文本数据主题分析及演化分析

陈柏彤
上海大学 图书情报档案系
baitongchen@shu.edu.cn

上海大学
Shanghai University

数据类型

- 文本数据
- 网络数据
- 时空数据
- 高维数据

数据类型

- 文本数据
- 网络数据
- 时空数据
- 高维数据

相关技术

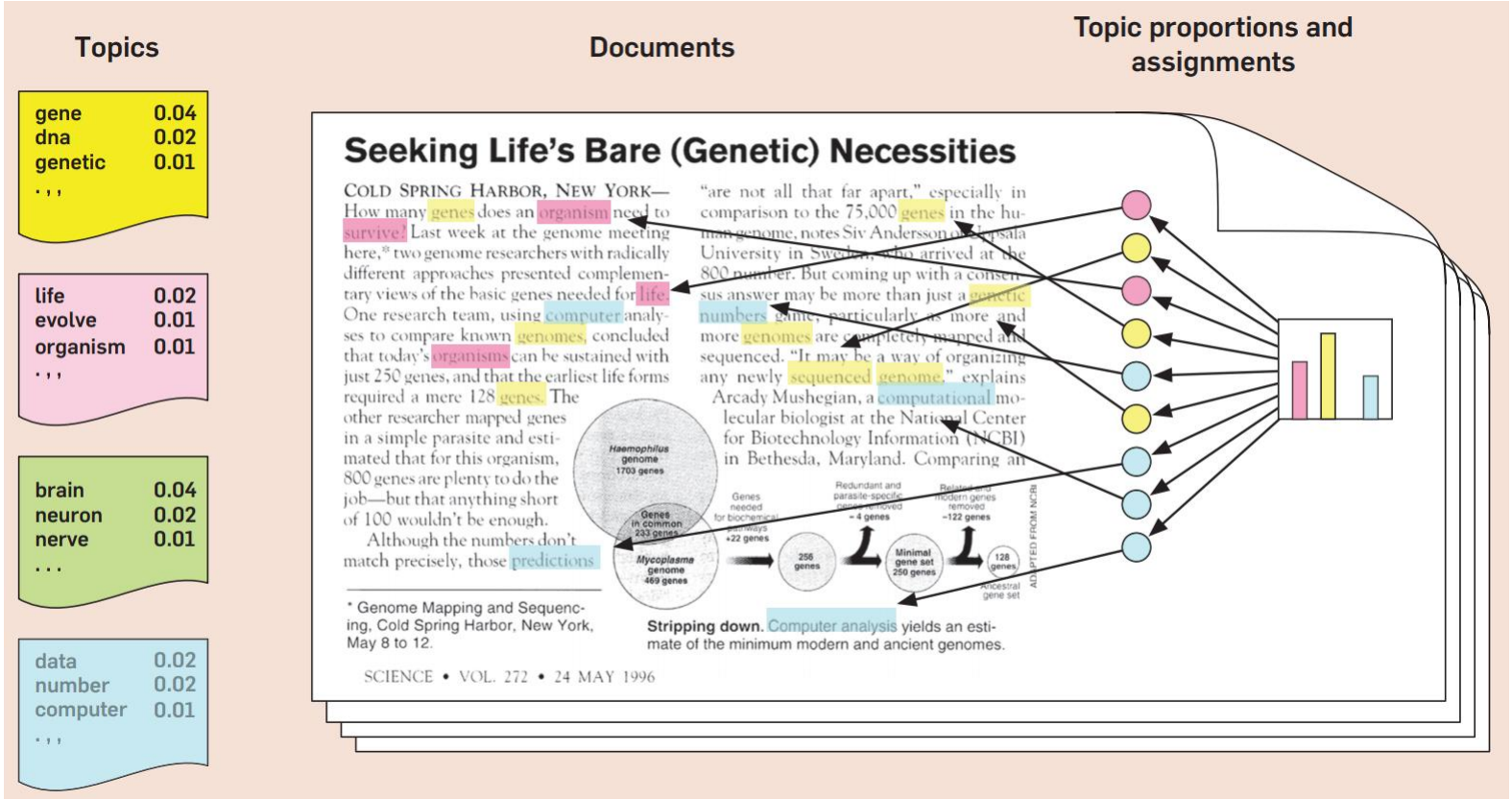
- 自然语言处理技术　Natural Language Processing
- 文本挖掘技术　Text Mining：
  - 主题模型　Topic Models
  - 词嵌入模型　Word Embeddings

