Text Data Management and Analysis: Opportunities and Challenges

ChengXiang ("Cheng") Zhai
Department of Computer Science
University of Illinois, Urbana-Champaign
Motivation ("big text data"): Intelligent Text Information Systems

Text data is ubiquitous and growing rapidly
Examples: Web pages, blogs, news articles, email, tweets, and literature...

Many opportunities for developing intelligent text information systems to help people manage and exploit large amounts of text data
Example applications:

- all kinds of search engines
- recommender systems (e.g., news recommender, literature recommender)
- categorization systems (e.g., spam detection, …)
- intelligent education systems (e.g., intelligent MOOC, …)
- summarization systems (e.g., product opinion summarization)
- text analytics applications (e.g., business intelligence, researcher’s workbench, …)
- … …
General Research Questions:
Applied NLP, Information Retrieval, and Text Mining

Text Mining

- How to analyze text data to discover knowledge to support user tasks?

Information Retrieval

- How to help users access relevant information with minimum effort?

Content/Semantic Representation

Applied NLP

- How to extract and represent content from text data?

Input Sources:

- WWW
- Blog
- News
- Email
- Literature
- Twitter
A Closer Look at TIMAN

Models
- Machine Learning
- Pattern Recognition
- Data Mining
- Statistics
- Optimization

Algorithms
- Natural Language Processing

Text Information Management & Analysis

Applications
- Applications
- Web, Bioinformatics
- Human Computer Interaction
- Library & Information
- Databases
- Software engineering
- Computer systems

Systems
Where Are We Today?
(Big Picture)

**ML:** √ Supervised learning is mature
But ineffective without sufficient training data

**DM:** √ Lots of mining algorithms
But only effective for salient patterns

**IR:** √ keyword search (models, feedback, evaluation)
√ application-ready retrieval toolkits
But many queries don’t perform well

**NLP:** √ shallow (POS, partial syntactic/semantic analysis)
But deep NLP is fragile

**HCI:** √ Understanding of users
But no operational guidelines for designing interface

**DB:** √ Mature technologies and industry
But limited to structured data

**SYS/APP:** √ Large-scale commercial systems
But lack of intelligence
State of the Art of TIM

Text Acquisition/Crawling

- General crawlers available, but not that much published research

Basic search engine technology is mature, but browsing and recommendation aren’t.

Subject categorization is mature, but others aren’t.

“bag of words” representation, lexical semantics, shallow parsing, fragmental semantics (entity/relation recognition, …)

Technology for storing/indexing large amounts of text is available, but only for simple processing

Extractive summarization is mature, but abstractive isn’t.

Many attempts, but no success yet.

Visualization Engines

Visualization

Extraction of simple entities/relations is mature

Standard clustering is mature

Topic mining is relatively mature, but deeper mining isn’t.

Text Acquisition/Crawling

- Many attempts, but no success yet.

Subject categorization is mature, but others aren’t.

“bag of words” representation, lexical semantics, shallow parsing, fragmental semantics (entity/relation recognition, …)

Technology for storing/indexing large amounts of text is available, but only for simple processing

General crawlers available, but not that much published research

Engines

Engines

Information Access

Information Organization

Knowledge Acquisition

Engines
Opportunity Analysis

Intelligence

- Intelligent information agent
- Decision/Task support
- Information analysis

Information Access
- Recommendation
- Search
- Browsing

Scalability
- Gigabytes
- Terabytes
- Petabytes

Application impact
- Productivity (Web, email, …)
- Decision making (government, business, personal)
- Health/Medical/Biology

Education
High-Impact TIMAN Challenges

• Search Engine Technologies
  – Open IR system:
    • novel evaluation paradigm, evolutionary IR system
  – Intelligent personal information agent:
    • optimal user modeling/personalization, task support
  – Large-scale browsing system:
    • seamless integration of querying and browsing
  – Interactive search engine
    • search engines that talk to users, long-tail search
  – Optimal retrieval models
    • broad impact on all search engines
High-Impact TIMAN Challenges

• Analysis Engine Technologies
  – Scalable semantic (topic) analysis of text
    • first topic analyzer toolkit
  – Integrative analysis of text and structured data:
    • novel statistical models for joint analysis
  – Opinion explorer:
    • new paradigm for opinion exploration and digestion
  – Inference-based text mining:
    • discovery assistant that speeds up scientific discovery
  – Information quality & knowledge provenance:
    • broad impact on all analysis tasks
High-Impact TIMAN Challenges

• Novel Applications
  – Medical/Health information system:
    • unified medical forums, medical entity search, medical literature search, medical record mining, precision medicine, ...
  – Intelligent education system
    • Intelligent MOOC
  – Business intelligence system:
    • online opinion monitoring and analysis
  – Intelligent biological discovery assistant:
    • logic-based integration of knowledge and inferences
  – Text-based forecasting applications:
    • Prediction of stock markets based on text data
    • Prediction of HIV rates based on social media