Detecting Topic Evolution in Bibliographic Database Exploiting Citations

Graduate School of
Systems and Information Engineering,
University of Tsukuba

Hirotoshi Ito
Toshiyuki Amagasa
Hiroyuki Kitagawa
Outline

- Background
- Non-negative matrix factorization (NMF)
- Proposed method
- Experiments
- Conclusion and future work
Bibliographic DBs

- Google scholar, MS Academic Search, DBLP, CiteSeerX, ADS, Medline/PubMed, CiNii, …

- Huge academic information accumulated.

- Extract inherent academic knowledge to support researchers.
Topic evolution

- Changes of major research topics over time.

Researchers can get:

- major research topics,
- how they evolved,
- etc.
Related work (1/2)

- **Topic detection**
  - Probabilistic generative model
    - p-LSI [Hofmann. 1999]
    - LDA [Blei et al. 2003]
  - Matrix Factorization
    - Non-negative Matrix Factorization [Lee et al. 1999]
  - Graph analysis
    - term-graph [Jo et al. 2007]

- **Non-negative matrix factorization (NMF) is attracting much attentions.**
  - Lower computational cost than probabilistic model
  - High ability of topic detection as well as probabilistic models
  - Relatively simple algorithm
Related work (2/2)

Detecting topic evolutions

- Using probabilistic model considering textual data and citations [He et al. 2009]
- Using NMF introduced the topic transition matrix that explicitly connects past and present topics [Vaca et al. 2014]
- Scheme of detecting a tendency of topic transitions in whole of the bibliographic database [Masada et al. 2012]

Detecting community evolution

- Scheme of detecting research community introduced new similarity measure of cluster similarity [Tajeuna et al. 2015]
Approach

- Partition DB by fixed-size time windows, and form doc-term matrices using title and abst.
- Detect topics in each matrix.
- Link similar topics in consecutive time windows.
- Exploit CITATIONS for better results.
Applying NMF to doc-term matrix

- NMF is to approximate a non-negative matrix by two matrices with lower rank by optimizing loss function.
  - doc x term $\rightarrow$ doc x topic $\times$ topic x term
Matrix $W$: term distribution in each doc

- $W$: Ratio of each topic of each document contains

- Each doc is approximated with a topic vector of fewer dimensions.

- **Topic-based clustering** of docs can be performed.

Source: W. Xu et al. “Document Clustering Based On Non-negative Matrix Factorization” In SIGIR’03
Dataset partitioning

- Overlap time intervals
  - To connect topics smoothly
  - To use as clue of connecting topics

\[ D^{(t)} : \text{set of documents in time interval } t \]
Dataset partitioning → Transforming into matrices → Detecting topics using NMF → Connecting topics based on similarity

- Transform $D^{(t)}$ into matrix
  - Using title and abstract
  - Transform these textual data into vector (bag-of-words)
  - Align document vectors

- Transform documents that are cited by documents in time interval $t$ into matrix
  - When detect topics, cited papers have important informations

$D^{(t)}$ → $X^{(t)}$ → # of feature terms

$D^{(t)}_c$ → $C^{(t)}$ → # of feature terms
Detect topics by applying NMF to matrix that is connected $X(t)$ and $C(t)$

Loss function:

$$L = \arg \max_{W_X^{(t)},W_C^{(t)},H^{(t)}} \left\| X^{(t)} - W_X^{(t)}H^{(t)} \right\|_F^2 + \delta \left\| C^{(t)} - W_C^{(t)}H^{(t)} \right\|_F^2$$

- $\delta$: Influence of cited papers for topics
**Linking similar topics**

**Approach 1**
- Link topics if their word dist is similar.

**Approach 2**
- Link topics if they share many docs in common.

Source: W. Xu et al. “Document Clustering Based On Non-negative Matrix Factorization” In SIGIR’03
Experiments

Dataset:
- CiteSeerX: 701,686 papers from 1996 to 2014.

![Graphs showing data volume in each year](image)

*Figure 4.* The data volume in each year

Environment
- Python 2.7 + numpy / spicy
Topic evolution: CiteSeerX

- A square represents topics detected in each year
- Terms in square describes each topic
- Size of terms indicates strength of term in topic
  - These are not labels by human tagging
Topic evolution: arXiv (1/4)

- **Neutrino**
  - 1995: neutrino, solar, oscil, supernova, mix, flux, decay, process, core
  - 1996: neutrino, solar, flux, mix, decay, supernova, neutrino
  - 1997: neutrino, solar, mix, decay, supernova, exper, mix
  - 1998: neutrino, solar, mix, decay, supernova, exper, mix
  - 1999: neutrino, quark, decal, hadron, nucleon, nuclear

- **Quark**
  - 1995: quark, heavy, decal, hadron, meson, gcd, product, gluon, baryon
  - 1996: quark, heavy, decal, calcul, hadron, gcd, product, order, nucleon
  - 1997: quark, symmetri, coupl, group, invar, fermion, non, chiral
  - 1998: quark, heavy, nucleon, product, decay, hadron, gcd, scatter, meson

- **Gauge theory**
  - 1995: gaug, invar, fermion, symmetri, non, coupl, abelian, loop, boson
  - 1996: gaug, symmetri, invar, fermion, non, coupl, lactic, abelian, space

- **Merge**
  - Neutrino merge into Neutrino
  - Quark merge with Gauge theory
Topic evolution: arXiv (2/4)

2002
- neutrino
  - solar
  - mix
  - decal
  - experi
  - lepton
  - violet
  - atmosphere

2003
- neutrino
  - solar
  - mix
  - decal
  - experi
  - flux
  - violet
  - lepton

2004
- neutrino
  - solar
  - mix
  - experii
  - decal
  - violet
  - flux

2005
- neutrino
  - solar
  - mix
  - decal
  - experi
  - supernova
  - violet

2006
- neutrino
  - solar
  - mix
  - decal
  - experi
  - univers
  - cold
  - cosmic
  - background

Dark Matter emerge

merge into Neutrino
Topic evolution: arXiv (3/4)

1998
- Quantum Mechanics
  - quantum
  - classic
  - mechan
  - oper

1999
- Quantum Mechanics
  - quantum
  - classic
  - mechan
  - dot
  - space

2000
- Quantum Computer
  - quantum
  - classic
  - mechan
  - comput

2001
- Quantum Computer
  - quantum
  - classic
  - mechan
  - dot
  - gener
  - comput
Topic evolution: arXiv (4/4)

Molecular Cloud merge with Star Formation

Molecular Cloud merge with GA
Conclusion and Future Work

**Conclusion**
- We propose a scheme detecting topic evolution based on NMF exploiting citations
- Our scheme successfully detect topic evolution
- In a view point of diversity, our scheme greatly improve from a prior work

**Future work**
- Discuss about validity of topic and topic evolution
- More efficient algorithms so that we can deal with large datasets