FACULTÉ DES SCIENCES DHAR EL MAHRAZ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH

Text Mining and Web Mining: Overview

El Habib NFAOUI (elhabib.nfaoui@usmba.ac.ma)
LIIAN Laboratory, Faculty of Sciences Dhar Al Mahraz, Fes
Sidi Mohamed Ben Abdellah University, Fes

2018-2019

Outline

1. What is Data Mining?
2. Text Mining
3. Web Mining
4. Web Content Mining
5. Web Structure Mining

1. What is Data Mining?

- Data mining is also called **knowledge discovery in databases (KDD)**. It is commonly defined as the process of discovering useful **patterns** or **knowledge** from data sources, e.g., **databases**, **texts**, **images**, the **Web**, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization.

- There are many data mining tasks. Some of the common ones are **supervised learning** (or **classification**), **unsupervised learning** (or **clustering**), **association rule mining**, and **sequential pattern mining**.

1. What is Data Mining?

- A data mining application usually starts with an understanding of the application domain by **data analysts** (**data miners**), who then identify suitable data sources and the target data. With the data, data mining can be performed, which is usually carried out in three main steps:
  - **Pre-processing**: The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute or feature selection. Details about data pre-processing can be found in any standard data mining textbook.
  - **Data mining**: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.
  - **Post-processing**: In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various evaluation and visualization techniques are used to make the decision.
- The whole process (also called the **data mining process**) is almost always iterative. It usually takes many rounds to achieve the final satisfactory result, which is then incorporated into real-world operational tasks.

2. Text Mining

- **Text mining** and text analytics are broad umbrella terms describing a range of technologies for analyzing and processing semi structured and unstructured text data. The unifying theme behind each of these technologies is the need to "**turn text into numbers**" then powerful algorithms can be applied to large document databases. Converting text into a structured, numerical format and applying analytical algorithms require knowing how to both use and combine techniques for handling text, ranging from individual words to documents to entire document databases.

- The origin of text mining as a field is twofold. The name exists as a homage to data mining; it has been suggested (Hearst, 1999) that an appropriate name for text mining would be "text data mining", implying that text data mining is a variation on the general field of data mining and exists as a subfield of that more generic field. Text mining is defined by Tuffery (2011) as "**the automatic processing of natural language text data available in reasonably large quantities in the form of computer files, with the aim of extracting and structuring their contents and themes, for the purposes of rapid (nonliterary) analysis, the discovery of hidden data, or automatic decision making**".

2.1 Practice Areas of Text Analytics

- Text mining **can be divided** into "**seven practice areas"**, based only on the practical distinctions in data and goal for an analyst trying to solve a given problem. Though distinct, these areas are **highly interrelated**; a typical text mining project will **require techniques from multiple areas**. The seven practice areas are as follows:

1. **Search and information retrieval (IR):** Storage and retrieval of text documents, including search engines and keyword search.
2. **Document clustering:** Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
3. **Document classification:** Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
4. **Web mining:** Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.
5. **Information extraction (IE):** Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi structured text.
6. **Natural language processing (NLP):** Low-level language processing and understanding tasks (e.g., part of speech tagging); often used synonymously with computational linguistics.
7. **Concept extraction:** Grouping of words and phrases into semantically similar groups.

2.1 Practice Areas of Text Analytics

- These seven practice areas exist at the key intersections of text mining and the six major other **fields that contribute to it**. Figure 1 depicts, as a Venn diagram, the overlap of the seven fields of text mining, data mining, statistics, artificial intelligence and machine learning, computational linguistics, library and information sciences, and databases; it also locates the seven practice areas at their key intersections. For example, the practice area of text classification draws from the field of data mining, and the practice area of information retrieval draws from the two fields of databases and library and information sciences.

- Tables 1 and 2 provide alternative methods for identifying the practice areas based on algorithms and desired products.

