Higher-order Relation Schema Induction using Tensor Factorization with Back-off and Aggregation

Madhav Nimishakavi[1] Manish Gupta[2] Partha Talukdar[1]

[1] Indian Institute of Science (IISc), Bangalore [2] Microsoft

Objective

Higher-order Relation Schema Induction (**HRSI**) is the task of identifying schemas of n-ary relations from a set of documents specific to a given domain, without any supervision.

Corpus	Induced Schemas
"..Federer won Wimbledon finals against Nadal in 2007.."	<player> win <championship> <opponent> <time>
"... Serena williams will play Muguruza on saturday in the french open.."	<player> play <opponent> <time> <tournament>

Approach

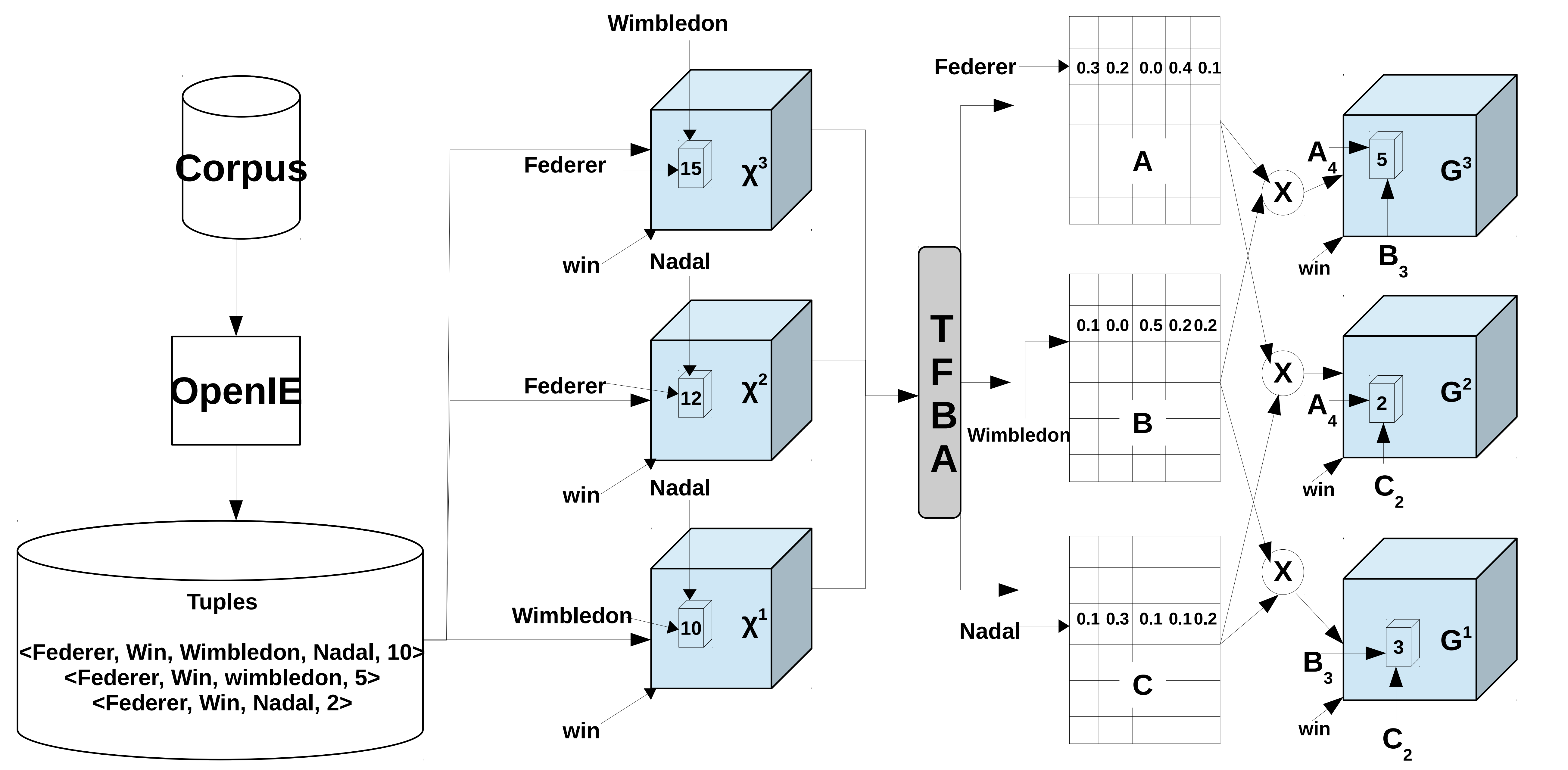
1. Open IE Tensor: Extract surface level triples from documents using OpenIE including the Location and Time arguments.

