A Markov Random Field Model for Entity-Relationship Retrieval

Pedro Saleiro, Nataša Milić-Frayling
Eduarda Mendes Rodrigues, Carlos Soares

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
University of Nottingham UK | CHINA | MALAYSIA

Entity-Relationship (E-R) Retrieval: given a query containing types of multiple entities and relationships connecting them, search for relevant tuples of related entities.

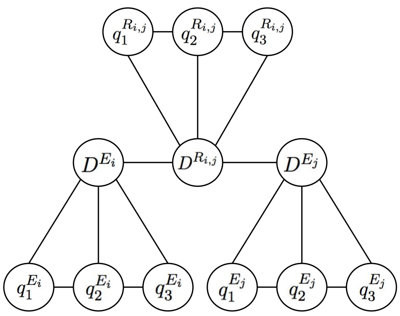
Example: *Silicon Valley companies founded by Harvard graduates* expects a list of tuples <*company, founder*> as results, such as <*Facebook, Mark Zuckerberg*>.

Problem: IR-centric approach to E-R retrieval without pre-defined entity types and relationships.

Approach: we propose the Entity-Relationship Dependence Model (ERDM) that models complex queries about entities that are connected through a relationship using the Markov Random Field model for retrieval.

Entity-Relationship Dependence Model (ERDM)

ERDM creates a composite model allowing the computation of a joint posterior of multiple documents given multiple queries, instead of one document given one query.



Suppose that we have a relationship query of the format $Q = \{Q^{E_i}, Q^{R_{i,j}}, Q^{E_j}\}$ we want to rank a relationship document $D^{R_{i,j}}$ and two entity documents D^{E_i} and D^{E_j} by descending order of the following joint posterior:

$$\begin{aligned} & P_\Lambda(D^R, D^E|Q) \\ & \stackrel{rank}{=} \log P_\Lambda(D^{R_{i,j}}, D^{E_i}, D^{E_j}, Q^{R_{i,j}}, Q^{E_i}, Q^{E_j}) \\ & \stackrel{rank}{=} \log \prod_{c \in C(G)} \psi(c; \Lambda) \\ & \stackrel{rank}{=} \sum_{c \in C(G)} \log \exp[\lambda_c f(c)] \\ & \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f(c) \end{aligned} \quad (1)$$

(2)

The potential functions are computed for five types of cliques: $\psi(Q^{E_i}, D^{E_i}; \Lambda)$; $\psi(Q^{E_j}, D^{E_j}; \Lambda)$; $\psi(Q^R, D^R; \Lambda)$; $\psi(D^{E_i}, D^R; \Lambda)$; $\psi(D^{E_j}, D^R; \Lambda)$;

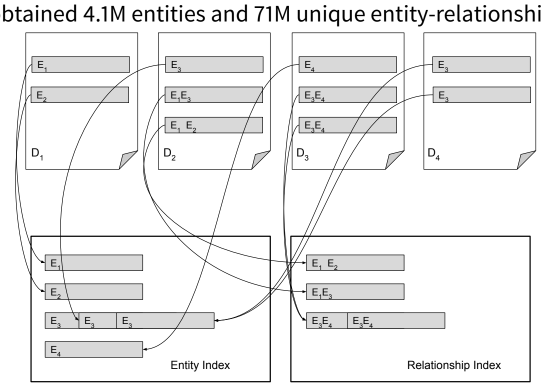
We use unigram and bi-gram Language Models as feature functions for every clique with a query and a document node.
The potential function for the 2-cliques composed by an entity document and a relationship document is the following:

$$f_T^{ER}(D^{E_i}, D^{R_{i,j}}) = \left[(1-\alpha) tf_{\{1,0\},(D^{E_i},D^{R_{i,j}})} + \alpha \frac{df^R_{D^{E_i}}}{df^R}\right] \quad (3)$$

where tf indicates whether an entity E_i is present in the relationship document $D^{R_{i,j}}$ or not. The background model employs the notion of entity frequency as the following: $df^R_{D^{E_i}}$ is the total number of relationship documents containing the entity E_i and df^R is the entity-pair frequency in the relationship corpus.
Learning to rank is performed using the Coordinate Ascent algorithm under the sum normalization and non-negativity constraints.

Data and Indexing

E-R retrieval requires collecting evidence for both entities and relationships that can be spread across multiple documents.
Our design pattern basically can be thought as creating a meta-document D^{E_i} for each entity, as well as, a meta-document $D^{R_{i,j}}$ for each entity-pair (relationship).
These meta-documents are created by extracting entity and entity-pairs contexts from the corpus of raw documents. For each raw document D we extract entity or entity-pair associated terms.
We use ClueWeb-09-B corpus with FACC1 entity linking as dataset.
We obtained 4.1M entities and 71M unique entity-relationships.



Test Collections

Collection	Amount	Example NL query	Example relational format
ERQ	28	*Find novels written by Jane Austen.*	{*novel, written by, Jane Austen*}
COMPLEX	60	*Economists influenced by Karl Marx*	{*Economist, influenced by , Karl Marx*}
RELink	100	*Dog breeds and country of origin*	{*dog breed, original from, country*}

Results

	ERQ			
	MAP	P@10	MRR	NDCG@10
BaseE	0.0469	0.0109	0.0489	0.038
BaseR	0.1041	0.0509	0.1089	0.1104
ERDM	**0.3107**	0.1903	0.3761	0.3175
	COMPLEX			
	MAP	P@10	MRR	NDCG@10
BaseE	0.0264	0.005	0.0318	0.1223
BaseR	0.0585	0.0184	0.0748	0.0778
ERDM	**0.2879**	0.1417	0.3296	0.3323
	RELink			
	MAP	P@10	MRR	NDCG@10
BaseE	0.0395	0.019	0.0679	0.0395
BaseR	0.0451	0.021	0.0663	0.0726
ERDM	**0.1249**	0.048	0.1726	0.1426