An Introduction to Text Data Mining

Adam Zimmerman

The Ohio State University

Data Mining and Statistical Learning Discussion Group
September 6, 2013

Outline

1. Basics

2. Algorithms

3. Future directions

Outline

1. Basics
2. Algorithms
3. Future directions

Outline

1. Basics
2. Algorithms
3. Future directions

Basics

- Because of its lack of structures, text data is typically managed via a search engine instead of a database.
- There's plenty of research in search engines and information retrieval, but the goal of text mining is information analysis.
- Web-based applications (e.g., social media) have encouraged *joint mining* of text data in the context of heterogeneous data domains, such as with multimedia (Instagram images, Youtube videos) or cross-lingual linkages (between documents written in different languages).

Key Characteristics of Text Data

- Text data is *sparse* and *high dimensional*.
  - A corpus might come from a lexicon of 100,000 words, but any one document might contain only a few hundred words.
  - A corpus can be represented as a *sparse term-document matrix* of size $n \times d$, where n is the number of documents, and d is the size of the lexicon vocabulary.
  - The (i, j)th entry is the (normalized) frequency of the jth word in the lexicon in document i.
- Text data data can be analyzed at different levels of representation.
  - Bag-of-words
  - String of words
  - Semantically

Algorithms for text mining

- **Information extraction from text data**; e.g., extracting entities and their relationships
- **Text summarization**
  - **Extractive summarization**—summary consists of information units extracted from the original text
  - **Abstractive summarization**—summary can contain synthesized information units that don't necessarily occur in the documents
- **Unsupervised learning methods from text data**—require neither training data nor manual effort
  - **Clustering**—partitioning a corpus of documents into topical clusters; each document belongs to one cluster
  - **Topic modeling**—each document has a membership probability of the cluster; works well with dimension reduction techniques

- **Dimensionality reduction for text mining**—reducing dimensionality from the size of the lexicon vocabulary d
  - **Latent semantic indexing** (**LSI**) is a common approach, useful because it brings out semantic aspects; e.g., reduction in noise from *synonymy* and *polysemy*.
  - **Probabilistic topic models** such as **probabilistic latent semantic analysis** (**PLSA**) and **latent Dirichlet allocation** (**LDA**) reduce dimensionality in a probabilistic way, giving topic representations based on word distributions.
- **Supervised learning methods for text data**—use training data (often for classification); many of these are extensions from machine learning, including **rule-based classifiers**, **decision trees**, **nearest neighbor classifiers**, **maximum-margin classifiers**, and **probabilistic classifiers**.

- **Transfer learning with text data**—transferring knowledge between domains with heterogeneous attributes (a special type of supervised learning)
  - Cross-lingual mining; e.g., easy to find labeled English documents, but hard to find labeled Chinese documents
  - Cross-media transfer (text, multimedia)
- **Probabilistic techniques for text mining**
  - Unsupervised topic models such as **PLSA** and **LDA**
  - Supervised learning methods such as **conditional random fields**
- **Mining text streams**—*one-pass constraint*: difficulty storing data for processing requires continuous mining as data comes in (e.g., Twitter, Google news)

- **Cross-lingual mining of text data**—In addition to transfer learning, includes clustering documents in different languages, machine translation, and analyzing comparable/parallel corpora
- **Text mining in multimedia networks**—enriching the mining process by simultaneous use of data from other domains together with text collection (related to transfer learning)
- **Text mining in social media**
  - Quick, free expression in context of a wide range of subjects
  - Commercial applications for influencing users and targeted marketing
  - Mining dynamic data often containing poor and non-standard vocabulary
  - Usage of links between individuals to improve quality of mining process (as opposed to methods based on only content or links)

- **Opinion mining from text data**—supporting consumer decisions, business intelligence, eliminating spam and noise
- **Text mining from biomedical data**
  - Efficient location and access to knowledge buried in huge amounts of literature
  - Supplementation of other biomedical data such as genome sequences, gene expression data, and protein data

Future directions for research which show promise

- **Scalable and robust methods for natural language understanding**; e.g., “the semantic web"
- **Domain adaptation and transfer learning**—Many text mining tasks are supervised learning, depending on available training data. Research is needed to overcome limitations and inadequacies of domain adaptation and transfer learning methods for adapting training data from other domains or tasks.
- **Contextual analysis**—incorporating information such as authors, sources, and time
- **Parallel text mining**—increasing size of data may require parallelization due to single-machine storage and processing limits