PatentMiner: Topic-driven Patent Analysis and Mining

Jie Tang, Bo Wang, Yang Yang, Po Hu, Yanting Zhao, Xinyu Yan, Bo Gao, Minlie Huang
Department of Computer Science and Technology, Tsinghua University, China
jietang@tsinghua.edu.cn

Peng Xu, Weichang Li, and Adam K. Usadi
ExxonMobil Research and Engineering Company, New Jersey, USA

When a Company Develops IP Strategies...

- What are the *hot topics* in recent years?
- What are the *most influential* works, researchers, and organizations for a specific topic?
- Who are my *competitors* for a specific topic?

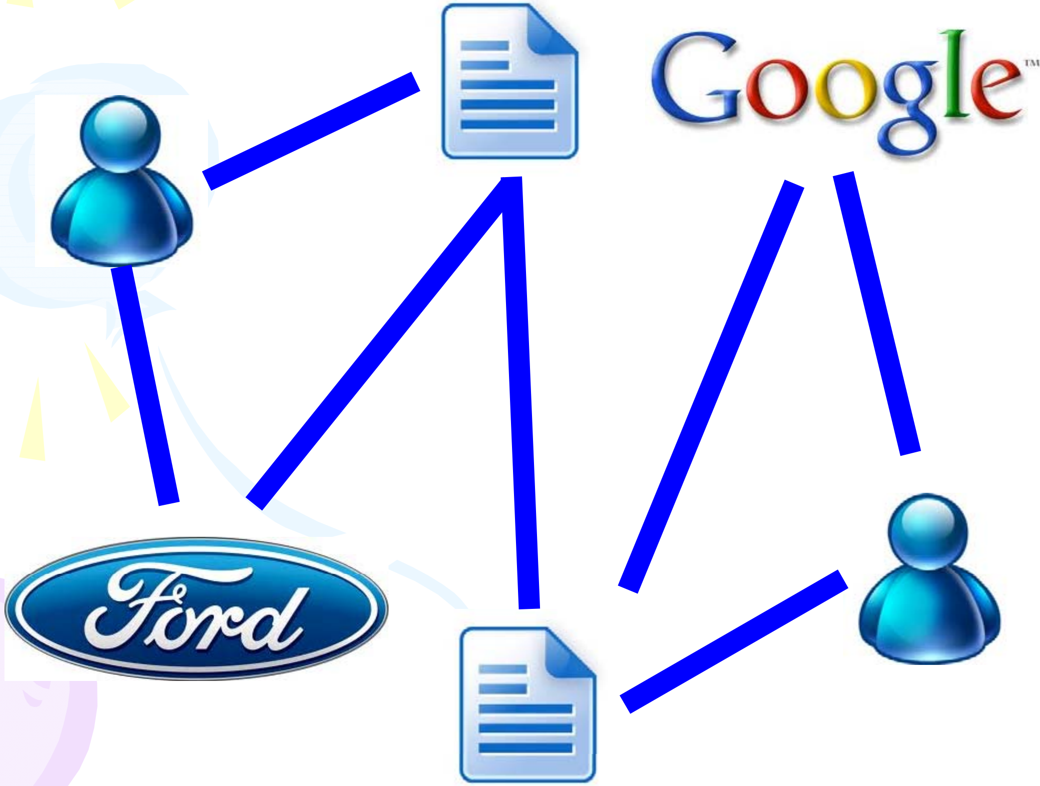
What is PatentMiner?

- Existing automated patent analysis systems only focus on the search function
  - Google Patent, WikiPatent, FreePatentsOnline

- PatentMiner is designed for an ***in-depth*** analysis of patent activity at the topic-level
  - Topic-driven modeling
  - Heterogeneous network co-ranking
  - Intelligent competitive analysis
  - Patent summarization