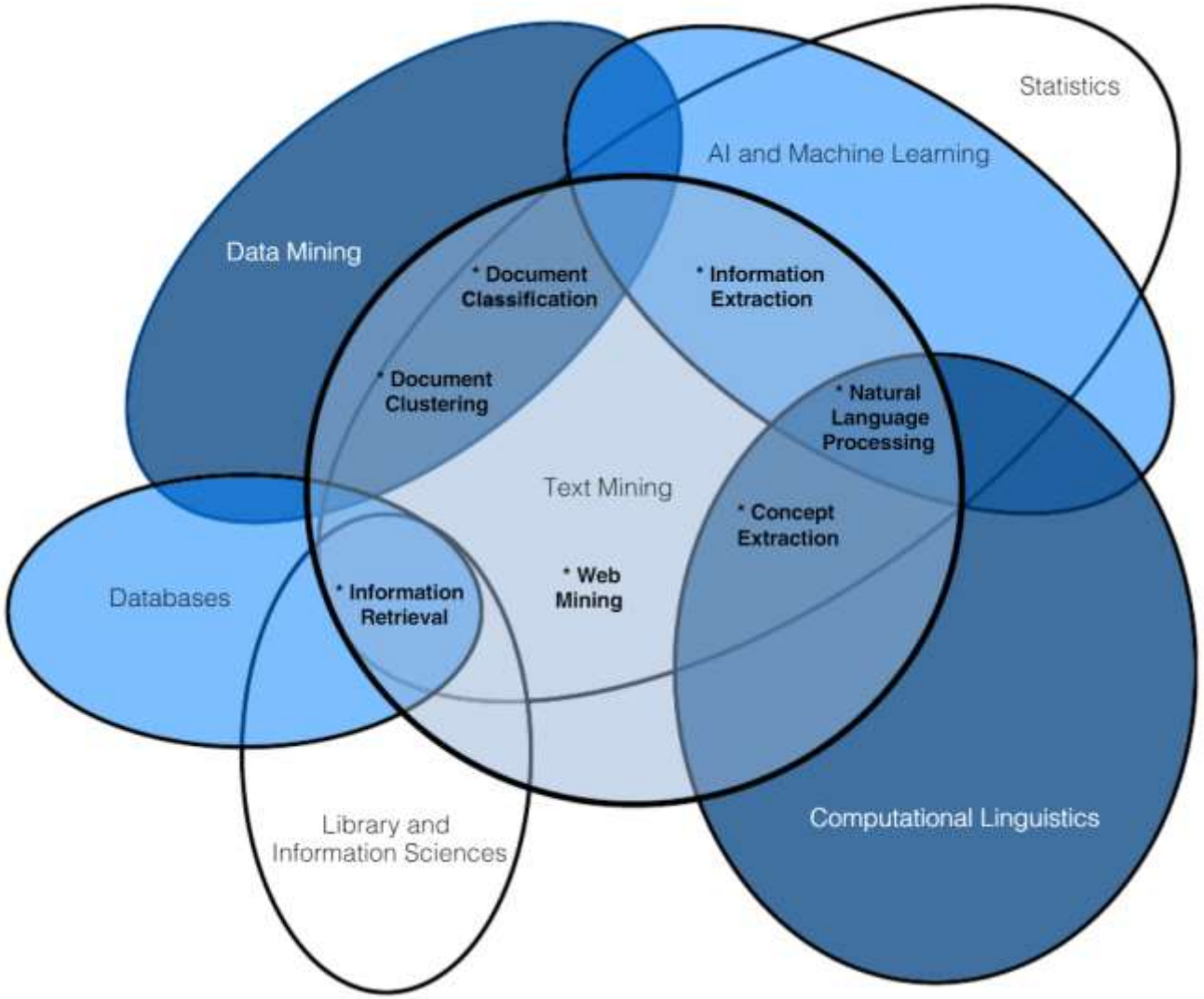


Figure 1. A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields. [1]

Table 1: Text Mining Topics and Related Practice Areas

Topic	Practice Area (Number)
Keyword search	Search and information retrieval
Inverted index	Search and information retrieval
Document clustering	Document Clustering
Document similarity	Document Clustering
Feature selection	Document classification
Sentiment analysis	Document classification
	Web mining
Dimensionality reduction	Document classification
eDiscovery	Document classification
Web crawling	Web mining
Link analytics	Web mining
Entity extraction	Information extraction
Link extraction	Information extraction
Part of speech tagging	Natural language processing
Tokenization	Natural language processing
Question answering	Natural language processing
	Search and information retrieval
Topic modeling	Concept extraction
Synonym identification	Concept extraction

Table 2: Common Text Mining Algorithms and the Corresponding Practice Area

Algorithm	Area
Naïve Bayes	Document classification
Conditional random fields	Information extraction
Hidden Markov models	Information extraction
k-means	Clustering
Singular value decomposition (SVD)	Document classification, clustering
Logistic regression	Document classification
Decision trees	Document classification
Neural network	Document classification
Support vector machines	Document classification
MARSplines	Document classification
Link analysis	Concept extraction
k-nearest neighbors	Document classification
Word clustering	Concept extraction
Regression	Classification

2.2 Practice Areas: Brief Descriptions

The following are brief descriptions of the problems faced in each practice area.

- **Search and Information Retrieval**

Search and information retrieval covers indexing, searching, and retrieving documents from large text databases with keyword queries. With the rise of powerful Internet search engines, including Google, Yahoo!, and Bing, search and information retrieval has become familiar to most people. Nearly every computer application from email to word processing includes a search function.

- **Document Clustering (unsupervised technique)**

Document clustering uses algorithms from data mining to group similar documents into clusters. Clustering algorithms are widely available in many commercial data and text mining software packages.

Formally, given a set of documents and a similarity measure among documents find clusters such that:

- documents in one cluster are more similar to one another
- documents in separate clusters are less similar to one another

2.2 Practice Areas: Brief Descriptions

- **Document Classification (supervised technique)**

Document classification assigns a **known set of labels** to **untagged** documents, using a model of text learned from documents with known labels. Like document clustering, document classification draws from an enormous field of work in data mining, statistics, and machine learning. It is one of the most prominent techniques used in text mining.

