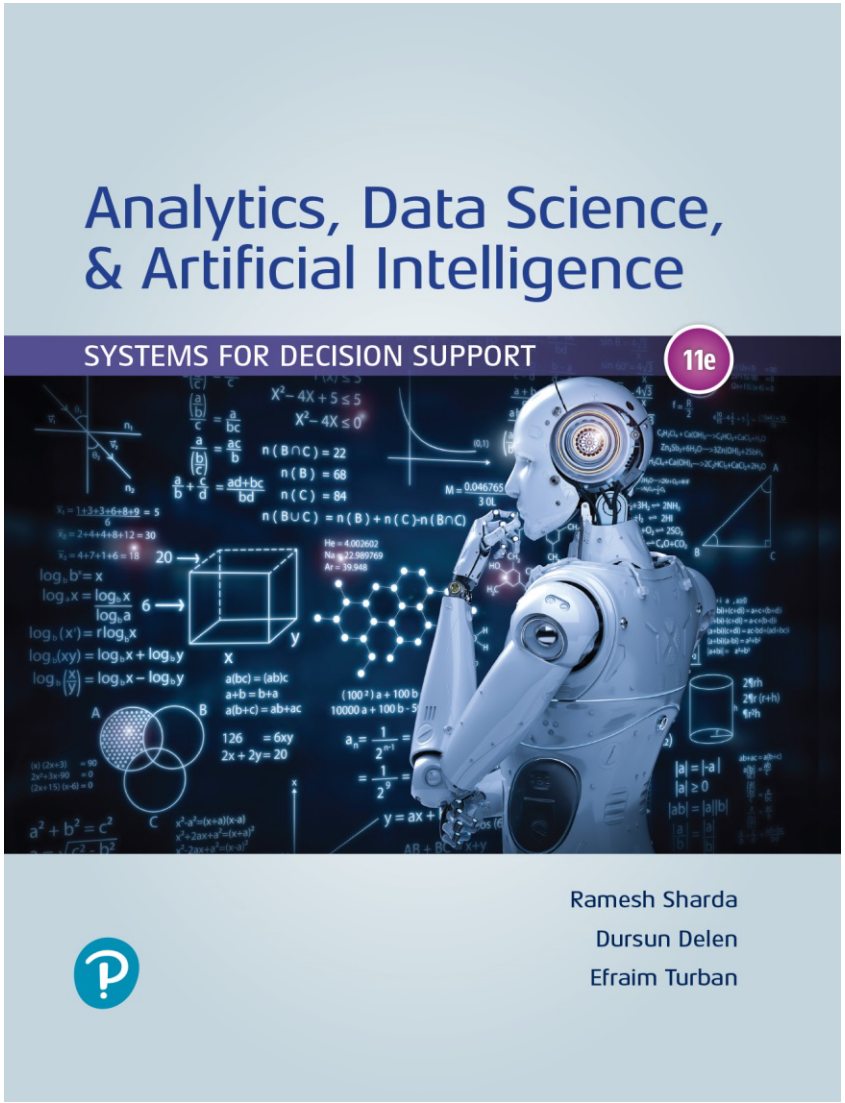
Analytics, Data Science and AI: Systems for Decision Support

Eleventh Edition



Chapter 7

Text Mining, Sentiment Analysis, and Social Analytics

Learning Objectives (1 of 2)

7.1 Describe text mining and understand the need for text mining

7.2 Differentiate among text analytics, text mining and data mining

7.3 Understand the different application areas for text mining

7.4 Know the process of carrying out a text mining project

7.5 Appreciate the different methods to introduce structure to text-based data

7.6 Describe sentiment analysis

Learning Objectives (2 of 2)

7.7 Develop familiarity with popular applications of sentiment analysis

7.8 Learn the common methods for sentiment analysis

7.9 Become familiar with speech analytics as it relates to sentiment analysis

7.10 Learn three facets of Web analytics—content, structure, and usage mining

7.11 Know social analytics including social media and social network analyses

Text Mining Concepts (1 of 2)

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Text Mining Concepts (2 of 2)

