Increasing Relevance of Rank by Price

Vaughan Kitchen

July 9, 2019

What is IR?

" **Information retrieval (IR)** is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds. "

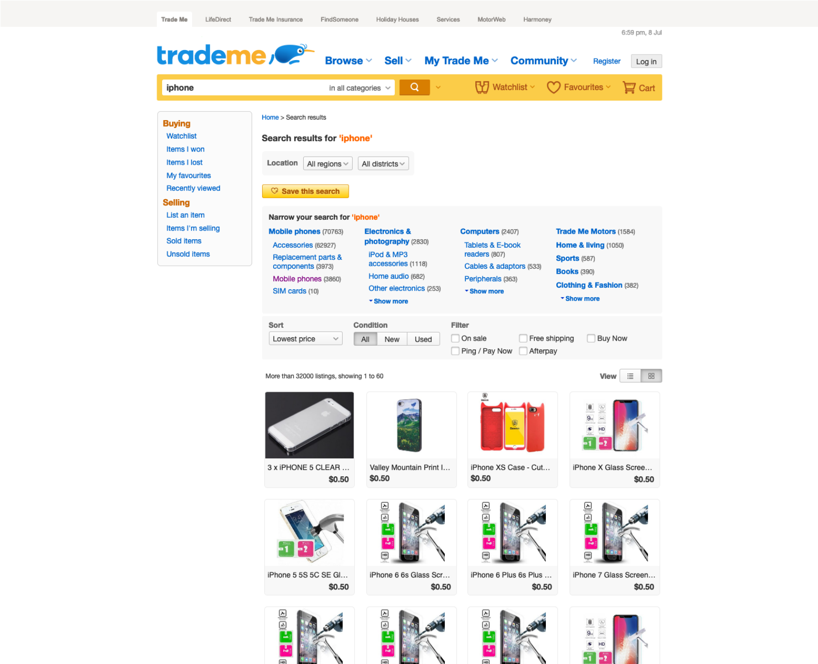
— wikipedia

(fulfilling a users information needs)

eCommerce Search

A search engine integrated into an online store helping users find products. Users perform queries where the search engine tries to match keywords to products that it has stored. Can have a deeper semantic understanding of text, do query rewriting, or fuzzy matching which seperates it from a database (which just does look ups).

The Problem with eCommerce Search



The Problem with eCommerce Search

- These are not iPhones
- Many of the results are not relevant
- Often the non-relevant results are cheaper
- Non-relevant results feature as the top results

Some Definitions

" **recall** (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances "

— wikipedia

" **precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances "

— wikipedia

- Recall = Found/Total No. Relevant
- Precision = Found/Total Returned
- Often seen as Precision at n, P@n, e.g. P@10

Precision or Recall?

- P@10 result poor when price ordered
- High precision needed
- User always wants cheapest item
- Will use competitor if they're cheaper
- High recall needed
- High accuracy needed

Metrics

- P@n: Precision at n
- Mean Average Precision (MAP): Sum(P-at-reldoc-n) / Relevant, averaged
- Discounted Cumulative Gain: Relevance gain of extra documents discounted for index in list
- Normalized Discounted Cumulative Gain: DCG normalized against ideal result
- (Collaboration with eBay and IR community)

SIGIR Data Challenge 2019

- High Accuracy Recall Task
- Opened May 27
- Closes July 18
- Three phases: Unsupervised, Supervised, Combined
- Approximately 900,000 documents with price, title, category
- 150 Queries
- Binary classification

SIGIR Data Challenge 2019

B - Baseline submission

Rank	Participant team	precision	recall	f1	tpr	fpr	Last submission at
1	Gotta Recall em All	0.74	0.88	0.81	0.88	0.22	1 month ago
2	Uplab	0.76	0.85	0.80	0.85	0.20	1 month ago
3	ir_h	0.73	0.87	0.80	0.87	0.23	1 month ago
4	QUTDataScience	0.69	0.90	0.78	0.90	0.29	1 month ago
5	Otago	0.77	0.80	0.78	0.80	0.18	1 month ago
6	Mercari-search	0.73	0.77	0.75	0.77	0.21	1 month ago
7	JediOrder (v1)	0.62	0.87	0.72	0.87	0.39	1 month ago
8	f-group-team (Wordcount)	0.50	0.96	0.66	0.96	0.68	1 month ago
9	RIIID (v3)	0.52	0.89	0.66	0.89	0.60	1 month ago
10	MLML	0.52	0.89	0.66	0.89	0.60	1 month ago
11	CaptainTeemo	0.50	0.93	0.65	0.93	0.67	1 month ago
12	alphaIR	0.42	1.00	0.59	1.00	1.00	1 month ago
13	Last of Us	0.00	0.00	0.00	0.00	0.00	1 month ago

SIGIR Data Challenge 2019

- 5th/12 successful submissions in Unsupervised phase
- Unable to submit in supervised phase due to busyness
- Attempting final phase over next two weeks
- Search engine written from scratch
- Conjunctive search with Porter's stemming
- Final attempt will be with Entity Linking

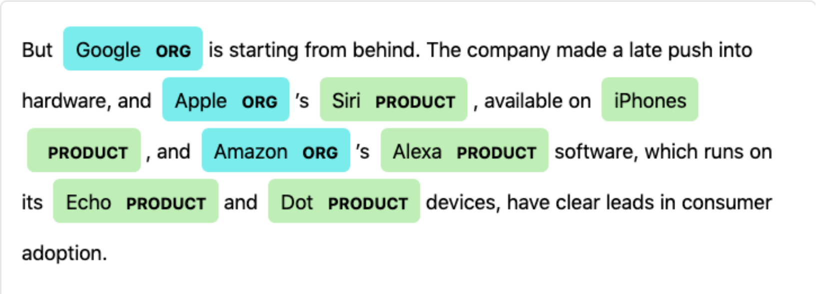
spaCy Named Entity Recognition

```python
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp(u"Apple is looking at buying U.K. startup for $1 billion")

for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

spaCy Named Entity Recognition



Entity Linking

Fast and Space-Efficient Entity Linking in Queries (Blanco et al.)

- Fast and Space-Efficient Entity Linking in Queries (Blanco et al.)
- Hash table to entities with thesaurus mined from query logs weighted to preferred results
- Successful for query segmentation
- "clinton falls asleep" $\rightarrow$ "clinton falls | asleep", "clinton | falls asleep"

A Multi-View-Based Collective Entity Linking Method (Liu et al.)

- Collective Entity Linking
- Considers *description, background knowledge, references* and *entity graph structure*
- *description* gives best performance to complexity ratio

Entity Linking

- Requires large scale preprocessing of Wikipedia
- Performance considerations in practice
- Unsure on overall effect on result quality
- Short text (titles only) may affect results

References

- Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and Space-Efficient Entity Linking in Queries. Proceedings of the Eigth ACM International Conference on Web Search and Data Mining, Pages 179-188. http://dx.doi.org/10.1145/2684822.2685317
- Ming Liu, Gu Gong, Bing Qin, and Ting Liu. 2019. A Multi-View–Based Collective Entity Linking Method. ACM Transactions on Information Systems 37, 2, Article 23 (Feb.2019),29pages. https://doi.org/10.1145/3300197

Questions?