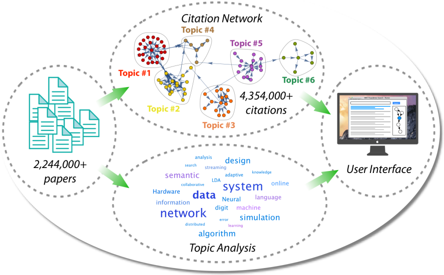
AceRec: Academic Paper Recommendation System

Zengwen Yuan
Jiaming Shen, Zhaowei Tan, Yunqi Guo

December 1, 2014



Outline

Introduction

Motivation

- Online Social Network is booming
- Recreational Social Network
- Academic Social Network?

State-of-Art Academic Engines

- Google Scholar, Microsoft Academic Search
- DBLP, CiteSeer^{X}, etc.
- ArnetMiner, ResearchGate, etc.

However, none of above provides a comprehensive search suggestion of the research topic evolution tendency as time goes by.
What about an academic search engine for layman?

Our Goal

- To build an academic search engine which can:
  1. Return paper search results based on topic similarity with user's query
  2. Analyse the latent topic distribution and topic development over time
  3. Visualize the "topic tree" starting from a particular paper
  4. and more.

General Considerations

Dataset

For an excellent and accurate search engine/recommendation system, it is essential that the system have a dataset which contains large volume of authentic data.

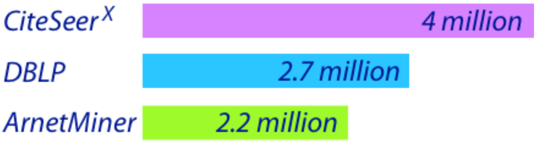


Figure: Volume of dataset (# publication)

Dataset (Cont'd)

- We want to obtain the following metadata of papers:
  title, author, abstract, keywords, citation, reference, year, venue

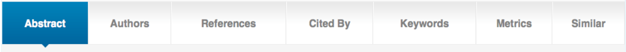


Figure: Metadata Example

Data Preprocessing

For further processing, we need to preprocess the corpus, forming one entry for each paper.

This procedure includes:

- stripping puncutations, space and stop words
- converting all the words to lower-case (and stem processing)
- repeating the title three times (weight factor) and appending it to the abstract

Then the data can be used for Topic Analysis and LDA.

Topic Model

Why we need topic model? (Unsupervised)
Extract the latent topic from papers

computer	chemistry	cortex	orbit	infection
methods	synthesis	stimulus	dust	immune
number	oxidation	fig	jupiter	aids
two	reaction	vision	line	infected
principle	product	neuron	system	viral
design	organic	recordings	solar	cells
access	conditions	visual	gas	vaccine
processing	cluster	stimuli	atmospheric	antibodies
advantage	molecule	recorded	mars	hiv
important	studies	motor	field	parasite

Figure: Five topics from a 50-topic LDA model fit to *Science* (David M. Blei)

Latent Dirichlet Allocation

PLSA (Probabilistic Latent Semantic Analysis) $\Rightarrow$ LDA

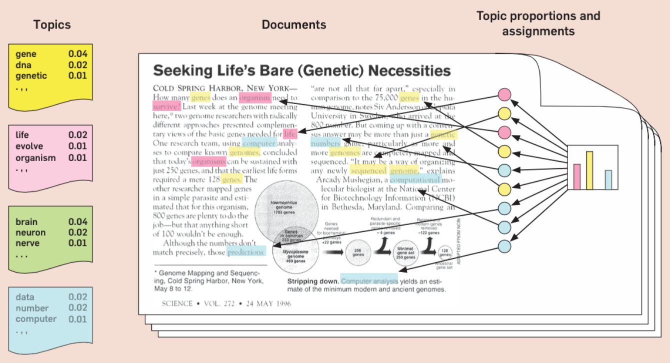


Figure: The intuitions behind latent Dirichlet allocation

Latent Dirichlet Allocation (Cont'd)

- Assume that some "topics" exist for the whole collection
- Each document is assumed to be generated as follows:
  1. Choose a distribution over the topics
  2. For each word in the document, randomly choose:
     a) a topic from the distribution over topics in step #1
     b) a word from the corresponding distribution over the vocabulary

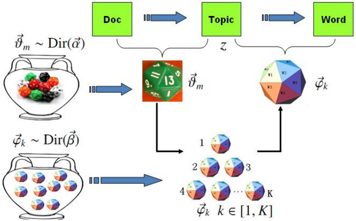


Figure: A Brief Illustration of LDA Model (Rickjin, 2013)

Query Interface and Mapping

What happens when user starts a query:

1. Resolve the user's input, and find the keywords
2. Find the most close topic (keyword mapping)
3. Trace back to the paper
4. Return the search result

Network Analysis

- Complex citation network (4.4 million citation relationships)
- Citation and reference suggests latent time orders in publication and topic development
- Authors have their collective credit weights (Hua-Wei Shen and Barabási, PNAS 2014)
- Matthew effect

