From Keywords to Concepts:

A Late Interaction Approach to Semantic Product Search on IKEA.com

Authors:
Amritpal Singh Gill, Sannikumar Patel, Péter Varga, Patrick Miller, Sakis Athanasiadis

IKEA Retail (Ingka Group)
Amsterdam, The Netherlands

IKEA Search

- 3M daily users
- 30% of online users use search
- 50M searches/week
- 2000 searches per second (peak)
- 30+ markets

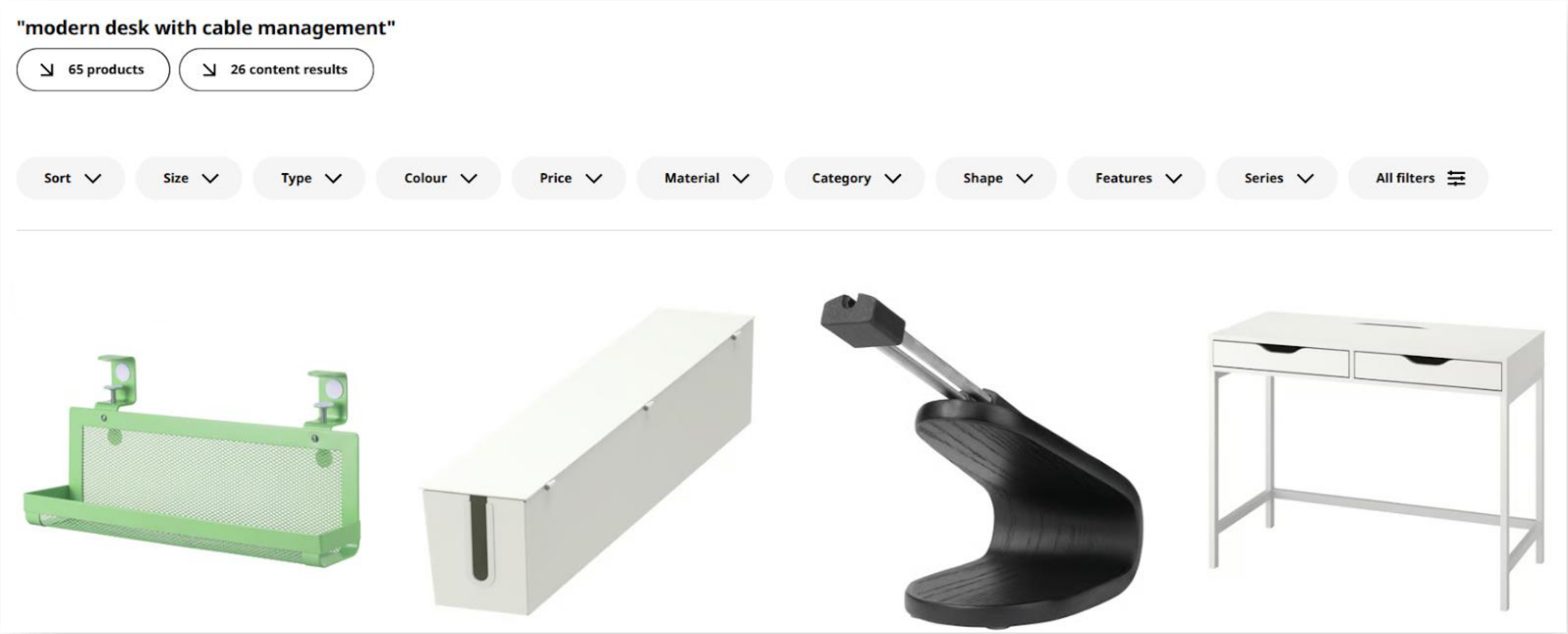
Why Change Search At IKEA.com?