Heterogeneous Patent Network

- $G = (V_d, V_a, V_c, E_{da}, E_{dc}, E_{dd'}, E_{ac})$

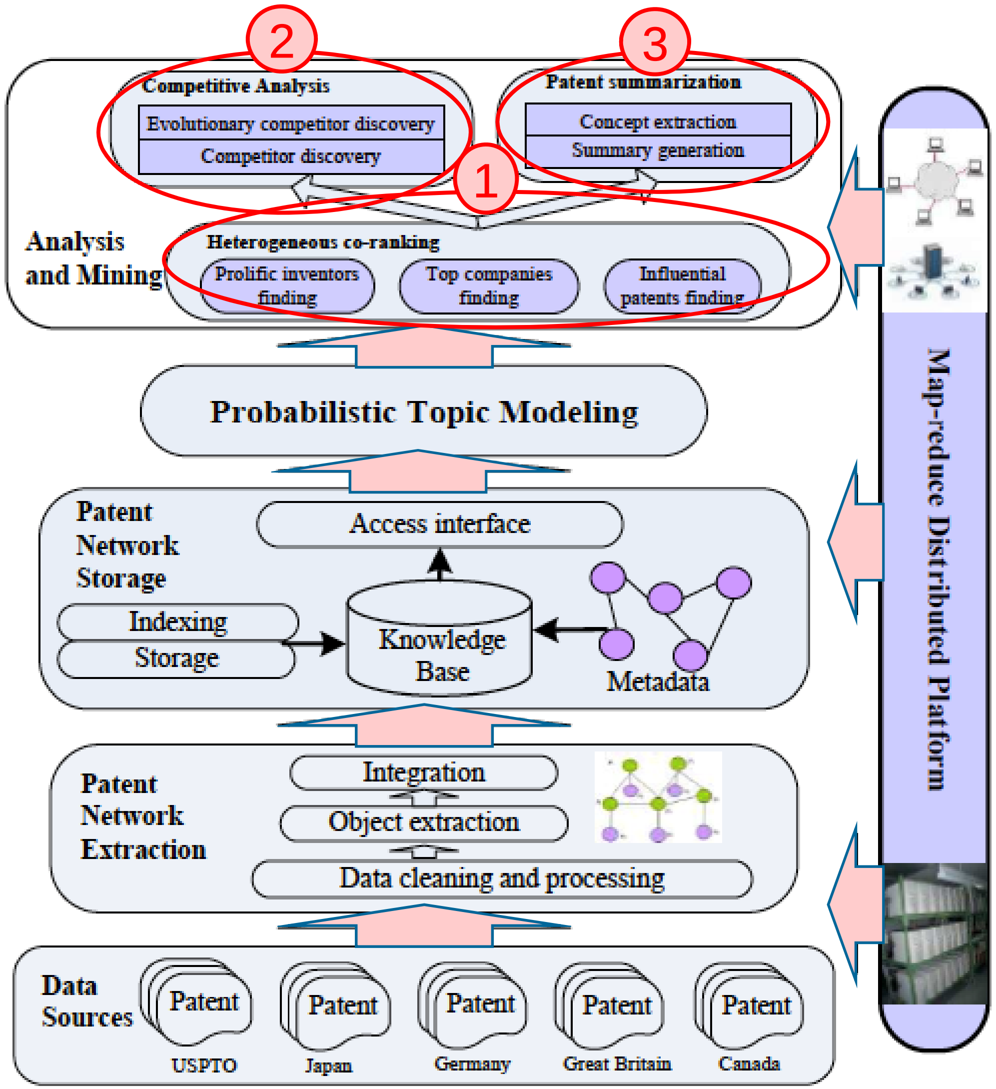


V_d: set of patents

V_a: set of inventors

V_c: set of companies

Architecture of PatentMiner



Modeling Patent Network

- Inventor-Company-Topic (ICT) model
  - Incorporate *patents*, *companies* and *inventors*
  - Three major distributions:
    - inventor-topic distribution θ_{xz}
    - company-topic distribution ψ_{cz}
    - word-topic distribution $\phi_{z_{di} w_{di}}$
  - Log-Likelihood of a collection of patents D:

$$\mathcal{L}(\mathbf{D}) = P(\mathbf{x}, \mathbf{z}, \mathbf{w}, \mathbf{c} | \Theta, \Phi, \Psi, \mathbf{a}) =$$

$$\prod_{d=1}^{M} \prod_{i=1}^{N_d} \frac{1}{A_d} \times \prod_{z=1}^{K} \left(\prod_{x=1}^{A} \theta_{xz}^{m_{xz}} \prod_{j=1}^{W} \phi_{zw_j}^{n_{zw_j}} \prod_{c=1}^{C} \psi_{zc}^{n_{zc}} \right)$$

  - Parameter estimation: Gibbs sampling
    - Calculate posterior of z and sample the topic for each word

Modeling Patent Network (cont.)