- Formally, given a collection of labeled documents (training set) find a model for the class as a function of the values of the features.

- Goal: Previously unseen documents should be assigned a class as accurately as possible.

- **Web Mining**

Web mining is its own practice area due to the unique structure and enormous volume of data appearing on the web. Web documents are typically presented in a structured text format with hyperlinks between pages. These differences from standard text present a few challenges and many opportunities. As the Internet becomes even more ingrained in our popular culture with the rise of Facebook, Twitter, and other social media channels, web mining will continue to increase in value. Though it is still an emerging area, web mining draws on mature technology in document classification and natural language understanding.

2.2 Practice Areas: Brief Descriptions

- **Information Extraction**

The goal of information extraction is to construct (or extract) structured data from unstructured text. Information extraction is one of the mature fields within text mining, but it is difficult for beginners to work in without considerable effort, since it requires specialized algorithms and software. Furthermore, the training and tuning of an information extraction system require a large amount of effort. There are a number of commercial products available for information extraction, but all of them require some customization to achieve high performance for a given document database.

- Subtasks
  - Named Entity Recognition and Disambiguation
    - "**M. Smith** likes fishing"
    - Which **M. Smith**?
  - Co-reference Resolution
    - "**M. Smith** likes fishing. But **he** doesn't like biking."
    - Does he refer to M. Smith?
  - Relationship Extraction
    - PERSON works for ORGANIZATION
    - PERSON located in LOCATION

2.2 Practice Areas: Brief Descriptions

- **Natural Language Processing**

Natural language processing (NLP) has a relatively long history in both linguistics and computer science. NLP is a powerful tool for providing useful input variables for text mining such as part of speech tags and phrase boundaries.

- **Concept Extraction**

Extracting concepts is, in some ways, both the easiest and the hardest of the practice areas to do. The meaning of text is notoriously hard for automated systems to “understand.” However, some initial automated work combined with human understanding can lead to significant improvements over the performance of either a machine or a human alone.

As shown in Figure 2, text mining draws upon many techniques in the broader field of text analytics.

