Adaptive Transformer Search: Large Language Vector Search

Ranjth Kumar, Patrick DiLoreto
James Taylor, Muzammel Hoque

hello@particularaudience.com
www.particularaudience.com

Abstract. Particular Audience's Adaptive Transformer Search (ATS) addresses the most challenging problems of ecommerce search today. In this paper we will explore ATS, a proprietary technological approach to search that leverages the novel application of advanced Large Language Models (LLMs) and vector embedding technology to understand meaning and intent behind user queries. The impact of poor search experiences on retailers is significant, with billions of dollars lost annually. Adaptive Transformer Search solves the underlying problems causing search abandonment; built purposely to create intuitive search experiences, increase conversion, drive additional purchase, and foster customer loyalty, while removing the need for manual intermediation by website owners.

1. Introduction

Discovery on the Internet has come to rely on search and recommendation technologies for fast and intuitive information retrieval. While legacy keyword search has worked well enough, it still suffers from inherent flaws associated with exact token matching and a tangle of rules that overfit for specific outcomes, rules that invariably create downstream problems for alternate use cases. Each of these issues are exacerbated by messy and/or incorrect metadata in a product feed. The cost of this problem is estimated by Google to be worth $300bn per annum in the USA alone [1]. Despite the fact that 21% of all retail sales occur online [2], with ~43% of website users opting to use the onsite search bar [3], 94% of consumers say they receive irrelevant results when searching a retail website [1].

Long tail queries are susceptible to poor or zero search results, and synonym based approaches to relate edge queries to indexed product data fall short of understanding the context and intent behind search queries. In a research report conducted by Baymard Institute in 2014 [4], findings claimed the state of ecommerce search was ‘broken’. Despite the nine years that have passed since this report was published, the state of ecommerce search has not materially improved, perhaps only to have worsened relative to consumer expectations born of technological improvements in web search, recommendation systems, and chat based interfaces to artificial intelligence.

When a customer can’t find what they’ve searched for on a website, whether or not it exists, they assume the website does not have it (online or in-store), often leaving with no guarantee they will return. 76% of customers report they abandon a retailer after failing to find what they are searching for, with 48% then purchasing the item elsewhere. More than half report they typically abandon their entire shopping cart after failing to find a single item on a website [5]. Eighty-five percent of consumers say they view a brand differently after experiencing search difficulties [1] and 77% avoid websites where they’ve had poor search experiences [6]. Customers are not alone in acknowledging the extensive problem of bad site search; retailers agree, 90% of US based website managers surveyed are concerned about the cost of search abandonment to their business [1].

What is needed is an artificially intelligent large language based search system with semantic comprehension of query and possible results, allowing users to immediately discover what they are looking for without the need for intermediation by website owners with complex rule sets that attempt to correct site search results. The objective of this paper is to clearly define how Particular Audience’s Adaptive Transformer Search (ATS) addresses the most challenging problems of ecommerce search. We will explore ATS, a proprietary technological approach to search that leverages the novel application of advanced Large Language Models (LLMs) and vector embedding technology to understand meaning and intent behind user queries. The impact of poor search experiences on retailers is significant, with billions of dollars lost annually. Adaptive Transformer Search directly addresses the issues leading to search abandonment by prioritising intuitive search experiences. In doing so, ATS increases conversion, drives additional purchases, and promotes customer loyalty.

2. Legacy Search

The cornerstones of data retrieval in search technology have been relational databases and keyword search engines. Keyword search operates on a principle of literal matching, pairing query terms with those in an indexed database, with techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) being the most common for ranking relevancy. The TF-IDF value, for example, increases in direct proportion to the frequency of a word occurring in a document, balanced by the presence of the word in other documents within the corpus to help adjust for the fact that some words appear more frequently in general. This matching does not take into account any meaning or context of either the query or the index, rather is merely capable of identifying exact match similarity of tokens, or words.

