Paper ID : 548

Entity Matching across Heterogeneous Sources

Yang Yang[1], Yizhou Sun[2], Jie Tang[1], Bo Ma[1], and Juanzi Li[1]

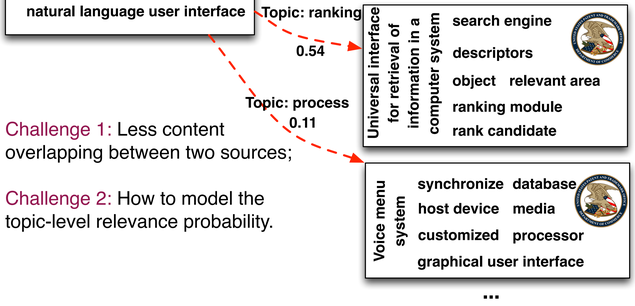
[1]Department of Computer Science and Technology, Tsinghua University
[2]College of Computer and Information Science, Northeastern University

{sherlockbourne, mabox}@gmail.com, {jietang, lijuanzi}@tsinghua.edu.cn, yzsun@cs.neu.edu

KDD2015

Given an entity from a source domain, how to find its matched entities from target domain?

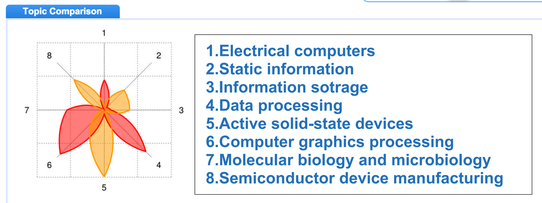
Product-patent matching: core problem behind Apple VS Samsung



Challenge 1: Less content overlapping between two sources;

Challenge 2: How to model the topic-level relevance probability.

Application: Competitor Analysis (pminer.org)



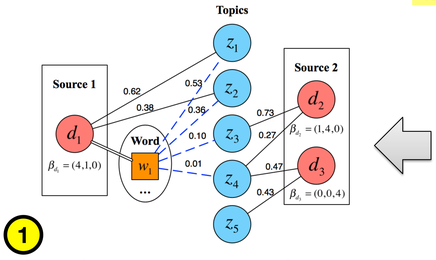
Proposed Model

Baseline method (CS+LDA, RW+LDA)

1. Learning LDA on the source domain and target domain respectively.
2. Given an entity, ranking others as candidates according to the topic similarity.

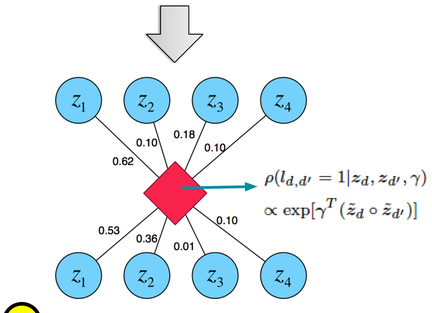
Integrate topic extraction and entity matching into a unfired framework

Proposed model (CST)



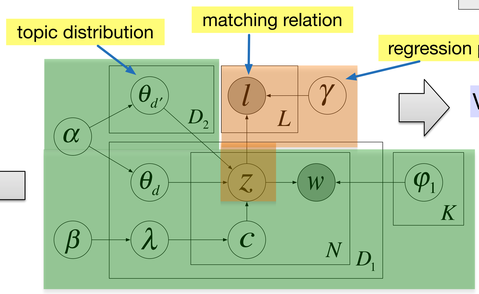
1

Cross-Sampling-Based Entity Generation



2

Matching Relation Generation



Variational inference for model learning

E-Step: update variational parameters:

$$\eta_{d,c} = \beta_{d,c} + N_d \times \epsilon_{d,c}$$

$$\tau_{d,k} = \alpha_k + \sum_{n=1}^{N_d} \vartheta_{d,n,k}$$

$$\epsilon_{d,n,c} \propto \exp\{\Psi(\eta_{d,c}) - \Psi(\sum_{i \in R(d)} \eta_{d,i})\}$$

$$\vartheta_{d,n,k} \propto \sum_{d' \in \{R(d),d\}} (\exp\{\sum_{d'' \neq d'} \frac{\gamma_k \sum_{i=1}^{N_{d''}} \vartheta_{d'',i,k}}{N_{d'} N_{d''}} + \Psi(\tau_{d',k}) - \Psi(\sum_{j=1}^{K} \tau_{d',j})\} \epsilon_{d,n,d'} \times \varphi_{t,k,v}$$

