An Information Retrieval Framework for Contextual Suggestion Based on Heterogeneous Information Network Embeddings

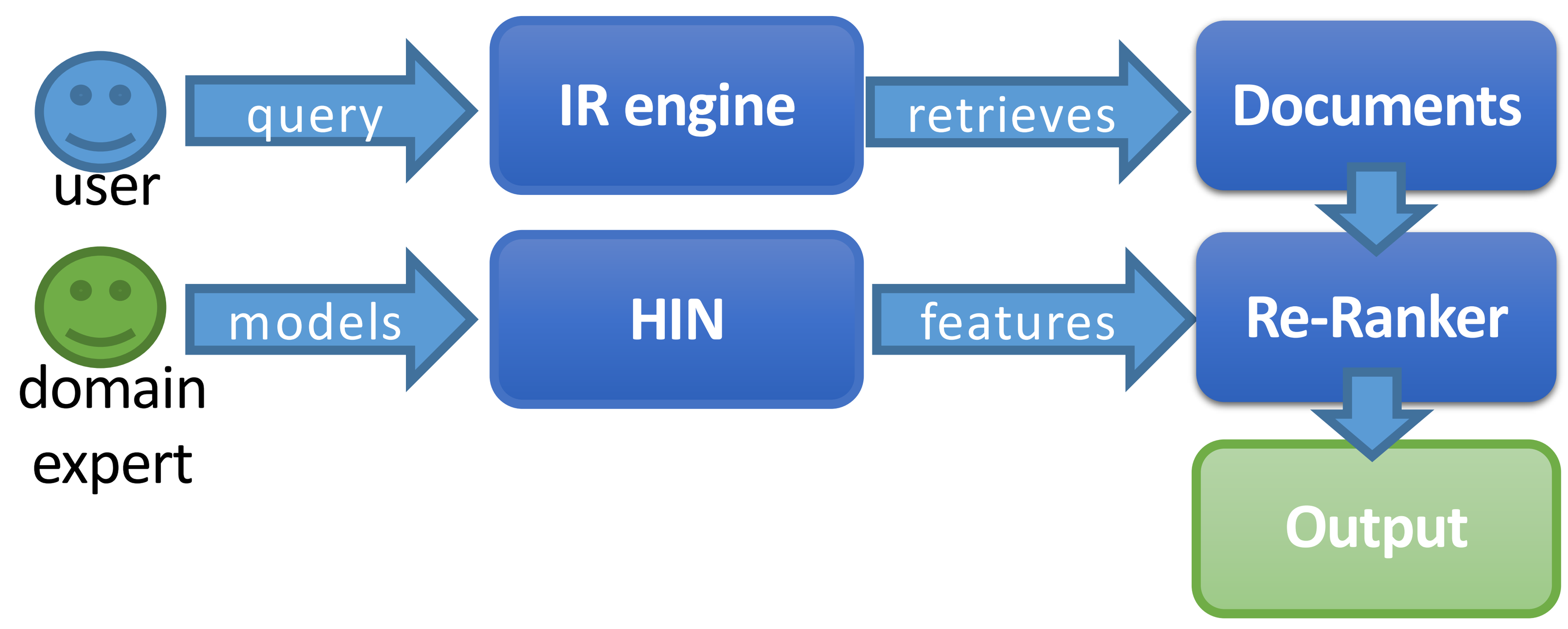
IBM Research

Dominic Seyler, Praveen Chandar, Matthew Davis

dseyler2@illinois.edu, praveenr@spotify.com, matthew.davis@invitae.com

Overview

- Contextual suggestion: User query is augmented by user model (i.e. the context of the query)
- User model can be previously rated (or viewed) documents that will be considered at query time
- Idea: Domain expert models query context using Heterogeneous Information Network (HIN) embeddings.
- Application: 1) run query using "regular" IR engine (e.g. OKAPI BM 25) 2) re-rank retrieved documents by taking HIN embeddings into account



Problem Formulation

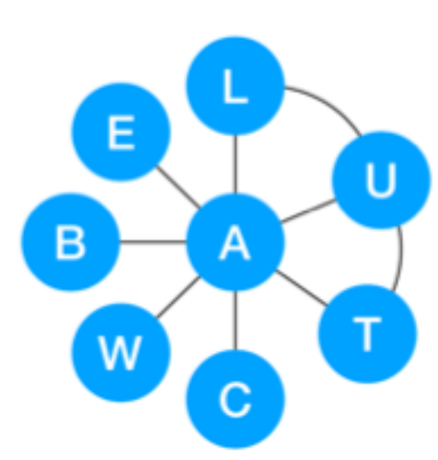
- Focus on TREC Contextual Suggestion task, where the IR system is assisting a user in planning a trip to a target city.
- input to the system is a list of requests (R) and user profiles (U), where user profiles are a list of rated attractions (preferences), gender and age.
- The output is a ranked list of attractions not in the preference list, ordered by their posterior probability conditioned on the user profile and request

$$input = \{R, U = \{info, pref\}\}$$
$$R = \{group, season, trip\_type, duration, location\}$$
$$info = \{gender, age\}$$
$$pref = \left\{(attraction, rating, tags)^{1,\dots,k}\right\}$$
$$output = \{P(attraction|R,U)|\forall attraction \notin pref\}$$