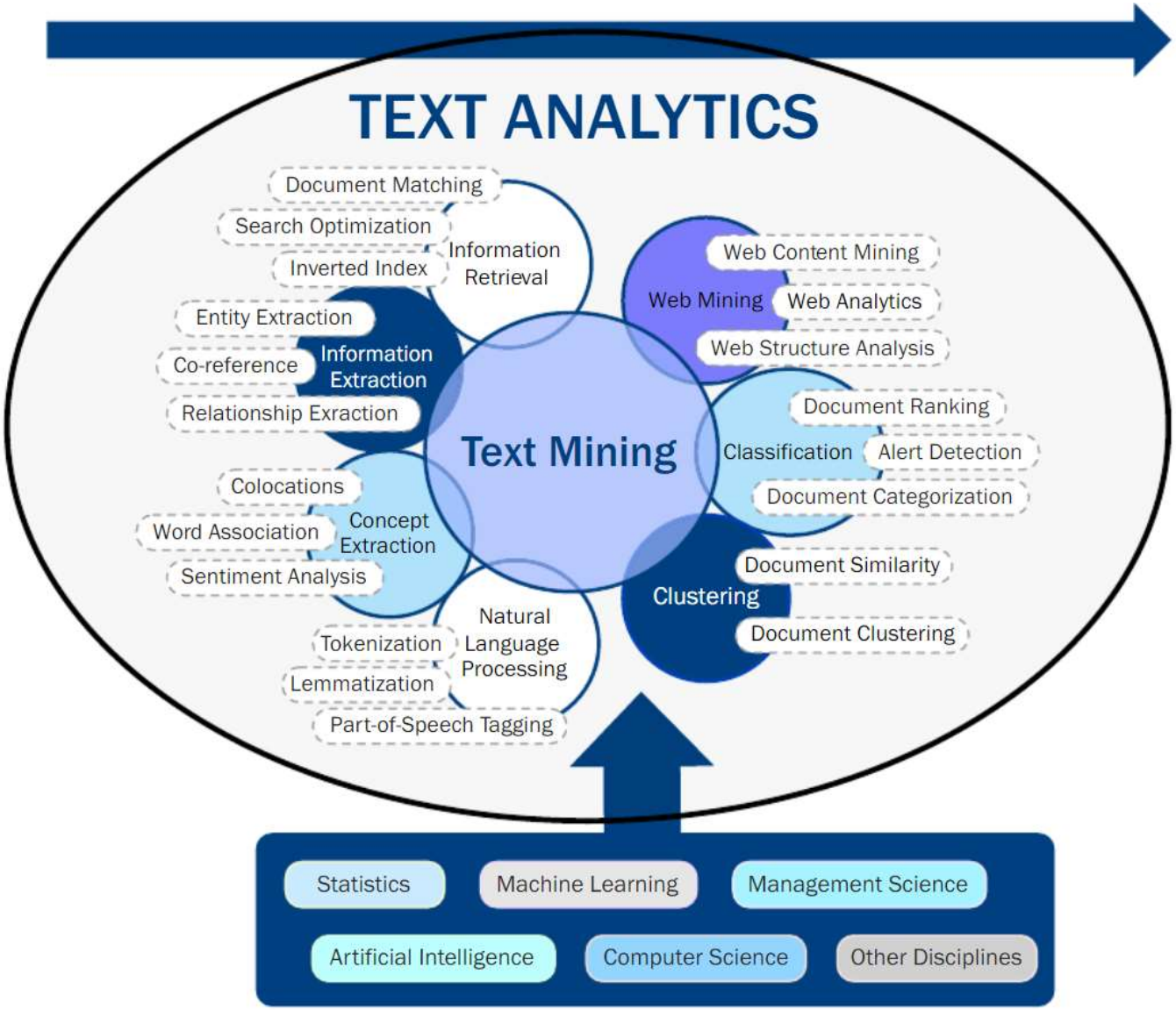


Figure 2. Text mining is proving to be extremely useful, drawing upon contributions of many text analytical components and knowledge from many external disciplines (shown in blue at the bottom), which result in directional decisions affecting external results (shown by the blue arrow at the top) [1].

2.3 Interactions between the Practice Areas

- The seven practice areas **overlap considerably**, since many practical text mining tasks sit at the intersection of multiple practice areas. For example, entity extraction draws from the practice areas of information extraction and text classification, and document similarity measurement draws from the practice areas of document clustering and information retrieval.

- Postscript

  A common claim among data miners is that 80 to 90 percent of the project time is consumed by data preparation steps. The same is true for text mining. In contrast to data mining, where some of the data are in text format, all of the data for text mining are in text format. The initial challenge is to transform these text data into a numerical format for subsequent analysis.

3. Web Mining

- Web mining aims to discover useful information or knowledge from the **Web hyperlink structure**, **page content**, and **usage data**. Although Web mining uses many data mining techniques, but it is not purely an application of traditional data mining techniques due to the heterogeneity and semi-structured or unstructured nature of the Web data. So web mining has developed in a rather confined niche.

- *Web* is a collection of inter-related files on one or more *Web servers.*
  - Web mining aims to extract knowledge from Web data
  - Web data is
    - Web content – text, image, records, etc.
    - Web structure – hyperlinks, tags, etc.
    - Web usage – http logs, app server logs, etc.

3.1 Sub-Fields of Web mining

- Based on the primary **kinds of data used** in the mining process, Web mining tasks can be categorized into three types: **Web structure mining**, **Web content mining** and **Web usage mining**.

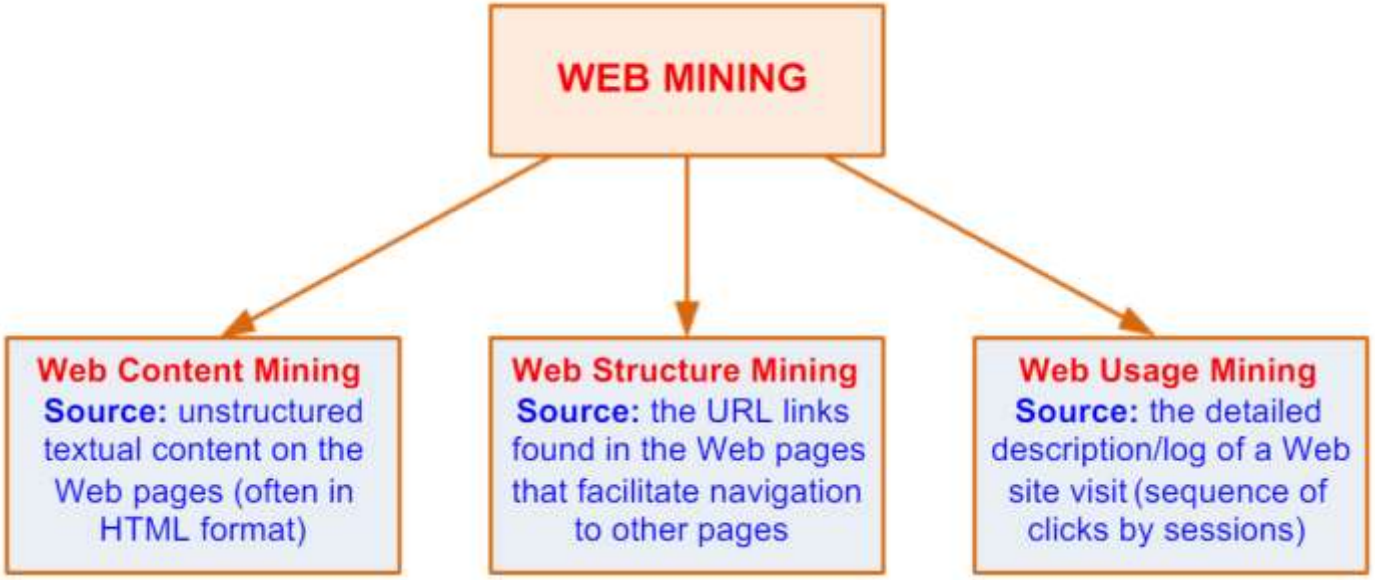


Figure 4. Three components of web mining [1].

