Entity Matching across Heterogeneous Sources

Yang Yang*, Yizhou Sun+, Jie Tang*, Bo Ma#, and Juanzi Li*

*Tsinghua University +Northeastern University #Carnegie Mellon University

Data&Code available at: http://arnetminer.org/document-match/

Apple Inc. VS Samsung Co.

- A patent infringement suit starts from 2012.
  - Lasts 2 years, involves $158+ million and 10 countries.
  - 7 / 35546 patents are involved.

Apple's patent

How to find **patents** relevant
to a specific **product**?

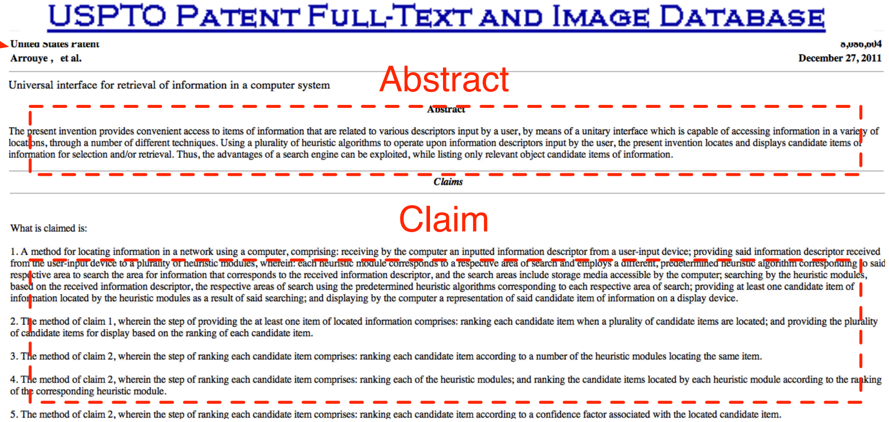
Galaxy S II Skyrocket	X*	X	X*	X	X
Galaxy S III		X	X*		X
Galaxy Tab 2 (10.1)		X	X*		X
Stratosphere	X*	X	X*	X*	X

SAMSUNG devices accused by APPLE.

Cross-Source Entity Matching

- Given an entity in a **source** domain, we aim to find its matched entities from **target** domain.
  - Product-patent matching;
  - Cross-lingual matching;
  - Drug-disease matching.

Siri


USPTO Patent Full-Text and Image Database

United States Patent
Arrouye , et al.
December 27, 2011

Universal interface for retrieval of information in a computer system

Abstract

The present invention provides convenient access to items of information that are related to various descriptors input by a user, by means of a unitary interface which is capable of accessing information in a variety of locations, through a number of different techniques. Using a plurality of heuristic algorithms to operate upon information descriptors input by the user, the present invention locates and displays candidate items of information for selection and/or retrieval. Thus, the advantages of a search engine can be exploited, while listing only relevant object candidate items of information.

Claims

What is claimed is:

1. A method for locating information in a network using a computer, comprising: receiving by the computer an inputted information descriptor from a user-input device; providing said information descriptor received from the user-input device to a plurality of heuristic modules, wherein: each heuristic module corresponds to a respective area of search and employs a different, predetermined heuristic algorithm corresponding to said respective area to search the area for information that corresponds to the received information descriptor, and the search areas include storage media accessible by the computer; searching by the heuristic modules, based on the received information descriptor, the respective areas of search using the predetermined heuristic algorithms corresponding to each respective area of search; providing at least one candidate item of information located by the heuristic modules as a result of said searching; and displaying by the computer a representation of said candidate item of information on a display device.

2. The method of claim 1, wherein the step of providing the at least one item of located information comprises: ranking each candidate item when a plurality of candidate items are located; and providing the plurality of candidate items for display based on the ranking of each candidate item.

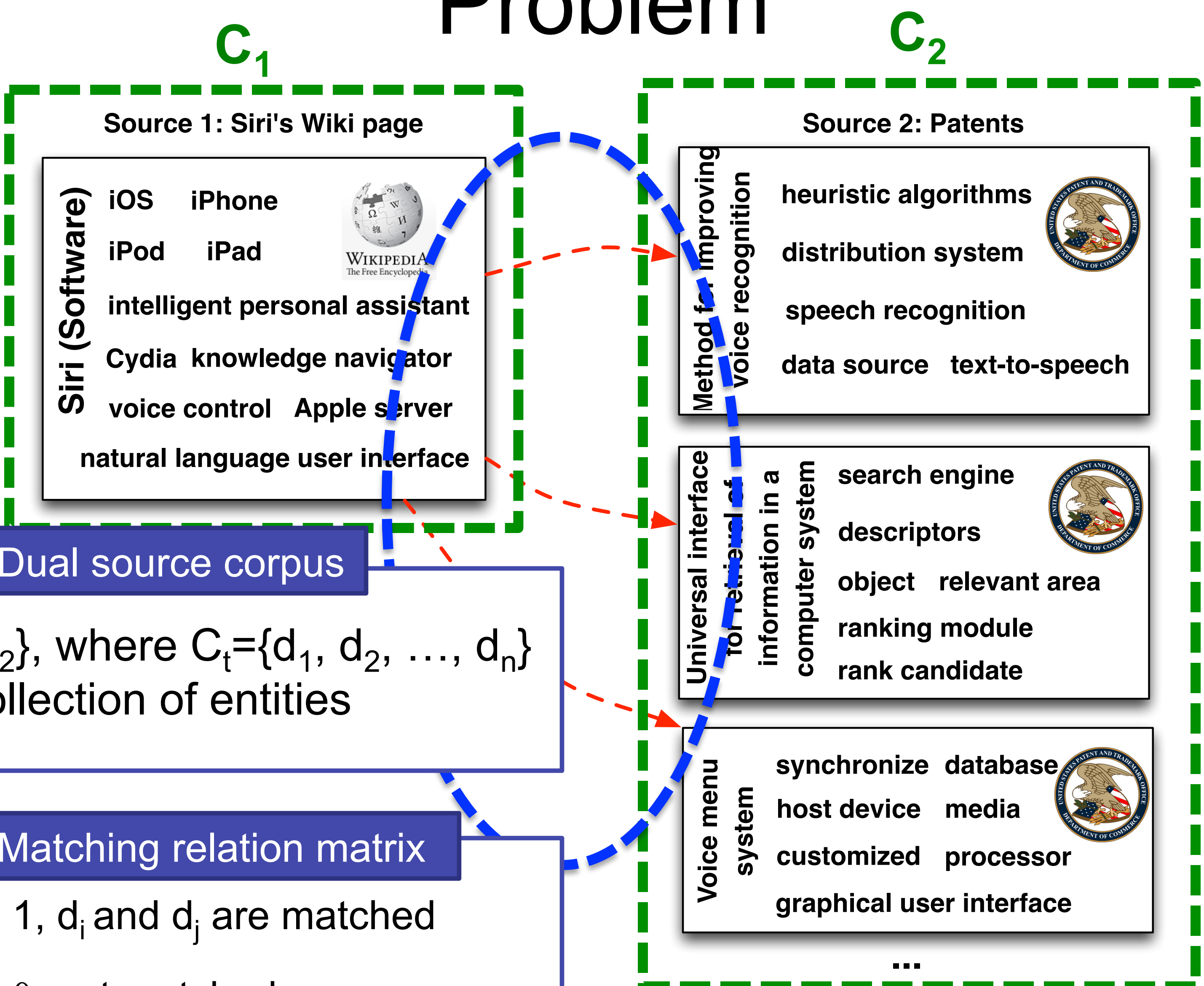
3. The method of claim 2, wherein the step of ranking each candidate item comprises: ranking each candidate item according to a number of the heuristic modules locating the same item.