HIN Modeling

Node Type	Description
$\mathcal{U}$	User
$\mathcal{L}$	Location
$\mathcal{A}$	Attraction
$\mathcal{T}$	User tags/endorsements
$\mathcal{B}$	Token in attraction's business name
$\mathcal{W}$	Token on attraction's homepage
$\mathcal{C}$	Category tags from attraction's profile page
$\mathcal{E}$	Named entities in attraction's profile page

T1: Node Types



F1: Topology

Meta-Path	Semantics
$\mathcal{A}-\mathcal{U}$	Attractions were rated by a user.
$\mathcal{A}-\mathcal{T}-\mathcal{U}$	Attractions were tagged/endorsed by a user.
$\mathcal{A}-\mathcal{T}-\mathcal{A}-\mathcal{U}$	Attractions share tags/endorsements with other attractions that were rated by a user.
$\mathcal{A}-\mathcal{B}-\mathcal{A}-\mathcal{U}$	Attractions share business tokens with other attractions that were rated by a user.
$\mathcal{A}-\mathcal{W}-\mathcal{A}-\mathcal{U}$	Attractions share words on web page with other attractions that were rated by a user.
$\mathcal{A}-\mathcal{C}-\mathcal{A}-\mathcal{U}$	Attractions belong to the same category as other attractions that were rated by a user
$\mathcal{A}-\mathcal{E}-\mathcal{A}-\mathcal{U}$	Attractions mentioning the same entities as other attractions that were rated by a user

T2: Meta-Path Semantics

LTR Framework

- After HIN embeddings are trained for each meta-path, the similarity of objects within the HIN can be used as features in a learning to rank (LTR) framework.
- Since each of the meta-paths capture different semantics we decided to learn a parameter for each meta-path separately.

$$similarity(n_1, n_2|M) = cos(v^M_{n_1}, v^M_{n_2}) = \frac{v^M_{n_1} * v^M_{n_2}}{||v^M_{n_1}||_2 ||v^M_{n_2}||_2} \quad (1)$$
$$f(n_1, n_2) = \{similarity(n_1, n_2|M_i)\}, \forall i \in \{1 \dots N\} \quad (2)$$
$$F(a_i|r_i, u_i) = f(a_i, u_i), \forall a_i \in A^{candidates} \quad (3)$$

Experiments

Node Types	NDCG@5
$\{\mathcal{U},\mathcal{L},\mathcal{A}\}$	.2400(±.0005)
$\{\mathcal{U},\mathcal{L},\mathcal{A},\mathcal{T}\}$	.2565(±.0010)
$\{\mathcal{U},\mathcal{L},\mathcal{A},\mathcal{T},\mathcal{B}\}$	.2932(±.0006)
$\{\mathcal{U},\mathcal{L},\mathcal{A},\mathcal{T},\mathcal{B},\mathcal{W}\}$	.2986(±.0003)
$\{\mathcal{U},\mathcal{L},\mathcal{A},\mathcal{T},\mathcal{B},\mathcal{W},\mathcal{C}\}$	.3081(±.0004)
$\{\mathcal{U},\mathcal{L},\mathcal{A},\mathcal{T},\mathcal{B},\mathcal{W},\mathcal{C},\mathcal{E}\}$	**.3206(±.0003)**

T3: More Fine-grained Representations of Documents Improve Performance.

Top-k	TFIDF	χ^2	MI
10	**.3309(±.0004)**	.2900(±.0003)	.3176(±.0005)
50	.3157(±.0004)	.3183(±.0004)	.2982(±.0003)
100	.3246(±.0003)	.3105(±.0005)	.3159(±.0005)
500	.3294(±.0001)	.2937(±.0005)	.2996(±.0007)
1000	.3163(±.0006)	.3135(±.0006)	.3079(±.0002)

T4: Reduction of Graph Sparsity using Feature Selection Methods Improves Performance.

System	NDCG@5	P@5	MRR
DUTH_knn (debugged) [4]	.3388	.4690	.6697
This work	**.3309**	**.4476**	**.6475**
Laval_batch_3 [5]	.3281	.5069	.6501
USI5 [1]	.3265	.5069	.6796
bupt_pris_2016_cs.2_.4_max [11]	.2936	.4483	.6255
UAmsterdamDL [2]	.2824	.4448	.5924

T5: Comparison to Other Systems

REFERENCES

Jingbo Shang et al. *Meta-Path Guided Embedding for Similarity Search in Large-Scale Heterogeneous Information Networks (2016).*

Chuan Shi et al. A survey of heterogeneous information network analysis. IEEE Trans. Know. and Data Eng (2017).