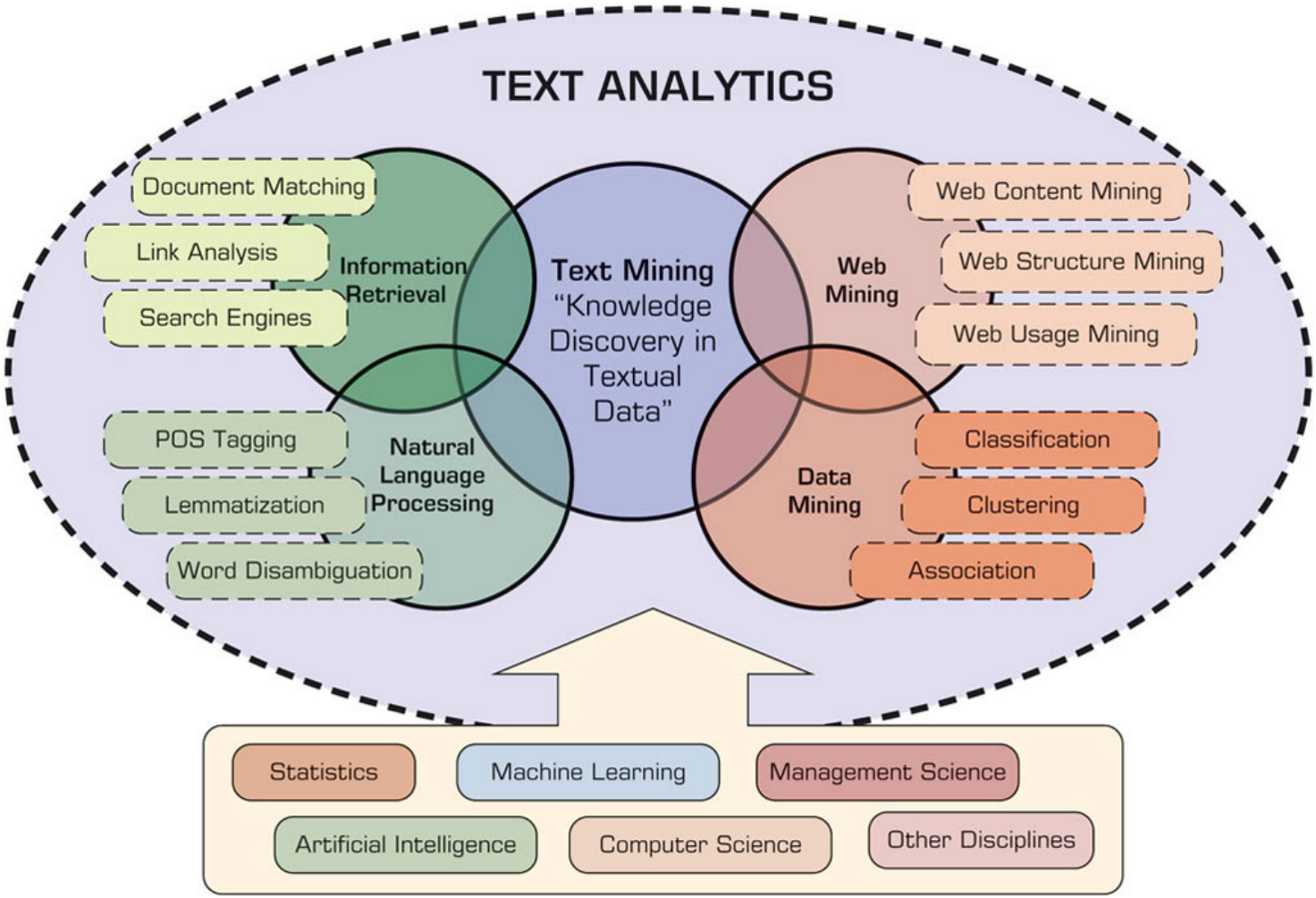
- Benefits of text mining are obvious especially in text-rich data environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communication records (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

Text Analytics (1 of 2)

- Text analytics a broader concept that that includes information retrieval, text mining, data mining, web mining, and NLP.
- Information retrieval is searching and identifying relevant documents for a given set of key terms.

Text Analytics (2 of 2)

Figure 7.2 Text Analytics, Related Application Areas, and Enabling Disciplines.



Data Mining versus Text Mining

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data:
  - Structured versus unstructured data
  - Structured data: in databases
  - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on
- To perform text mining – first, impose structure to the data, then mine the structured data.

Text Mining Application Areas (1 of 2)

- Information extraction
  - Identifying key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
- Topic tracking
  - Identifying key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
- Summarization
  - Summarizing a document to save the reader time.
- Categorization
  - Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.

Text Mining Application Areas (2 of 2)

- Clustering
  - Grouping similar documents without having a predefined set of categories.
- Concept linking
  - Grouping similar documents without having a predefined set of categories.
- Question answering
  - Finding the best answer to a given question through knowledge-driven pattern matching.

Text Mining Terminology (1 of 5)

- Unstructured or semistructured data
- Corpus (and corpora)
  - a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- Terms
  - A *term* is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of NLP methods.
- Concepts
  - *Concepts* are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher-level abstraction.

Text Mining Terminology (2 of 5)

- Stemming
  - The process of reducing inflected words to their stem (or base or root) form. For instance, *stemmer, stemming, stemmed* are all based on the root *stem*.
- Stop words (*or, noise words*)
  - Are words that are filtered out prior to or after processing natural language data (like: a, an, the, of, on, etc.)
- Synonyms
  - Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., *movie, film*, and *motion picture*).

Text Mining Terminology (3 of 5)

- Polysemes (or, homonyms):
  - Syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows."
- Tokenizing
  - A *token* is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs.
- Term dictionary
  - a collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.

Text Mining Terminology (4 of 5)

- Word frequency
  - The number of times a word is found in a specific document.
- Part-of-speech tagging
  - the process of marking the words in a text as corresponding to a particular part of speech (nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.
- Morphology
  - This is the branch of the field of linguistics and a part of NLP that studies the internal structure of words (patterns of word formation within a language or across languages).

Text Mining Terminology (5 of 5)

- Term-by-document matrix (or, Occurrence matrix)
  - the common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in columns, documents are listed in rows, and the frequency between the terms and documents is listed in cells as integer values.
- Singular value decomposition (or, Latent semantic indexing)
  - This dimensionality reduction method is used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method.