4. The method of claim 2, wherein the step of ranking each candidate item comprises: ranking each of the heuristic modules; and ranking the candidate items located by each heuristic module according to the ranking of the corresponding heuristic module.

5. The method of claim 2, wherein the step of ranking each candidate item comprises: ranking each candidate item according to a confidence factor associated with the located candidate item.

Product-Patent matching

Problem



Input 1: Dual source corpus

$\{C_1, C_2\}$, where $C_t=\{d_1, d_2, \ldots, d_n\}$ is a collection of entities

Input 2: Matching relation matrix

$$L_{ij} = \begin{cases} 1, & d_i \text{ and } d_j \text{ are matched} \\ 0, & \text{not matched} \\ ?, & \text{unknown} \end{cases}$$

Challenges

Source 1: Siri's Wiki page

Siri (Software)

iOS iPhone
iPod iPad

WIKIPEDIA
The Free Encyclopedia

intelligent personal assistant
Cydia knowledge navigator
voice control Apple server
natural language user interface

Source 2: Patents

Method for improving voice recognition

heuristic algorithms
distribution system
speech recognition
data source text-to-speech

Universal interface for retrieval of information in a computer system

search engine
descriptors
object relevant area
ranking module
rank candidate

Voice menu system

synchronize database
host device media
customized processor
graphical user interface

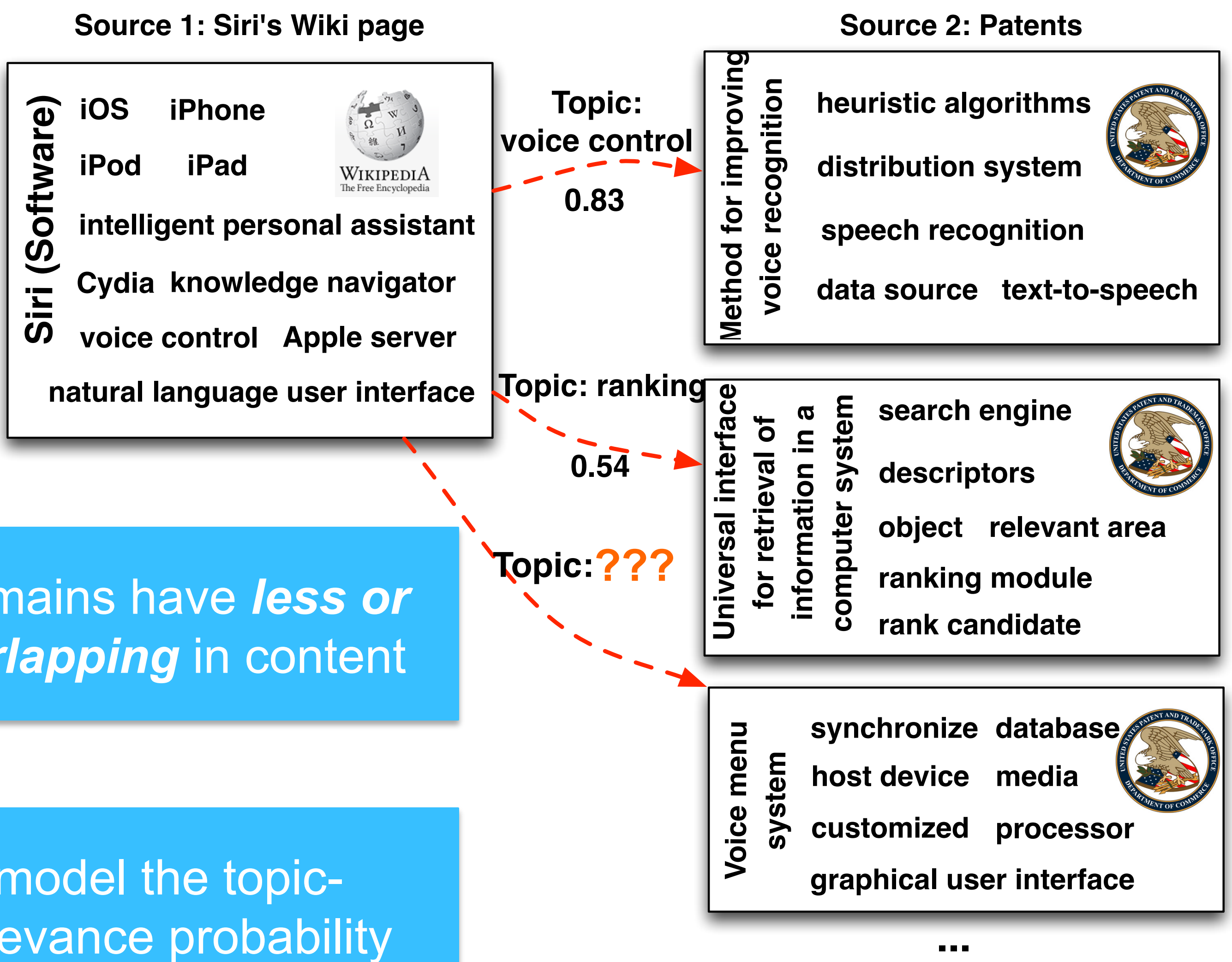
...

