
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
    - company-topic distribution
    - word-topic distribution
  - Log-Likelihood of a collection of patents D:
    \[
    \mathcal{L}(D) = P(x, z, w, c|\Theta, \Phi, \Psi, \alpha) = \prod_{d=1}^{M} \prod_{i=1}^{N_d} \frac{1}{A_d} \times \prod_{z=1}^{K} \left( \prod_{x=1}^{A} \theta_{xz}^{m_{xz}} \prod_{j=1}^{W} \phi_{zwj}^{n_{zwj}} \prod_{c=1}^{C} \psi_{zc}^{n_{zc}} \right)
    \]
  - 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
    \[ \mathcal{O}(D) = -\mathcal{L}(D) + \gamma_1 \Omega_1 + \gamma_2 \Omega_2 + \gamma_3 \Omega_3 \]
Generative Process

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>P@1</th>
<th>P@5</th>
<th>MAP</th>
<th>N@1</th>
<th>N@5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBS</td>
<td>0.209</td>
<td>0.1087</td>
<td>0.2904</td>
<td>0.2009</td>
<td>0.2841</td>
</tr>
<tr>
<td>TopCom+TBD</td>
<td>0.1731</td>
<td>0.0846</td>
<td><strong>0.3078</strong></td>
<td>0.1731</td>
<td>0.2871</td>
</tr>
<tr>
<td>TopCom+PBC</td>
<td><strong>0.2098</strong></td>
<td><strong>0.1161</strong></td>
<td>0.2920</td>
<td><strong>0.2098</strong></td>
<td><strong>0.3085</strong></td>
</tr>
<tr>
<td><strong>Topic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM+LDA</td>
<td>0.1536</td>
<td>0.1221</td>
<td>0.2643</td>
<td>0.1536</td>
<td>0.2524</td>
</tr>
<tr>
<td>TopCom+DBC</td>
<td>0.1369</td>
<td>0.1270</td>
<td>0.2388</td>
<td>0.1469</td>
<td>0.2446</td>
</tr>
<tr>
<td>TopCom+HBC</td>
<td><strong>0.1620</strong></td>
<td><strong>0.1366</strong></td>
<td><strong>0.2781</strong></td>
<td><strong>0.1620</strong></td>
<td><strong>0.2874</strong></td>
</tr>
</tbody>
</table>

### Cisco (Network Device) vs. AT&T Corp. (Communication)

<table>
<thead>
<tr>
<th>Cisco 1996-2000</th>
<th>AT&amp;T Corp. 2006-2010</th>
<th>AT&amp;T Corp. 2001-2005</th>
<th>AT&amp;T Corp. 2006-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>3Com</td>
<td>Lucent</td>
<td>Lucent</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Juniper</td>
<td>IBM</td>
<td>NEC</td>
</tr>
<tr>
<td>Lucent</td>
<td>Broadcom</td>
<td>NEC</td>
<td>Motorola</td>
</tr>
<tr>
<td>AT&amp;T Corp.</td>
<td>Intel</td>
<td>Verizon</td>
<td>IBM</td>
</tr>
<tr>
<td>Intel</td>
<td>Nortel</td>
<td>Microsoft</td>
<td>Intel</td>
</tr>
<tr>
<td>Sun</td>
<td>Intel</td>
<td>Samsung</td>
<td>Samsung</td>
</tr>
<tr>
<td>3Com</td>
<td>Canon</td>
<td>Microsoft</td>
<td>Motorola</td>
</tr>
<tr>
<td>DEC</td>
<td>IBM</td>
<td>Motorola</td>
<td>Motorola</td>
</tr>
<tr>
<td>HP</td>
<td>Fujitsu</td>
<td>Ericsson</td>
<td>Cisco</td>
</tr>
<tr>
<td></td>
<td>Sony</td>
<td>ALCATEL</td>
<td>Samsung</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AOL</td>
</tr>
</tbody>
</table>
Experiments on Patent Summarization

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

<table>
<thead>
<tr>
<th>Data</th>
<th>Metrics</th>
<th>DP</th>
<th>MMR</th>
<th>ILP</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAC2008</td>
<td>ROUGE-1</td>
<td>0.349</td>
<td>0.348</td>
<td>0.371</td>
<td>0.414</td>
</tr>
<tr>
<td></td>
<td>ROUGE-2</td>
<td>0.097</td>
<td>0.096</td>
<td>0.103</td>
<td>0.116</td>
</tr>
<tr>
<td>TAC2009</td>
<td>ROUGE-1</td>
<td>0.334</td>
<td>0.343</td>
<td>0.372</td>
<td>0.444</td>
</tr>
<tr>
<td></td>
<td>ROUGE-2</td>
<td>0.091</td>
<td>0.096</td>
<td>0.105</td>
<td>0.126</td>
</tr>
</tbody>
</table>
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!