 1805–1814. ACM, 2018.

Saar Kuzi, Anna Shtok, and Oren Kurland. Query Expansion Using Word Embeddings. In *CIKM*, pages 1929–1932. ACM, 2016.

Bram van den Akker, Ilya Markov, and Maarten de Rijke. ViTOR: Learning to Rank Webpages Based on Visual Features. In *The Web Conference*, pages 3279–3285. ACM, 2019.

Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Improving Outfit Recommendation with Co-Supervision of Fashion Generation. In *The Web Conference*, pages 1095–1105. ACM, 2019.

Xiao Li, Ye-Yi Wang, and Alex Acero. Learning Query Intent from Regularized Click Graphs. In *SIGIR*, pages 339—-346. ACM, 2008.

Ali Ahmadvand, Surya Kallumadi, Faizan Javed, and Eugene Agichtein. JointMap: Joint Query Intent Understanding for Modeling Intent Hierarchies in E-commerce Search. In *SIGIR*. ACM, 2020.

Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen. Query Understanding through Knowledge-based Conceptualization. In *IJCAI*, pages 3264–3270, 2015.

Ricardo Baeza-Yates. Semantic Query Understanding. In *SIGIR*, page 1357. ACM, 2017.

Zhuyun Dai and Jamie Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR*, pages 985–988. ACM, 2019.

David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *The Web Conference*, pages 373–383, 2020b.

Liangjie Hong and Mounia Lalmas. Tutorial on Online User Engagement: Metrics and Optimization. In *The Web Conference*, pages 1303–1305. ACM, 2019.

Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. News Comments: Exploring, Modeling, and Online Prediction. In *ECIR*, pages 191–203. Springer, 2010.

Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application. In *KDD*, pages 368–377. ACM, 2018.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. Learning Latent Vector Spaces for Product Search. In *CIKM*, pages 165–174. ACM, 2016.

Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. On Application of Learning to Rank for E-Commerce Search. In *SIGIR*, pages 475–484. ACM, 2017.

Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke. To Model or to Intervene: A Comparison of Counterfactual and Online Learning to Rank from User Interactions. In *SIGIR*, pages 15–24. ACM, July 2019.

David Carmel, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. Promoting Relevant Results in Time-Ranked Mail Search. In *WWW*, pages 1551–1559, 2017.

Na Dai, Milad Shokouhi, and Brian D Davison. Learning to Rank for Freshness and Relevance. In *SIGIR*, pages 95–104, 2011.

Stephen E Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33 (4):294–304, 1977.

Bhaskar Mitra and Nick Craswell. An Introduction to Neural Information Retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Transactions on Internet Technology*, 17(1), 2017.

Allen Goran, Giuseppe Riccardi, and Jeremy Wright. How May I Help You. *Speech Communications*, 23(1–2):113–127, 1997.

Allen Goran, Alicia Abella, Tirso Alonso, Giuseppe Riccardi, and Jeremy Wright. Automated Natural Spoken Dialog. *Computer*, 35:51–56, April 2002.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You? In *NAACL*, pages 210–217. ACL, 2000.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating Clarifying Questions for Information Retrieval. In *The Web Conference*, pages 418–428. ACM, 2020.

Helia Hashemi, Hamed Zamani, and W. Bruce Croft. Neural Representation Learning for Clarification in Conversational Search. In *SIGIR*, 2020.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards Conversational Recommender Systems. In *SIGKDD*, pages 815–824. ACM, 2016.

Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. How do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *CHIIR*, pages 325–328. ACM, 2017.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskénazi. Let's Go Public! Taking a Spoken Dialog System to the Real World. In *Ninth European Conference on Speech Communication and Technology*, pages 885–888, 2005.

Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. User Intent Prediction in Information-Seeking Conversations. In *CHIIR*, pages 25–33. ACM, 2019.