Web Mining is a Multi-Disciplinary Field, it draws ideas and techniques from: Machine Learning, Natural Language Processing, Social Network Analysis, Database Systems, …

3.1 Sub-Fields of Web mining

- ***Web content mining*** extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc., for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer opinions. This process tabulates information **from text, image, audio, or video data on the web**. Web content mining sometimes is also called web text mining because text content is used quite often.

- ***Web usage mining*** aims to study user clicks and their applications to e-commerce and business intelligence. The objective is to capture and model behavioral patterns and profiles of users who interact with a Web site. Such patterns can be used to better understand the behaviors of different user segments, to improve the organization and structure of the site, and to create personalized experiences for users by providing dynamic suggestions of products and services using recommender systems.

- ***Web structure mining*** discovers knowledge from hyperlinks, which represent the structure of the Web. It uses graphing methods to illustrate connection structures of websites.

3.2 The Web Mining Process

The **Web mining process** is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages.

Once the data is collected, we go through the same three-step process: data pre-processing, Web data mining and post-processing. However, the techniques used for each step can be quite different from those used in traditional data mining.

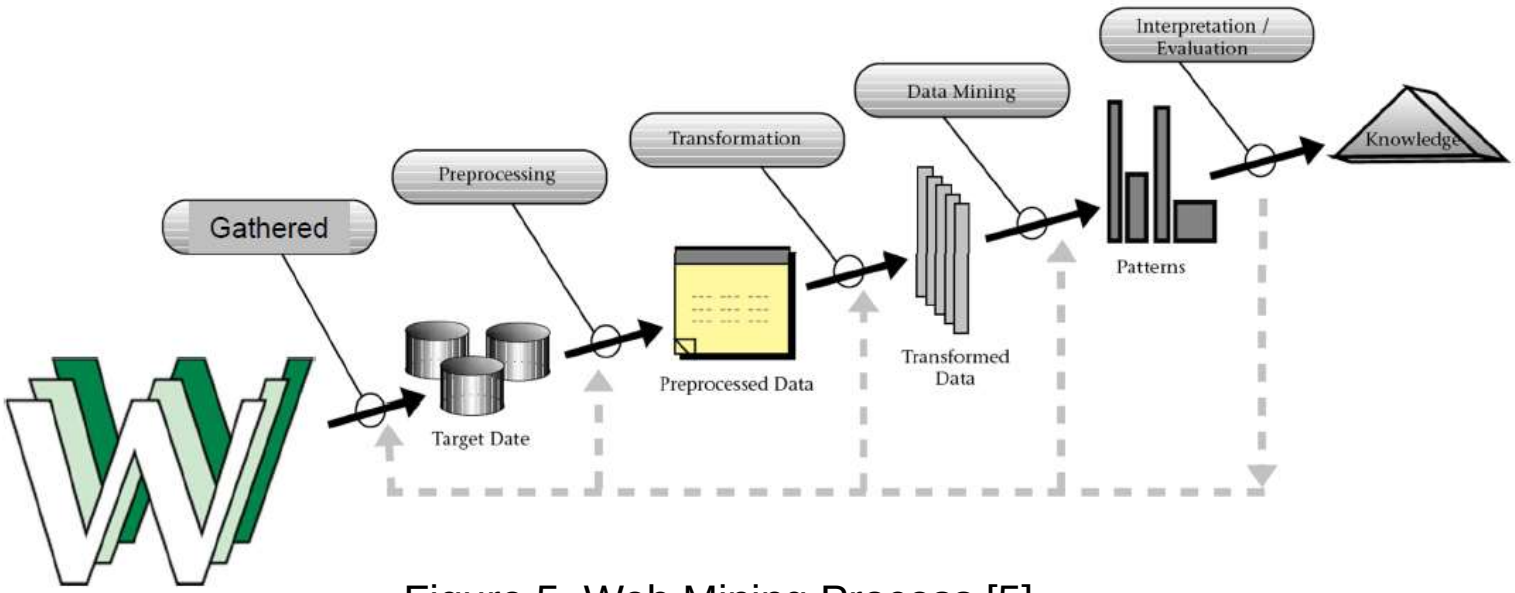


Figure 5. Web Mining Process [5]

3.2 The Web Mining Process

- **Gathering of Web Data**
  - Crawl documents or data
  - Retrieve data via Web API
  - Download pre-gathered data sets
- **Exploration**
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records

3.2 The Web Mining Process

- **Preprocess** and **Transform data into a representation that is suitable for the chosen data mining methods**
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)
- **Methods**
  - Aggregation, sampling
  - Dimensionality reduction / feature subset selection
  - Attribute transformation / text to term vector
  - Discretization and binarization
- **Good data preparation is key to producing valid and reliable models.**
- **Data preparation estimated to take 70-80% of the time and effort of a data mining project!**

3.2 The Web Mining Process

- **Data Mining**
  - Input: Preprocessed Data
  - Output: Model / Patterns

  Steps:

  1. **Apply** data mining method.
  2. **Evaluate** resulting model / patterns.
  3. **Iterate**
     - Experiment with **different parameter settings**.
     - Experiment with **different alternative methods**.
     - Improve **preprocessing** and **feature generation**.
     - **Combine** different methods.

