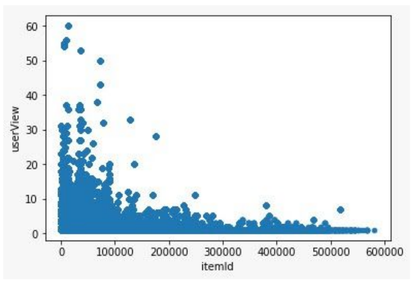
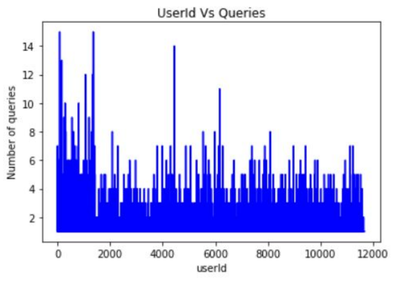
Personalized Search Ranking

Abindu Dhar, Ashwin Krishna, Rishabh Singla, Srividhya Balaji

MOTIVATION:

With the rapid growth of data on the Internet, it has become important to help users alleviate the problem of information overload and select interesting and necessary information in web applications. We aim at providing improved Personalized search results to each user based on their activity by leveraging the recent advances in graph embedding techniques to exploit graph structured data for automatic feature extraction.

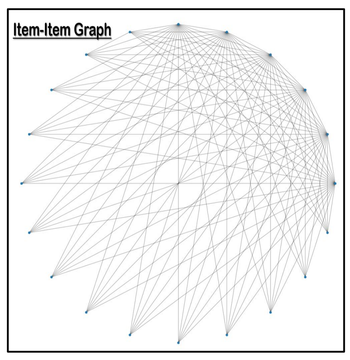




DATASET:

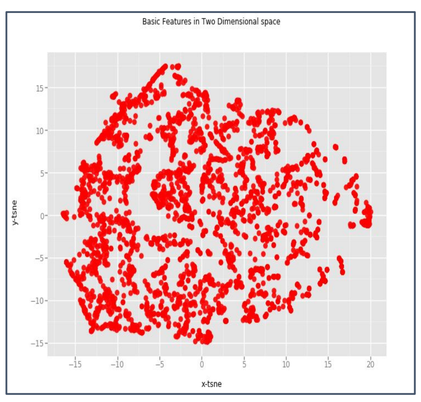
We used the Ecommerce Search Dataset from CIKM Cup 2016 Track 2 . The dataset includes user sessions extracted from e-commerce search engine logs, with anonymized user ids, hashed query terms, hashed product descriptions and meta-data, log-scaled prices, clicks, and purchases.

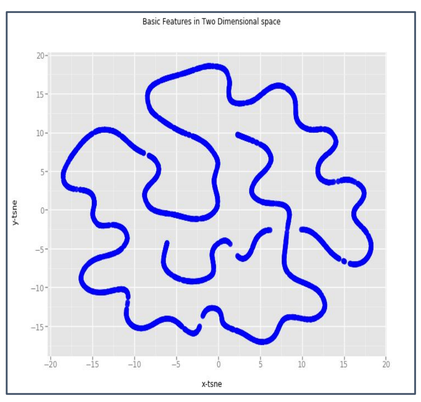
Statistics	Value	
# products	184047	
# unique queries	26137	
Vocabulary size	181194	
Length of queries	2.66 ± 1.77	
Length of product descriptions	5.12 ± 2.04	
# items per order	1.34 ± 0.96	
	Original	Chronological
# queries (train/test)	35615/16218	28380/3011
# search sessions	26880	21505
# browsing sessions	349607	242852
# clicks	37705	30160
# views	1235380	857008
# orders	13506	9130



METHODOLOGY:

We implemented different models to understand the improvement at each step and ranked the items using a Pairwise Ranking SVM approach . We consider only queries that contain search tokens and non null user Ids.





Graph Based Model : Building a product graph by embedding the basic features of each product in every node and drawing an edge between products that occur together in the same query. Using a Graph Autoencoder to extract a 16 /32 length Graph Embeddings which captures the relationship between products. User features are obtained as a function of query embeddings and graph embeddings averaged across all sessions . Finally we Combine the embeddings and obtain a relevance score to re-rank the items in descending order of their scores.

RESULTS:

Integrating the graph based features provided the best result

RESULTS	NDCG for Query-full Data
Baseline	0.22767
CIKM Cup 2016 First Place	0.55
Our Graph Best Model	0.63095

MODEL	nDCG	nDCG IMPROVEMENT
Random Shuffle(Baseline)	0.22767	
Basic Features(10)	0.25595	12.4214%
Basic Features +Token Embeddings	0.26397	15.944129%
Basic+ Cosine of Token Embeddings	0.28906	26.96446%
Graph Embedding(16)	0.31546	38.56019%
Graph Embeddings(32)	0.33333	46.4092%
Basic Features + Token Embeddings(100)	0.35624	56.47208%
Graph Embeddings(16) + Cosine	0.38685	69.91698%
Basic Features + Token Embeddings + Cosine	0.43456	90.87275%
Graph Embeddings(16) +Token Embeddings(50)	0.63095	177.1335%

CHALLENGES :

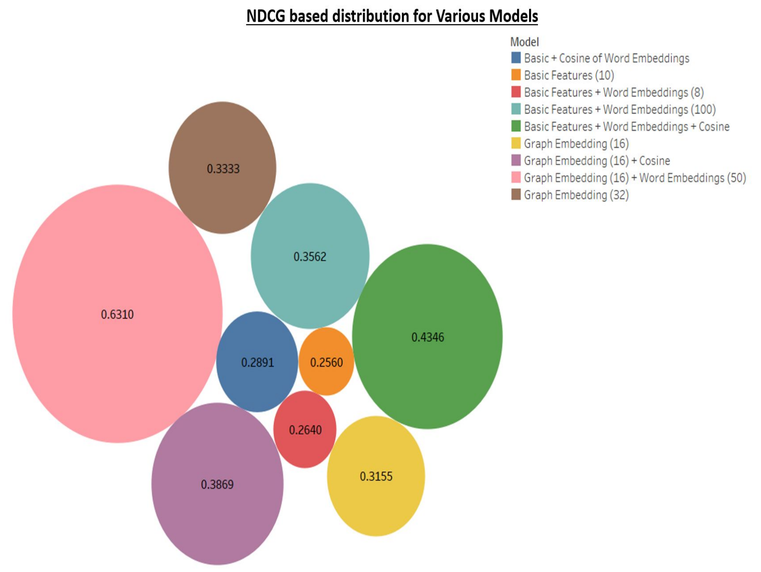
- The Dataset was anonymized which removed semantic features.
- Graph Embedding required a lot of compute resources.

TAKEAWAYS :

From the tsne graph and ndcg results we can conclude that graph embeddings provide us a better model by providing intrinsic relationships.

WHAT'S NEXT:

- To build more user personalized features based on sessions.
- To apply neural models and provide comparisons with RankSVM.



REFERENCES :

[1]https://competitions.codalab.org/competitions/11161
[2] Kipf, Thomas N., and Max Welling. "Variational graph auto-encoders." arXiv preprint arXiv:1611.07308 (2016).

MODEL ARCHITECTURE