2. Back-off : Instead of contsructing one very high dimensional tensor from the triples, construct three 3-d tensors. In all the three tensors mode-3 corresponds to relations.

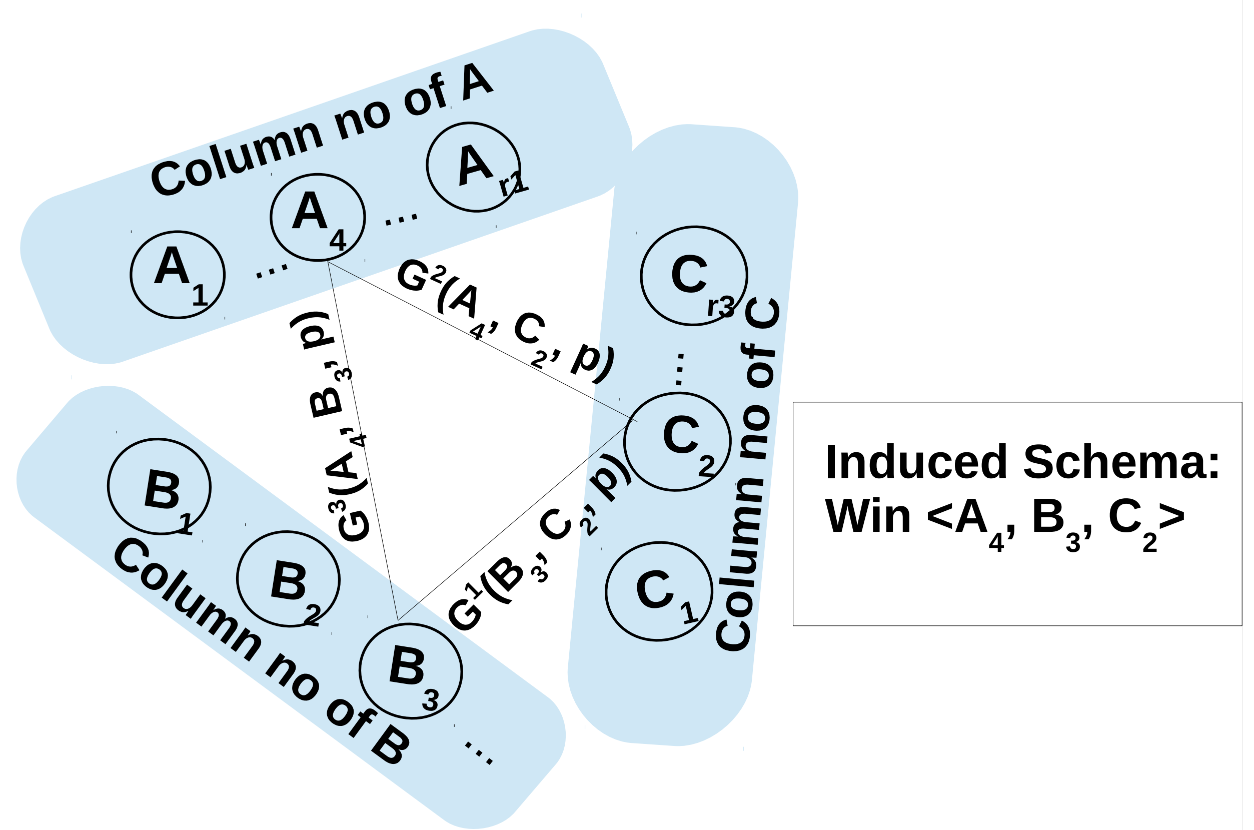
3. Factorization: Perform joint factorization of all the three tensors with non-negative constraints.