3.3 Recurring Challenges

- huge amount of available data → requires sampling or multiple machines
- un-/semi-structured nature of data
- heterogeneity of data → data integration might be a challenge
- distributed nature of data → often requires large-scale crawling

4. Web Content Mining

- Web Content Mining is the process of extracting useful information from the contents of Web documents (data available online).
  - Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables.

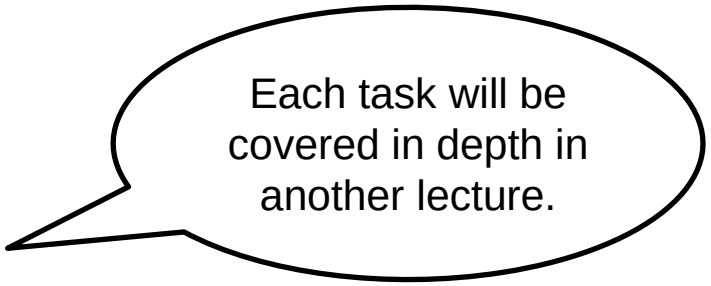
- Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). For example, the content of a collection of web pages can be analyzed using some natural language processing techniques, such as Latent Dirichlet Allocation or Sentiment Analysis tools. These techniques are especially important for extracting subjective information about web users, and so they are widely found in many commercial applications, from marketing to consultancy.

4.1 Web Content Mining Applications

- Identify the topics represented by a Web Documents
- Categorize Web Documents
- Find Web Pages across different servers that are similar
- Applications related to relevance
  - Queries –Enhance standard Query Relevance with User, Role, and/or Task Based Relevance
  - Recommendations –List of top “n” relevant documents in a collection or portion of a collection.
  - Filters –Show/Hide documents based on relevance score
- Collaborative Question Answering
- Event Detection

4.2 Content Mining Tasks

- Content Classification
- Content Clustering
- Associations
- Concept Hierarchy creation
- Content Relevance
- Topic Identification
- Sentiment Analysis
- ….



4.3 Distinct Aspects of Text in Social Media

Textual data in social media have their own unique features. They should be considered when conducting text and web analytics methods.

Time Sensitivity

- An important and common feature of many social media services is their real-time nature. Particularly, bloggers typically update their blogs every several days, while microblogging and social networking users may post news and information several times daily. The large number of real-time updates contain abundant information, which provides a lot of opportunities for detection and monitoring of an event.
- With the rapid evolution of content and communication styles in social media, text is changing too. Different from traditional textual data, the text in social media is not independent and identically distributed data anymore. A comment or post may reflect the user's interest, and a user is connected and influenced by his friends. People will not be interested in a movie after several months, while they may be interested in another movie released several years ago because of the recommendation from his friends; reviews of a product may change significantly after some issues, like the comments on Toyota vehicles after the break problem. All these problems originate from the time sensitivity of textual data in social media.

4.3 Distinct Aspects of Text in Social Media

- **Short Length (Shortness)**

Certain social media web sites restrict the length of user-created content such as microblogging messages, product reviews, QA passages and image captions, etc.

- Twitter allows users to post news quickly and the length of each tweet is limited to 280 characters.
- Picasa comments are limited,
- Personal status messages on Windows Live Messenger are restricted to 128 characters.

As we can see, data with a short length is ubiquitous on the web at present. As a result, these short messages have played increasing important roles in applications of social media. Successful processing short texts is essential to text analytics methods.

Unlike standard text with lots of words and their resulting statistics, short messages consist of few phrases or sentences. They cannot provide sufficient context information for effective similarity measure, the basis of many text processing methods.

4.3 Distinct Aspects of Text in Social Media

- **Unstructured Phrases, Informality and Implicitness:**
  - An important difference between the text in social media and traditional media is the variance in the quality of the content.
    - First, the variance of quality originates from people's attitudes when posting a micro blogging message or answering a question in a forum. Some users are experts for the topic and post information very carefully, while others do not post as high of quality. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content. This makes the tasks of filtering and ranking in such systems more complex than in other domains.
    - Second, when composing a message, users may use or coin new abbreviations or acronyms that seldom appear in conventional text documents. For example, messages like "How r u?", "Good 9t" are not really words, but they are intuitive and popular in social media. They provide users convenience in communicating with each other, however it is very difficult to accurately identify the semantic meaning of these messages.
  - Free usage of creative language, high contextualization, informal writings, and the use of alias/morphs.

4.3 Distinct Aspects of Text in Social Media

- **Noiseness** (noisy content):

Besides the unstructured expressions, the text is sometimes “noisy” for a specific topic. For instance, one QA passage in Yahoo! Answers “I like sony” should be noisy data to a post that is talking about iPad 2 release. It is difficult to classify the passage into corresponding classes without considering its context information.