Natural Language Processing (NLP) (1 of 4)

- NLP is …
  - a very important concept in text mining
  - a subfield of artificial intelligence and computational linguistics
  - the studies of "understanding" the natural human language

Natural Language Processing (NLP) (2 of 4)

- What is “Understanding” ?
  - Human understands, what about computers?
  - Natural language is vague, context driven
  - True understanding requires extensive knowledge of a topic
  - Can/will computers ever understand natural language the same/accurate way we do?

Natural Language Processing (NLP) (3 of 4)

- Challenges in NLP
  - Issues related to spoken language and different meanings of words and the context in which the words are spoken.
- Dream of AI community
  - to have algorithms that are capable of automatically reading and obtaining knowledge from text

Natural Language Processing (NLP) (4 of 4)

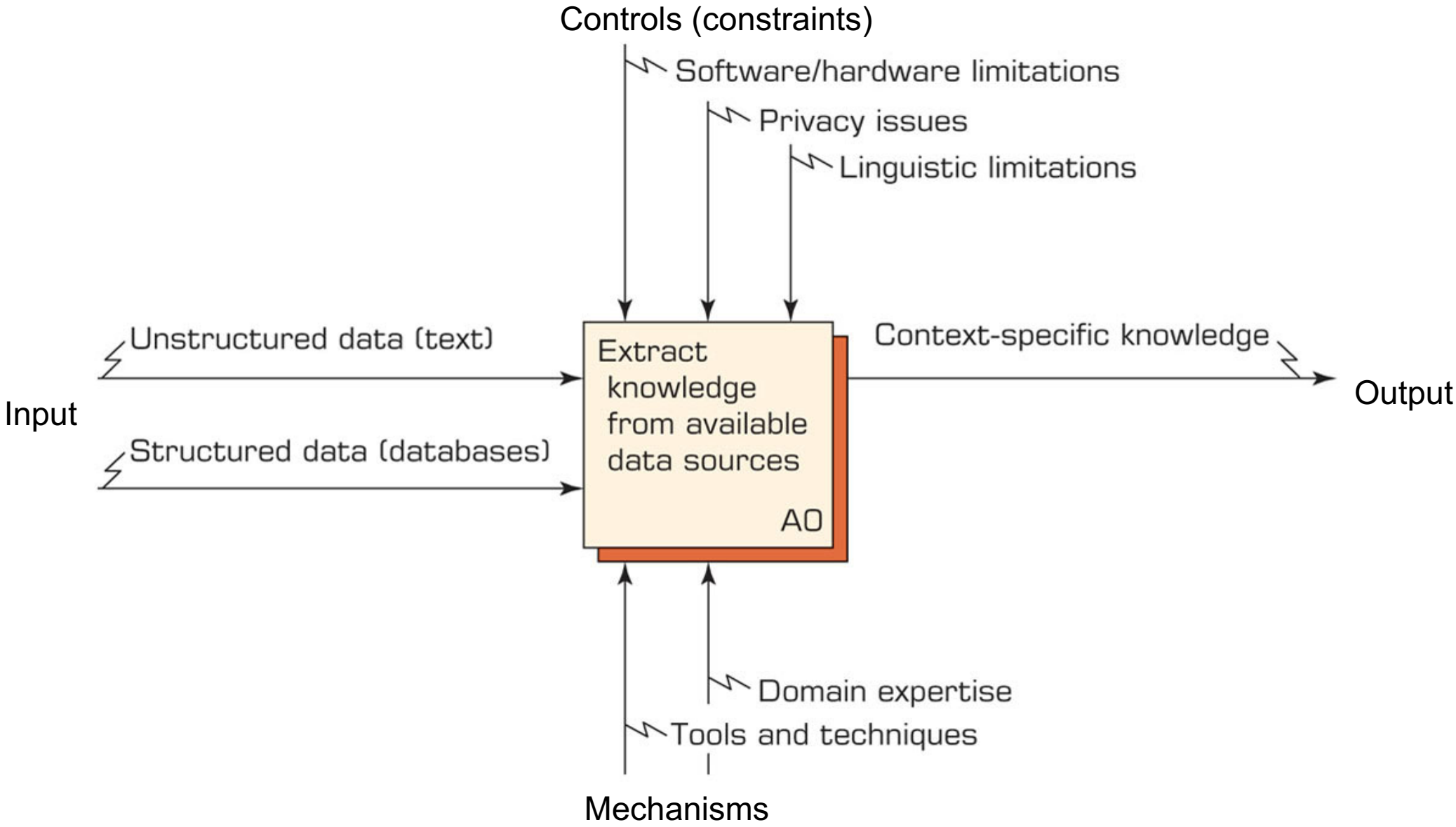
- WordNet
  - A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets.
  - Very expensive to build and maintain manually
  - A major resource for NLP applications.
  - Need automation to be completed.
- Area where WordNet has shown impact is in CRM and sentiment analysis.
- Sentiment Analysis
  - A technique used to detect favorable and unfavorable opinions toward specific products and services

NLP Task Categories

- Question answering
- Automatic summarization
- Natural language generation
- Natural language understanding
- Machine translation
- Foreign language reading & writing
- Speech recognition
- Text to Speech
- Text proofing
- Optical character recognition

