Lecture 2: Text Mining and Web Mining

Jiaheng Xie
Department of Management Information Systems
Eller College of Management
University of Arizona

Learning objectives

- Understand the major text mining techniques
- Understand what is Web mining and its techniques

Text mining techniques

Text Classification

Sentiment Analysis

Topic Modeling

Named Entity Recognition

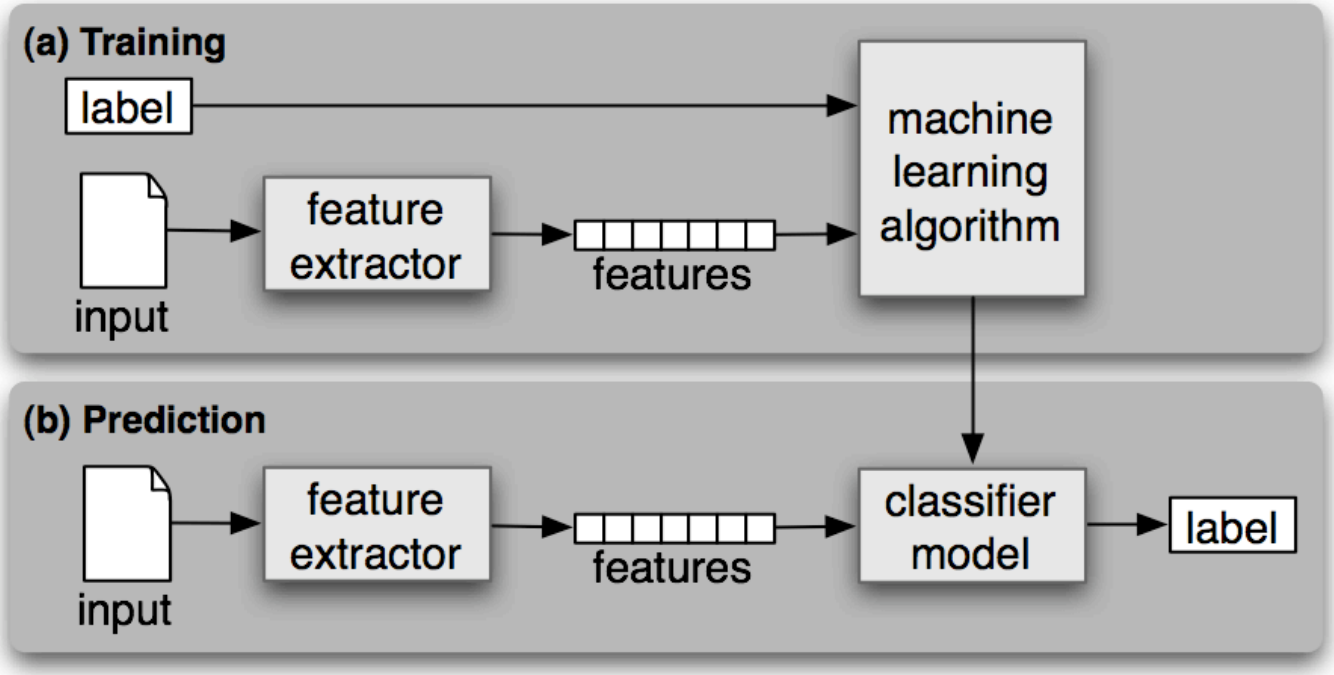
Text Classification

- Text Classification or text categorization is a problem in library science, information science, and computer science. Text classification is the task of choosing correct class label for a given input.

- Some examples of text classification tasks are
  - Deciding whether an email is a spam or not (**spam detection**) .
  - Deciding whether the topic of a news article is from a fixed list of topic areas such as “sports”, “technology”, and “politics” (**document classification**).
  - Deciding whether a given occurrence of the word *bank* is used to refer to a river bank, a financial institution, the act of tilting to the side, or the act of depositing something in a financial institution (**word sense disambiguation**).

Text Classification

- Text classification is a **supervised machine learning** task as it is built based on training corpora containing the correct label for each input. The framework for classification is shown in figure below.



(a) During training, a feature extractor is used to convert each input value to a feature set. These feature sets, which capture the basic information about each input that should be used to classify it, are discussed in the next section. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model.
(b) During prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels.

Text Classification

- Common features for text classification include: bag-of words (BOW), bigrams, tri-grams and part-of-speech(POS) tags for each word in the document.

- The most commonly adopted machine learning algorithms for text classifications are **naïve Bayes**, **support vector machines**, and **neural networks** classifications.

- Recent development: deep learning (RNN, LSTM…)

Sentiment Analysis

- Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source material.

- The rise of social media such as forums, micro blogging and blogs has fueled interest in sentiment analysis.
  - Online reviews, ratings and recommendations in social media sites have turned into a kind of virtual currency for businesses looking to market their products, identifying new opportunities and manage their reputations.

  - As businesses look to automate the process of filtering out the noise, identifying relevant content and understanding reviewers' opinions, sentiment analysis is the right technique.

Sentiment Analysis

- The main tasks, their descriptions and approaches are summarized in the table below:

Task	description	Approaches	lexicons/ algorithms
Polarity Classification	classifying a given text at the document, sentence, or feature/aspect level into positive, negative or neutral	lexicon based scoring	SentiWordNet, LIWC
		machine learning classification	SVM,NB, RNN
Affect Analysis	Classifying a given text into affect states such as "angry", "sad", and "happy"	lexicon based scoring	WordNet-Affect
		machine learning classification	SVM,NB, RNN
Subjectivity Analysis	Classifying a given text into two classes: objective and subjective	lexicon based scoring	SentiWordNet, LIWC
		machine learning classification	SVM,NB, RNN
Feature/Aspect Based Analysis	Determining the opinions or sentiment expressed on different features or aspects of entities (e.g., the screen[feature] of a cell phone [entity])	Named entity recognition + entity relation detection	SentiWordNet, LIWC, WordNet
			SVM,NB,RNN,CRF
Opinion Holder /Target Analysis	Detecting the holder of a sentiment (i.e. the person who maintains that affective state) and the target (i.e. the entity about which the affect is felt)	Named entity recognition + entity relation detection	SentiWordNet, LIWC, WordNet
			SVM,NB,RNN,CRF

