NOVA IMS Information Management School

Introduction to Text Mining

Roberto Henriques

Filipa Peleja

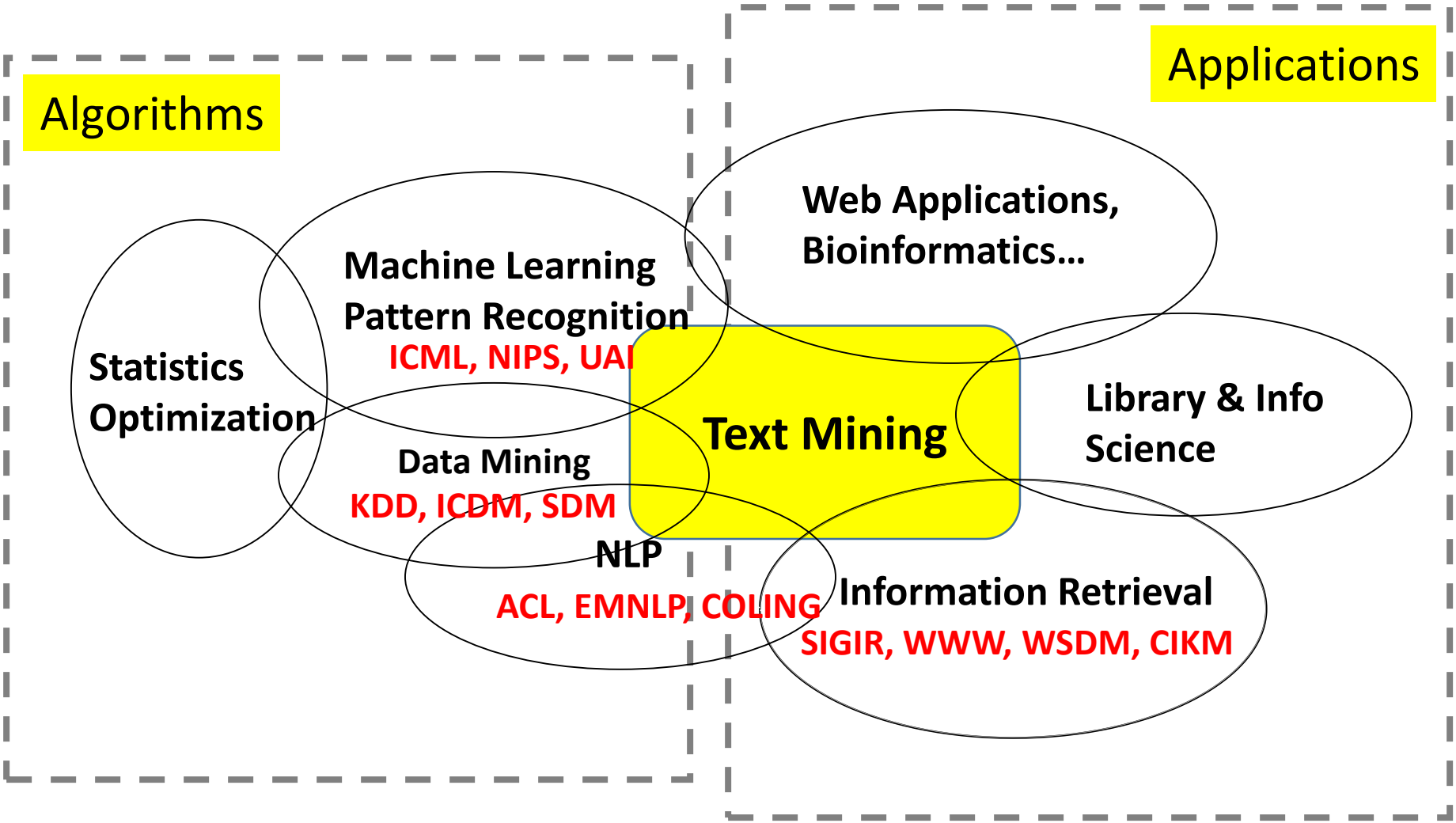
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Text Mining

- "*Text mining, also referred to as* ***text data mining****, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.*" - Wikipedia

- "*Another* way to view text data mining is as a process of **exploratory** data analysis that leads to **heretofore unknown** information, or to answers for questions for which the answer is not currently known." – Hearst, 1999