1

Two domains have ***less or no overlapping*** in content

Daily expression
vs
Professional expression

Challenges



1

Two domains have ***less or no overlapping*** in content

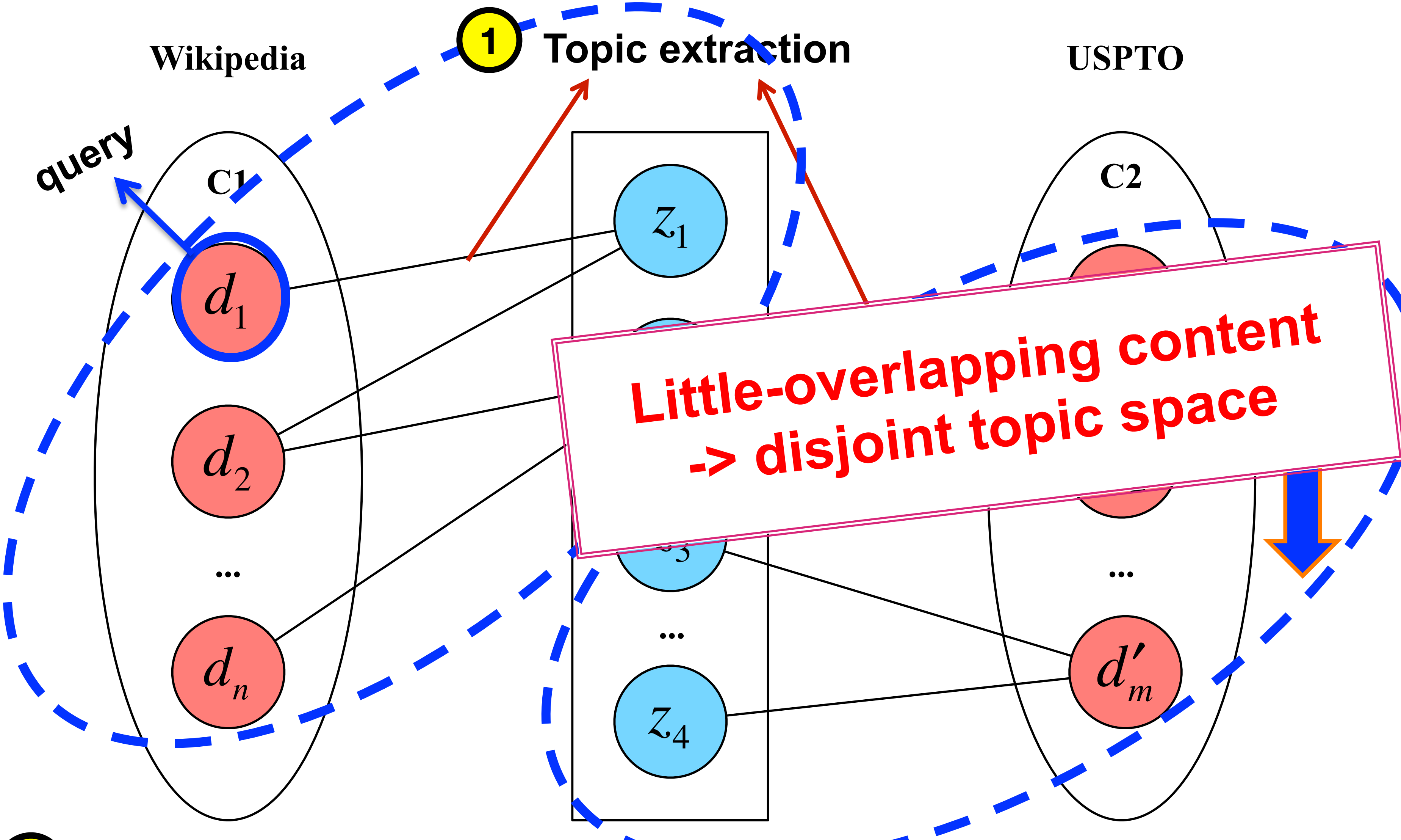
2

How to model the topic-level relevance probability

Our Approach

Cross-Source Topic Model

Baseline

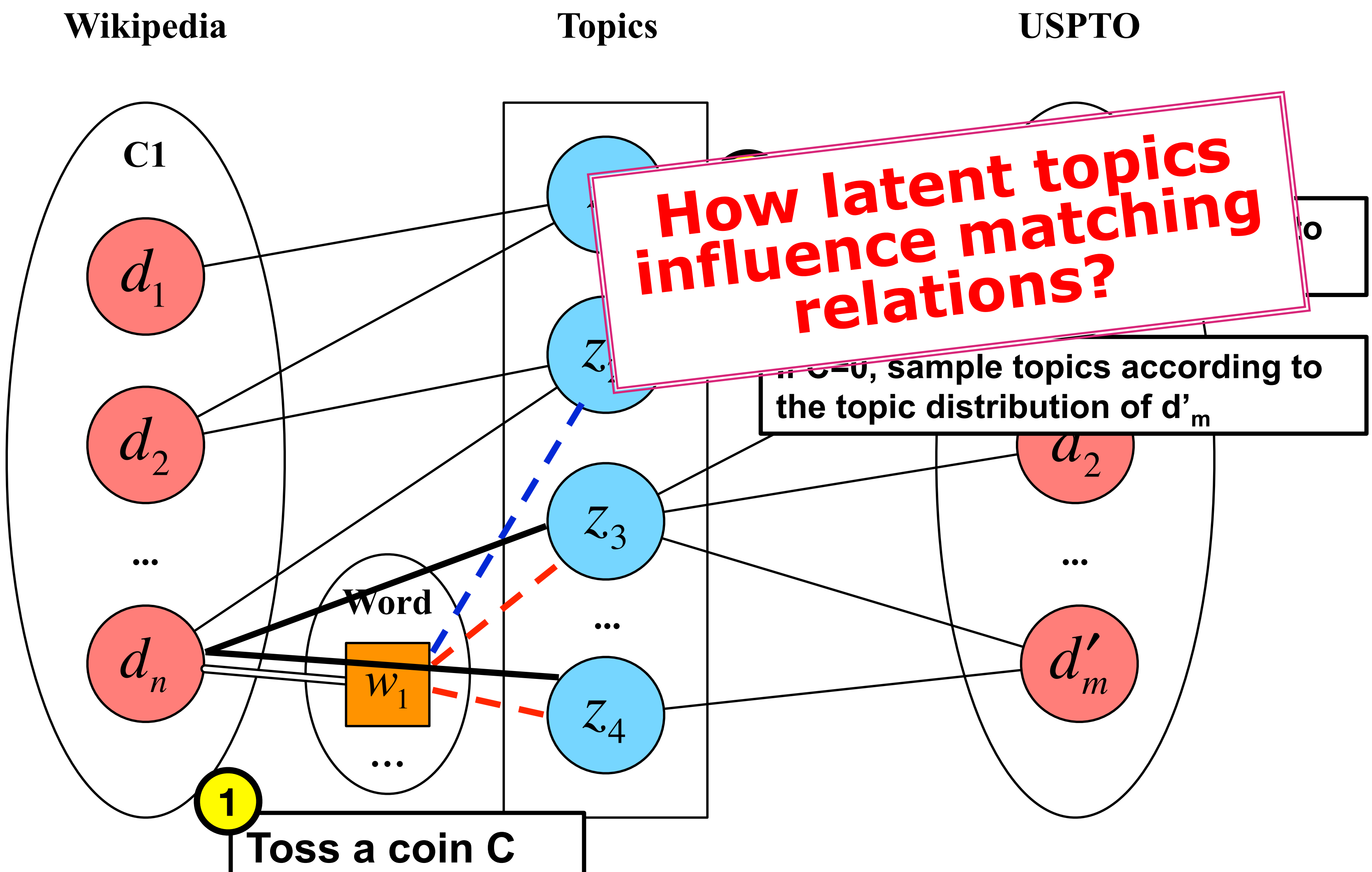