For this technology to work most effectively two things must be true:
1. The retailer data set must be uniform, consistent, and complete
2. The language, vernacular, or jargon used to query must match that of the index

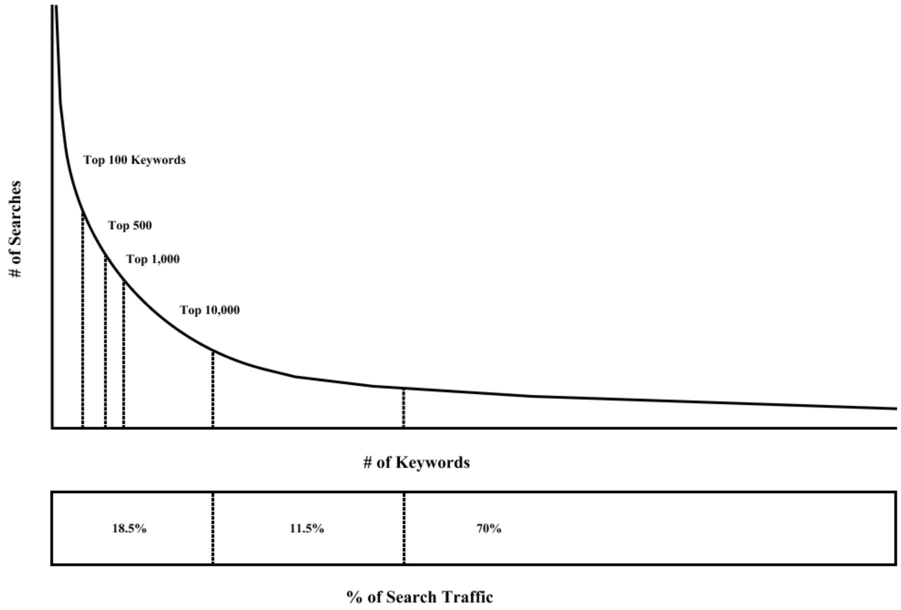
This sounds simple. However, more often than not, retailers' data is not uniform, consistent, or complete. In addition, as we've already discussed, users have become familiar with conducting their searches in colloquial language which often does not match the retailer data. Without the inference of meaning, matching of words is impossible.

Keyword search configuration is a laborious task, requiring understanding of the website's unique user search behaviour, knowledge of the corpus of data fields and tags in context of their importance to a product, and technical understanding of how certain normalisation, tokenization, and weighing should be applied to fields within the search index. Despite the amount of configuration that is done, there is still reliance on token matching, and in some cases tokenization methods implemented to solve for one problem or use case creates new problems for other queries. Oftentimes weights or 'boost' and 'searchandising' tactics are applied to specific categories or groups of products to combat poor results, conversely disappointing the retailer and the consumer when these products are ranked high in the results of what seems like an irrelevant query.

Due to the principle that keyword search relies upon matching of tokens, this approach is incapable of grasping the nuanced meaning, context, or intent inherent in a user's query. Synonyms libraries are the most common tool to circumvent this problem. Of course 'pants' and 'trousers' are synonymous with one another, however that simple relationship will need to be manually mapped in keyword search. This extends to more complex relationships between words that are not synonyms but rather share meaning with each other. For example, 'lightweight', 'portable', 'small', 'mini' and 'travel' all might share the same meaning when referring to a 'camera tripod'. These same words however might not share meaning when referring to other items in this same retailer catalogue, an example such as 'travel bag' and 'mini bag' might be very different items, one designed with lots of compartments for a long trip and the other being compact and miniature. The extent to which synonyms must be manually configured for keyword search is far underestimated, and in the case of words that share meaning only in specific context, a synonym library will only confuse queries where the meaning isn't shared.

URL redirects are largely used to manage for explicit results to poor keyword matches, directing a user to a listing page or landing page that relates to their query. This practice can be a good quick fix when a retailer is struggling to accurately match and rank search results for specific keyword, however in almost all cases, doing this does not provide a list of results that match the query by relevance, rather a static page of specific products that requires the user to further filter and sort.

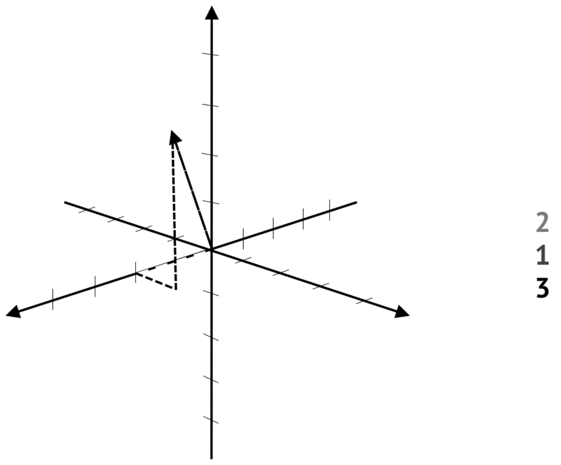
Both url redirects and synonyms libraries are helpful to fix many of the underlying problems with relevancy in keyword search today. Oftentimes these tools are used to combat null search result keywords or the top ranking keywords on a website, however managing a large catalogue at scale across thousands of search keywords creates manual limitations. Most ecommerce retailers will attempt to optimise for the highest-volume queries, effectively giving up on more than 70% of requests that constitute the ‘long tail’ of search terms due to the sheer inability to manually manage them all.