- The Pear Analytics report* on 2000 sample tweets demonstrated that :
  - 40.55% of the tweets are pointless babble,
  - 37.55% are conversations,
  - and **only 8.7%** have pass-along value.

4.3 Distinct Aspects of Text in Social Media

- **Abundant Information**

Social media in general exhibit a rich variety of information sources. In addition to the content itself, there is a wide array of non-content information available. For example:

- Twitter allows users to utilize the “#” symbol, called hashtag, to mark keywords or topics in a Tweet (tag information);
- an image is usually associated with multiple labels which are characterized by different regions in the image;
- users are able to build connection with others (link information) in Facebook and other social network sites;
- Wikipedia provides an efficient way for users to redirect to the ambiguity concept page or higher level concept page (semantic hierarchy information).

The text analytics in social media is able to derive data from various aspects, which include user, content, link, tag, time stamp etc.

4.4 Content Preparation

- **Gathering of Web Data**
  - Retrieve data via Web API
  - Download pre-gathered data sets
  - Crawl documents or data (as an example, Forums do not provide programmatic interfaces (APIs) to capture data). The process of data extraction automatically from websites is called '**web scraping**'.

- Each web page is usually gathered and organized (using a **parsing** technique), processed to remove the unimportant parts from the text (Natural Language Processing), and then analyzed.

- Note that in text mining, authors use the term **document** to describe the unit of text under analysis. This is a broader definition. In practice, this could be mean typical documents, paragraphs, sentences, “tweets” on social media, or other defined sections of text.

4.4 Content Preparation

- **Parsing**

A web page is written in HTML format, so the first operation is to extract the relevant pieces of information. An HTML parser builds a tree of tags from which the content can be extracted (figure below). Nowadays, there are many parsers available. For example, **Python Scrapy** is an application framework for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like web mining, text mining, data mining, information processing or historical archival.

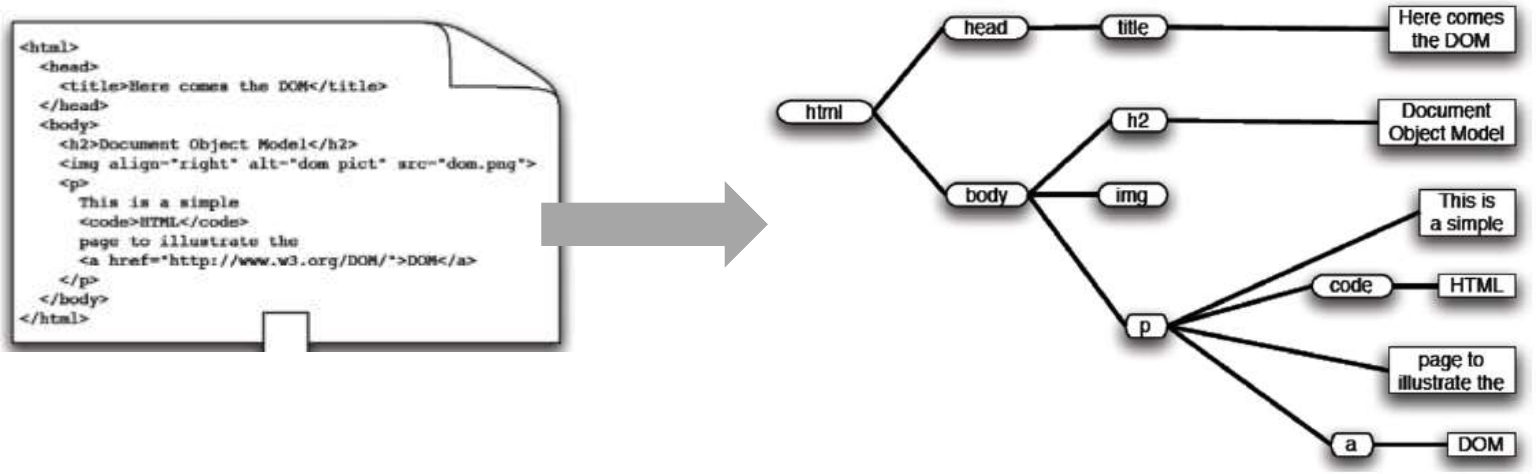


Figure 6. Illustration of the DOM (or tag) tree built from a simple HTML page. Internal nodes (shown as ovals) represent HTML tags, with the <html> tag as the root. Leaf nodes (shown as rectangles) correspond to text chunks.

4.4 Content Preparation

Example: Python Scrapy

Here's an example of a typical Scrapy shell session where we start by scraping the main page of Python language, https://www.python.org/.

First, we launch the shell (figure below):

```
scrapy shell "https://www.python.org/" --nolog
```

Then, the shell fetches the URL (using the Scrapy downloader) and prints the list of available objects and useful shortcuts (you'll notice that these lines all start with the [s] prefix). After that, we can start using those objects for parsing the page.