Text Mining



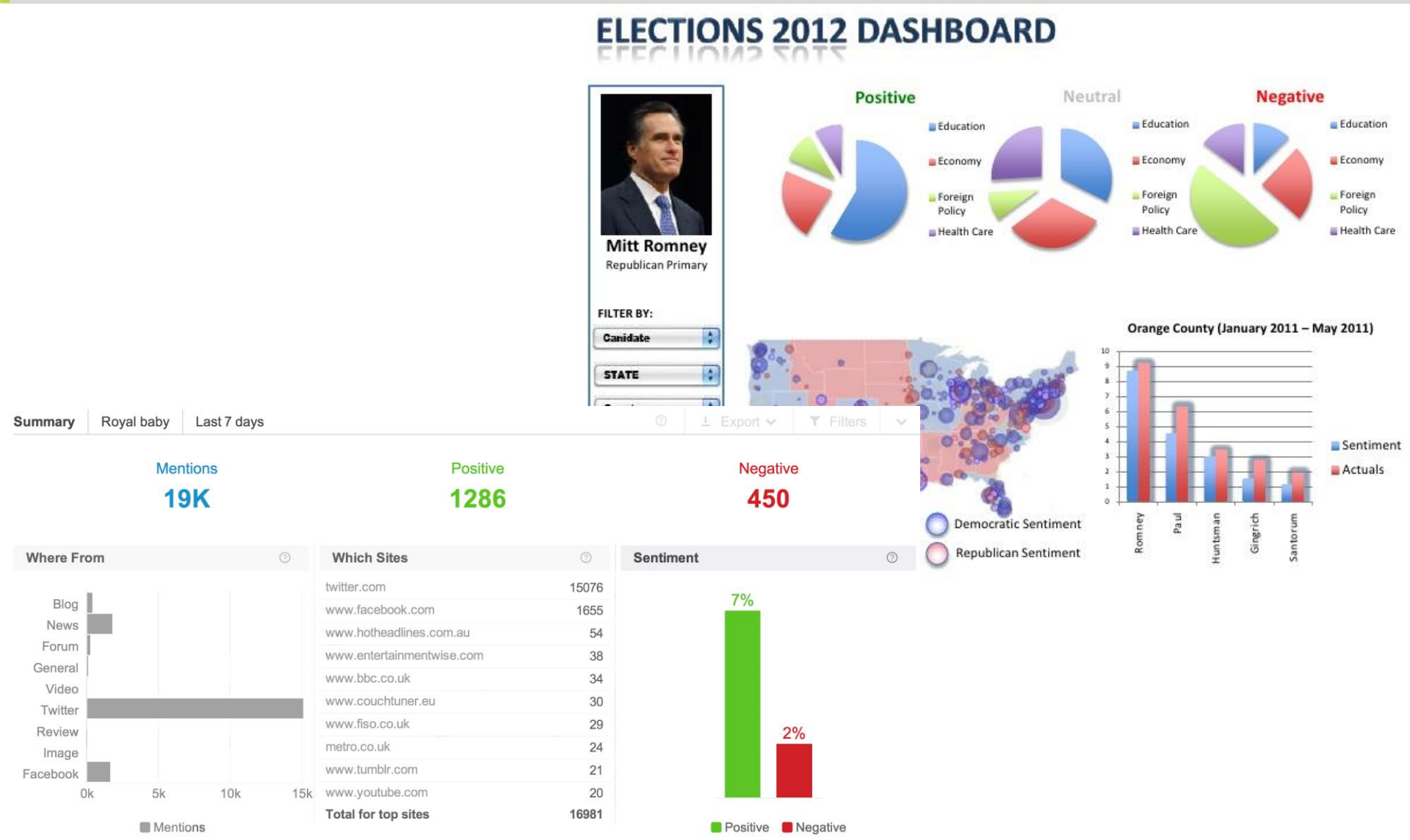
[1] http://www.cs.virginia.edu/~hw5x/Course/TextMining-2018Spring/_site/lectures/ .

IBM Watson won Jeopardy!

- IBM Watson on February 2011 won Jeopardy
  - A game that needs answers to questions posed in natural language and require speed, accuracy and confidence

"Political Jeopardy" by Nate Beeler, published in PoliticallyIllustrated.com

Text Mining for Sentiment Analysis



Defining Text Mining

- Knowledge-intensive process in which a user interacts with a document collection
- Text mining: extract useful information from data sources by exploring data and identify different patterns
- Unlike data mining, text mining data sources are always document collections and patterns are not found in database records but in unstructured textual data in documents from these collections

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Text Mining and Data mining

- Data mining assumes that data is already stored in a structured format

- Data mining mainly focus in two tasks:
  - Data cleansing and normalizing data
  - Create extensive numbers of tables joins

- Pre-processing operations are the centre of text mining system:
  - Identification and extraction of representative features from textual documents
  - Pre-processing techniques are responsible for transforming unstructured data into a more explicitly structured format – **this is concern that is not relevant to most data mining systems**

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Text Mining challenges

- Data is natural language “free text”
  - Data is not well-organized
  - Natural language text contains ambiguities on many levels
    - Lexical, syntactic, semantic and pragmatic
  - Many times we need annotated data
    - Expensive to obtain for large datasets
- What do we want to look at?

Centrality of Natural Language

- Due to its nature text mining draws advances in other computer science disciplines which are concerned with processing natural language:

  - Information Retrieval
  - Information Extraction
  - Corpus-based computational linguistics

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Document collection

- Text mining focus on looking at the *document collection*
  - Document collection: can be any group of text-based documents

- Many text mining solutions aim to discover patters **across** very large documents collections

- The number of documents in each collection can go from many thousands to tens of millions

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Static or Dynamic Document Collections

- **Static:** the initial set of documents remains unchanged

- **Dynamic:** term applied to document collections characterized by their inclusion of new or updated documents over time
  - Large document collections with high rates of documents change can be highly demanding in terms of performance optimization for text mining systems

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

PubMed

PubMed

PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

- Real-word document collection suitable as initial input for text mining
- Online service that contains text-based document abstracts for more than 12 million research papers on topics of the life sciences
- Collects papers from 1966 to the present

The size of document collections that are represented by PubMed makes manual attempts to correlate data across documents, map complex relationships or identify trends

extremely labour-intensive

and, many times,

impossible to achieve

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

PubMed

PubMed

PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

- Automatic methods for identifying and exploring inter-document data dramatically enhance the speed and efficiency of research activities
- Automatic exploration techniques are many times a baseline requirement for researchers to be able to recognize subtle patterns across large numbers of natural language documents
- Text mining systems do not usually run their knowledge algorithms on unprepared document collections:
  - An important task in text mining is known as ***preprocessing operations***

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

A basic element in text mining: the document

- A document can be informally defined as a **unit of discrete textual data within a collection**
- Usually (but not always) correlates with some real-world document: e-mail, research paper, news article, business report etc.
- A document:
  1. Does not necessarily exists only within the context of one particular collection
  2. Can, and generally does, exist in any number or type of collections (from very formally organized to very ad hoc)
  3. Can be a member of different document collections or subsets of the same document collection and can exist in these at the same time

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

A basic element in text mining: the document

- A d… …ithin a c…
- U… …mail, r…
- A…
  1… …ion
  2… …very
  3… …same

A document about Microsoft's antitrust litigation:

- Could exist in a Microsoft' legal documents collection
- But also, in a completely non-related to Microsoft documents. In a document collections oriented toward current affairs, legal affairs, antitrust-related affairs
- Or/and in a collection of software company news