Topic Modeling

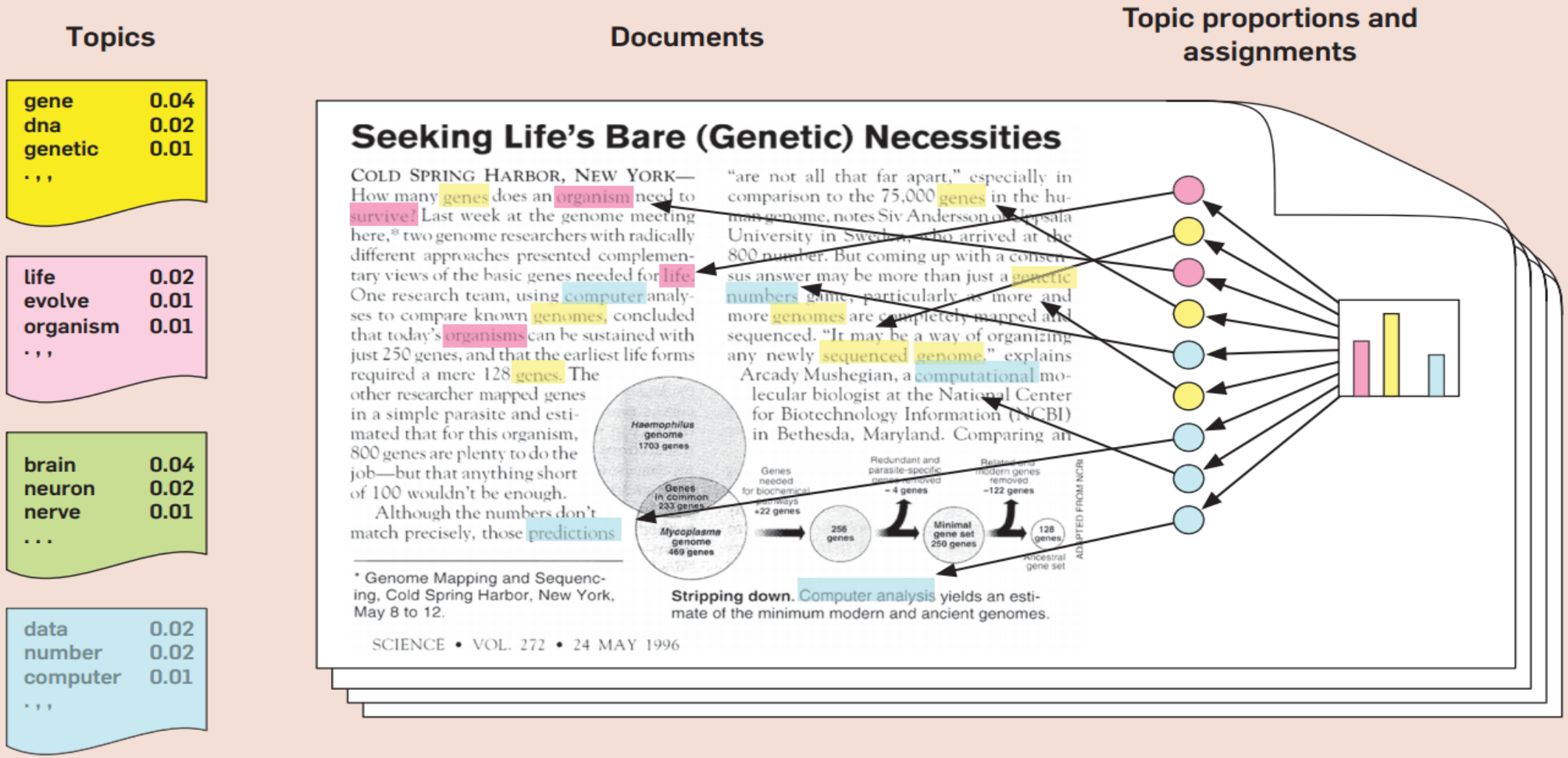
- Topic models are a suite of algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.

- Topic modeling algorithms include Latent Semantic Analysis(LSA), Probability Latent Semantic Indexing (PLSI), and Latent Dirichlet Allocation (LDA).
  - Among them, **Latent Dirichlet Allocation** (**LDA**) is the most commonly used nowadays.

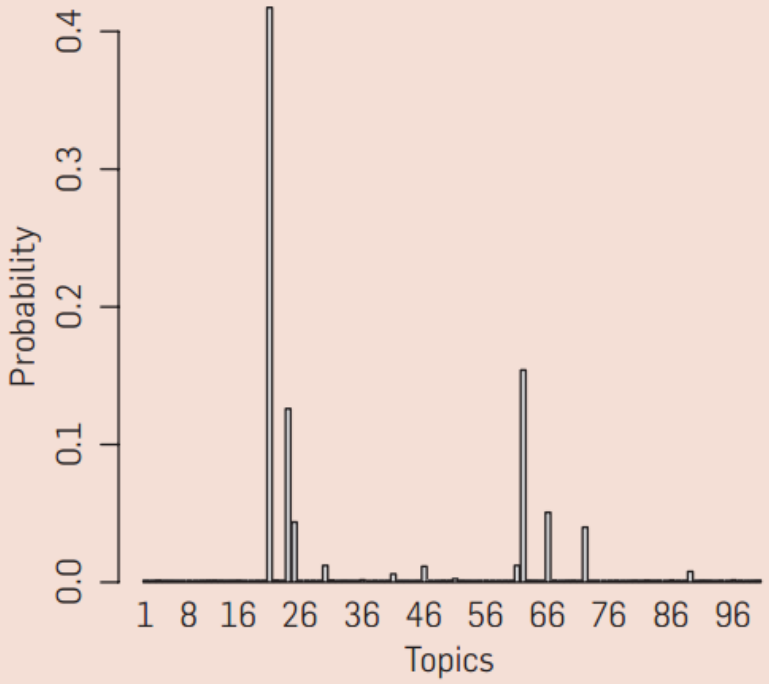
- Topic modeling algorithms can be applied to massive collections of documents.
  - Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API.