- Dynamic ICT (DICT) model
  - To capture the *temporal information*
  - Three smoothing requirements
    - Inventor-topic smoothing

      $$\Omega_1 = \sum_z (\theta_{az}^t - \theta_{az}^{t-1})^2$$

    - Company-topic smoothing

      $$\Omega_2 = \sum_z (\psi_{cz}^t - \psi_{cz}^{t-1})^2$$

    - Topic smoothing

      $$\Omega_3 = \sum_z (P(z)^t - P(z)^{t-1})^2$$

  - Objective function

$$\mathcal{O}(\mathbf{D}) = -\mathcal{L}(\mathbf{D}) + \gamma_1\Omega_1 + \gamma_2\Omega_2 + \gamma_3\Omega_3$$

Generative Process

Initialize $\alpha^0 = 50/K$, $\beta^0 = 0.01$, and $\mu^0 = 0.01$;

foreach *time-stamp* t **do**

- Draw $\alpha^t | \alpha^{t-1} \sim \mathcal{N}(\alpha^{t-1}, \delta^2 I)$;
- Draw $\beta^t | \beta^{t-1} \sim \mathcal{N}(\beta^{t-1}, \sigma^2 I)$;
- Draw $\mu^t | \mu^{t-1} \sim \mathcal{N}(\mu^{t-1}, \epsilon^2 I)$;
- For each topic z^t, draw ϕ_z^t and ψ_z^t respectively from Dirichlet prior β^t and μ^t;
- **foreach** *word* w_{di} *in patent* d **do**
  - Draw an inventor x_{di} from $\mathbf{a}_d$ uniformly;
  - Draw a topic z_{di}^t from a multinomial distribution $\theta_{x_{di}}^t$ specific to inventor x_{di}, where θ^t is generated from the Dirichlet prior α^t;
  - Draw a word w_{di}^t from multinomial $\phi_{z_{di}}^t$;
  - Draw a company stamp c_{di}^t from multinomial $\psi_{z_{di}}^t$;
- **end**

end

Algorithm 1: Probabilistic generative process in DICT.

Heterogeneous Co-Ranking

- Rank patents, companies, and inventors by leveraging the power of ***textual*** and ***network*** information
- Basic idea: propagate the relevance score (to the query) between the linked objects
  - Intuition: an inventor with higher quality patents ranks higher
  - Using ICT model and language model to calculate the relevance score

Competitive Analysis

- Quantitatively characterize the competitive relations between companies
- Global competitor discovery
  - Word-based similarity
  - Topic-based divergence
  - Probability-based correlation (based on ICT)
- Topic-level competitor discovery
  - Utilize topic distribution associated with each company
- Evolutionary competitor discovery

Patent Summarization

- Automatically generate a concise and informative summary for a set of patents
- Basic idea: choose a set of representative sentences as the summary

Data Set

- A patent network includes
  - 3,880,211 patents
  - 2,134,211 inventors
  - 421,032 companies
- We conduct three experiments to evaluate our methods