In recent years more modern search engines have incorporated knowledge graphs to enhance results by organising and connecting structured data from various sources to create a comprehensive representation of information. While they can help to improve organisation of data for context of relationships between attributes and products, knowledge graphs do not understand meaning, and the configuration and ongoing management of nodes is an arduous task requiring full time resource as well as consistent and complete data.

3. Vector

A vector can be explained using the analogy of coordinates on a map. Cell 'A2' on a two dimensional map informs a user of where to look for something listed in an index. Consider a three dimensional map; you might require a third instruction to locate that same item: [out,2] [across,1] [up,3]; from a central axis point.



Vectors need not be restricted to 2D or even 3D space, rather a hyper dimensional space represented as a string of numbers. Adaptive Transformer Search benefits from more than seven hundred dimensions.

Understanding the distance between two vectors, whether through deduction or the angle between two vector locations from the central axis point, allows us to understand how 'similar' two entities are. The closer the entities, the more similar they are likely to be, assuming the correct data points are used to define the components of their vector identity.

Imagine a search corpus that includes both car models and brands as well as animal species. In vector space, trained correctly, the animals and the cars ought to locate within their own distinct clusters, and within those clusters you might expect Lions to cluster apart from Tigers, and so forth.

Vectors are therefore a far richer approach to understanding and retrieving data, as compared with token matching approaches in keyword search.

4. Transformer

LLMs are powerful artificial intelligence models designed to understand and generate human language. They operate by learning patterns and structures from vast amounts of text data which involves predicting missing words in sentences, next character, and understanding the relationships between different words and phrases. Transformers are a type of LLM architecture used in deep learning, particularly in natural language processing tasks. In simple terms, transformers take the words of an input sentence and create a ‘summary’ that tells us what the sentence is about. Sentence transformers first tokenize and convert sentences into word embeddings which are then passed through transformer layers utilising an attention mechanism to output embeddings that represent the entire sentence. The attention layer is intelligence capable of allowing the tokenized words of a sentence to interact with one another, thereby assigning weighting to what matters most in that sentence. An example would be in language translation, if using an attention model to translate a sentence from one language to another, the model would select the most important words and assign them a higher weight. With their ability to understand and generate human-like text, transformers serve great use in generating vector embeddings for semantic search results.

The pinnacle of exceptional search is recall, the ability to retrieve all positive matches and then accurately rank results to a query in a ‘logical’ order. Logical, being a subjective term that infers true human-like semantic understanding of a query, its context, and the index data which is being searched. If given enough time, a human mind could find and rank the products in a catalogue of 14,000 items relative to a search query. Leveraging the power of advanced LLMs, Particular Audience’s Adaptive Transformer Search (ATS) is engineered to do just this. ATS has the ability to replicate that same understanding and logic in matching and ranking of results, completing tasks with more precision in milliseconds.

PA’s Adaptive Transformer Search is built using transformer models, converting sequential long form text (retailer catalogue and website data) into dense vectors and indexing them in high-dimensional space. The conversion of a sequence of words into a vector is known as sentence embeddings, a concept popularised by large language models such as Google's BERT and OpenAI's GPT. The transformer model used to generate the sentence embedding determines the vector coordinates of each product or page that can be searched on a retailer website. The accuracy and precision of these embeddings is most critical to ensuring a vector search retrieves the most logical results.

5. Vertical Tuning

Beginning with the same open source transformers that power both BERT and GPT, we built proprietary Vertical Tuned Models (VTM) by fine tuning specific data sets for most accurate vector embeddings unique to each retail vertical. Particular Audience VTMs are fine tuned on synthetically generated data from models trained on >500k real internet search queries, as well as real internet scale positive sentence pairs. Our VTMs have demonstrated over 37% improvement in relevancy score ranking when compared to open source models pre-trained on over one billion sentence pairs. This process ensures the sentence embeddings created from VTMs will generate the most relevant search results out of the box, even before the model begins to learn and adapt on a retailer's own dataset. This solves one of site search's biggest pain points: the long, complex and meticulous configuration of a search index.