- Boolean keyword search fails on complex intent

- Customers adapt their behavior (shorter, less natural queries)
  - Top 5 search queries in the U.S. → *"desk", "dresser", "kallax", "shelves", "curtains"*

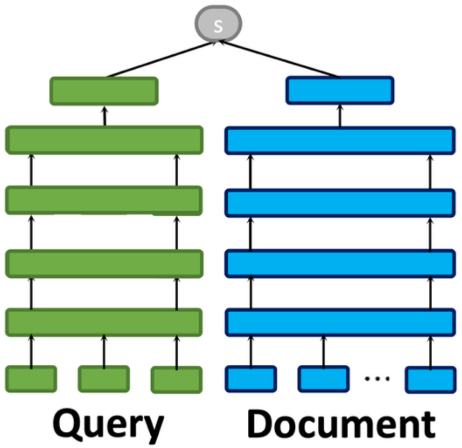
- <u>**Opportunity**</u>: Offer **semantic search** to better capture complex user intent

Boolean search



Semantic Search

Dense vector



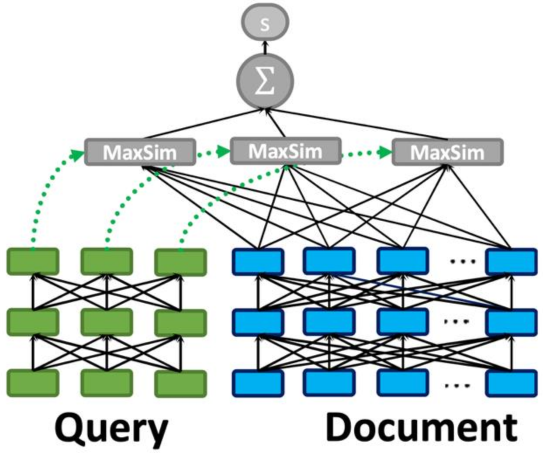
- One global embedding blurs token importance

Difficult to distinguish <=

- Larger embedding size => Higher latency

"200 cm tall wardrobe"
"230 cm tall wardrobe"

Multi vector *(Late interaction)*



- Keeps token granularity and importance

=> Easy to distinguish

- Smaller embedding size => Low latency

<u>Image source</u>: Khattab & Zaharia (2020), *ColBERT: Efficient Passage Search via BERT*, SIGIR '20

Semantic Search - Late interaction model

- BERT encoder – 110 million parameters
- Token-level matching → captures fine-grained semantics
- End-to-end retrieval and ranking in < **30 ms**
- Deployed over 31 k products in the U.S. market

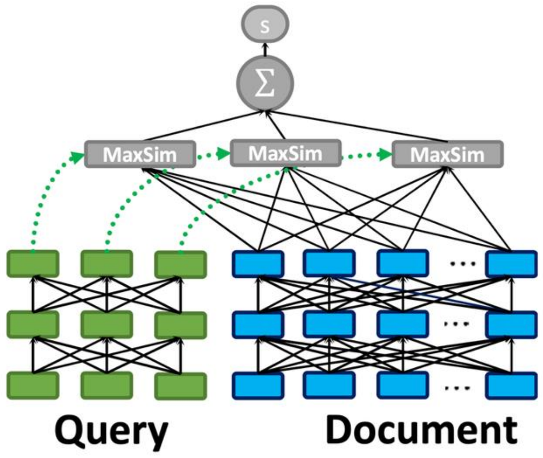


Image source: Khattab & Zaharia (2020), *ColBERT: Efficient Passage Search via BERT*, SIGIR '20