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

"Weakly structured" or semistructred Documents

- Documents that have little on strong typographical, layout or markup indicators to denote structure (e.g. research papers, business reports or news stories) are often referred to as ***free-format or weakly structured***

A typographical element can be:
- Punctuation marks
- Capitalization
- Numeric
- Special characters

- Documents with extensive and consistent format elements in which field-type metadata can be more easily inferred (e.g. e-mail, HTML pages, PDF files, word files with templates or style-sheet constraints) are usually described as ***semistructured documents***

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Document features

- Preprocessing operations aim to support text mining:
  - Identify elements that are contained in a natural language document
  - Transform a document from non-structured representation into a structured representation

- But, given the potentially large number of words, phrases, sentences, typographical elements and layout artifacts an essential task for most text mining systems is the **identification of simplified subset of document features**

- This feature dimensionality issue is typically of much greater magnitude in text mining systems than in classic data mining systems

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Features Dimensionality - Reuters

- Even for small document collections the word-level features required to represent such documents can be exceedingly large
- A very small news collection from Reuters with 15,000 documents enclose 25,000 nontrivial word stems
- Tens of thousands of concept-level features may be relevant for a specific domain
- In comparison with a data mining task the number of attributes are significantly smaller

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Dimensionality and Sparsity

- Features **dimensionality** problems is a driving factor in the development of text mining preprocessing techniques

- Feature **sparsity**: many features only appear in a very small percentage of the documents collection, hence, the support for many patterns is low

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Characters, Words, Terms and Concepts (1)

- **Characters**:
  - Full set of all characters or some filtered subset
  - Common character-based representation are bigrams or trigrams

N-grams:

a pair of consecutive written units such as letters, syllables, or words

- **Words**:
  - Specific words selected directly from the original document
  - Can be described as the basic level of semantic richness
  - Phrases, multiword expressions or multiword hyphenates do not constitute single word-level features

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Characters, Words, Terms and Concepts (2)

- **Terms**:
  - single words and multiword phrases selected directly from the corpus of a native document by means of *term-extraction* methodologies
  - Specific words and expressions found in the original document for which they are meant to be generally representative

"President Abraham Lincoln experienced a

career that took him from log cabin to White House (…)"

A list of terms to represent this example could include:

- Single word forms such as: "Lincoln", "took", "career" and "cabin"
- Multiwords "President Abraham Lincoln", "log cabin" and "White House"

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Characters, Words, Terms and Concepts (3)

- **Concepts**:
  - Generated by means of manual, statistical, rule-based or hybrid categorization algorithms
  - Can be manually generated however this might be highly labour-extensive
  - Complex preprocessing routines are commonly adopted
  - Can be single words, multiword expressions, whole clauses or larger syntactical units which are related to specific concept identifiers

A document collection that includes reviews of sports cars **may not actually include the specific word "automotive" or the specific phrase "test drives"** but these concepts might be nevertheless **found among the set of concepts used to identify and represent the collection**

Characters, Words, Terms and Concepts (4)

- **Categorization methodologies** many times require some degree of cross-referencing with external knowledge source
  - **Statistical methods**: many times its used annotated training documents
  - **Manual or rule based methods**: many times involves interaction with *gold-standard* such as a pre-existing domain ontology, lexicon, formal concept hierarchy or, even, a human domain expert
- **Term- and concept-based:**
  - Features in the most condensed and expressive levels of semantic value
  - In overall size exhibit roughly the same efficiency and are generally much more efficient than character- or word-based document models

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Characters, Words, Terms and Concepts (5)

- **Term-level** can sometimes be easier to generate automatically than **concept-level** as quite often the latest has entailed some level of human interaction
- **Concept-level** can be better than other representations when handling synonymy and polysemy and clearly best at relating a given feature to its various hyponyms and hypernyms

Synonymy: a word or phrase that means exactly or nearly the same as another word or phrase in the same language, for example *shut* is a synonym of *close*.
Polysemy: the coexistence of many possible meanings for a word or phrase.
Hyponym: a word of more specific meaning than a general or superordinate term applicable to it. For example, *spoon* is a hyponym of *cutlery*.
Hypernym: a word with a broad meaning constituting a category into which words with more specific meanings fall; a superordinate. For example, *colour* is a hypernym of *red*.

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Concepts-level drawbacks

- During preprocessing the relative complexity of applying heuristics required to extract and validate concept-type features

- Domain-dependence of many concepts. Manual generated concept-level representations are fixed and labour-intensive to assign

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Search for Patterns and Trends

- The core functionality of a text mining system resides in the analysis of *concept co-occurrence* patterns across documents in a collection
- Text mining systems rely on algorithmic and heuristic approaches to consider distributions, frequent sets and various associations of concepts at an inter-document level
- The objective is to enable a use to discover the nature and relationships of concepts as reflected in the collection as a whole

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Examples: Search for Patterns and Trends

In a collection of news articles

A large number of articles on politician X and "scandal" maybe indicate negative image of the character of X

A growing number of news articles

That company Y co-occurs with product Z might indicate a shift of focus in company Y's interests (competitors might like to know this)

In a collection of biomedical data

A potential relationship might be inferred between two proteins P1 and P2 by the pattern of:

(a) Several articles mentioning the protein P1 in relation to the enzyme E1

(b) A few articles describing functional similarities between enzymes E1 and E2 without referring to any protein names

(c) Several articles linking enzyme E2 to protein P2

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Text mining methods

- Seek to discover co-occurrence relationships as reflected by the documents collection
- Many times based on large-scale, brute-force search directed at large, high-dimensionality features sets generally produce very large number of patterns
- Need to provide not only relevant but also manageable results sets (e.g. patterns identified) to the user