相关技术

- 自然语言处理技术　Natural Language Processing
- 文本挖掘技术　Text Mining：
  - 主题模型　Topic Models
  - 词嵌入模型　Word Embeddings

主题模型

- 一篇文档由多个主题混合而成
- 每个主题是一个关于词项的概率分布

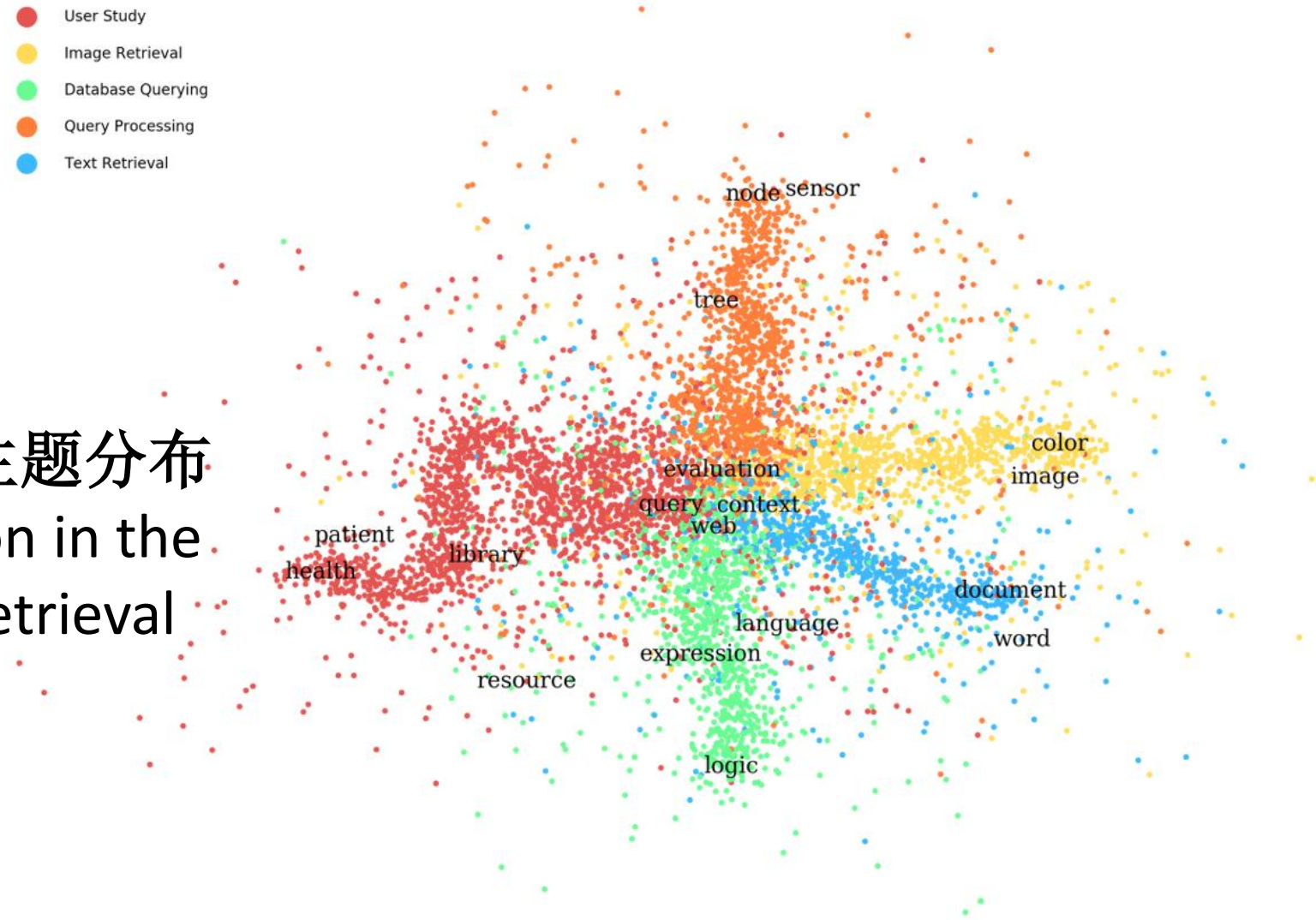


Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM, 55(4):77–84.