4. Constrained Clique Mining : Construct schemas from the columns of factor matrices by mining constrained cliques.

Tensor Factorization with Back-off and Aggregation (TFBA)



Schema Construction



A_4 : <Federer, Djokovic, Nadal , ...>
B_3 : <Wimbledon, Rogers Cup, RolandGarros , ...>
C_2 : <London, Paris, Montreal, ...>

Model

$$\min_{A,B,C,\mathcal{G}} f(\mathcal{X}^3,\mathcal{G}^3,A,B)+f(\mathcal{X}^2,\mathcal{G}^2,A,C)$$
$$+\, f(\mathcal{X}^1,\mathcal{G}^1,B,C)$$
$$+\,\lambda_a \parallel A \parallel_F^2 + \lambda_b \parallel B \parallel_F^2 + \lambda_c \parallel C \parallel_F^2$$

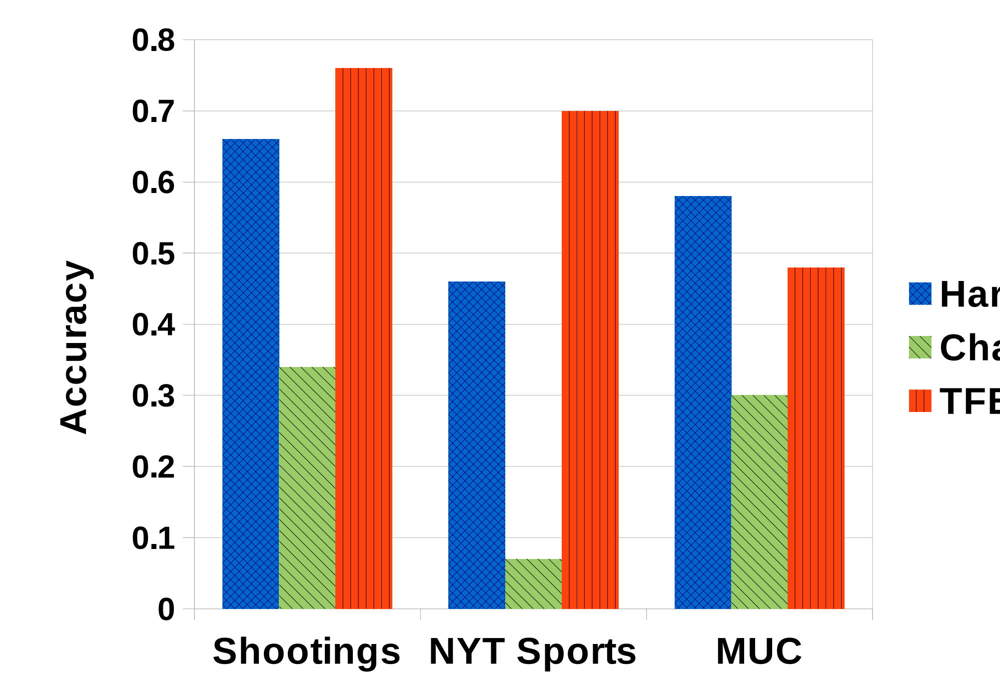
where,

$$f(\mathcal{X}^i,\mathcal{G}^i,P,Q) = \parallel \mathcal{X}^i - \mathcal{G}^i \times_1 P \times_2 Q \times_3 I \parallel_F^2$$
$$A\in\mathbb{R}_+^{n_1\times r_1},\, B\in\mathbb{R}_+^{n_2\times r_2},\, C\in\mathbb{R}_+^{n_3\times r_3}$$
$$\mathcal{G}^1\in\mathbb{R}_+^{r_2\times r_3\times m},\, \mathcal{G}^2\in\mathbb{R}_+^{r_1\times r_3\times m},\, \mathcal{G}^3\in\mathbb{R}_+^{r_1\times r_2\times m}$$

Schemas Induced by TFBA

Schemas Induced	Valid?
Shootings	
<shooting, shooting incident, double shooting> leave <one person, two people, three people> <dead, injured, on edge>	✓
<police, officers, huntsville police> identify<man, victims, four victims> <sunday, shooting saturday, wednesday afternoon> <apartment, bedroom, building in the neighborhood>	✓
<police, officers, huntsville police> say<man, victims, four victims> <sunday, shooting saturday, wednesday afternoon>	✗
NYT Sports	
<yankees, mets, jets> spend <$ num million, $ num, $ num billion> <last season, year num>	✓
<red sox, yankees, mets> win <world series, title, world cup> <last season, year num>	✓
<umpire, mike cameron, andre agassi> get <ball, lives, grounder> <back, forward, num-yard line>	✗

Results



Codes

Takeaway

1. Mining schemata for relations is an important problem.
2. So far, the focus has been only on binary relations in the literature.
3. To the best of our knowledge, this is the first ever attempt to mine schemata for n-ary relations.