When creating embeddings from data on a retail website, the information must be represented in a sentence, considering not only the meaning of individual words but also their order and context amongst others. In our experimentation, we tested many different structures when composing sentences with a hypothesis that the more logical the sentence is in regards to a hierarchy of information, the better precision our index vectors will have. True to LLM understanding of natural language, as we improve the logical structuring of our sentences, precision of our vector embeddings improved rank relevancy by up to 43%.

In real time, customer queries from a website are converted into dense vector embeddings that use PA-VTMs, thereby allowing our Search Vector Retrieval (SVR) to measure similarity between query and index vectors at production level speed (<300ms). PA-SVR captures the angles between vectors to measure similarity, an effective metric for comparing embeddings in high-dimensional vector space.

In the simplest terms, while not synonymous, the words 'animals' and 'nature' will be located closer together in vector space than 'animals' and 'cars'. This intuitive correlation, while clear to a human mind, would not be recognised by keyword search without an explicit mapping of synonyms or the use of a knowledge graph to associate entities. If we return to our example of shared meaning in the context of camera tripods, our VTMs understand that 'lightweight', 'portable', 'small', 'mini' and 'travel' are related to one another in this context. Upon searching for a 'compact tripod' ATS will return results for all the aforementioned types of tripods, solving one of the biggest problems today related to the vernacular used in querying a website. Where long tail queries make up more than half, and in some cases 70%+ of a website's search, ATS provides better customer experience from search to conversion.

Given understanding of shared contexts, PA-VTMs solve another problem: retrieving substitute results for an otherwise failed or 'at-risk' user query. Consider a consumer electronics retailer who sells select brands of Android smartphones. If a customer searches for 'Huawei phone', but that brand is not carried by our example retailer, PA-VTMs are capable of understanding the context and will return the similar items from alternate brands, e.g. Samsung phones are likely to display in the results, saving the customer from zero search results and reducing risk of abandonment. While substitute brands and products could be set up as synonyms or redirects, the resource constraints of a human team limits rule building to only the top search terms, without the ability to tackle every long tail problem nor the ever changing product catalogue. ATS has the ability to eliminate zero search results from a retailer website by as much as 70%.

6. Synthetic Data

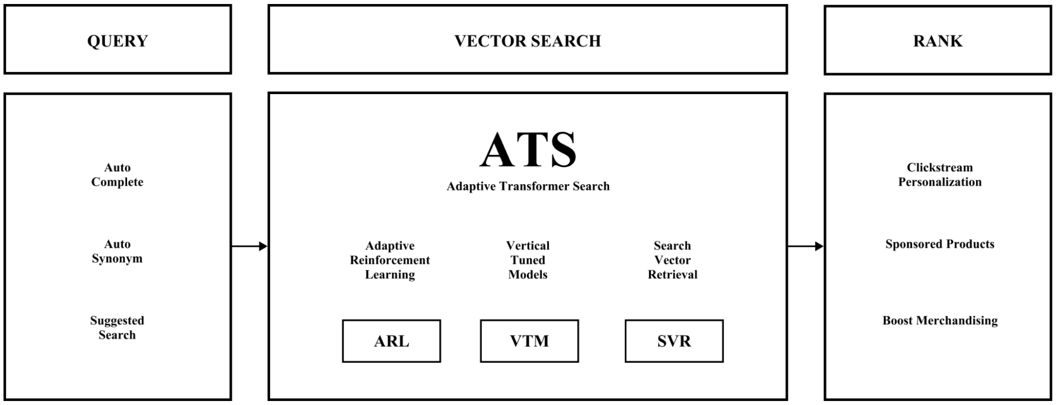
To further enable PA-VTMs we have succeeded in the reverse generation of synthetic search queries to help train query-click-pairs. This application of generative artificial intelligence allows us to fill gaps in pre-existing data leveraging internet scale query-click-pairs. Additional approaches to rank order optimization for local and vertical search applications include Rank Proxy inference that interprets mainstream web search engine results via web crawler to train query-click-pairs, and cross-domain query-click-pairs via our browser extension clickstream ingestion.