主题分析及演化分析

- **实际情况：**获取到一些文本数据，其他信息未知
- **预期目标：**
  - 基于文本数据，推测主题分布
  - 基于时间标签，考察主题演化

数据 ➡ 知识



基于文本数据，探测主题分布

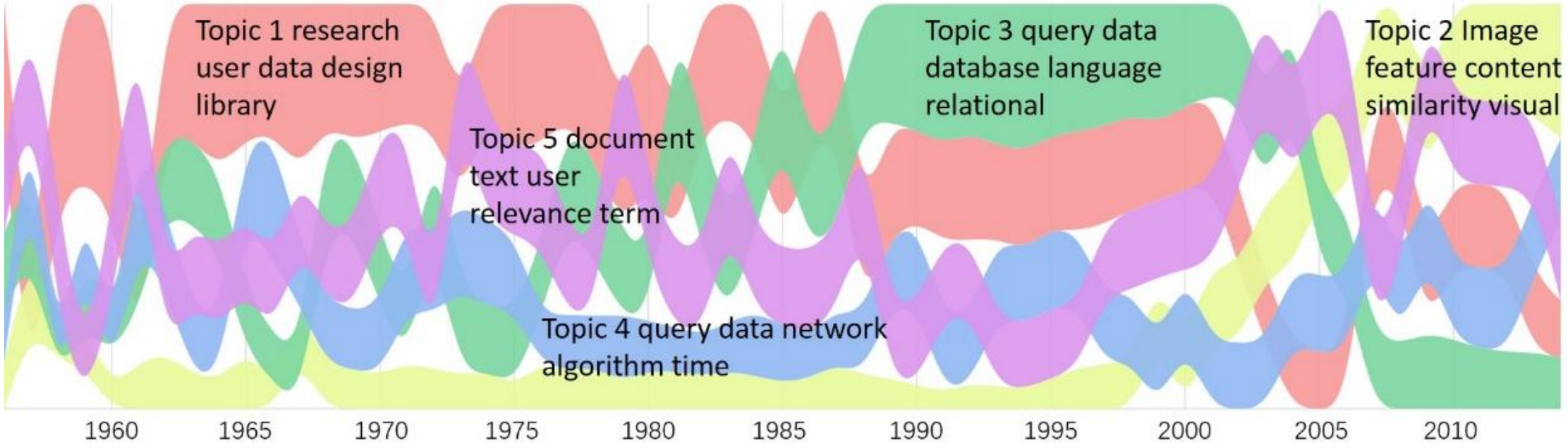
- Word-topic distribution in the field of Information Retrieval

上海大学 Shanghai University

Chen Baitong, et al. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. Journal of Informetrics, 2017, 11(4): 1175-1189.