M-Step: update model parameters:

$$\varphi_{t,k,v} \propto \sum_{d=1}^{D_t} \sum_{n=1}^{N_d} \vartheta_{d,n,k} \mathbf{1}(w_{d,n}^t = v)$$

$$\gamma_k = \frac{\sum_{d,d'} 1}{2 \sum_{d,d'} l_{d,d'} [(\Upsilon_d - \Upsilon_{d'}) \circ (\Upsilon_d - \Upsilon_{d'})]_k}$$

Cross-Source Topic Model

Input: a dual source corpus C, a matching relation matrix L, and hyper-parameters α and β
foreach *entity* d **do**
  Generate $\theta_d \sim \text{Dir}(\alpha)$;
end
% cross-sampling-based entity generation
foreach d *in each source* t **do**
  Set β according to L_d;
  Generate $\lambda_d \sim \text{Dir}(\beta)$;
  for $n = 1$ *to* N_d **do**
    Generate $c_{d,n} \sim \text{Mult}(\lambda_d)$, $c_{d,n}$ can be d or the index of matched entities with d;
    Draw a topic $z_{d,n} \sim \text{Mult}(\theta_{c_{d,n}})$ from the topic distribution of the entity c;
    Draw a word $w_{d,n} \sim \text{Mult}(\varphi_{t,z_{d,n}})$ from $z_{d,n}$-specific word distribution;
  end
end
% matching relation generation
foreach (d, d') *with possible links* **do**
  Generate $l_{d,d'} \sim \rho(\cdot | z_d, z_{d'}, \gamma)$;
end

Experimental Results

Product-Patent Matching

Task: given a Wiki article describing a product, finding all relevant patents.

Dataset:
- 13,085 Wiki articles;
- 15,00 patents from USPTO;
- 1,060 matching relations in total.

Method	P@3	P@20	MAP	R@3	R#20	MRR
CS+LDA	0.111	0.083	0.109	0.011	0.046	0.053
RW+LDA	0.111	0.117	0.123	0.033	0.233	0.429
RTM	0.501	0.233	0.416	0.057	0.141	0.171
RW+CST	**0.667**	0.167	0.341	**0.200**	0.333	0.668
CST	**0.667**	**0.250**	**0.445**	0.171	**0.457**	**0.683**

Content Similarity based on LDA (CS+LDA): cosine similarity between two articles' topic distribution extracted by LDA;
Random Walk based on LDA (RW+LDA): random walk on a graph where edges indicate the topic similarity between articles;
Relational Topic Model (RTM): used to model links between documents;
Random Walk based on CST (RW+CST): uses CST instead of LDA comparing with RW+LDA.

Cross-lingual Matching

Task: given an English Wiki article, finding all Chinese article reporting the same content.

Dataset:
- 2,000 English articles from Wikipedia;
- 2,000 Chinese articles from Baidu Baike;
- Each English article corresponds to one Chinese article.

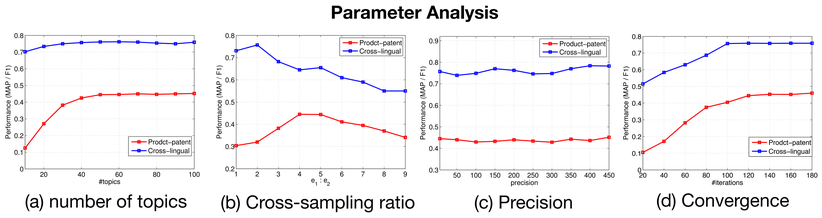
Method	Precision	Recall	F1-Measure	F2-Measure
Title Only	**1.000**	0.410	0.581	0.465
SVM-S	0.957	0.563	0.709	0.613
LFG	0.661	0.820	0.732	0.782
LFG+LDA	0.652	0.805	0.721	0.769
LFG+CST	0.682	**0.849**	**0.757**	**0.809**

Title Only: only considers the (translated) title of articles.
SVM-S: famous cross-lingual Wikipedia matching toolkit.
LFG: mainly considers the structural information of Wiki articles.
LFG+LDA: adds content feature (topic distributions) to LFG by LDA.
LFG+CST: adds content feature to LFG by employing CST.

Topics relevant to both Apple and Samsung

Title	Top Patent Terms	Top Wiki Terms
Gravity Sensing	rotational, gravity, interface, sharing, frame, layer	gravity, iPhone, layer, video, version, menu
Touchscreen	recognition, point, digital, touch, sensitivity, image	screen, touch, iPad, os, unlock, press
Application Icons	interface, range, drives, icon, industrial, pixel	icon, player, software, touch, screen, application

Parameter Analysis



(a) number of topics (b) Cross-sampling ratio (c) Precision (d) Convergence