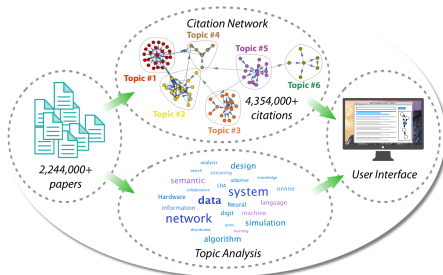


# AceRec: Academic Paper Recommendation System

Zengwen Yuan  
Jiaming Shen, Zhaowei Tan, Yunqi Guo

December 1, 2014



# Outline

## ① Introduction

- Motivation

- State-of-Art Academic Engines

- Our Goal

## ② General Considerations

- Dataset and Preprocessing

- Topic Analysis

- Network Analysis

## ③ Implementation

- Data Acquisition

- LDA Analysis

- Citation Analysis

# Introduction

## ① Introduction

Motivation

State-of-Art Academic Engines

Our Goal

## ② General Considerations

Dataset and Preprocessing

Topic Analysis

Network Analysis

## ③ Implementation

Data Acquisition

LDA Analysis

Citation Analysis

# Motivation

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



# State-of-Art Academic Engines

- Google Scholar, Microsoft Academic Search
- DBLP, CiteSeer<sup>X</sup>, 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

## ① Introduction

Motivation

State-of-Art Academic Engines

Our Goal

## ② General Considerations

Dataset and Preprocessing

Topic Analysis

Network Analysis

## ③ Implementation

Data Acquisition

LDA Analysis

Citation Analysis

# 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.

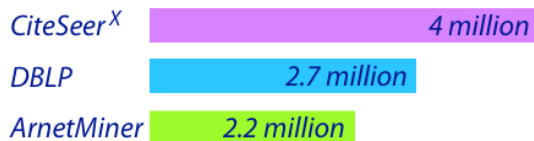


Figure: Volume of dataset (# publication)



# Dataset (Cont'd)

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

Abstract	Authors	References	Cited By	Keywords	Metrics	Similar
----------	---------	------------	----------	----------	---------	---------

Figure: Metadata Example

# Data Preprocessing

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

This procedure includes:

- stripping punctuations, 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

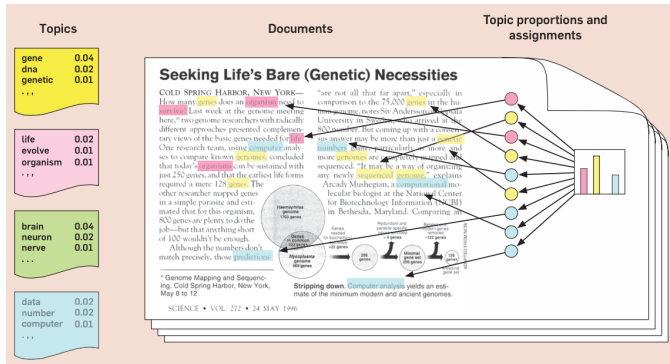


Figure: The intuitions behind latent Dirichlet allocation

# Latent Dirichlet Allocation (Cont'd)

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

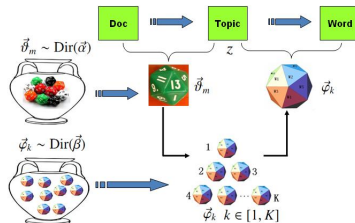


Figure: A Brief Illustration of LDA Model (Rickjin, 2013)