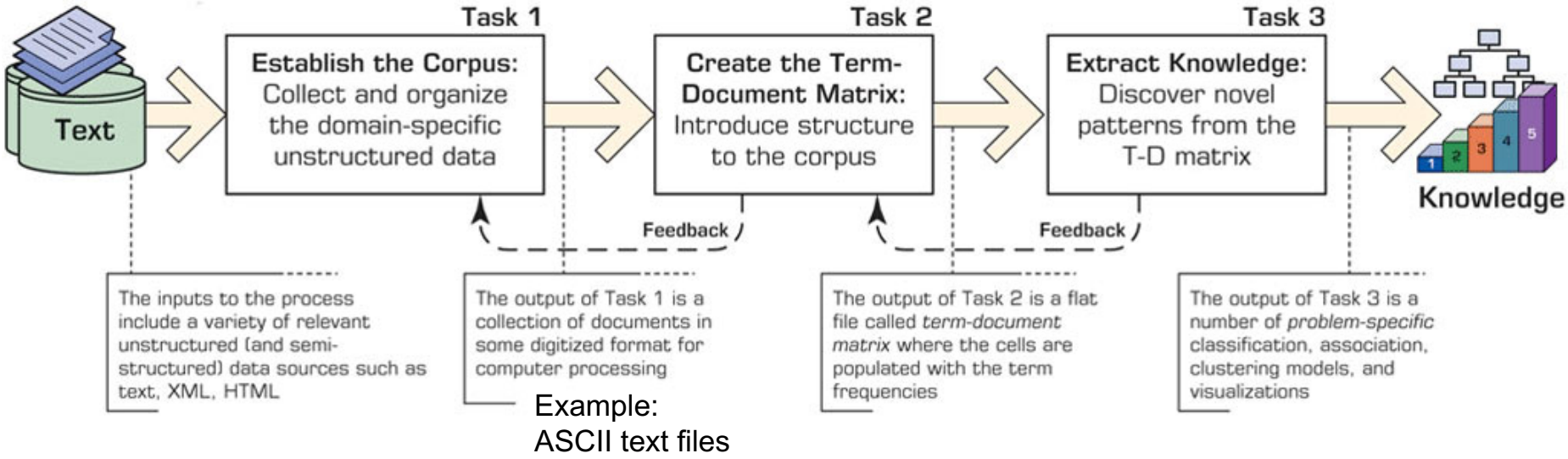
Text Mining Process (1 of 7)

- A Context Diagram for Text Mining Process



Text Mining Process (2 of 7)

Figure 7.6 The Three-Step/Task Text Mining Process.



Text Mining Process (3 of 7)

- Step 1: Establish the corpus
  - Collect all relevant unstructured data (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings…)
  - Digitize, standardize the collection (e.g., all in ASCII text files)
  - Place the collection in a common place (e.g., in a flat file, or in a directory as separate files)

Text Mining Process (4 of 7)

- Step 2: Create the Term–by–Document Matrix

Terms / Documents	Investment Risk	Project Management	Software Engineering	Development	SAP	…	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
…							

Text Mining Process (5 of 7)

- Step 2: Create the Term–by–Document Matrix (TDM) (Cont.)
  - Should all terms be included?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming

Text Mining Process (6 of 7)

- Step 2: Create the Term–by–Document Matrix (TDM) (Cont.)
  - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    - Manual - a domain expert goes through it
    - Eliminate terms with very few occurrences in very few documents
    - Transform the matrix using singular value decomposition (SVD)

Text Mining Process (7 of 7)

- Step 3: Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
  - Association
  - Trend Analysis (…)