- Topic modeling algorithms can be adapted to many kinds of data.
  - They have been used to find patterns in genetic data, images, and social networks.

The figure below shows the intuitions behind **latent Dirichlet allocation.** We assume that some number of "topics", which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic .



The figure below show real inference with LDA. 100-topic LDA model is fitted to 17,000 articles from journal *Science*. At left are the inferred topic proportions for the example article in previous figure. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Named Entity Recognition

- Named entity refers to anything that can be referred to with a proper name.
- Named entity recognition aims to
  - Find spans of text that constitute proper names
  - Classify the entities being referred to according to their type

Type	Sample Categories	Example
People	Individuals, fictional Characters	***Turing*** is often considered to be the father of modern computer science.
Organization	Companies, parties	***Amazon*** plans to use drone copters for deliveries.
Location	Mountains, lakes, seas	The highest point in the ***Catalinas*** is ***Mount Lemmon*** at an elevation of 9,157 feet above sea level.
Geo-Political	Countries, states, provinces	The Catalinas, are located north, and northeast of ***Tucson***, ***Arizona***, ***United States.***
Facility	Bridges, airports	In the late 1940s, ***Chicago Midway*** was the busiest airport in the United States by total aircraft operations.
Vehicles	Planes, trains, cars	The updated ***Mini Cooper*** retains its charm and agility.

In practice, named entity recognition can be extended to types that are not in the table above, such as temporal expressions (time and dates), genes, proteins, medical related concepts (disease, treatment and medical events) and etc..

Named Entity Recognition

- Named entity recognition techniques can be categorized into knowledge-based approaches and machine learning based approaches.

Category	Advantage	Disadvantage	Tools /Ontology
Knowledge-based approach	Require little training data	Creating lexicon manually is time-consuming and expensive; encoded knowledge might be importable across domains.	**General Entity Types**
			• WordNet
			• Lexicons created by experts
			Medical domain:
			• GATE (University of Sherfield)
			• UMLS (National library of Medicine)
			• MedLEE (Originally from Columbia University, commercialized now)
Machine learning approach - Conditional Random Field (CRF) - Hidden Markov Model (HMM)	Reduced human effort in maintaining rules and dictionaries	Prepared a set of annotated training data	**Conditional Random Field tools**
			• Stanford NER
			• CRF++
			• Mallet
			Hidden Markov Model tools
			• Mallet
			• Natural Language Toolkit(NLTK)

Ontology

- Ontology represents knowledge as a set of concepts with a domain, using a shared vocabulary to denote types, properties, and interrelationships of those concepts.
- In text mining, ontology is often used to extract named entities, detect entity relations and conduct sentiment analysis. Commonly used ontologies are listed in the table below:

Name	Creator	Description	Application
WordNet	Princeton University	A large lexical database of English.	Word sense disambiguation Text summarization Text similarity analysis
SentiWordNet	Andrea Esuli, Fabrizio Sebastian	SentiWordNet a lexical resource for opinion mining.	Sentiment analysis
Linguistic Inquiry and Word Count(LIWC)	James W. Pennebaker, Roger J. Booth, Martha E. Francis	LIWC is a lexical resource for sentiment analysis.	Sentiment analysis Affect analysis Deception detection
Unified Medical Language System (UMLS)	US National Library of Medicine	The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences.	Medical entity recognition
Consumer Health Vocabulary (CHV)	University of Utah	Mapping consumer health vocabulary to standard medical terms in UMLS.	Medical entity recognition, Health social media analytics

Talk with your neighbors

- Discuss with your neighbors for 5 min

- Given you have the product reviews of the new iPhone. You would like to know
  - If people like the new iPhone or not
  - What aspects of the iPhone do people like/dislike

- What text mining techniques you will use?
- How do you use them?

A-Z list of Open Source NLP toolkits

Text Mining Frameworks in R

- tm provides a comprehensive text mining framework for R. The Journal of Statistical Software article Text Mining Infrastructure in R gives a detailed overview and presents techniques for count-based analysis methods, text clustering, text classification and string kernels.
  - tm.plugin.dc allows for distributing corpora across storage devices (local files or Hadoop Distributed File System).
  - tm.plugin.webmining allow importing news feeds in XML (RSS, ATOM) and JSON formats. Currently, the following feeds are implemented: Google Blog Search, Google Finance, Google News, NYTimes Article Search, Reuters News Feed, Yahoo Finance, and Yahoo Inplay.
- RcmdrPlugin.temis is an Rcommander plug-in providing an integrated solution to perform a series of text mining tasks such as importing and cleaning a corpus, and analyses like terms and documents counts, vocabulary tables, terms co-occurrences and documents similarity measures, time series analysis, correspondence analysis and hierarchical clustering.
- openNLP provides an R interface to OpenNLP , a collection of natural language processing tools including a sentence detector, tokenizer, pos-tagger, shallow and full syntactic parser, and named-entity detector, using the Maxent Java package for training and using maximum entropy models.
- RWeka is a interface to Weka which is a collection of machine learning algorithms for data mining tasks written in Java. Especially useful in the context of natural language processing is its functionality for tokenization and stemming.
- tidytext provides means for text mining for word processing and sentiment analysis using dplyr, ggplot2, and other tidy tools.
- monkeylearn provides a wrapper interface to machine learning services on Monkeylearn for text analysis, i.e., classification and extraction.
- For more R packages for text mining, please check: https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