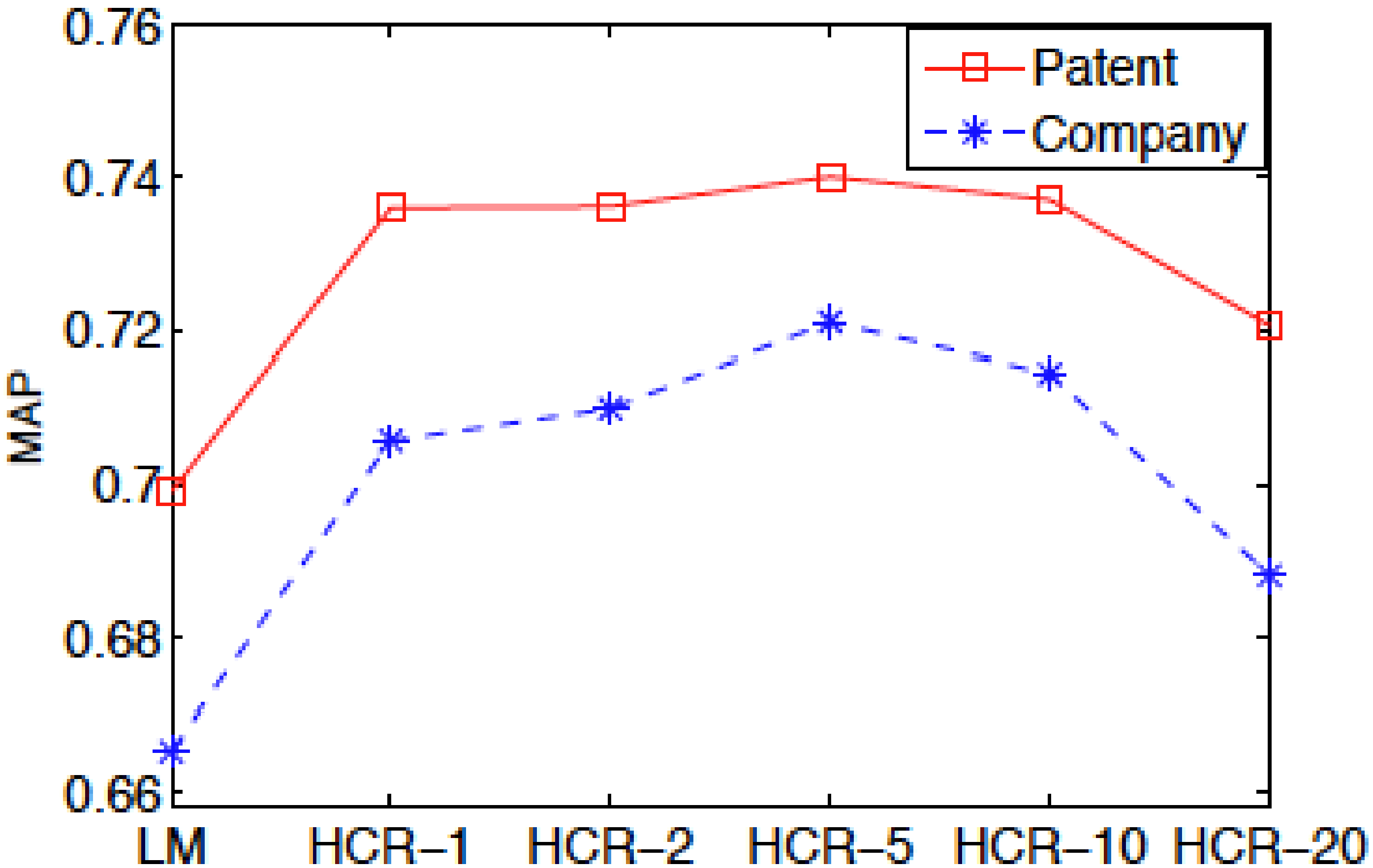
Experiments on Heterogeneous Co-Ranking

- 50 popular queries (e.g., ‘‘data mining’’)
- Label ‘‘like’’ and ‘‘dislike’’ on top 20 results by 5 annotators
- Use language model as baseline
- Vary # of propagation steps of our method

Ranking Performance

Object	Method	P@1	P@5	MAP	N@1	N@5
Patent	LM	.7001	.6900	.6991	.7021	.6833
	HCR-1	.7592	.7102	.7359	.7592	.7310
	HCR-2	.7598	.7201	.7361	.7600	.7300
	HCR-5	.7600	.7298	.7400	.7678	.7367
Company	LM	.6931	.6790	.6654	.6888	.6532
	HCR-1	.7167	.6833	.7058	.7167	.6934
	HCR-2	.7189	.6900	.7100	.7200	.7000
	HCR-5	.7201	.6999	.7210	.7201	.7031

Propagation Steps Analysis



Experiments on Competitive Analysis

- Obtain the ground truth from Yahoo! Finance
- Two baseline methods
  - WBS: represent each company as a bag of words and rank candidates according to Cosine similarity
  - LM+LDA: generate topic-word distribution by LDA and combine language model for competitor discovery
- Vary scoring measures in our method

Performance of Competitor Analysis

	Methods	P@1	P@5	MAP	N@1	N@5
Global	WBS	.2009	.1087	.2904	.2009	.2841
	TopCom+TBD	.1731	.0846	**.3078**	.1731	.2871
	TopCom+PBC	**.2098**	**.1161**	.2920	**.2098**	**.3085**
Topic	LM+LDA	.1536	.1221	.2643	.1536	.2524
	TopCom+DBC	.1369	.1270	.2388	.1469	.2446
	TopCom+HBC	**.1620**	**.1366**	**.2781**	**.1620**	**.2874**

