PERSONALIZATION, RECOMMENDATIONS AND SIMILARITY TOWARDS REAL-TIME, PERSONALIZED RECOMMENDATIONS
MOTIVATION
“SIMILARITIES ARE EXTENSIVELY USED BY ONLINE RETAILERS FOR MANY DIFFERENT RECOMMENDATION TASKS.”

Item2Vec: Neural Item Embedding for Collaborative Filtering - Barkan, Oren Koenigstein, Noam
WHAT'S MISSING IN THIS PICTURE?

...AM I SEEING THIS?
...ARE THESE IDENTICAL?
...ARE THESE RELEVANT TO ME?

SHOULD I BUY THESE?
“THERE IS MORE SIMILARITY IN ... A PRECIOUS PAINTING BY DEGAS AND A FROSTED MUG OF ROOT BEER THAN YOU EVER THOUGHT POSSIBLE.”

A. Alfred Taubman
MOTIVATION

NON-SENSE, PAINTING BY DEGAS AND ROOT BEER?

http://www.wga.hu/html_m/d/degas/2/1870s_04.html
http://www.caffeineinformer.com/caffeine-content/mug-root-beer
“THERE IS MORE SIMILARITY IN THE MARKETING CHALLENGE OF SELLING A PRECIOUS PAINTING BY DEGAS AND A FROSTED MUG OF ROOT BEER THAN YOU EVER THOUGHT POSSIBLE.”

A. Alfred Taubman
Concept of Similarity is ripe with subjectiveness.

- Similarity can be by: **Attributes, Function, Application, Class, Structure** OR **Representation**
- It is quite clear that every **Context** changes the interpretation of similarity
- In other words we can say that it changes by Semantics of the features we use to measure similarity by.

“Note that we focus on one main aspect of similarity—text content. This might not completely capture the human-judgement notion of similarity in all cases.”[1]
MOTIVATION

Consumer **Viewing TV**

Consumer **Buying TV**

Consumer **Bought TV**

Recommendation

Recommendation

Recommendation
“THERE IS MORE SIMILARITY IN {CONTEXT} {QUERY} AND {TARGET} THAN YOU EVER THOUGHT POSSIBLE.”

A. Alfred Taubman
COMPONENTS

- **Context** – defines the domain for interpreting similarity
- **Query** – predicate for similarity in the given context (vector representation)
- **Search Space** – target neighborhood (vector representation)
APPLICATIONS OF SIMILARITY

**Text** - Documents, Spreadsheets, Databases etc.

**Visual** - Images, Videos, Scanned Artifacts etc.

**Audio** - Speech, Sound, Music etc.

**Sensory** - Pressure, Temperature, Humidity etc.

**Geo/Astro/Spatial** - Geo-Mapping, Space, Astronomy, Drones etc.

**Time Series** - Sales Trends etc.
Big Data can be really Big and Annoying!

- Dimensionality Reduction - Embedding & Projection
- Locally Relevant Features vs. Globally Relevant Features
- Space-Time Complexity
- Computational or Arithmetic Intensity - Parallel computing systems are limited by memory bandwidth and communication overhead
SIMILARITY METRICS

▸ **Distance Metrics** - Manhattan, Euclidean, Mahalanobis, Chebyshev

▸ **Similarity Measures** - Cosine Similarity, Jaccard Similarity, Tanimoto Similarity

\[ \propto N \times d \]

*N – Size of Data, d – dimensions of data (based on representation)*
TYPICAL PRACTICAL SOLUTIONS

- Representation is Fixed
- Models are fixed and orchestrated as Strategies
- Recommendations Space is defined by historical data
- Most of the performance is derived by speeding up queries

APPLICATION

Engineering Effort to Setup, Standardize and Automate
REAL-TIME RECOMMENDATION ARCHITECTURE

Demands the following capabilities in Real-time

- Representation Change
- Model Selection
- Training
- Parameter Learning
- Optimization
APPLICATION

HOW DO WE APPROACH IT?

▸ **Real-time recommendations:**
  ▸ Context Change
  ▸ Representation Change
  ▸ Search
  ▸ Deliver Results

CONTEXT CHANGE

UPDATE SEARCH SPACE
UPDATE QUERY
SELECT DISTANCE MEASURE

PERFORM SEARCH

RETURN RESULTS
WHAT WE NEED

- Build a platform that supports the following seamlessly in real-time:
  - Model Blending – Strategies can be combined in Real-Time
  - Representation Change – Can drive model selection
  - Rescoring – Models based on recent interaction can be used to rescore similarity
  - Similarity Metrics – Can be chosen on the fly based on the model
APPLICATION

SOME COMMON TECHNIQUES

- **Local Space Exploration** – Locality Sensitive Hashing for NN
- **Bloom Filters** – For fast membership tests, used by Cassandra
- **Query Analysis and Efficient Indexes** – Lucene’s Indices/Analyzers form the basis for Solr and Elasticsearch which are heavily used in production.
APPLICATION

ARCHITECTURAL LAYERS

- Serve API Public Cloud - HA Cluster
- Cassandra Hybrid Cloud - HA Cluster
- Ingest API OpenStack HA Cluster
- Elasticsearch OpenStack HA Cluster
- Kafka Enterprise HA Cluster
- Spark Distributed Computing Cluster
- Storm Distributed Event Processing
- HDFS/Hadoop Private Hadoop Cluster
- Data Acquisition Layer to Ingest Data
- Raw Data Heterogenous Data Sources
FUTURE
Similarity Learning: A common approach for learning similarity, is to model the similarity function as a bilinear form. For example, in the case of ranking similarity learning, one aims to learn a matrix that parametrizes the similarity function.

Meta-Learning: We are quickly moving towards a world where meta-learning will hold central place in most of industries using Machine Learning to solve business problems.

Small Data Training: Next step in science is using cognitive skills of human beings and bringing them to machines. So, machines can use a few examples to learn a new concept instead of millions rows of data.


WEB


QUESTIONS

Thank You!