Name	Main Features	Language	Creators	Website
Antelope framework	Part-of-speech tagging, dependency parsing, WordNet lexicon	C#, VB.net	Proxem	[1]
Apertium	Machine translation for language pairs from Spanish, English, French, Portuguese, Catalan and Occitan	C++, Java	(various)	[2]
ClearTK	Wrappers for machine learning libraries(SVMlight, LibSVM, OpenNLP MaxEnt) and NLP tools (Snowball Stemmer, OpenNLP, Stanford CoreNLP)	Java	The Center for Computational Language and Education Research at the University of Colorado Boulder	[3]
cTakes	Sentence boundary detection, tokenization, normalization, POS tagging, chunking, context(family history, symptoms, disease, disorders, procedures) annotator, negation detection, dependency parsing, drug mention annotator	Java	Children's Hospital Boston, Mayo Clinic	[4]
DELPH-IN	Deep linguistic analysis: head-driven phrase structure grammar (HPSG) and minimal recursion semantic parsing	LISP, C++	Deep Linguistic Processing with HPSG Initiative	[5]
Factorie	scalable NLP toolkit for named entity recognition, relation extraction, parsing, pattern matching, and topic modeling(LDA)	Java	University of Massachusetts Amherst	[6]
FreeLing	Tokenization, sentence splitting, contradiction splitting, morphological analysis, named entity recognition, POS tagging, dependency parsing, co -reference resolution	C++	Universitat Politècnica de Catalunya	[7]
General Architecture for Text Engineering (GATE)	Information extraction(tokenization, sentence splitter, POS tagger, named entity recognition, coreference resolution), machine learning library wrapper(Weka, MaxEnt, SVMLight, RASP, LibSVM), Ontology (WordNet)	Java	GATE open source community	[8]
Graph Expression	Information extraction (named entity recognition, relation and fact extraction, parsing and search problem solving)	Java	Startup huti.ru	[9]

Name	Main Features	Language	Creators	Website
Learning Based Java	POS tagger, Chunking, coreference resolution, named entity recognition	Java	Cognitive Computation Group at UIUC	[10]
LingPipe	Topic classification, named entity recognition, clustering, POS tagging, spelling correction, sentiment analysis, logistic regression, word sense disambiguation	Java	Alias-i	[11]
Mahout	Scalable machine learning libraries (logistic regression, Naïve Bayes, Random Forest, HMM, SVM, Neural Network, Boosting, K-means, Fuzzy K-means, LDA, Expectation Maximization, PCA)	Java	Online community	[12]
Mallet	Document classification(Naïve Bayes, Maximum Entropy, decision trees), sequence tagging (HMM, MEMM, CRF), topic modeling (LDA, Hierarchical LDA)	Java	University of Massachusetts Amherst	[13]
MetaMap	Map biomedical text to the UMLS Metathesaurus and discover Metathesaurus concepts referred to in text.	Java	National Library of Medicine	[14]
MII nlp toolkit	de-identification tools for free-text medical reports	Java	UCLA Medical Imaging Informatics (MII) Group	[15]
MontyLingua	Tokenization, POS tagging, chunking, extractors for phrases and subject/verb/object tuples from sentences, morphological analysis, text summarization	Python, Java	MIT	[16]
Natural Language Toolkit (NLTK)	Interface to over 50 open access corpora, lexicon resource such as WordNet, text processing libraries for classification, tokenization, stemming, POS tagging, parsing and semantic reasoning.	Python	Online community	[17]
NooJ (based onINTEX)	Morphological analysis, syntactic parsing, named entity recogntion	.NET Framework-based	University of Franche-Comté, France	[18]

Name	Main Features	Language	Creators	Website
OpenNLP	Tokenization, sentence segmentation, POS tagging, named entity extraction, chunking, parsing, coreference resolution	Java	Online community	[19]
Pattern	Wrapper for Google, Twitter and Wikipedia API, web crawler, HTML DOM parsing, POS tagging, n-gram search, sentiment analysis, WordNet, machine learning algorithms for clustering and classification, network analysis and visualization	Python	Tom De Smedt, CLiPS,University of Antwerp	[20]
PSI-Toolkit	Text preprocessing, sentence splitting, tokenization, lexical and morphological analysis, syntactic/ semantic parsing, machine translation	C++	Adam Mickiewicz University in Poznań	[21]
ScalaNLP	Tokenization, POS tagging, sentence segmentation, sequence tagging (CRF, HMM), machine learning algorithms(linear regression, Naïve Bayes, SVM, K-Means, LDA, Neural Network)	Scala	David Hall and Daniel Ramage	[22]
Stanford NLP	Tokenization, POS tagging, named entity recognition, parsing, coreference, topic modeling, classification (Naïve Bayes, logistic regression, maximum entropy), sequence tagging(CRF)	Java	The Stanford Natural Language Processing Group	[23]
Rasp	Tokenization, POS tagging, lemmatization, parsing	C++	University of Cambridge, University of Sussex	[24]
Natural	Tokenization, stemming, classification (Naïve Bayes, logistic regression),morphological analysis, WordNet	JavaScript, Node Js	Chris Umbel	[25]
Text Engineering Software Laboratory (Tesla)	Tokenization, POS tagging, sequence alignment	Java	University of Cologne	[26]
Treex	Machine translation	Perl	Charles University in Prague	[27]

Name	Main Features	Language	Creators	Website
UIMA	Industry standard for content analytics, contains a set of rule based and machine learning annotators and tools	Java / C++	Apache	[28]
VisualText	Tokenization, POS tagging, named entity recognition, classification, text summarization	NLP++ / compiles to C++	Text Analysis International, Inc	[29]
WebLab-project	Language identification, named entity recognition, semantic analysis, relation extraction, text classification and clustering, text summarization	Java / C++	OW2	[30]
UniteX	Tokenization, sentence boundary detection, parsing, morphological analysis, rule-based named entity recognition, text alignment, word sense disambiguation	Java & C++	Laboratoire d'Automatique Documentaire et Linguistique	[31]
The Dragon Toolkit	tools for accessing PubMed, TREC collection, NewsGroup articles, Reuters Articles, and Google Search Engine, ontologies(UMLS, WordNet, MeSH), tokenization, stemming, POS tagging, named entity recognition, classification(Naïve Bayes, SVM-light, LibSVM, logistic regression), clustering(K-Means, hierarchical clustering), topic modeling(LDA), text summarization,	Java	Drexel University	[32]
Text Extraction, Annotation and Retrieval Toolkit	Tokenization, chunking, sentence segmenting, parsing, ontology(WordNet), topic modeling(LDA), named entity recognition, stemming, machine learning algorithms(decision tree, SVM, neural network)	Ruby	Louis Mullie	[33]
Zhihuita NLP API	Chinese text segmentation, spelling checking, pattern matching,	C	Zhihuita.org	[34]

Introduction to Web Mining

How big is the Web ?