时间 ➡ 演化：基于时间标签，考察主题演化

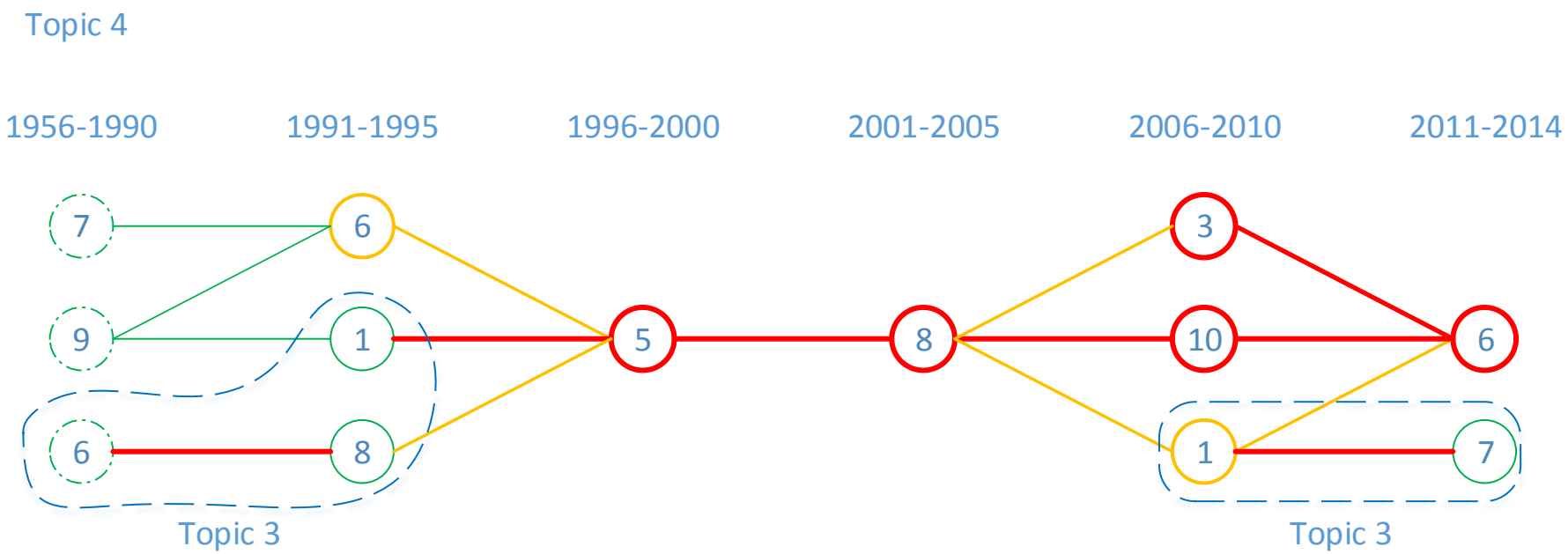
- Topic trends



上海大学 Shanghai University

时间 ➡ 演化

- Topic merging and splitting



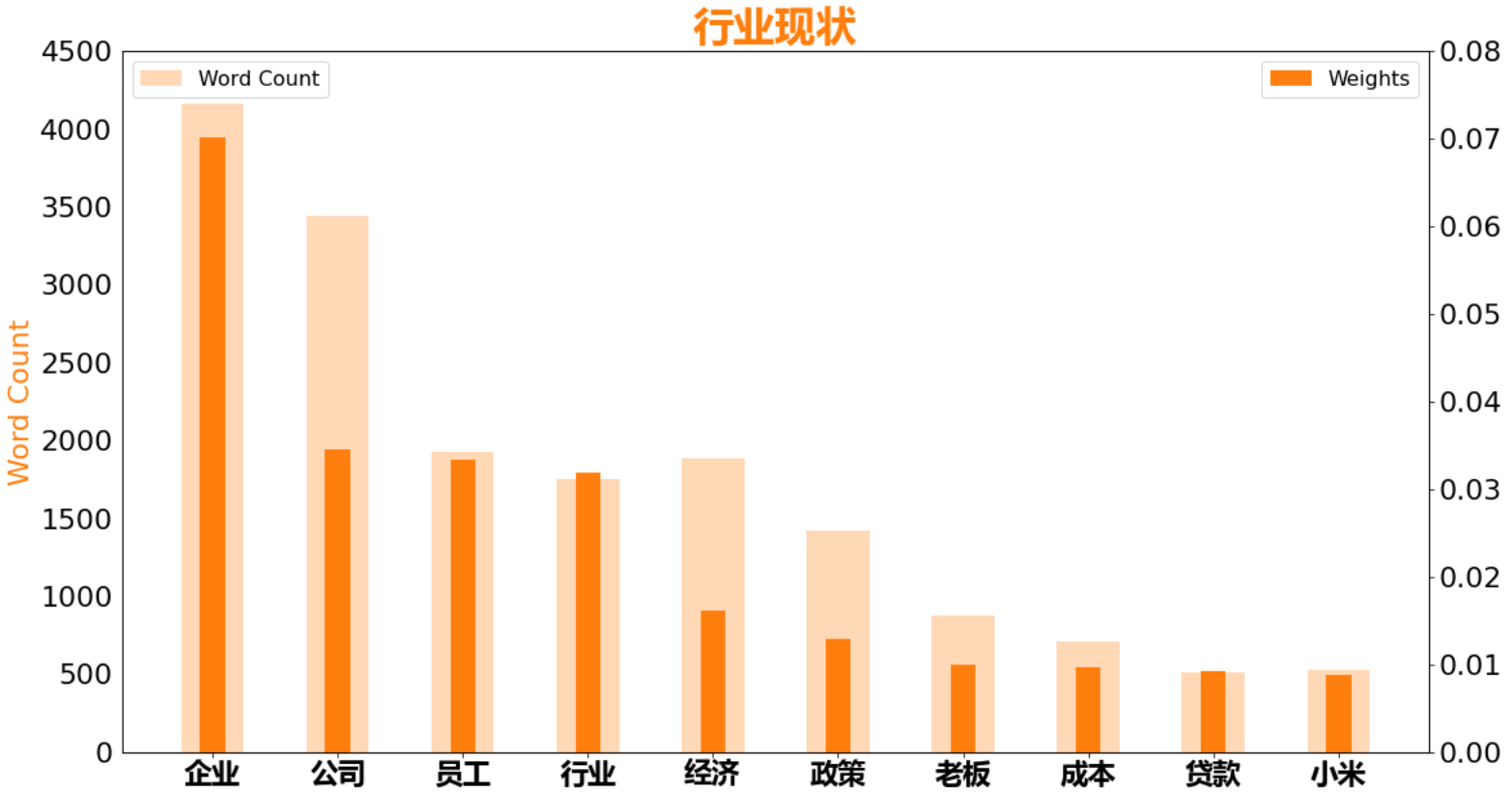