4.4 Content Preparation

For example, we can obtain the title's text using the response object and the xpath language.

```
In [1]: response.xpath('//title/text()').extract()
Out[1]: ['Welcome to Python.org']
```

Or we want to extract all the embedded links in page (this operation is needed for the crawler to work), which are usually put on <a>, and the URL value is on an href attribute:

```
In [2]: response.xpath("//a/@href").extract()
Out[2]:
['#content',
 '#python-network',
 '/',
 '/psf-landing/',
 'https://docs.python.org',
 'https://pypi.python.org/',
```

```
Invite de commandes - scrapy  shell "https://www.python.org/" --nolog
Microsoft Windows [version 10.0.16299.371]
(c) 2017 Microsoft Corporation. Tous droits réservés.

C:\Users\HP>scrapy shell "https://www.python.org/" --nolog
[s] Available Scrapy objects:
[s]   scrapy     scrapy module (contains scrapy.Request, scrapy.Selector, etc)
[s]   crawler    <scrapy.crawler.Crawler object at 0x000001E74C381940>
[s]   item       {}
[s]   request    <GET https://www.python.org/>
[s]   response   <200 https://www.python.org/>
[s]   settings   <scrapy.settings.Settings object at 0x000001E74D597C50>
[s]   spider     <DefaultSpider 'default' at 0x1e74d837860>
[s] Useful shortcuts:
[s]   fetch(url[, redirect=True]) Fetch URL and update local objects (by default, redirects are followed)
[s]   fetch(req)                  Fetch a scrapy.Request and update local objects
[s]   shelp()           Shell help (print this help)
[s]   view(response)    View response in a browser
In [1]: response.xpath('//title/text()').extract()
Out[1]: ['Welcome to Python.org']

In [2]: response.xpath("//a/@href").extract()
Out[2]:
['#content',
 '#python-network',
 '/',
 '/psf-landing/',
 'https://docs.python.org',
 'https://pypi.python.org/',
 '/jobs/',
 '/community/',
 '#top',
 '/',
 '#site-map',
 '#',
 'javascript:;',
 'javascript:;',
 'javascript:;',
 '#',
 'http://plus.google.com/+Python',
 'http://www.facebook.com/pythonlang?fref=ts',
 'http://twitter.com/ThePSF',
```

Figure 7. Scraping the main page of Python language and extracting useful information

5. Web Structure Mining

- **Definition**

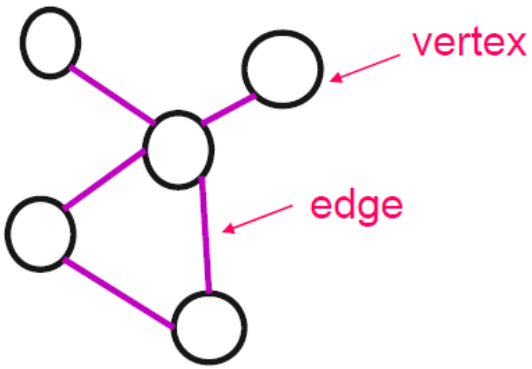
  Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, discover communities of users who share common interests. We can also discover the social ties (relations) among actors that interact on the Web. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

- Focuses on the **structure**, but can of course also be **combined with content or usage mining techniques**.
- The research at the hyperlink level is also called ***Hyperlink Analysis.***

- **Typical Sources of Data**
  - Web crawls including HTML pages and hyperlinks
  - crawls of the blogosphere
  - social networks including explicit relations between actors (your Facebook friend network)
  - other types of community data (discussion forums, email conversations, …)

5. Web Structure Mining

- Web as a graph. A Graph is a collection of vertices that are connected by lines.
  - Nodes = websites
  - Edges = links
    - Use algorithms from graph theory



Community	Points	Lines
Math	vertices	lines: edges, arcs
Computer Science	nodes	links
Physics	sites	bonds
Sociology	actors	ties, relations

Tasks and applications:

- Find out popular Websites (Google Page Rank)
- Web Communities Detection
- Prominence: Who are the “most important” actors in a social network?
- ……..

References

[1] John Elder, Dursun Delen, Thomas Hill, Gary Miner and Bob Nisbet. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Pub. Date: 2012, pages: 1093, ISBN: 978-0-12-386979-1. Publisher: Elsevier Science

[2] Charu C. Aggarwal, Chengxiang Zhai. MINING TEXT DATA. ISBN: 9781461432234 1461432235. Springer US, 2012.

[3] Isoni Andrea. Machine Learning for the Web. Pub. Date: 2016, pages: 299, ISBN: 978-1-78588-660-7. Publisher: Packt Publishing

[4] Bing Liu. Web Data Mining. Pub. Date: 2011, Second Edition, pages: 622. ISBN: 978-3-642-19459-7. Publisher: Springer-Verlag Berlin Heidelberg

[5] Christian Bizer, Cäcilia and Zirn Oliver Lehmberg. Web Mining course.

[6] Jaideep Srivastava. Web Mining : Accomplishments & Future Directions, University of Minnesota, USA