- Number of pages
  - Technically, infinite
    - Because of dynamically generated content
    - Lots of duplication (30-40%)
  - Best estimate of “unique” static HTML pages comes from search engine claims
    - Google = 50 billion
    - Bing = 5 billion

The web as a graph

- Pages = nodes, hyperlinks = edges
  - Ignore content
  - Directed graph

- High linkage
  - 8-10 links/page on average

Web Mining

- What is Web Mining
  - Discovering useful information from the World-Wide Web and its usage patterns

- Applications
  - Web search: e.g., Google, Bing, YouTube, Twitter…
  - Vertical Search: e.g., FatLens, Become,…
  - Recommendations: e.g., Amazon.com, Netflix
  - Advertising: e.g., Google Adsense
  - Fraud detection: click fraud detection, …
  - Web site design: e.g., landing page optimization

Web Mining

- The Web is perhaps the single largest and distributed data source in the world that is easily accessible.

- It consists of:
  - Web usage mining: discover user access patterns from usage logs, e.g., clickstreams.
  - Web structure mining: discover knowledge from hyperlinks.
  - Web content mining: mine knowledge from page contents.

Web Mining

- Web Mining vs Data Mining
  - Structure: textual information and linkage structure
  - Scale: data generated per day is comparable to largest conventional data warehouses
  - Speed: often need to react to evolving usage patterns in real-time (e.g., merchandising)

Web Mining Topics

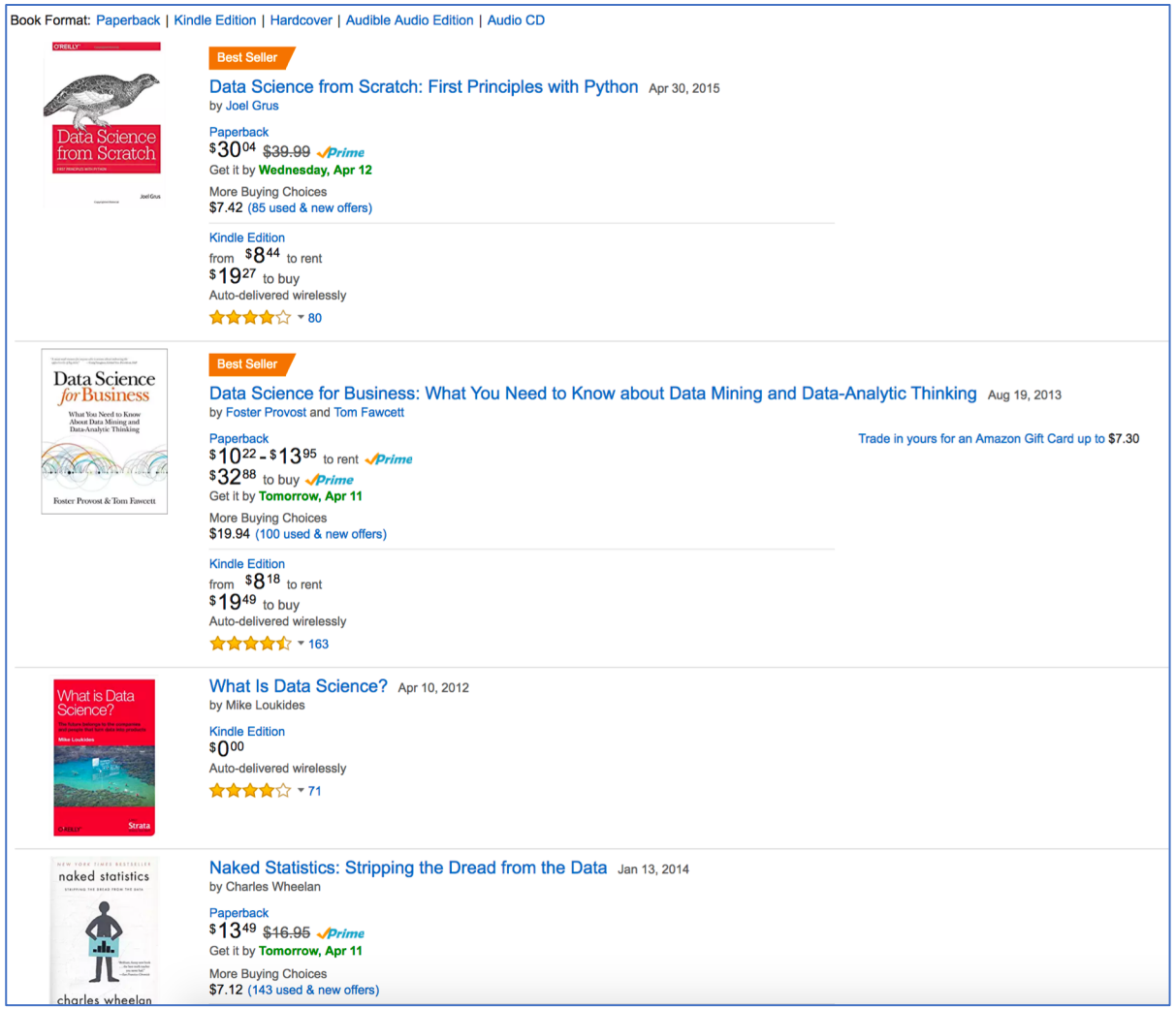
- Crawling the web
- Web graph analysis
- Structured data extraction
- Classification and vertical search
- Recommender systems
- Web advertising and optimization
- Mining web logs
- System issues