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

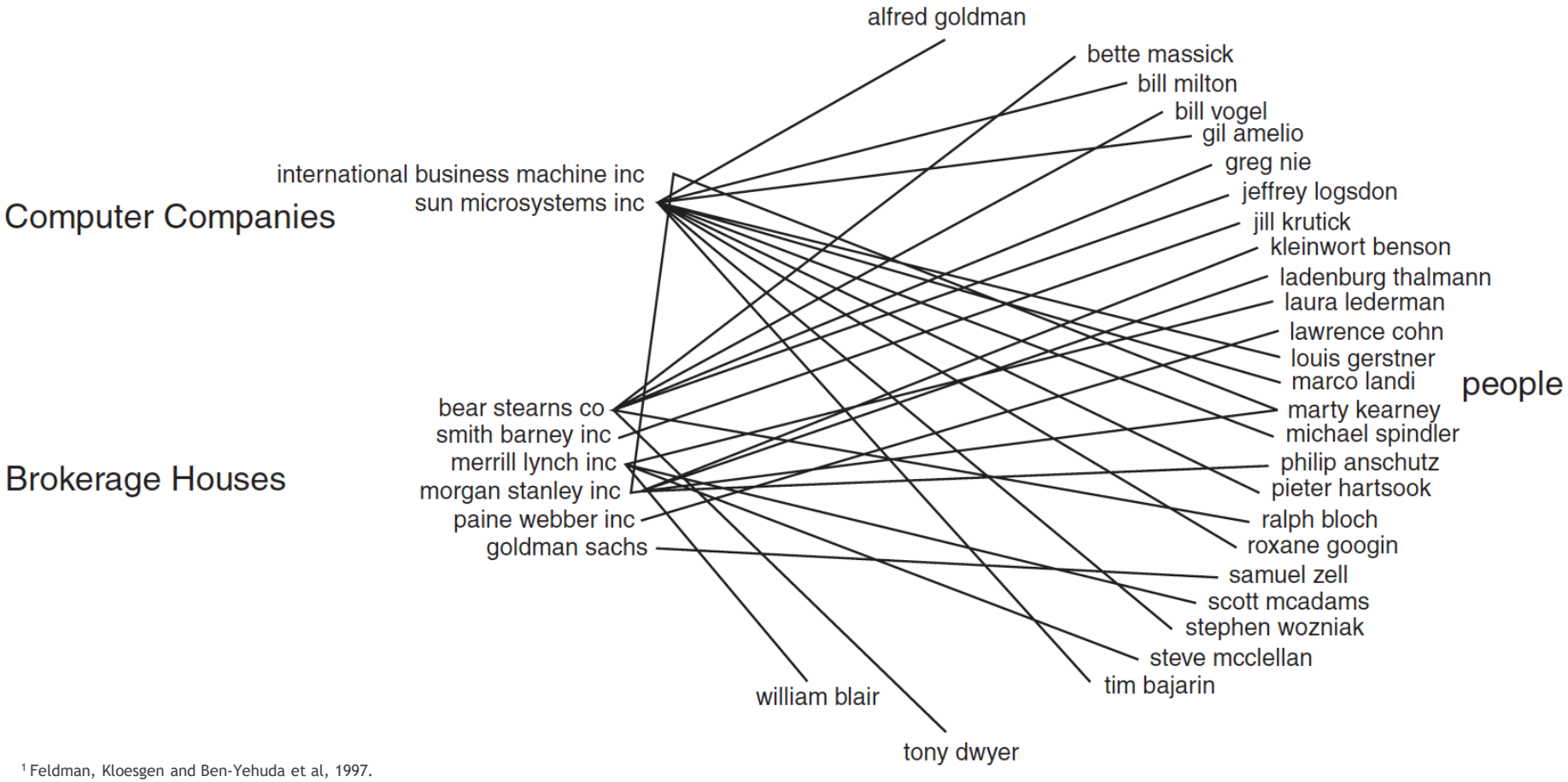
Visualization tools for Text Mining

- Text mining systems browsing can be both dynamic and content-based
- Browsing is guided by the actual textual content from documents within the collection
- Many times user browsing is facilitated by the graphical presentation of concept patterns in a the form of hierarchy
  - This helps in the user interactivity by organizing concepts
- Browsing is also navigational: enable user to move across concepts in a way that he might choose to either see the “big picture” view of the collection or drill down to more specific concept relationships

[1] R. Feldman and J. Sanger, The Text Mining Handbook, 2006.