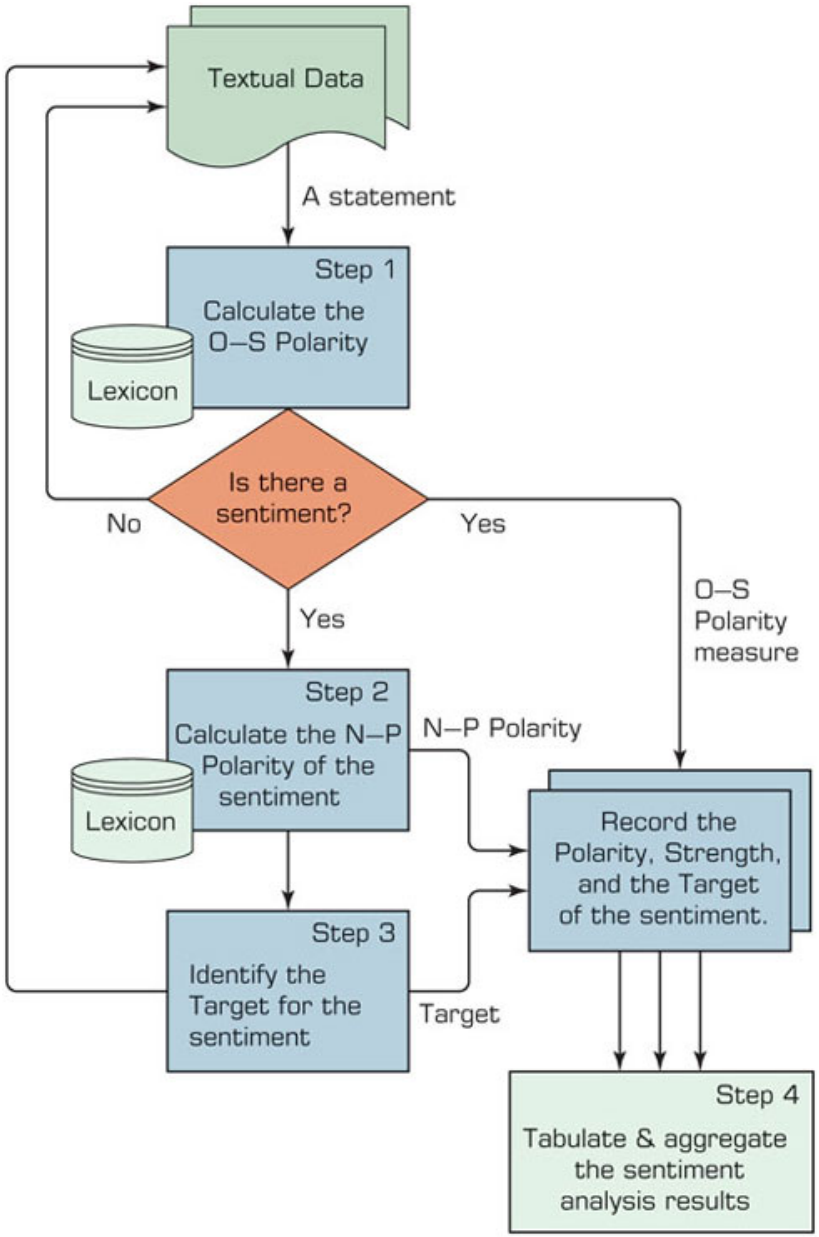
Sentiment Analysis

- Sentiment [?] belief, view, opinion, and conviction
- Sentiment analysis is trying to answer the question “What do people feel about a certain topic?”
- By analyzing data related to opinions of many using a variety of automated tools
- Used in variety of domains, but its applications in CRM are especially noteworthy (related to customers/consumers’ opinions)

Sentiment Analysis Applications

- Voice of the customer (VOC)
- Voice of the Market (VOM)
- Voice of the Employee (VOE)
- Brand Management
- Financial Markets
- Politics
- Government Intelligence
- E-commerce Site Design
- Email Filtering

Sentiment Analysis Process (1 of 3)



Sentiment Analysis Process (2 of 3)

- Step 1 – Sentiment Detection
  - Comes right after the retrieval and preparation of the text documents
  - It is also called detection of objectivity
    - Fact [= objectivity] versus Opinion [= subjectivity]
- Step 2 – N-P Polarity Classification
  - Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities
    - N [= negative] versus P [= positive]

Sentiment Analysis Process (3 of 3)

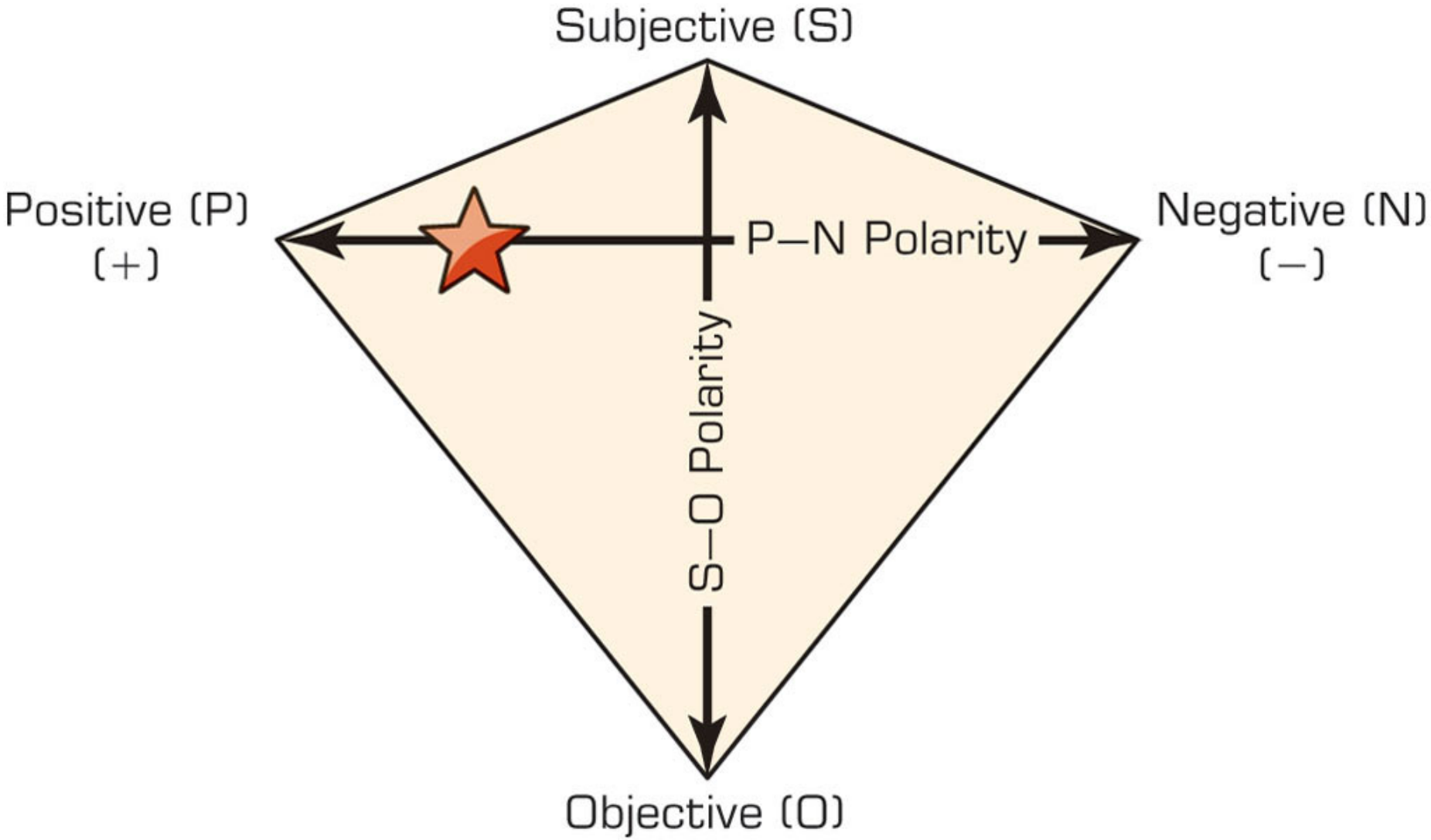
- Step 3 – Target Identification
  - The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, and event, etc.)
    - Level of difficulty the application domain
- Step 4 – Collection and Aggregation
  - Once the sentiments of all text data points in the document are identified and calculated, they are to be aggregated
    - Word Statement Paragraph Document