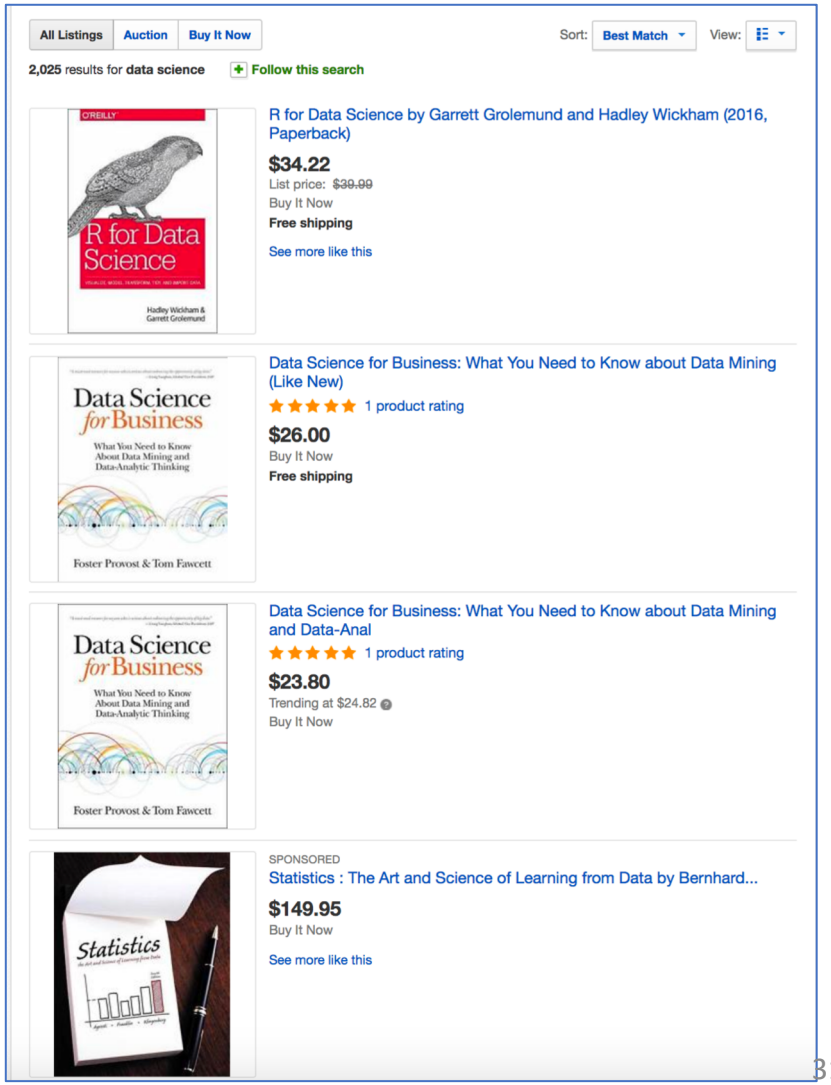
Web Content Mining

Structured Data Extraction

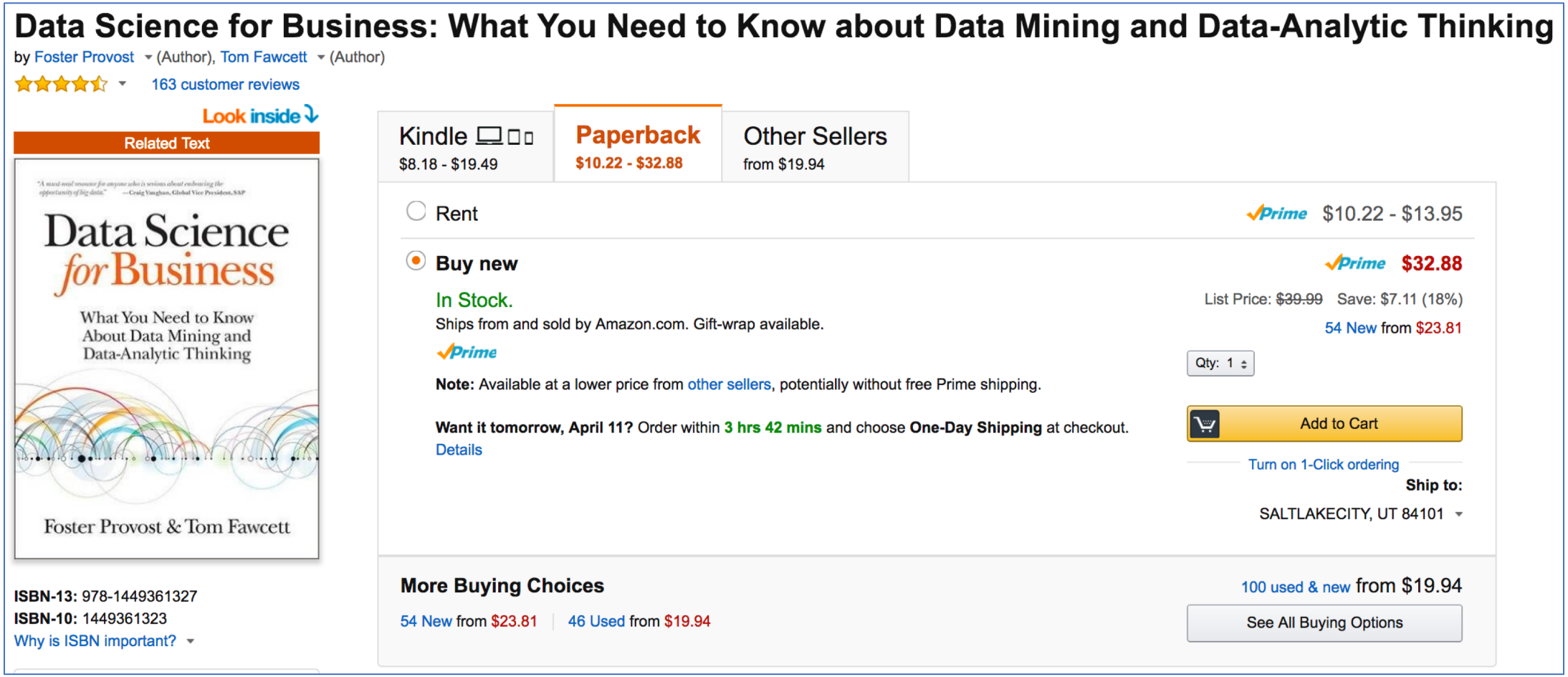
- A large amount of information on the Web is contained in regularly structured data objects.
  - often data records retrieved from databases.
- Such Web data records are important: lists of products and services.
- Applications: Gather data to provide valued added services
  - comparative shopping, object search (rather than page search), etc.
- Two types of pages with structured data:
  - List pages, and detail pages