Example of a visualization tool

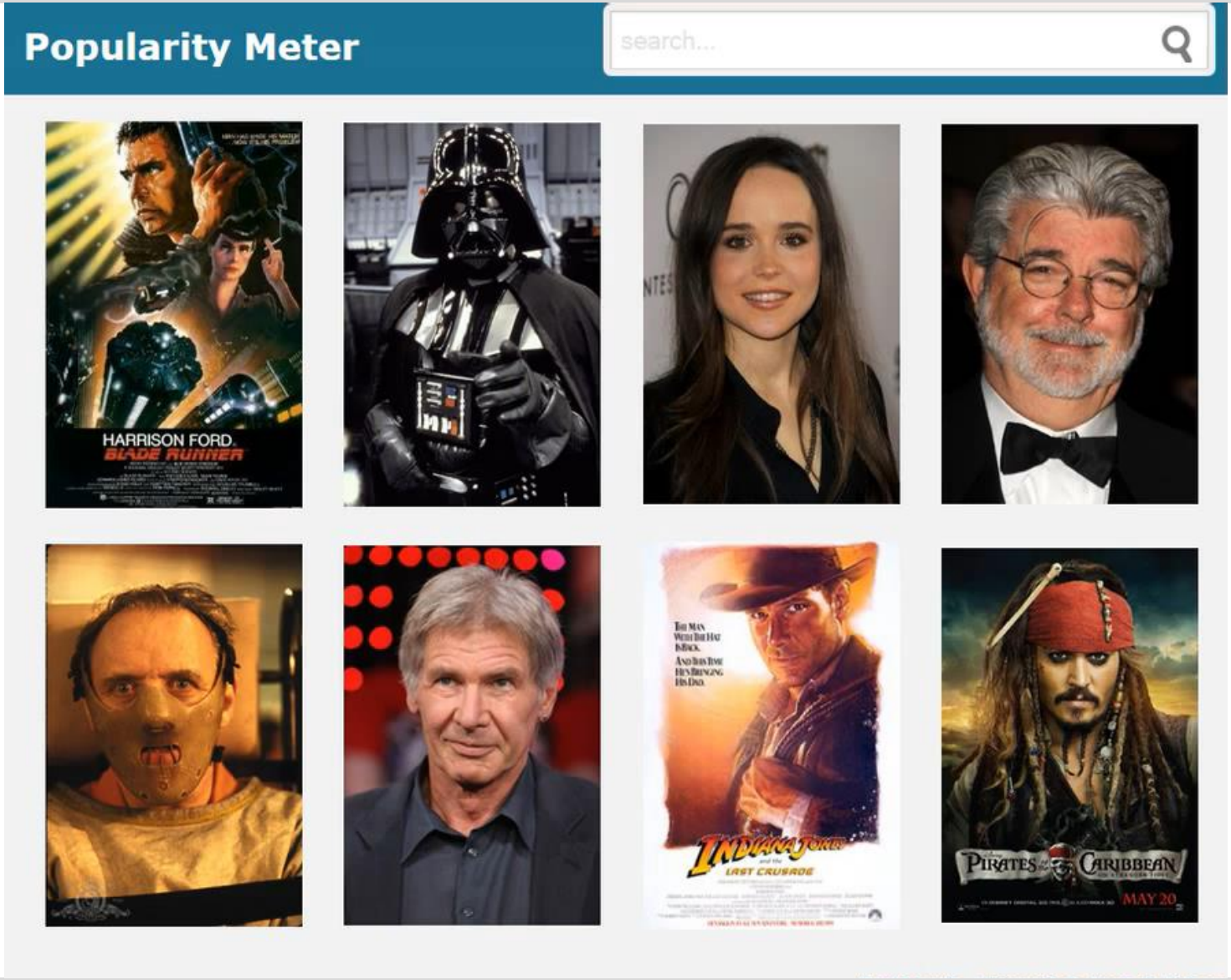
- Mapping concepts (keywords) within the context of all categories by means of a “category graph”



[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

NOVA IMS Information Management School

Entities reputation graph[1,2]

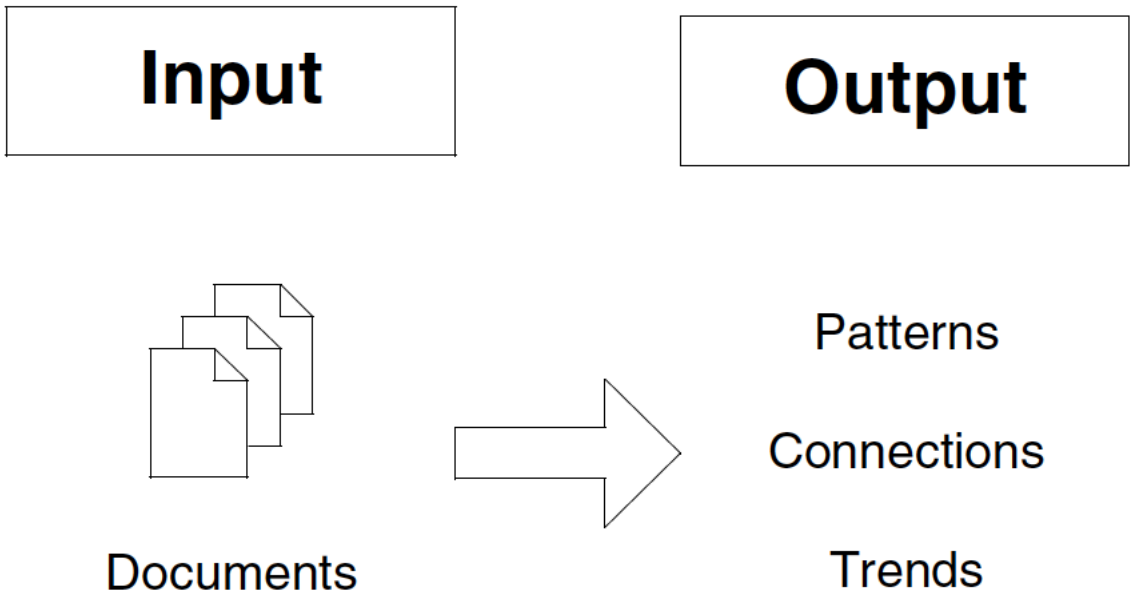


[1] F. Peleja, J. Santos and J. Magalhães, Reputation analysis with a ranked sentiment-lexicon, SIGIR, 2014.

[2] F. Peleja, J. Santos and J. Magalhães, Ranking linked-entities in a sentiment graph, IEEE/WIC/ACM, 2014.