2 Ranking candidates by topic similarity $Sim(d_1, d_2) = \frac{p_{d_1} \cdot p_{d_2}}{||p_{d_1}|| \times ||p_{d_2}||}$

Cross-Sampling

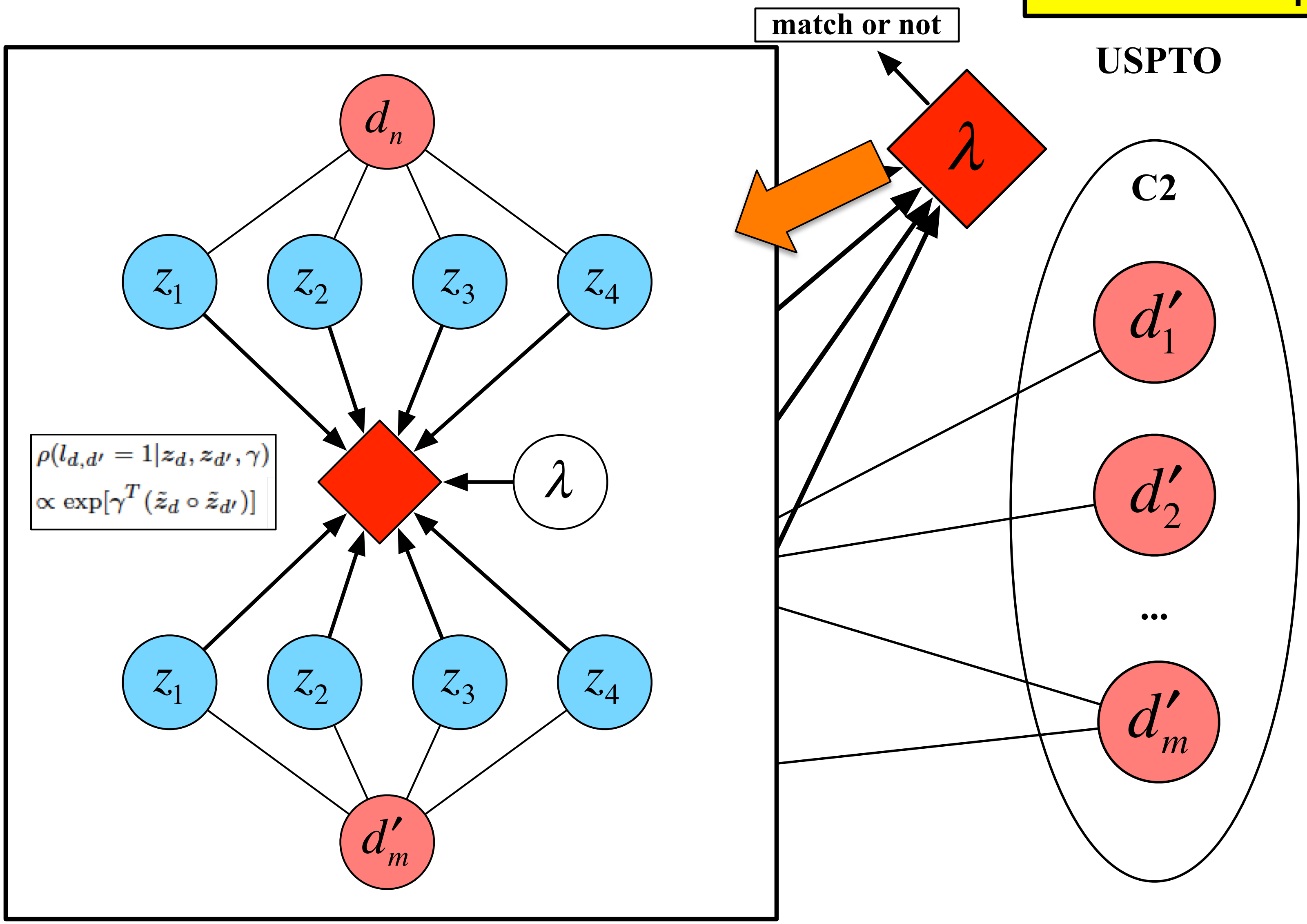
d_n is matched with d'_m

Bridge topic space by leveraging known matching relations.

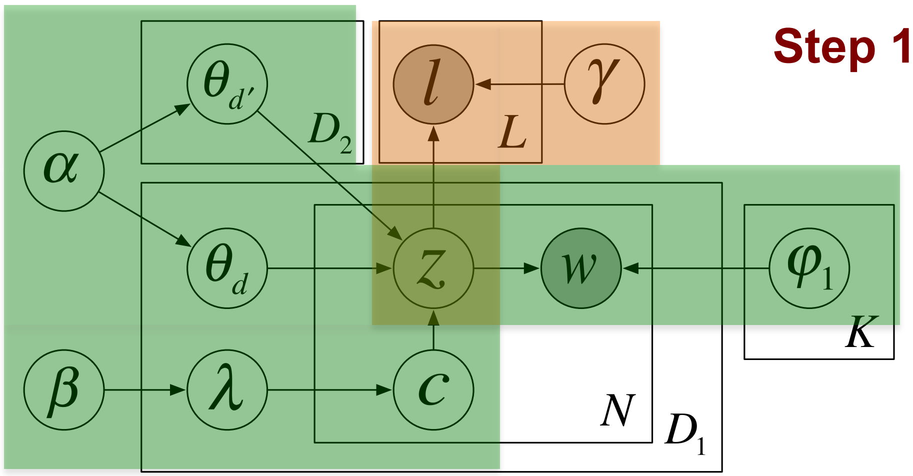


Inferring Matching Relation

Infer matching relations by leveraging extracted topics.