Visulization

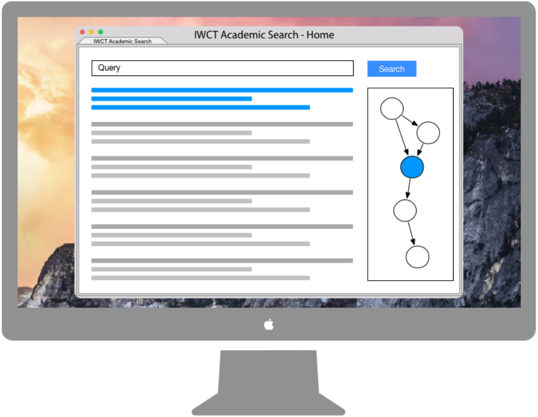
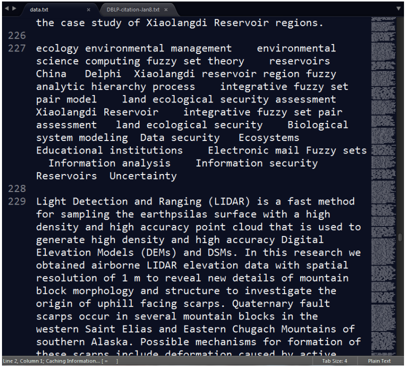


Figure: Website Prototype

Implementation

Paper Dataset

- We did experiments collecting paper metadata (Zhaowei).
- We already have a rather comprehensive dataset from ArnetMiner:



the case study of Xiaolangdi Reservoir regions.

ecology environmental management environmental science computing fuzzy set theory reservoirs China Delphi Xiaolangdi reservoir region fuzzy analytic hierarchy process integrative fuzzy set pair model land ecological security assessment Xiaolangdi Reservoir integrative fuzzy set pair assessment land ecological security Biological system modeling Data security Ecosystems Educational institutions Electronic mail Fuzzy sets Information analysis Information security Reservoirs Uncertainty

Light Detection and Ranging (LIDAR) is a fast method for sampling the earthpsilas surface with a high density and high accuracy point cloud that is used to generate high density and high accuracy Digital Elevation Models (DEMs) and DSMs. In this research we obtained airborne LIDAR elevation data with spatial resolution of 1 m to reveal new details of mountain block morphology and structure to investigate the origin of uphill facing scarps. Quaternary fault scarps occur in several mountain blocks in the western Saint Elias and Eastern Chugach Mountains of southern Alaska. Possible mechanisms for formation of these scarps include deformation caused by active

Figure: Comprehensive Dataset

Data Preprocessing

We preprocessed ∼1,600,000 entries (Jiaming):

reconstruct_corpus.py — PR-sjm-20141025

```python
    title_list = []
    abstract_list = []
    for i in range(0,M):
        # print "=========="
        if (i % 1000 == 0): print i
        raw_info = list_of_all_the_lines[i].split('\n')
        try:
            title = raw_info[0]
            if (title[0:2] == "#*" ):
                title = title[2:]
                title_list.append( seg(title) )

            abstract = raw_info[-1]
            if (abstract[0:2] == "#!" ):
                abstract = abstract[2:]
                abstract_list.append( seg(abstract) )
            else:
                # print "This paper:",i, "do not have abstract"
                abstract_list.append([])
```

Line: 81 | Python | Soft Tabs: 4

Figure: Data Preprocesser

Topic and LDA Anaysis

We used Gibbs Sampling and LDA algorithm with proper parameter:
topics = 100, iteration = 1000, top words = 200, words total = 827185
The experiment result shows high relavence within each topic:

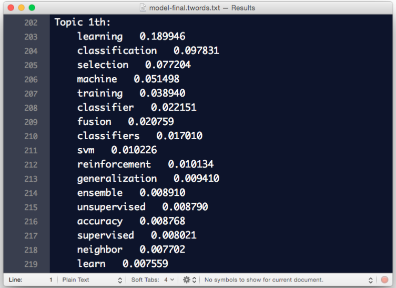


Figure: Word Distribution

Query Resolver

Query ⇒ Keyword ⇒ Topic ⇒ Paper

For each paper, we found the most relevant topics (Yunqi):

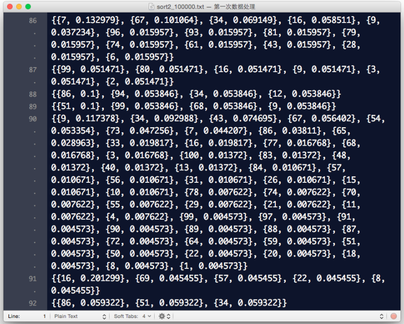


Figure: Possibility of Topics For Papers

Network Analysis

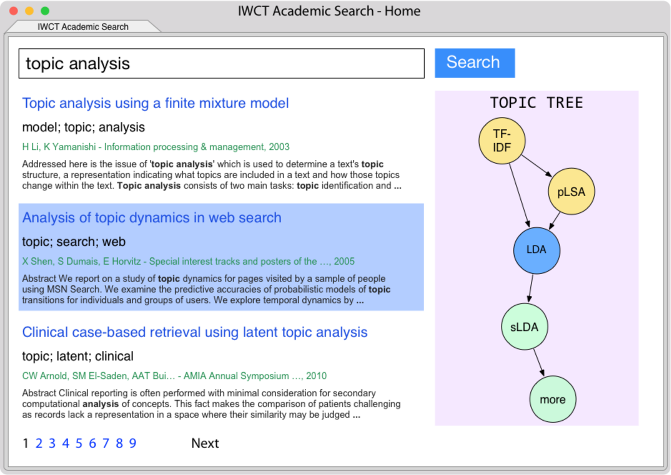


Figure: Topic Tree