General Architecture of Text Mining Systems

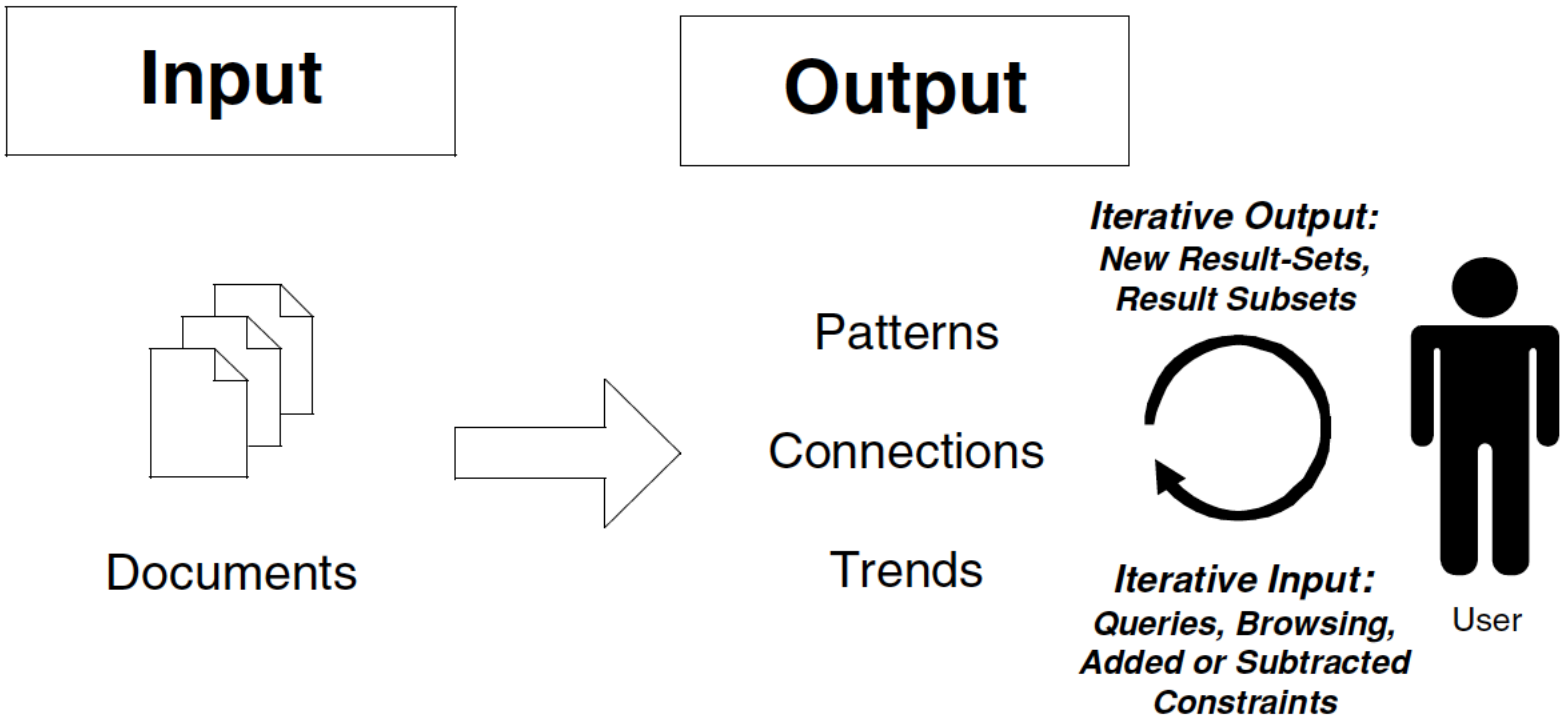
- At an abstract level, a text mining system takes in input (raw documents) and generates various types of output (e.g., patterns, maps of connections, trends)



[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

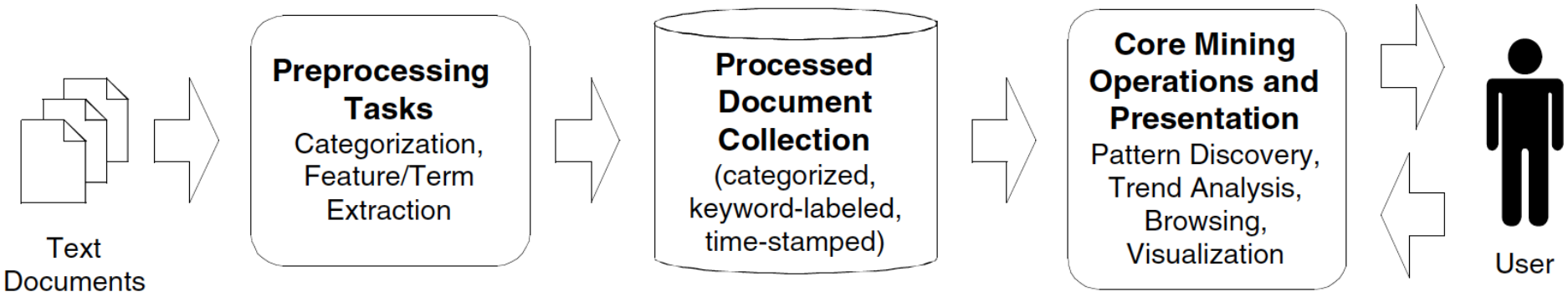
Iterative loop for user input and output

- A human-centered view of knowledge discovery
- User is part of an interactive loop of querying, browsing, and refining its query