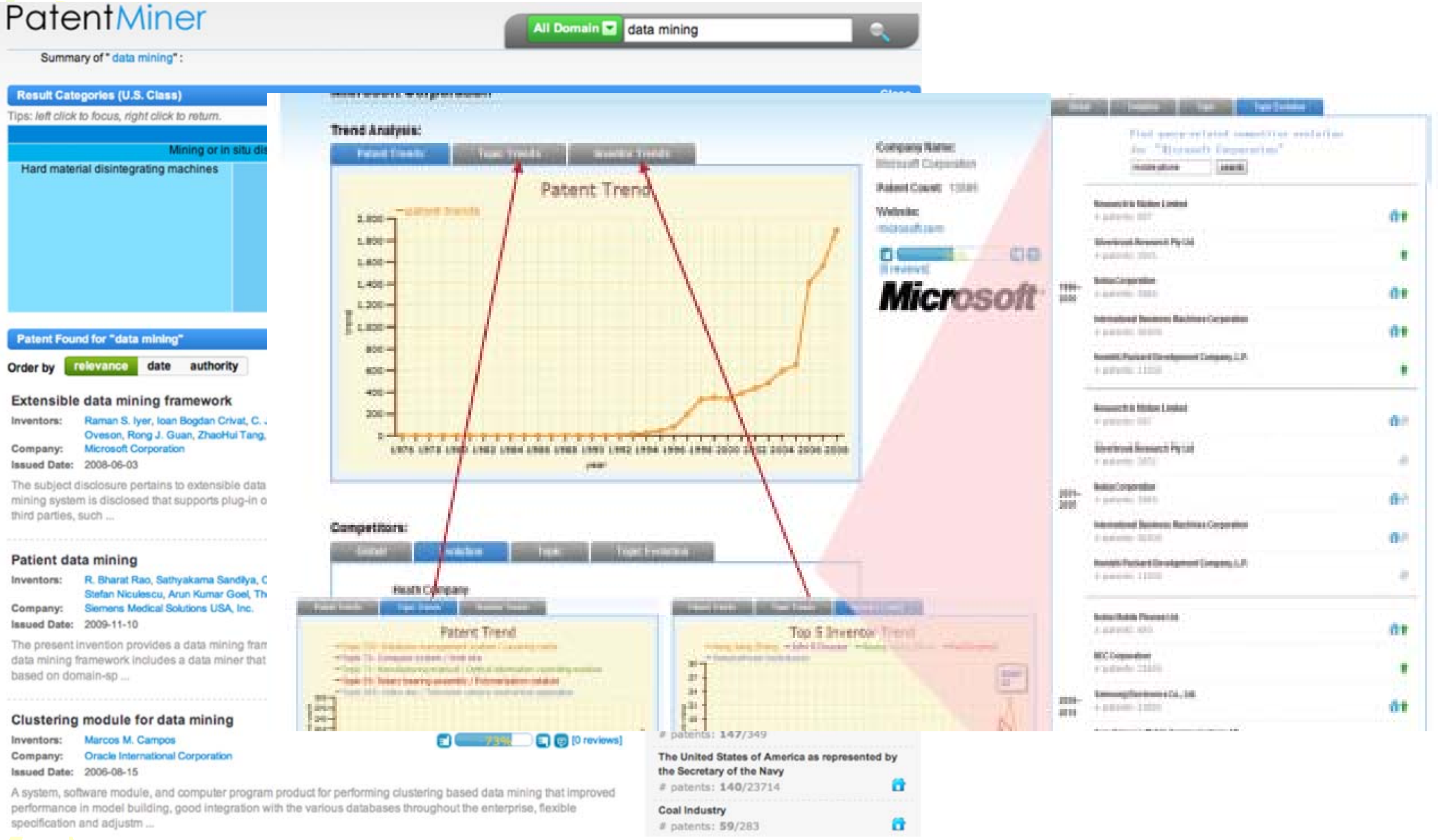
Cisco (Network Device)		AT&T Corp. (Communication)		
1996-2000	2006-2010	1996-2000	2001-2005	2006-2010
IBM	3Com	Lucent	Lucent	Lucent
Microsoft	Juniper	IBM	NEC	NEC
Lucent	Broadcom	NEC	Motorola	IBM
AT&T Corp.	Nortel	Verizon	IBM	Bell
Intel	Intel	Microsoft	Broadcom	Fujitsu
Sun	Canon	Samsung	Intel	Samsung
3Com	IBM	Motorola	Microsoft	Motorola
DEC	Fujitsu	Ericsson	Cisco	Verizon
HP	Sony	Alcatel	Samsung	AOL

Experiments on Patent Summarization

- Tested on benchmark data set TAC 2008 and 2009
- Two baselines
  - Maximal Marginal Relevance (MMR)
  - Diversity Penalty (DP)
- Performance

Data	Metrics	Methods			Gold Standard
		DP	MMR	ILP	
TAC2008	ROUGE-1	0.349	0.348	**0.371**	0.414
	ROUGE-2	0.097	0.096	**0.103**	0.116
TAC2009	ROUGE-1	0.334	0.343	**0.372**	0.444
	ROUGE-2	0.091	0.096	**0.105**	0.126

Online System



Conclusion

- Propose DICT to model topical evolution of different objects in heterogeneous networks
- Propose a heterogeneous co-ranking algorithm and a competitor analysis algorithm
- Validate the methods on a real-world patent database

THANK YOU!