Methods of Polarity Identification

- Using a lexicon (dictionary) as a reference library (developed manually or automatically)
  - WordNet
  - SentiWordNet
  - WordNet-Affect

- Using a collection of training documents as the source of knowledge about the polarity of terms within a specific domain.
  - Inducing predictive models from opinionated textual documents.

P-N Polarity and S-O Polarity

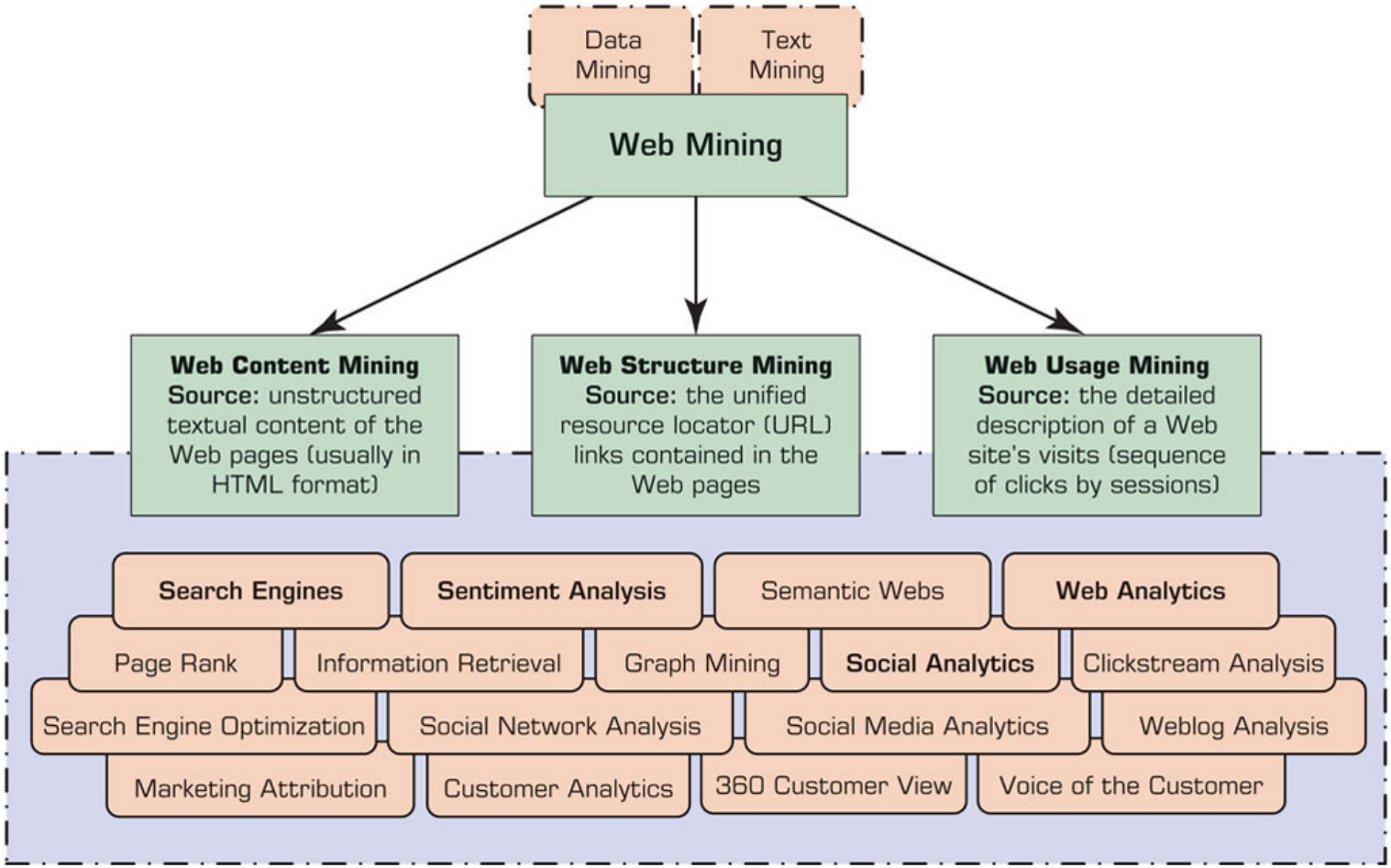


Web Mining Overview

- Web is the largest repository of data
- Data is in HTML, XML, text format
- Challenges (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- Opportunities and challenges are great!