Cross-Source Topic Model



Step 1:

Step 2:

Input: a dual source corpus C, a matching relation matrix L,
and hyper-parameters α and β

foreach *entity* d **do**
  Generate $\theta_d \sim \mathrm{Dir}(\alpha)$;
end
% cross-sampling-based entity generation
foreach d *in each source* t **do**
  Set β according to L_d;
  Generate $\lambda_d \sim \mathrm{Dir}(\beta)$;
  for $n = 1$ *to* N_d **do**
    Generate $c_{d,n} \sim \mathrm{Mult}(\lambda_d)$, $c_{d,n}$ can be d or the index of matched entities with d;
    Draw a topic $z_{d,n} \sim \mathrm{Mult}(\theta_{c_{d,n}})$ from the topic distribution of the entity c;
    Draw a word $w_{d,n} \sim \mathrm{Mult}(\varphi_{t,z_{d,n}})$ 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

Latent topics ⟷ Matching relations

Model Learning

- Variational EM
  - Model parameters: $\{\varphi, \gamma\}$
  - Variational parameters: $\{\vartheta, \tau, \eta, \epsilon\}$
  - E-step:

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

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

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

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

  - M-step:

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

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

Input: a dual source corpus C, a matching relation matrix L, and hyper-parameters α and β
Initialize $\{\vartheta, \tau, \eta, \epsilon, \varphi, \gamma\}$ randomly;
repeat
  % E-Step: optimize the ELBO;
  foreach *d in each source t* **do**
    for $c = 0$ *to* 1 **do**
      Update $\eta_{d,c}$ according to Eq. 6;
    end
    for $k = 1$ *to* K **do**
      Update $\tau_{d,k}$ according to Eq. 7;
    end
    for $n = 1$ *to* N_d **do**
      for $c = 0$ *to* 1 **do**
        Update $\epsilon_{d,n,c}$ according to Eq. 8;
      end
      for $k = 1$ *to* K **do**
        Update $\vartheta_{d,n,k}$ according to Eq. 9;
      end
    end
  end
  % M-Step: maximize the resulting ELBO;
  foreach *topic k in each source t* **do**
    foreach *term v* **do**
      Update φ according to Eq. 10;
    end
    Update γ_k according to Eq. 11;
  end
until *Convergence*;

Experiments

Task I: Product-patent matching

Task II: Cross-lingual matching

Task I: Product-Patent Matching

- Given a Wiki article describing a product, finding all patents relevant to the product.
- Data set:
  - 13,085 Wiki articles;
  - 15,000 patents from USPTO;
  - 1,060 matching relations in total.

Experimental Results

Training: 30% of the matching relations randomly chosen.

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 entities' topic distribution extracted by LDA.

Random Walk based on LDA (RW+LDA): random walk on a graph where edges indicate the hyperlinks between Wiki articles and citations between patents.

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.

Task II: Cross-lingual Matching

- Given an English Wiki article, we aim to find a Chinese article reporting the same content.
- Data set:
  - 2,000 English articles from Wikipedia;
  - 2,000 Chinese articles from Baidu Baike；
  - Each English article corresponds to one Chinese article.

Experimental Results

Training: 3-fold cross validation

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[1]: mainly considers the structural information of Wiki articles.

LFG+LDA: adds content feature (topic distributions) to LFG by employing LDA.

LFG+CST: adds content feature to LFG by employing CST.

[1] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. Cross-lingual Knowledge Linking Across Wiki Knowledge Bases. WWW'12. pp. 459-468.

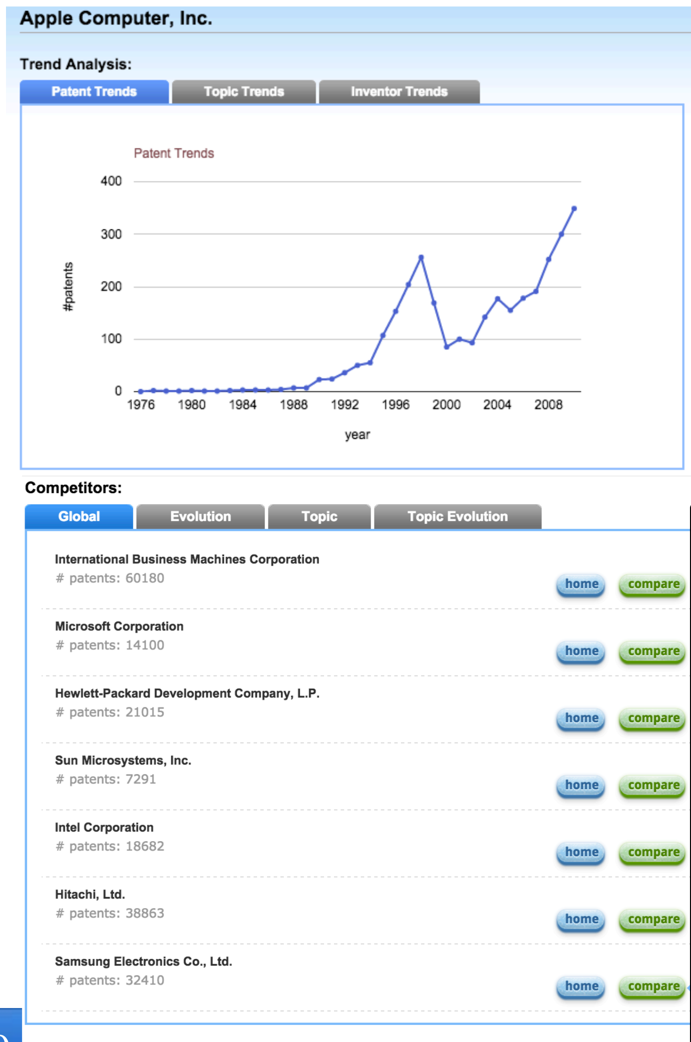
Topics Relevant to Apple and Samsung

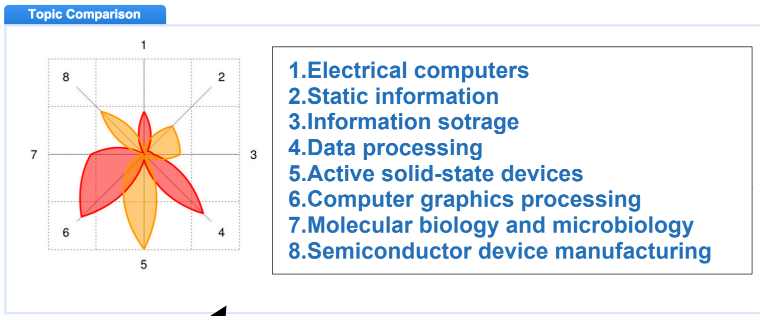
(Topic titles are hand-labeled)