List pages – Lists of Products





Detail Pages – detail description



Detail Pages

Product details

Paperback: 414 pages
Publisher: O'Reilly Media; 1 edition (August 19, 2013)
Language: English
ISBN-10: 1449361323
ISBN-13: 978-1449361327
Product Dimensions: 7 x 0.9 x 9.2 inches
Shipping Weight: 1.6 pounds (View shipping rates and policies)
Average Customer Review: ★★★★½ (163 customer reviews)
Amazon Best Sellers Rank: #6,027 in Books (See Top 100 in Books)
#3 in Books > Business & Money > Skills > **Business Mathematics**
#3 in Books > Computers & Technology > Databases & Big Data > **Data Mining**
#4 in Books > Textbooks > Computer Science > **Database Storage & Design**

If you are a seller for this product, would you like to **suggest updates through seller support**?

Customer Reviews

★★★★½ 163
4.5 out of 5 stars

5 star	77%
4 star	12%
3 star	7%
2 star	4%
1 star	0%

Share your thoughts with other

Write a customer review

See all verified purchase reviews

Top Customer Reviews

★★★★★ **The perfect balance**
By m l on August 10, 2013
Format: Kindle Edition

When trying to learn about a new field, one of the most common difficulties is to find books (and other materials) that have the right "depth". All too often one ends up with either a friendly but largely useless book that oversimplifies or a heavy academic tome that, though authoritative and comprehensive, is condemned to sit gathering dust in one's shelves. "Data Science for Business" gets it just right.

What I mean might become clearer if I point out what this book is *not*:

- It is *not* a computer science textbook with a focus on theoretical derivations and algorithms.

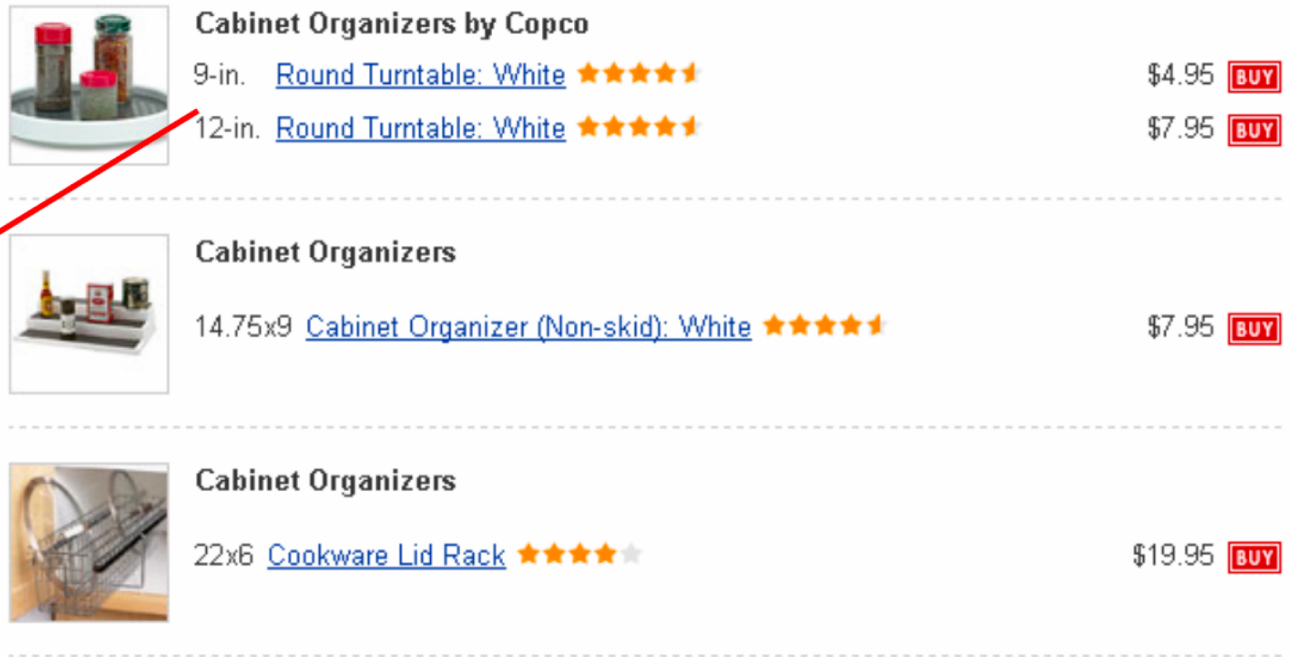
- It is *not* a "cookbook" that provides "step-by-step" guidance with little to no explanation of what one is doing.

- It is *not* your standard "management" title on the cool tech du jour available at airport stands and meant to be read in one sitting (buzzwords, hype and overly enthusiastic statements making up for the dearth of actual content).

Instead, it is close to being the perfect guide for the intelligent reader who -- regardless of whether s/he has a tech background -- has a sincere desire to learn how the tools and principles of data science can be used to extract meaningful information from huge datasets. Highly recommended.

2 comments | 226 people found this helpful. Was this review helpful to you? Yes No Report abuse

Extraction Task: an illustration



nesting

image 1	Cabinet Organizers by Copco	9-in.	Round Turntable: White	*****	$4.95
image 1	Cabinet Organizers by Copco	12-in.	Round Turntable: White	*****	$7.95
image 2	Cabinet Organizers	14.75x9	Cabinet Organizer (Non-skid): White	*****	$7.95
image 2	Cabinet Organizers	22x6	Cookware Lid Rack	*****	$19.95

Word-of-Mouth on the Web

- The Web has dramatically changed the way that people express their opinions. One can
  - post reviews of products at merchant sites, and
  - express opinions on almost anything in forums, discussion groups, and blogs, which are collectively called the user generated content.
- We only focus on mining product reviews here. Extract and summarize opinions in reviews.
- Benefits:
  - Potential Customer: No need to read many reviews
  - Product manufacturer: market intelligence, product benchmarking.