# 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

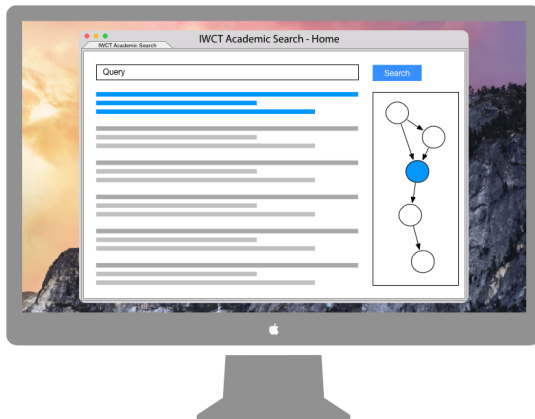


Figure: Website Prototype



# Implementation

## ① Introduction

Motivation

State-of-Art Academic Engines

Our Goal

## ② General Considerations

Dataset and Preprocessing

Topic Analysis

Network Analysis

## ③ Implementation

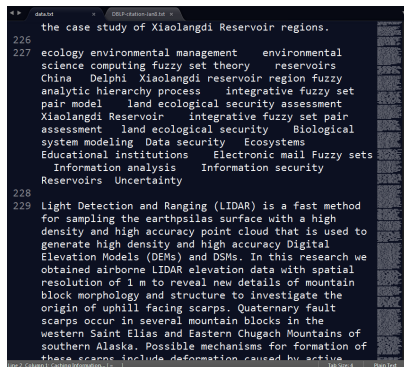
Data Acquisition

LDA Analysis

Citation Analysis

# Paper Dataset

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

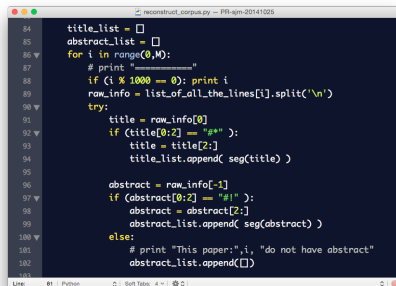


```
data.txt  OSL-Potential-land.txt
the case study of Xiaolangdi Reservoir regions.
226
227 ecology environmental management environmental
science computing fuzzy set theory reservoirs
China Delphi Xiaolangdi reservoir region fuzzy
analytic hierarchy process integrative fuzzy set
pair model land ecological 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
228
229 Light Detection and Ranging (LIDAR) is a fast method
for sampling the earth's 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  $\sim 1,600,000$  entries (Jiaming):

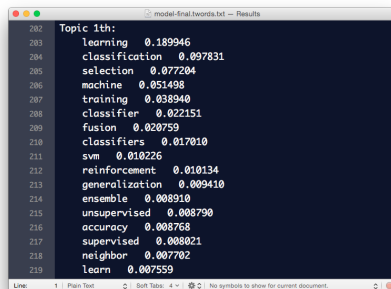
A screenshot of a code editor window titled 'reconstruct\_corpus.py - PFI-qjm-20141025'. The code is written in Python and processes a list of lines. It iterates through the lines, splitting each into title and abstract. Titles starting with '#' are added to 'title\_list', and abstracts starting with '#' are added to 'abstract\_list'. Lines without a title or abstract are printed with a message. The code is as follows:

```
84 title_list = []
85 abstract_list = []
86 for i in range(0,M):
87     # print "====="
88     if (i % 1000 == 0): print i
89     raw_info = list_of_all_the_lines[i].split('\n')
90     try:
91         title = raw_info[0]
92         if (title[0:2] == "#"):
93             title = title[2:]
94             title_list.append( seg(title) )
95
96         abstract = raw_info[-1]
97         if (abstract[0:2] == "#"):
98             abstract = abstract[2:]
99             abstract_list.append( seg(abstract) )
100     else:
101         # print "This paper:", i, "do not have abstract"
102         abstract_list.append([])
```

Figure: Data Preprocessor

# Topic and LDA Analysis

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 relevance within each topic:



```
202 Topic 1th:
203   learning  0.189946
204   classification  0.097831
205   selection  0.077204
206   machine   0.051498
207   training  0.038940
208   classifier 0.022151
209   fusion    0.020759
210   classifiers 0.017010
211   svm       0.010226
212   reinforcement 0.010134
213   generalization 0.009410
214   ensemble   0.008910
215   unsupervised 0.008790
216   accuracy   0.008768
217   supervised 0.008021
218   neighbor   0.007702
219   learn      0.007559
```

Figure: Word Distribution

# Query Resolver

Query  $\Rightarrow$  Keyword  $\Rightarrow$  Topic  $\Rightarrow$  Paper

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