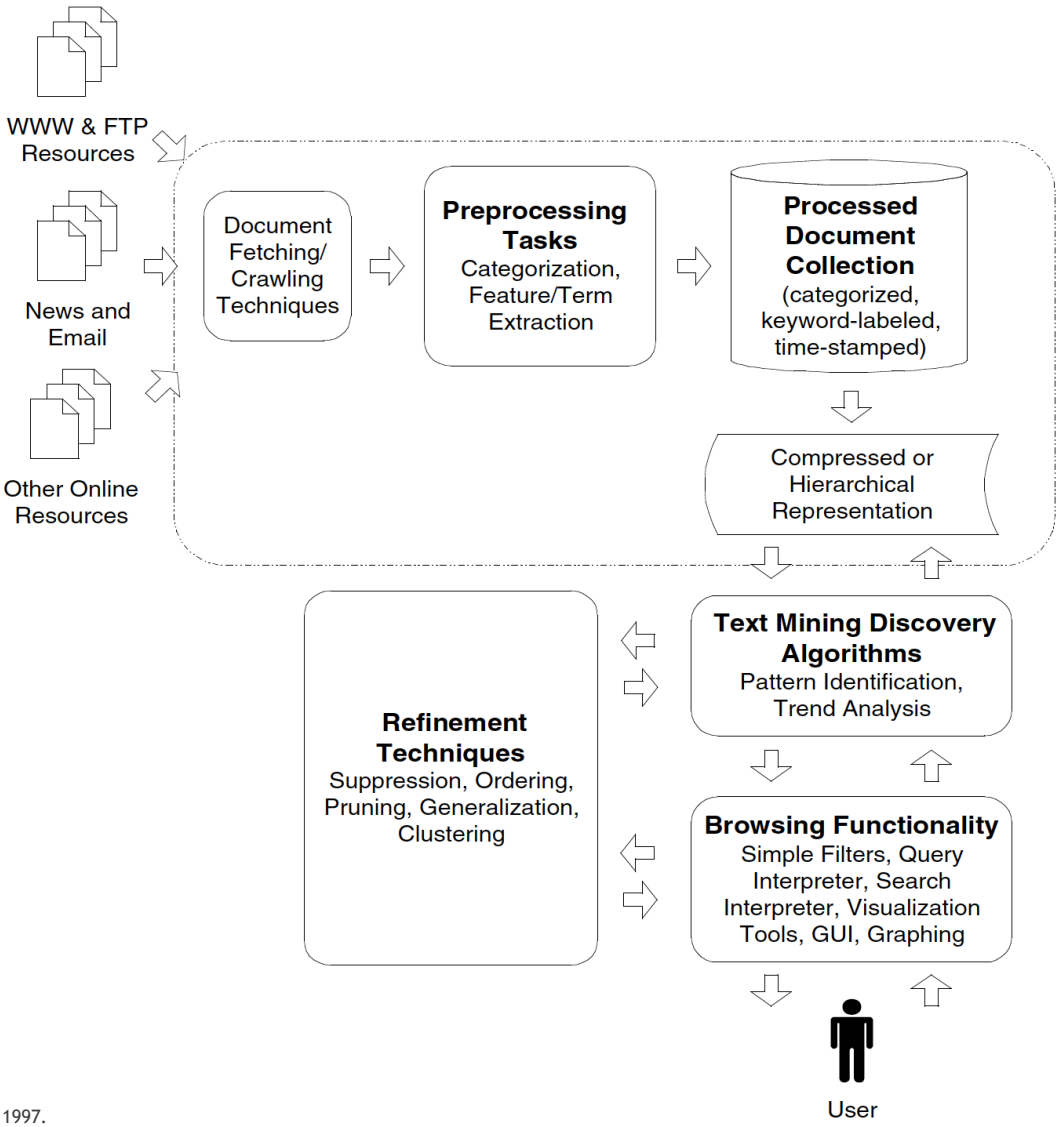
[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

High-level text mining functional architecture



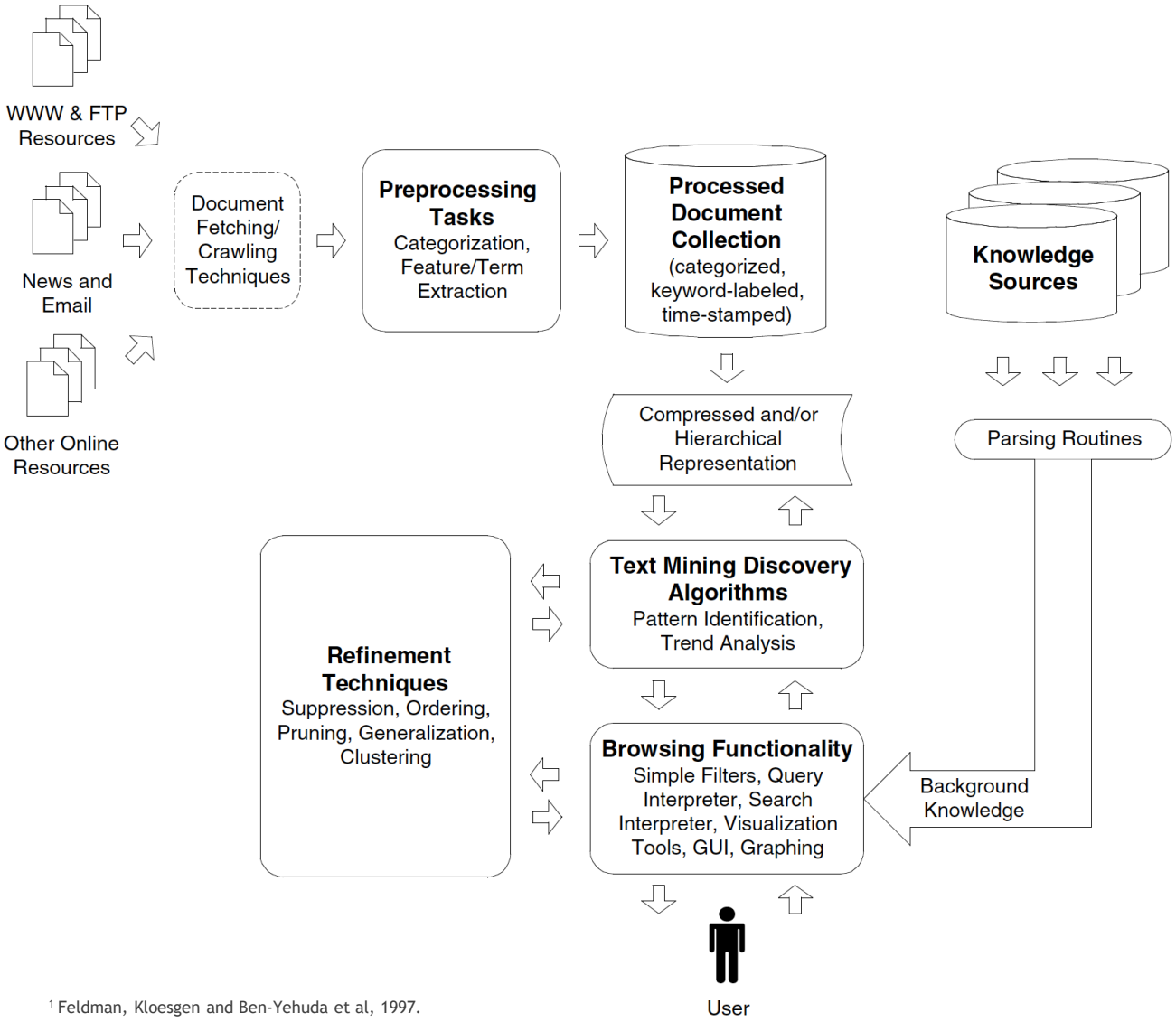
[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