7. Adaptive Reinforcement Learning

ATS is unique in its ability to continue to learn on a retailer's website. To ensure the Vertical Tuned Models continue to adjust with changes in user behaviour, preference, website and catalogue data, and language, we've built in Adaptive Reinforcement Learning (ARL) for our models to evolve in tandem, requiring no manual effort by the retailer. Through Adaptive Reinforcement Learning, our VTMs receive signals from live user behaviour, which over time continually shift embeddings in vector space, improving their precision and accuracy in the context of the retailer website. ARL ensures search results are most relevant in the context of the retailer's website, solving for the limitations retailers have in manually optimising their keyword search engines today.

8. Adaptive Transformer Search

Particular Audience's Adaptive Transformer Search combines vector search (SVR) and keyword search capabilities to deliver a modern approach to intuitive semantic search. Vertical tuning (VTMs) and local reinforcement learning (ARL) specific to each site powers significant improvements in relevancy, engagement and a drastic reduction in zero search results.



Particular Audience's sponsored products, clickstream personalisation, and merchandise boosting all play seamlessly into the user search experience.

9. Contextual Ranking

Adaptive Transformer Search requires no information of a user to comprehensively understand their demand. Demand may be inferred from explicit (search and navigation) and implicit (clickstream and purchase) signal data.

We are able to interpret the context of an anonymised user by their implicit signals, real-time clickstream and changing basket contents, to recommend relevant items in anticipation of their next action. This market leading recommendation technology can be leveraged in personalising the ranking of search results in ATS. While precise semantic recall is the most important part to getting search right on a retailer website, adaptation of the ranking of results to a user and their context drives click-through rates from search to increase.

10. Relevant Search Ads

Adaptive Transformer Search also contributes significantly to our product vision, leading the market with the most relevant sponsored product placements from Particular Audience Retail Media. Until now, suppliers using PA Retail Media search campaigns bid on specific keyword terms for their sponsored product placement. Although this has performed very well, manual configuration of keywords in a campaign is not scalable, and almost impossible to manage the keyword long tail of a retailer website. By integrating ATS into PA Retail Media search campaigns, we expect coverage and performance of sponsored products in search results to have a material commercial impact for retailers and suppliers alike.

11. Privacy

Privacy is foundational to our ethos, vehemently anti-segment, we collect zero personally identifiable information and do not utilise third party tracking.

Internet scale language data reinforces and adapts vector space to greatly improve relevance in information retrieval and recommendation, all while preserving user privacy.

12. Conversational Interfaces

Foundational LLMs mean search results are better able to fit to question and answer oriented queries. Particular Audience released its visual search technology in 2019 whereby users could upload images to search a retail website, leveraging related dense vector embeddings for similar item retrieval. This exercise taught us that user behaviour takes time to shift. While generative chat based interfaces have captured the zeitgeist, it might be premature to fully integrate them into onsite ecommerce search. Particular Audience's Adaptive Transformer Search is capable of interpreting natural language text inputs and returning applicable items, a capability that future proofs evolving consumer search behaviour.

13. Conclusion

We propose that Adaptive Transformer Search solves for the underlying problems impeding ecommerce search for 94% of consumers today. We started by investigating the limitations of conventional ecommerce search engines which rely on exact keyword matching and continuous manual updates. The legacy technology, incapable of understanding meaning and context of words in a query, underscores the requirements for a more advanced artificial intelligence search technology in ecommerce.

We have explained how Particular Audience's Adaptive Transformer Search leverages vertical specific LLMs, ensuring sentence embeddings created from PA-VTMs generate the most relevant search results out of the box. The models continue to adapt with localised reinforcement learning improving their precision and accuracy in the context of the retailer website. In solving the underlying problems in keyword search, ATS also eliminates significant manual configuration for website owners. Adaptive Transformer Search is built specifically to address the root causes of search abandonment, facilitating intuitive search experiences for every customer.

References

[1] https://www.googlecloudpresscorner.com/2021-07-27-Google-Clouds-Retail-Search-Equips-Retailers-with-Google-Quality-Search-Functionality-to-Improve-Product-Discovery-Reduce-Search-Abandonment

[2] https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/

[3] https://www.forrester.com/report/MustHave-eCommerce-Features/RES89561

[4] https://baymard.com/blog/ecommerce-search-report-and-benchmark

[5] https://www.techradar.com/news/poor-quality-websites-are-costing-businesses-billions-in-lost-sales

[6] https://cloud.google.com/blog/topics/retail/search-abandonment-impacts-retail-sales-brand-loyalty