Felipe Moraes, Jie Yang, Rongting Zhang, and Vanessa Murdock. The Role of Attributes in Product Quality Comparisons. In *CHIIR*, pages 253–262. ACM, 2020.

Damiano Spina, Johanne Trippas, Lawrence Cavedon, and Mark Sanderson. Extracting Audio Summaries to Support Effective Spoken Document Search. *Journal of the Association for Information Science and Technology*, 68(9):2101–2115, 2017.

Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W. Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N. Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D. Ekstrand, Preethi Lahoti, and Toshihiro Kamishima. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. *SIGIR Forum*, 53(2):20–43, December 2019.

Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and Addressing Algorithmic Bias in Practice. *Interactions*, 25(6):58—-63, October 2018.

Ashudeep Singh and Thorsten Joachims. Fairness of Exposure in Rankings. *arXiv preprint arXiv:1802.07281*, February 2018.

Dominik Kowald, Markus Schedl, and Elisabeth Lex. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *ECIR*, pages 35–42. Springer, 2020.

Harrie Oosterhuis, Rolf Jagerman, and Maarten de Rijke. Unbiased Learning to Rank: Counterfactual and Online Approaches. In *The Web Conference*, pages 299–300. ACM, 2020.

Steve Tadelis. Two-Sided e-Commerce Marketplaces and the Future of Retailing. In E. Baskar, editor, *Handbook on the Economics of Retail and Distribution*, pages 455–475. Edward Elgar Publishing, 2016.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *CIKM*, pages 2243–2251. ACM, 2018.

P. Resnick and R. Zeckhauser. Trust among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. *Advances in Applied Microeconomics*, 11:127–157, 2001.

A. Fradkin, E. Grewal, D. Holtz, and M. Pearson. Bias and Reciprocity in Online Reviews: Evidence from Field Experiments on Airbnb. In *Electronic Commerce*, page 641, 2015.

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, 31: 611–627, 2018.

Fei Cai and Maarten de Rijke. A Survey of Query Auto Completion in Information Retrieval. *Foundations and Trends in Information Retrieval*, 10(4):273–363, September 2016.

Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *Web Science*, pages 235–244. ACM, 2019.

Maartje Ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, Ron Mulder, Nick van der Wildt, and Maarten de Rijke. Do News Consumers Want Explanations for Personalized News Rankings? In *FATREC Workshop on Responsible Recommendation*, August 2017.

Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. Actionable Interpretability through Optimizable Counterfactual Explanations for Tree Ensembles. *arXiv preprint arXiv:1911.12199*, November 2019.

Esma Aïmeur, Gilles Brassard, José M. Fernandez, and Flavien Serge Mani Onana. ALAMBIC: A Privacy-Preserving Recommender System for Electronic Commerce. *International Journal of Information Security*, 7(5):307—334, September 2008.

Jon Degenhardt, Surya Kallumadi, Maarten de Rijke, Luo Si, Andrew Trotman, and Yinghui Xu. Workshop on ECommerce (ECOM17). In *SIGIR*. ACM, 2017.

Jon Degenhardt, Pino Di Fabbrizio, Surya Kallumadi, Mohit Kumar, Yiu-Chang Lin, Andrew Trotman, and Huasha Zhao. SIGIR 2018 Workshop on ECommerce (ECOM18). In *SIGIR*, pages 1407–1409. ACM, 2018.

Anne Schuth, Krisztian Balog, and Liadh Kelly. Overview of the Living Labs for Information Retrieval Evaluation LL4IR CLEF Lab 2015. In *CLEF*. Springer, 2015.

Krisztian Balog, Liadh Kelly, and Anne Schuth. Head First: Living Labs for Ad-hoc Search Evaluation. In *CIKM*, pages 1815–1818, 2014.

Rolf Jagerman, Krisztian Balog, and Maarten de Rijke. OpenSearch: Lessons Learned from an Online Evaluation Campaign. *Journal Data and Information Quality*, page Article 13, 2018.