Web Mining

Web mining (or Web data mining) is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage)



Web Content Mining

- Web content mining refers to the extraction of useful information from Web pages (unstructured textual content).
- Data collection via Web crawlers
- Can be used for things like:
  - Competitive intelligence (collecting intelligence about competitors' products, services, and customers).
  - Information/news/opinion collection.
  - Enhance the results produced by search engines.
  - Summarization.
  - Sentiment analysis.
  - Automated data collection and structuring for predictive modeling.

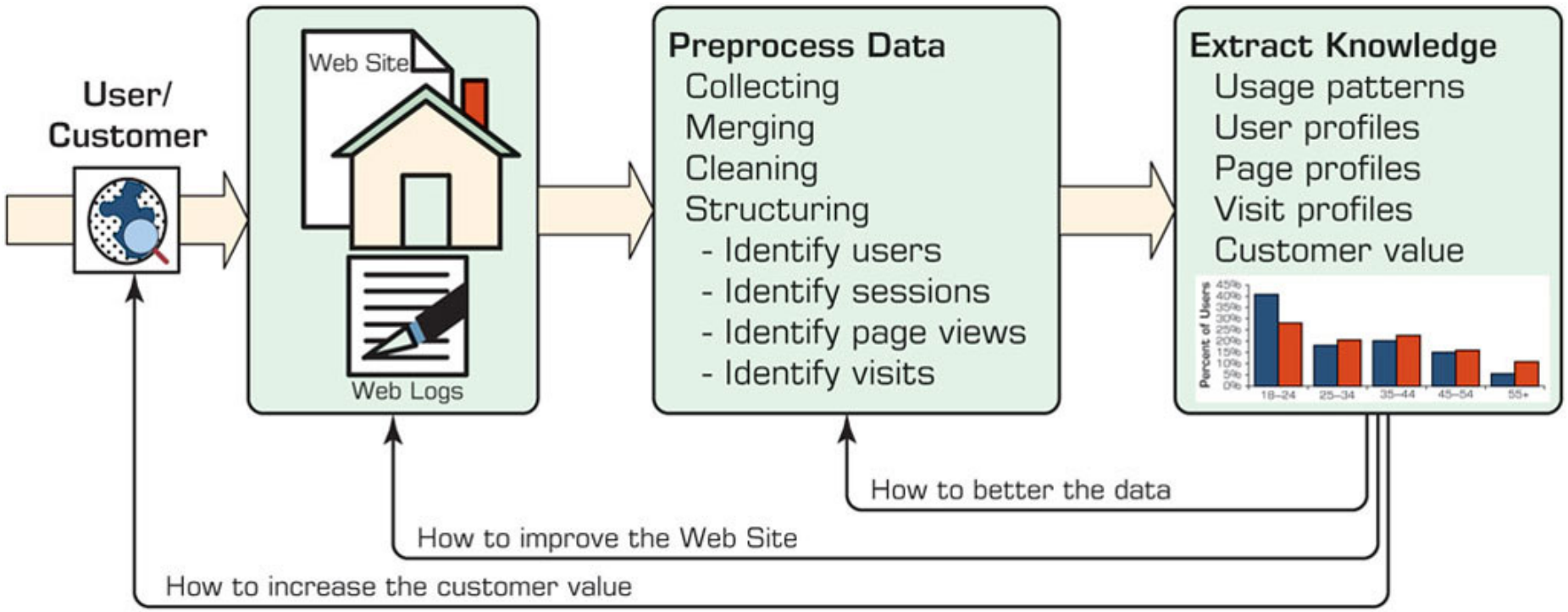
Web Structure Mining

- Web pages include hyperlinks
- Hyperlinks contain a significant amount of hidden human annotation
  - Authoritative pages
    - The collective endorsement of a given page by different developers on the Web might indicate the importance of the page and might naturally lead to the discovery of authoritative (centralized) Web pages
  - Hubs
    - One or more Web pages that provide a collection of links to authoritative pages (prominent sites on a specific topic of interest).
  - Hyperlink-induced topic search (HITS) algorithm
    - A link-analysis algorithm that rates Web pages using the hyperlink information contained within them.
    - Used to calculate hubs and authorities

Web Usage Mining (1 of 3)

- Also called Web Analytics
- Extraction of information from data generated through Web page visits and transactions.
- Clickstream data
  - data stored in server access logs, referrer logs, and client-side cookies
    - user characteristics and usage profiles
    - metadata, such as page attributes, content attributes, and usage data including time of access and duration
- Clickstream analysis
  - Analysis of clickstream data, using data mining and text mining techniques to find interesting patterns in customer activities and periods and times of visits.

Web Usage Mining (2 of 3) (Clickstream Analysis)



Web Usage Mining (3 of 3)

- Web usage mining applications
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products.
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups based on user access patterns
  - Predict user behavior based on previously learned rules and users' profiles
  - Present dynamic information to users based on their interests and profiles

Search Engines

- Google, Bing, Yahoo, …
- For what reason do you use search engines?
  - Search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word terms, or a complete sentence) that users have provided that have to do with the subject of their inquiry
- They are the workhorses of the Internet

Anatomy of a Search Engine

Two main cycles:

- Development Cycle
- Responding Cycle

1. Development Cycle Purpose
   - To create a huge database of documents/pages organized and indexed based on their content and information value.
   - Why? Due to its sheer size and complexity, searching the Web to find pages in response to a user query is not practical (or feasible within a reasonable time frame); therefore,
   - Search engines “cache the Web” into their database and use the cached version of the Web for searching and finding.
   - Once created, this database allows search engines to rapidly and accurately respond to user queries.