Web Usage Mining

Introduction

Preprocessing

Pattern Discovery

Pattern Analysis

Introduction

- **Web usage mining**: automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites.

- **Goal**: analyze the behavioral patterns and profiles of users interacting with a Web site.

- The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests.

Introduction

- Data in Web Usage Mining:
  - Web server logs
  - Site contents
  - Data about the visitors, gathered from external channels
  - Further application data
- Not all these data are always available.
- When they are, they must be integrated.
- A large part of Web usage mining is about processing usage/ clickstream data.
  - After that various data mining algorithm can be applied.

Data mining

Frequent Itemsets

- The “Home Page” and “Shopping Cart Page” are accessed together in 20% of the sessions.
- The “Donkey Kong Video Game” and “Stainless Steel Flatware Set” product pages are accessed together in 1.2% of the sessions.

Association Rules

- When the “Shopping Cart Page” is accessed in a session, “Home Page” is also accessed 90% of the time.
- When the “Stainless Steel Flatware Set” product page is accessed in a session, the “Donkey Kong Video” page is also accessed 5% of the time.

Sequential Patterns

- add an extra dimension to frequent itemsets and association rules - time
- “x% of the time, when A appears in a transaction, B appears within z transactions.”
- Example:The “Video Game Caddy” page view is accessed after the “Donkey Kong Video Game” page view 50% of the time. This occurs in 1% of the sessions.

Data mining (cont.)

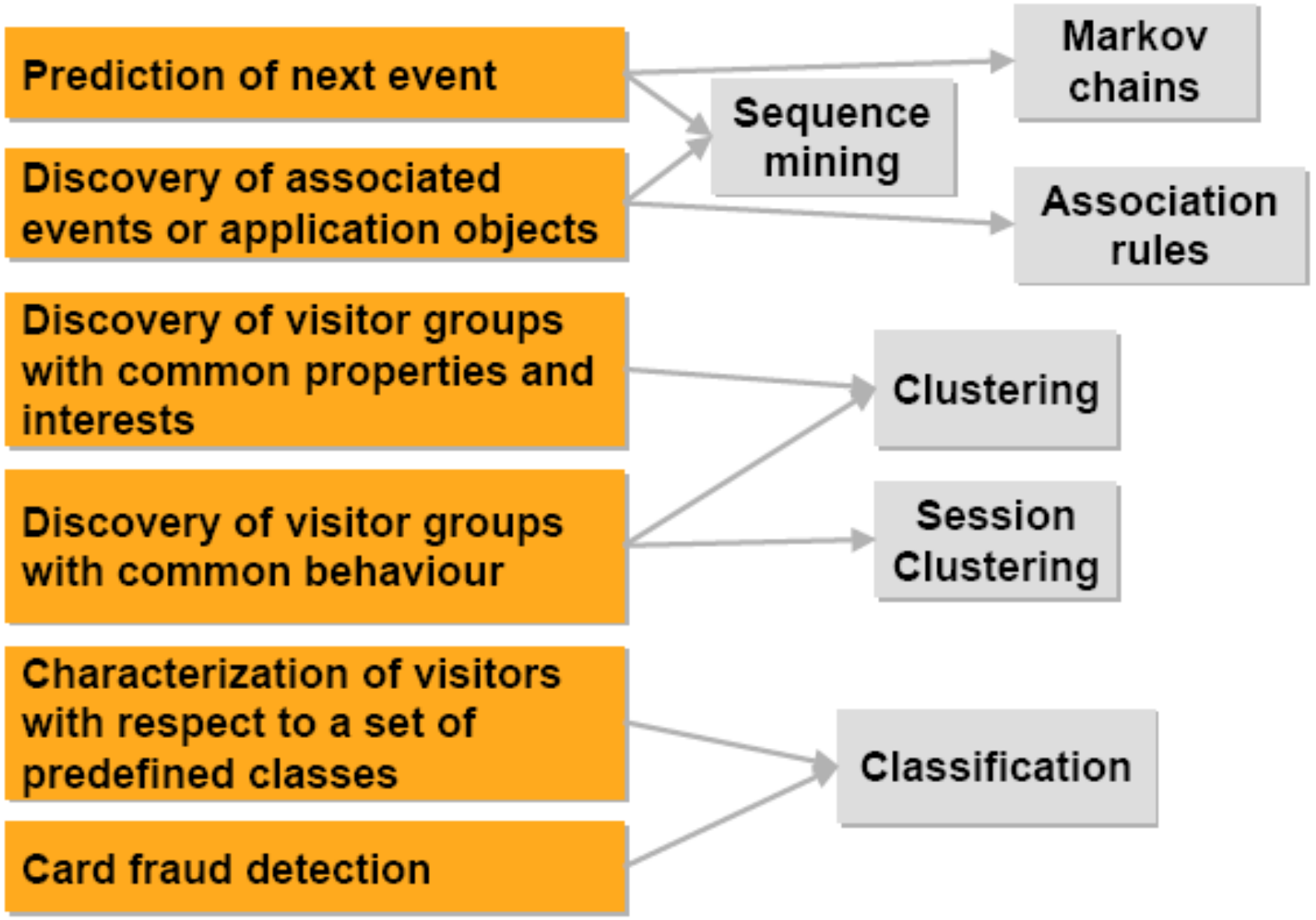
Clustering: Content-Based or Usage-Based

- Customer/visitor segmentation
- Categorization of pages and products

Classification

- "Donkey Kong Video Game", "Pokemon Video Game", and "Video Game Caddy" product pages are all part of the Video Games product group.
- customers who access Video Game Product pages, have income of 50K+, and have 1 or more children, should be get a banner ad for Xbox in their next visit.

Some usage mining applications



Talk with your neighbors

- Discuss with your neighbors for 5 min

- Imagine you are a data scientist at Google (YouTube).
- Based on what we discussed today and your prior knowledge
  - What would you do to improve the user satisfaction of recommended videos?