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

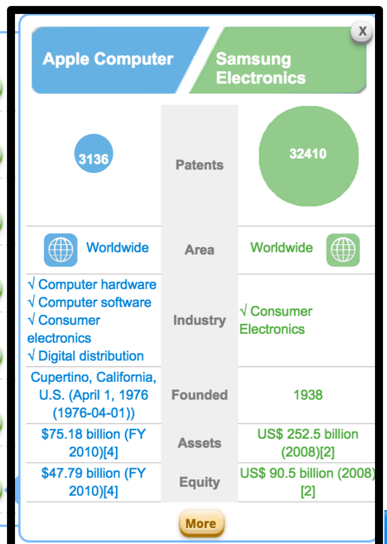
Prototype System

competitor analysis @ http://pminer.org





Radar Chart: topic comparison



Basic information comparison: #patents, business area, industry, founded year, etc.

Conclusion

- Study the problem of **entity matching across heterogeneous sources**.
- Propose the cross-source topic model, which integrates the **topic extraction** and **entity matching** into a unified framework.
- Conduct two experimental tasks to demonstrate the effectiveness of CST.

Thank You!

Entity Matching across Heterogeneous Sources

Yang Yang[*], Yizhou Sun[+], Jie Tang[*], Bo Ma[#], and Juanzi Li[*]

[*]Tsinghua University [+]Northeastern University [#]Carnegie Mellon University

Data&Code available at: http://arnetminer.org/document-match/

Apple Inc. VS Samsung Co.

- A patent infringement lawsuit starts from 2012.
  - Nexus S, Epic 4G, Galaxy S 4G, and the Samsung Galaxy Tab, infringed on Apple's intellectual property: its patents, trademarks, user interface and style.
  - Lasts over **2 years**, involves **$158+ million**.
- How to find patents relevant to a specific product?

Problem

- Given an entity in a **source** domain, we aim to find its matched entities from **target** domain.
  - Given a textural description of a **product**, finding related **patents** in a patent database.
  - Given an **English Wiki page**, finding related **Chinese Wiki pages**.
  - Given a specific **disease**, finding all related **drugs**.

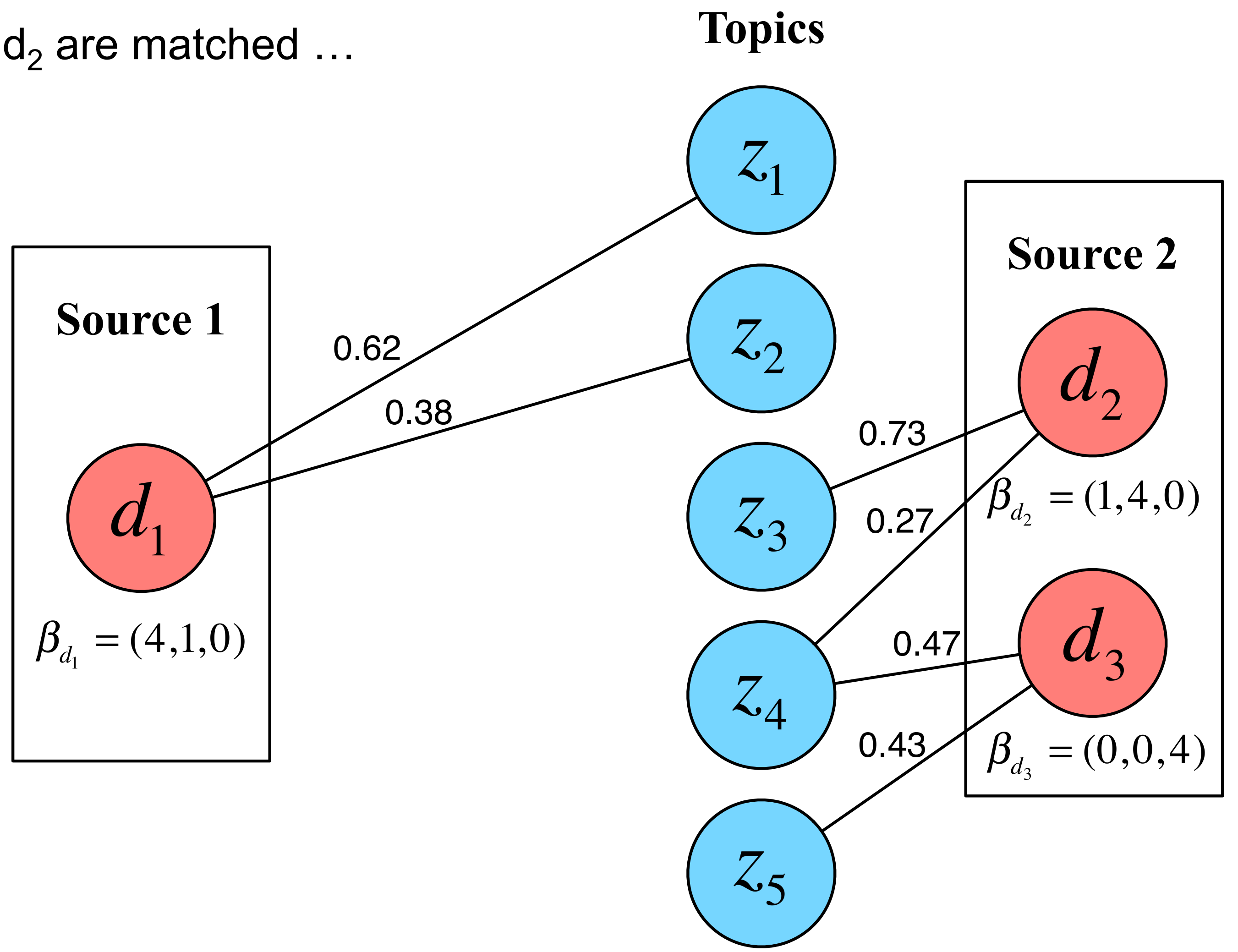
Basic Assumption

- For entities from different sources, their **matching relations** and **hidden topics** are influenced by each other.
- How to leverage the known matching relations to help **link** hidden topic spaces of two sources?