上海大学
Shanghai University

其他案例

- 社交媒体

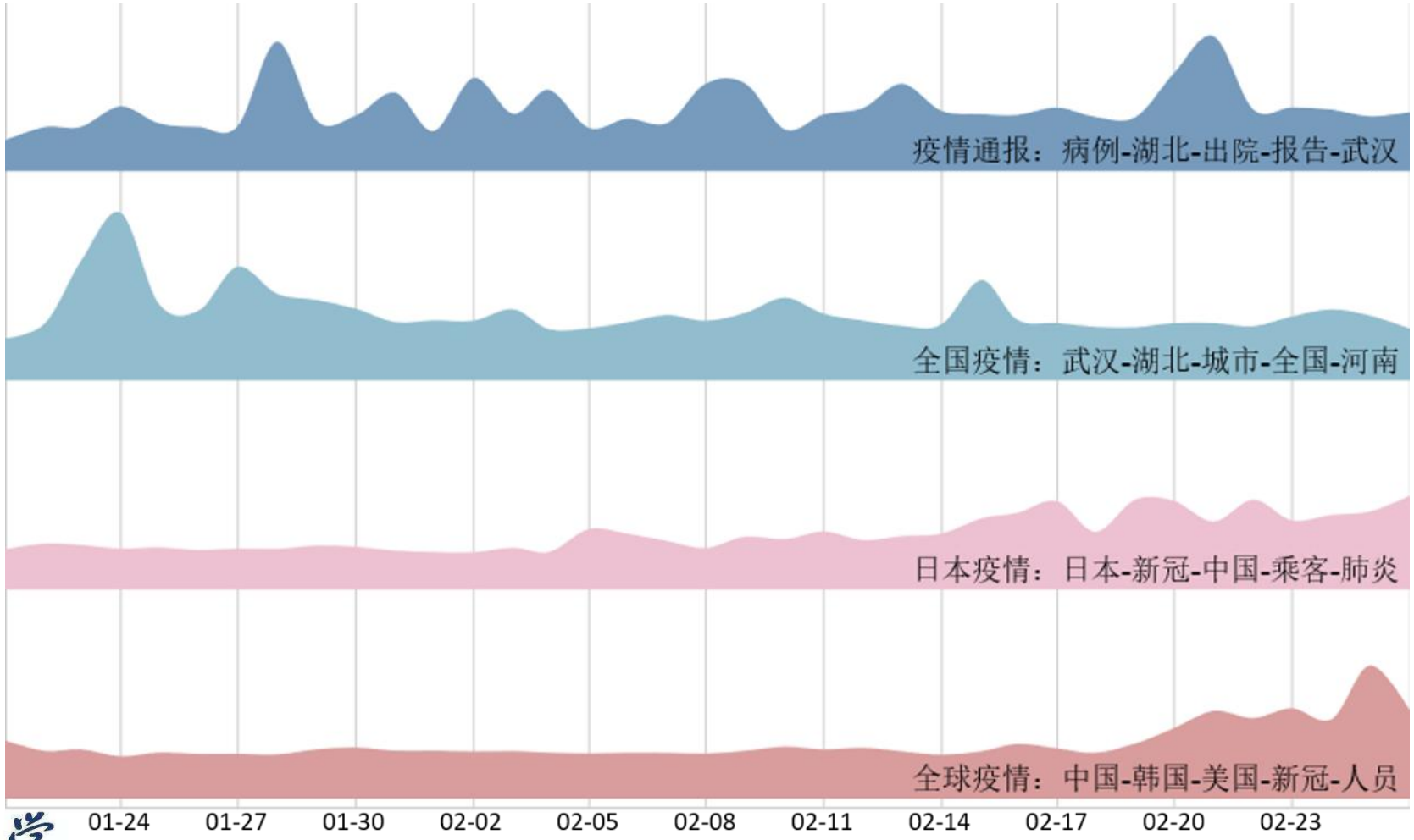
上海大学 Shanghai University

主题分析



上海大学
Shanghai University

演化分析



上海大学
Shanghai University