CS388 NLP Project Proposal

Text-based Mining for Hidden Relations and Trending

C. Vic Hu
vic@cvhu.org

Ali Unwala
aliunwala@gmail.com

I. Motivation

In economics and statistics, knowledge spillovers between researchers and inventors in close regions and similar domains have been studied to show strong influence [4] [5]. In this paper, we want to focus on implicit relations between US patents based on their terminology usage and writing styles. Furthermore, we will utilize the obtained knowledge relations to formalize hidden topics and make trend prediction.

II. Objective

Our main objectives are two folds. Firstly, we will build language models enhanced from ones proposed in [3] for each authors, years and patent classification numbers. Secondly, we wish to observe some hidden patterns correlated to years and classification numbers obtained from our first results, and construct a Markov Model to predict invention sequence and trending.

III. Data Source

Most of the text-based patents can be found on the US Patent and Trademark Office website [1]. Although their outdated markups and ambiguous formatting make it challenging to obtain clean and consistent data, we have spent some time on writing a simple web crawler and scraped around 700 documents. Each document has the following attributes of interest:

1) Patent number
2) Authors (with location)
3) Classification numbers
4) Abstract
5) Claims
6) Description

IV. Hypothesis

Based on the authorship prediction model presented in [3], we assume that similar patterns could be formulated in the context of US patents. In addition, we assume hidden topics exist and can be discovered over some patterns, and that their causal relationships can be established in a chronological order.

V. Method

To build language models for capturing authorship, chronological and topic patterns, we will adapt and enhance the existing Penn-treebank-based probabilistic context-free grammar and maybe some bag-of-word methods such as topic modeling. To establish hidden sequence models [2] and their relations with the observables (patent data), we will use some variations of the forward-backward algorithm to estimate relevant conditional probabilities and state transitions.

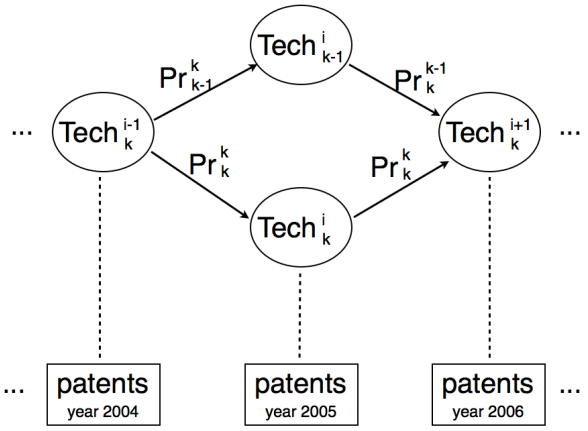

Fig. 1: Patents can be treated as observables for a hidden sequence of technology topics HMM

VI. Evaluation

To evaluate our work, we wish to focus on patents in solar and wind energy from 2003 to 2012. For the first part of attribute predictions (authorships, years and classification numbers), we can simply hold out a subset as our testing data and compare our predictions accordingly. For the second part of hidden relation discovery, we can train older patent data on our algorithm, and compare our generated predictions with the newer patent data.

References

[1] http://uspto.gov
[2] D. Ramage, *Hidden Markov Models Fundamentals*, CS299 Section Notes, 2007.
[3] S. Raghavan, A. Kovashka and R. Mooney, *Authorship Attribution Using Probabilistic Context-Free Grammars*, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 38-42, 2010.
[4] F. G. Braun, J. Schmidt-Ehmcke and P. Zloczysti, *Innovative Activity in Wind and Solar Technologies: Empirical Evidence on Knowledge Spillovers Using Patent Data*, CEPR Discussion Paper 7865, 2010.
[5] S. Breschi and F. Lissoni, *Knowledge Spillovers and Local Innovation Systems: A Critical Survey*, Industrial and Corporate Change, Volume 10 Number 4, 2001.