Generic text mining system



[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

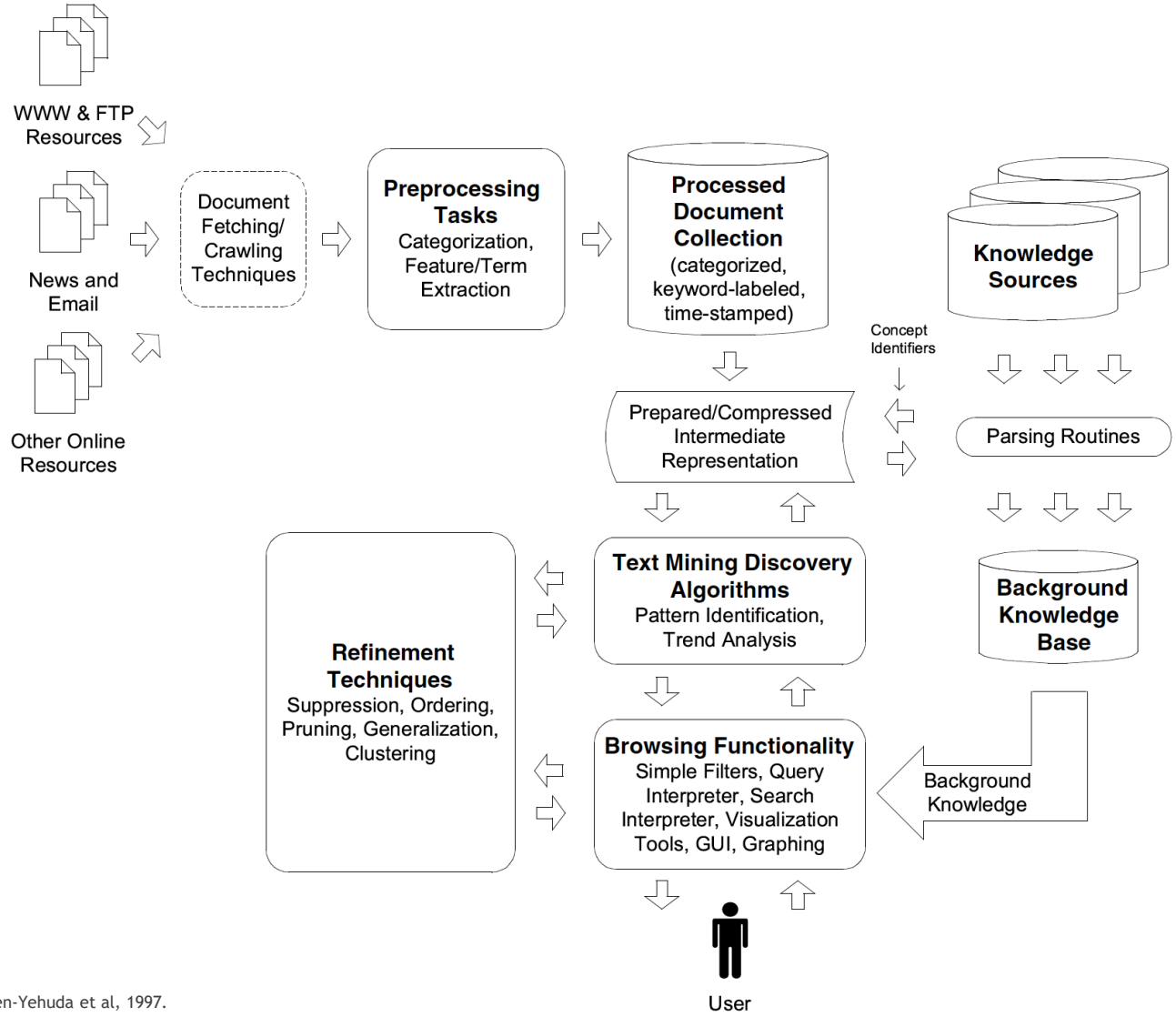
Domain-oriented text mining system



Many text mining systems such as medicine, financial services and many others can benefit from access to special background or domain-specific data sources

[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

Text mining system with knowledge base



[1] Feldman, Kloesgen and Ben-Yehuda et al, 1997.

Going back to IBM Watson… What is IBM Watson data collection?

- *"Watson had access to **200 million pages** of structured and unstructured content consuming four terabytes of disk storage including the full text Wikipedia."* – PC World

- *"The sources of information for Watson include **encyclopedias, dictionaries, thesauri, newswire articles** and **literary works**. Watson also used databases taxonomies, and ontologies. Specifically, **DBPedia, WordNet**, and **Yago** were used."* – Hearst, 1999

Questions

1. What is text mining?
   a) Text mining is an algorithm that takes unstructured text and organizes it
   b) Text mining is the process of distilling actionable insights from text
   c) Text mining is an evaluation metric used in data science for assessing learning algorithms on text
2. True or false? Word-based representation is not powerful.
   a) True
   b) False
3. What is an n-gram in the context of text mining?

[1] https://www.datacamp.com/courses/intro-to-text-mining-bag-of-words.

Questions

1. What is text mining?
   a) Text mining is an algorithm that takes unstructured text and organizes it
   b) Text mining is the process of distilling actionable insights from text
   c) Text mining is an evaluation metric used in data science for assessing learning algorithms on text
2. True or false? Word-based representation is not powerful.
   a) True
   b) False
3. **N-gram is a group of N words or characters which follow one another.**

[1] https://www.datacamp.com/courses/intro-to-text-mining-bag-of-words.