Cross-Sampling

1

d_1 and d_2 are matched …

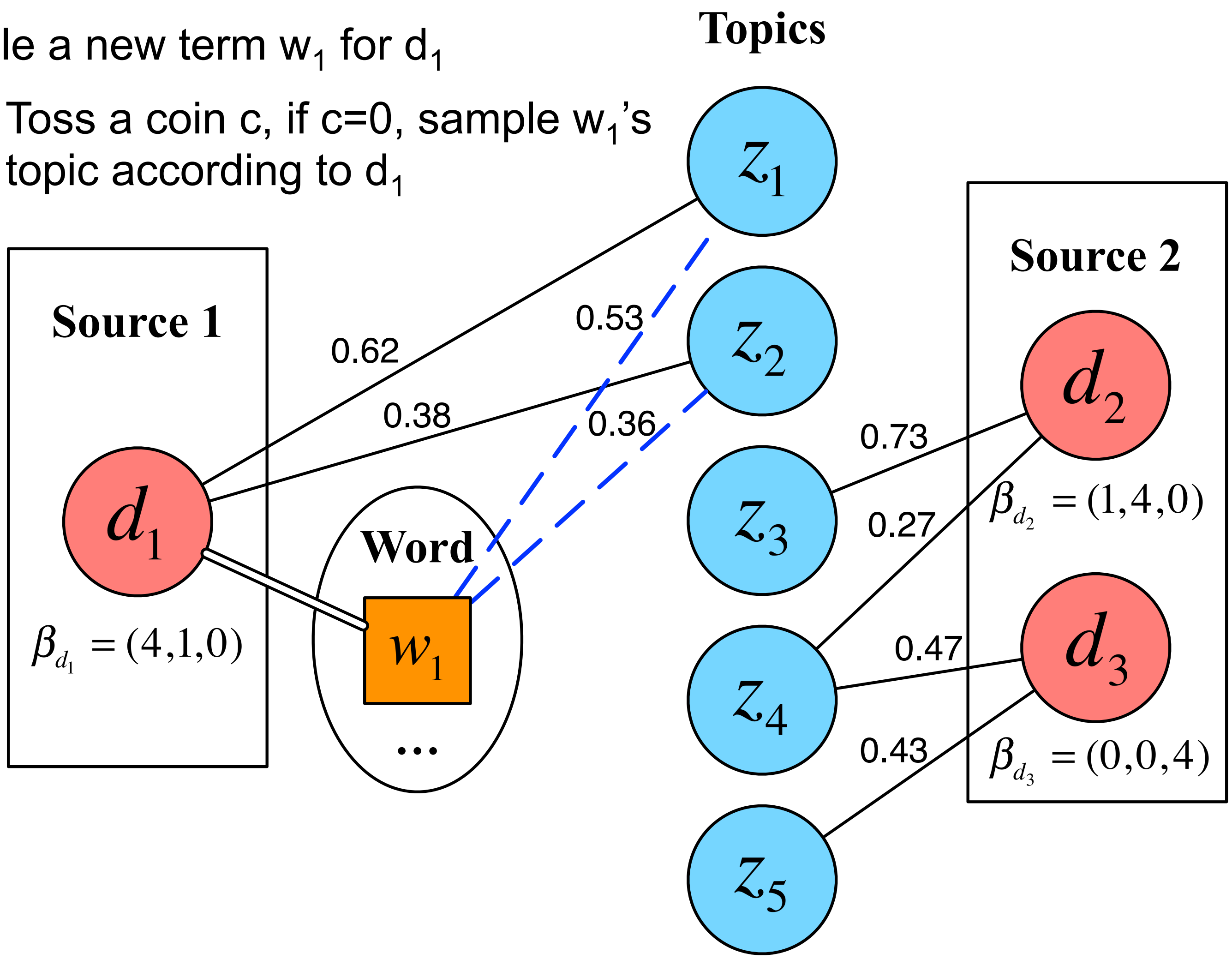


Cross-Sampling

2

Sample a new term w_1 for d_1

Toss a coin c, if c=0, sample w_1's topic according to d_1

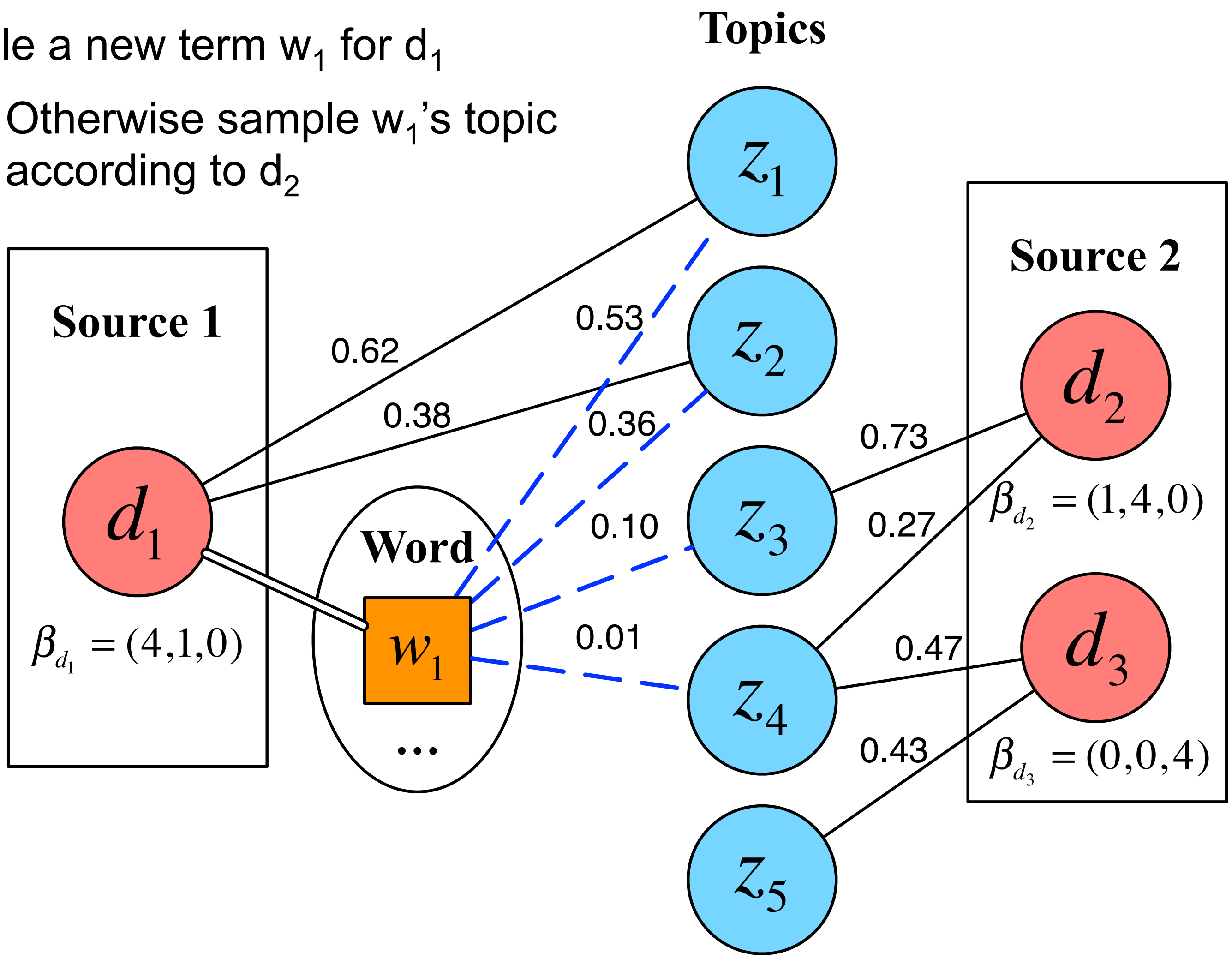


Cross-Sampling

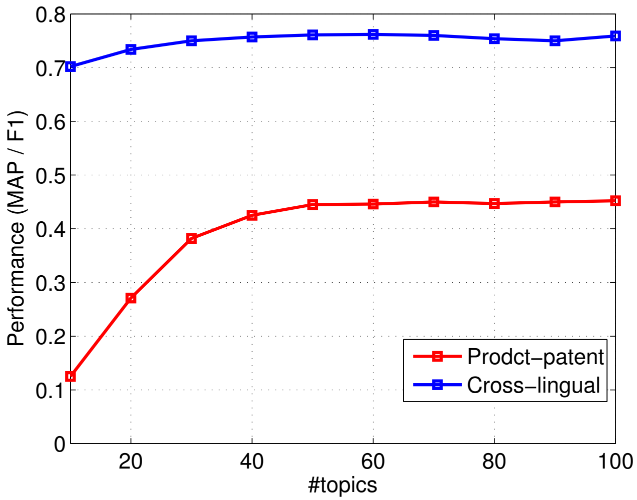
3

Sample a new term w_1 for d_1

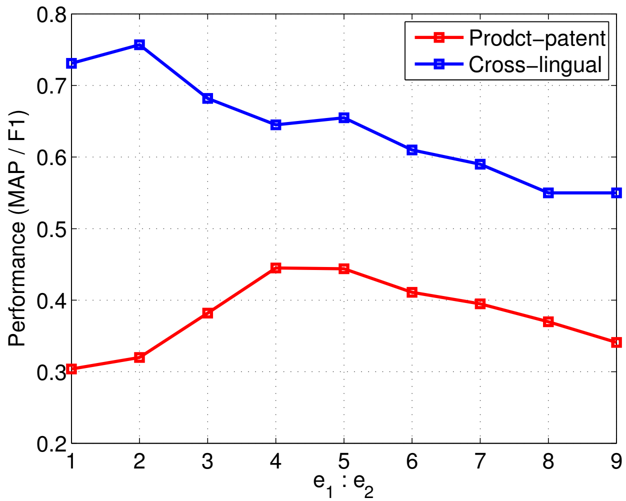
Otherwise sample w_1's topic according to d_2



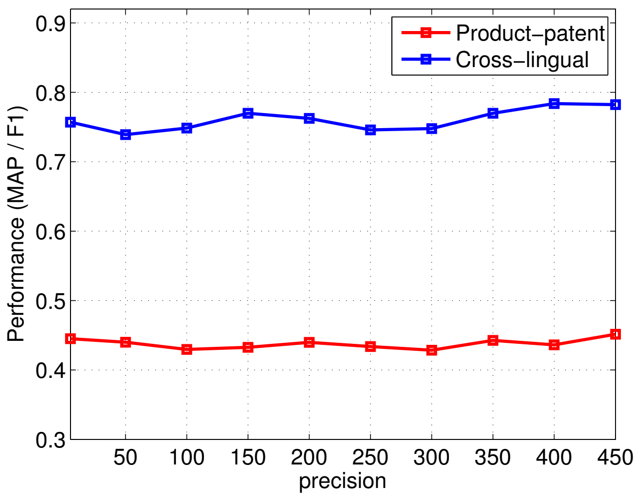
Parameter Analysis



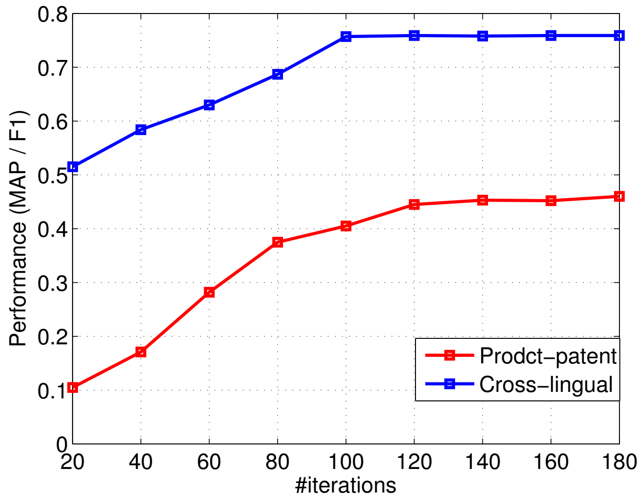
(a) Number of topics K



(b) Ratio



(c) Precision



(d) Convergence analysis

Experimental Results

Training: 30% of the matching relations randomly chosen.

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 hyperlinks between Wiki articles and citations between patents.

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.

Experimental Results

Training: 30% of the matching relations randomly chosen.

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 hyperlinks between Wiki articles and citations between patents.

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.