Semantic Search

modern desk with cable management

Boolean search




Semantic search




Semantic Search

table without chairs

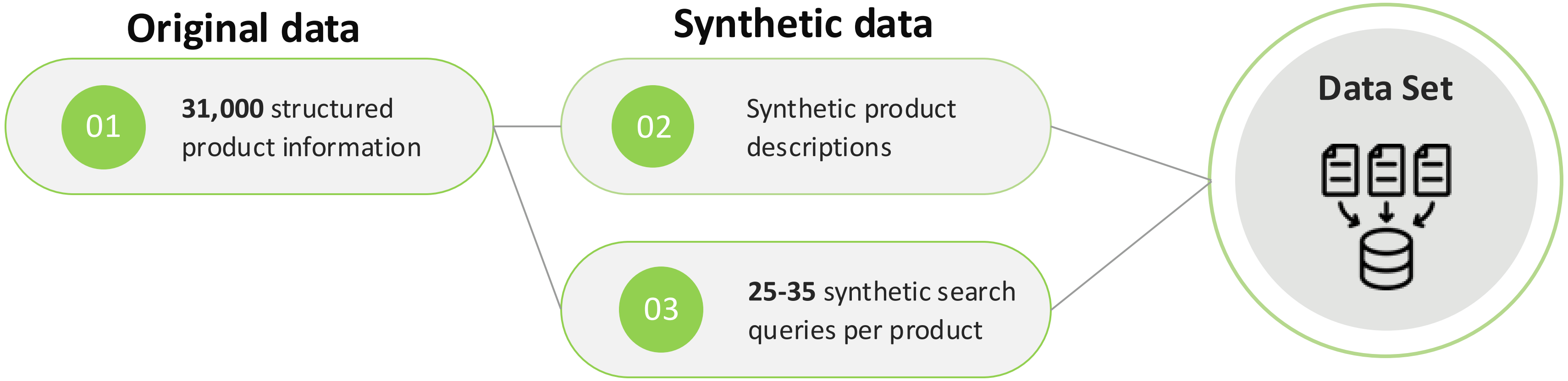
Boolean search

Semantic search

Training Data - Synthetic data generation

- **Customer queries were too simple**
  - Users searched with simple keywords (e.g., *“desk”, “sofa”*)
  - Reflect adaptation to Boolean search, not natural intent

- **LLM-generated queries captured rich intent**
  - Included synonyms, attributes, phrasing diversity
  - Prevented human labelling cost and promoted scalability



0.9 million (q, p+, p-) triplets

q	p^+	p^-
sofa	 *Friheten sofa*	 *Klippan sofa cover*
bed	 *Hemnes bed frame*	 *Vestmarka mattress*

Training Data - Strong negative sampling

- **Why not random negatives?**
  - Too easy — model learns to separate "*sofa*" from "*wardrobe*", not "*sofa*" from "*sofa cover*"
  - Leads to poor generalization and overconfidence

- **Dense vector embedding based hard negative selection**
  - Embed all product descriptions into dense vector embeddings
  - For each positive product, retrieve *100* nearest neighbors by cosine similarity
  - Filter based on category mismatch and type

- **Impact**
  - Boosts model's ability to reject subtle mismatches
  - Accelerates convergence and improves robustness

q	p^+	p^-
sofa	 *Friheten sofa*	 *Klippan sofa cover*
bed	 *Hemnes bed frame*	 *Vestmarka mattress*

Training & Loss

- Trained with contrastive in-batch softmax loss on token-similarity scores
  - High similarity score => relevant products
  - Low similarity score => irrelevant products

$$S(q, p^+) > S(q, p^-)$$

- Training setup
  - 0.9 million triplets
  - 150K training steps
  - Batch size 32
  - Learning rate 5e-6

q	p^+	p^-
sofa	 *Friheten sofa*	 *Klippan sofa cover*
bed	 *Hemnes bed frame*	 *Vestmarka mattress*

Adaptive Thresholding for Ranking

- **Problem:**
  - Semantic search retrieves top-N similar items in latent space S.
  - Due to ambiguous queries or low embedding granularity:
    - Results include marginally relevant items.
    - Similarity scores lack clear separation.
  - Fixed N → inconsistent relevance for different queries.

- **Observation:**
  - Synthetic product descriptions showed repeated phrasing.
  - Caused tight clustering in embedding space.
  - Small margins between similarity scores → no obvious threshold.

- **Solution (Compact):**
  - Compute score differences: $\Delta S = [s_2 - s_1, \ldots, s_n - s_{n-1}]$
  - Calculate mean (μ), std (σ), then Z-scores: $Z = (\Delta S - \mu) / \sigma$
  - Identify sharp drops: $Z_i < z\_threshold$
  - Validate drops using % change: $P = |s_{i+1} - s_i| / s_i < t_D$
  - Select optimal rank cutoff r_i dynamically per query

Evaluation

Offline model evaluation

Mean Average Precision: 0.65 → **0.82**
Human Evaluation: 0.8

A/B tests

+3.1% click-through rate
+1.96% conversion rate
+1.78% search interaction rate
+2.18% add-to-cart actions

Latency

< 30 ms

Business impact

Improved user engagement positively impacting business metrics

Conclusions

- Late-interaction search = keyword precision + semantic recall
- Achieved measurable business wins with minimal latency
- Synthetic data unlocked speed & coverage
- Hard negatives + thresholding critical for precision
- **Next**:
    - Expand to support more descriptive natural language queries with multiple user intents
    - Test and launch in more markets/languages

Thank you!

Correspondence contact:
Amritpal Singh Gill
amritpalsingh.gill@ingka.ikea.com