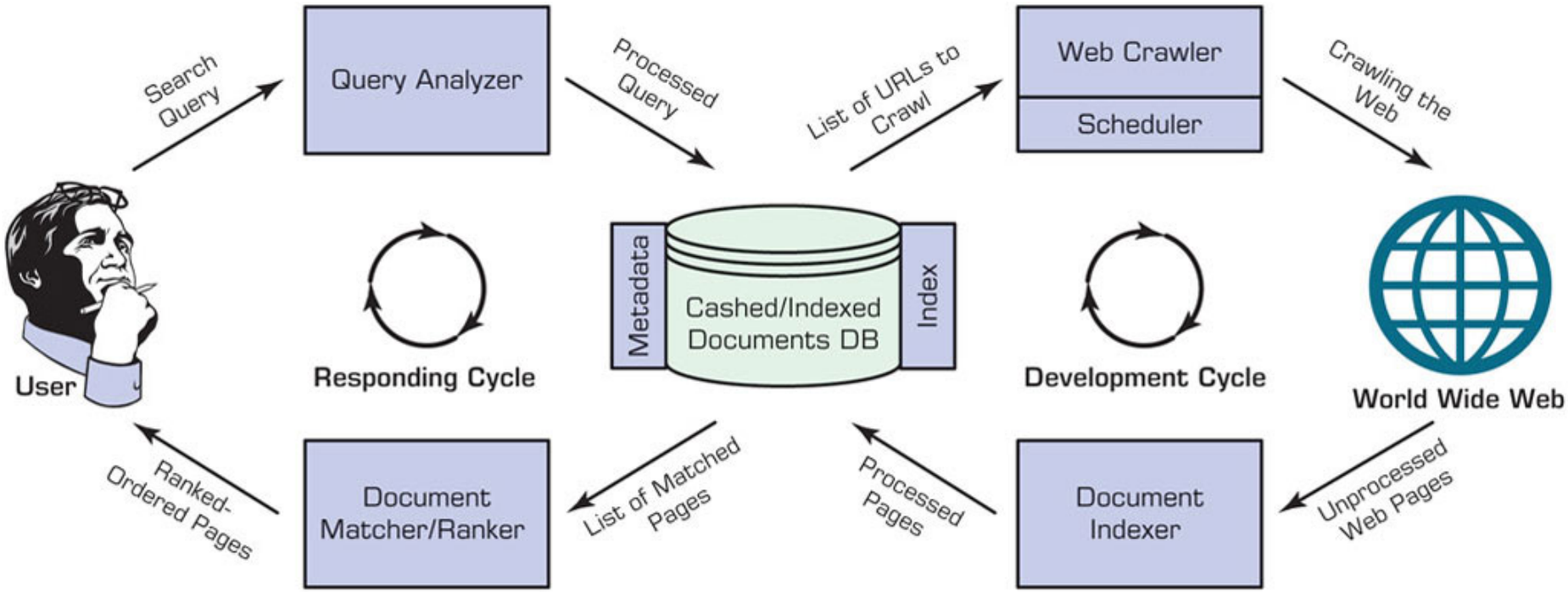
Anatomy of a Search Engine

- Development Cycle Components
  1. Web Crawler (spider)
     - a piece of software that systematically browses (crawls through) the Web for the purpose of finding and fetching Web pages.
  2. Document Indexer
     - As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process.
     - The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database.
     - Uses three steps:
       1. Preprocessing the document
       2. Parsing the document (text-mining tools & techniques)
       3. Creating the term-by-document matrix

Anatomy of a Search Engine

- Response Cycle Components
  1. Query Analyzer
     - Responsible for receiving a search request from the user (via the search engine's Web server interface)
     - Converting search request into a standardized data structure so that it can be easily queried/matched against the entries in the document database.
  2. Document Matcher/Ranker
     - The structured query data are matched against the document database to find the most relevant documents/pages
     - Rank matched results in the order of relevance/importance.
     - Return list to user.

Structure of a Typical Internet Search Engine



Top 15 Most Popular Search Engines

(by eBizMBA, August 2016)

Rank	Name	Estimated Unique Monthly Visitors
1	Google	1,600,000,000
2	Bing	400,000,000
3	Yahoo! Search	300,000,000
4	Ask	245,000,000
5	AOL Search	125,000,000
6	Wow	100,000,000
7	WebCrawler	65,000,000
8	MyWebSearch	60,000,000
9	Infospace	24,000,000
10	Info	13,500,000
11	DuckDuckGo	11,000,000
12	Contenko	10,500,000
13	Dogpile	7,500,000
14	Alhea	4,000,000
15	ixQuick	1,000,000

Search Engine Optimization

- It is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results
- Part of an Internet marketing strategy
- Based on knowing how a Search Engine works
  - Content, HTML, keywords, external links, …
- Indexing based on …
  - Webmaster submission of URL
  - Proactively and continuously crawling the Web

Copyright