```

86: {{7, 0.132979}, {67, 0.101064}, {34, 0.069149}, {16, 0.058511}, {9,
  0.037234}, {96, 0.015957}, {93, 0.015957}, {81, 0.015957}, {79,
  0.015957}, {74, 0.015957}, {61, 0.015957}, {43, 0.015957}, {28,
  0.015957}, {6, 0.015957}}
87: {{99, 0.051471}, {80, 0.051471}, {16, 0.051471}, {9, 0.051471}, {3,
  0.051471}, {2, 0.051471}}
88: {{86, 0.1}, {94, 0.053846}, {34, 0.053846}, {12, 0.053846}}
89: {{51, 0.1}, {99, 0.053846}, {68, 0.053846}, {9, 0.053846}}
90: {{9, 0.117378}, {34, 0.092988}, {43, 0.074695}, {67, 0.056402}, {54,
  0.053354}, {73, 0.047256}, {7, 0.044207}, {86, 0.038111}, {65,
  0.028963}, {33, 0.019817}, {16, 0.019817}, {77, 0.016768}, {68,
  0.016768}, {3, 0.016768}, {100, 0.01372}, {83, 0.01372}, {40,
  0.01372}, {40, 0.01372}, {13, 0.01372}, {84, 0.010671}, {57,
  0.010671}, {56, 0.010671}, {31, 0.010671}, {26, 0.010671}, {15,
  0.010671}, {10, 0.010671}, {78, 0.007622}, {74, 0.007622}, {70,
  0.007622}, {55, 0.007622}, {29, 0.007622}, {21, 0.007622}, {11,
  0.007622}, {4, 0.007622}, {99, 0.004573}, {97, 0.004573}, {91,
  0.004573}, {90, 0.004573}, {89, 0.004573}, {88, 0.004573}, {87,
  0.004573}, {72, 0.004573}, {64, 0.004573}, {59, 0.004573}, {51,
  0.004573}, {50, 0.004573}, {22, 0.004573}, {20, 0.004573}, {18,
  0.004573}, {8, 0.004573}, {1, 0.004573}}
91: {{16, 0.201299}, {69, 0.045455}, {57, 0.045455}, {22, 0.045455}, {8,
  0.045455}}
92: {{86, 0.059322}, {51, 0.059322}, {34, 0.059322}}
  
```

Figure: Possibility of Topics For Papers

# Network Analysis

IWCT Academic Search

topic analysis

Search

Topic analysis using a finite mixture model

model; topic; analysis

H Li, K Yamanishi - Information processing & management, 2003

Addressed here is the issue of 'topic analysis' which is used to determine a text's topic structure, a representation indicating what topics are included in a text and how those topics change within the text. **Topic analysis** consists of two main tasks: **topic** identification and ...

Analysis of topic dynamics in web search

topic; search; web

X Shen, S Dumais, E Horvitz - Special interest tracks and posters of the ..., 2005

Abstract We report on a study of **topic** dynamics for pages visited by a sample of people using MSN Search. We examine the predictive accuracies of probabilistic models of **topic** transitions for individuals and groups of users. We explore temporal dynamics by ...

Clinical case-based retrieval using latent topic analysis

topic; latent; clinical

CW Arnold, SM El-Saden, AAT Bui... - AMIA Annual Symposium ..., 2010

Abstract Clinical reporting is often performed with minimal consideration for secondary computational **analysis** of concepts. This fact makes the comparison of patients challenging as records lack a representation in a space where their similarity may be judged ...

1 2 3 4 5 6 7 8 9

Next

TOPIC TREE

```

graph TD
    TFIDF((TF-IDF)) --> pLSA((pLSA))
    TFIDF --> LDA((LDA))
    pLSA --> LDA
    LDA --> sLDA((sLDA))
    sLDA --> more((more))
  
```

Figure: Topic Tree