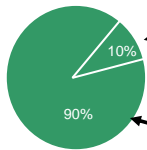


Mining semi-structured data

Goran Nenadic
University of Manchester
g.nenadic@manchester.ac.uk

Structured & unstructured data



Structured numerical or coded information

e.g. transactional databases

Unstructured or semi-structured information

e.g. company reports, market analyses, e-mails, videos, audio, legal documents, newspapers, Web pages, scientific papers, etc.

Semi-structured data

- No strict format; can still have some structure with optional entities, elements and attributes
 - some entities/elements/attributes may be missing, some repeated, and new ones can be added
 - objects with different structures, ad-hoc, diverse data sources
- Examples
 - HTML document with titles (e.g. <H1>) and paragraphs (<p>), but no information about author(s), date, etc.
 - LaTeX, Word documents



Textual data

- A lot of information is stored in documents, reports, e-mails, plans, articles, ...
 - Web, intranet, text archives, digital libraries
 - e.g. some estimations say that 80-90% of business information is in textual form
- Text representation is typically semi- or unstructured
 - structured: author, title, date, etc.
 - most of the content is unstructured



Textual data

- E.g. in biomedical sciences
 - ~600,000 scientific articles per year
 - ~2,000 articles per day
- WWW as a data source
 - 1997-2002: ~10 billion pages; 100 Tbytes
 - various data types with a lot of text
 - search engines

continues



Textual data

- Stored in documents (files) annotated for various purposes
 - type setting
 - visualisation (e.g. browsing)
 - document retrieval (search engines)
 - text analytics
- Various annotations supporting various purposes
- Most publishers use XML as internal format
 - XML databases and querying

continued

Main problems: ambiguity

- Different meanings/senses of words
 - e.g. *Apple* (the company) or *apple* (the fruit)
 - e.g. *Toyota* can be a *car* or a *company*
 - e.g. acronyms have different meanings in different contexts

USA =

United States Army	Union of South Africa
United States of America	Union Street Athletics
Ujhaanagar Sindh Association	Unionville-Sebewaing Area
Ultimate in Suspense and Action	Unique Settler Attributes
Unconditional Self-Acceptance	Unit Self Assessment
Unconventional Stellar Aspect	United Scenic Artists
Under Secretary of the Army	University of South Alabama
Underground Service Alert	University of South Australia
Underground Sewer Adapter	Unix System Admin
Underwriting Service Assistant	Unstable Angina
Unicycling Society of America	Unusually Sensitive Area

Tasks and techniques

- Information retrieval (IR)
 - select a set of relevant **documents**
- Information extraction (IE)
 - extract factual information (**facts**) from texts
- Question answering (QA)
 - find/generate an **answer** to a given question

Information retrieval (IR)

- Searching for relevant **documents**
 - whether in stand-alone or hypertext collections (both Internet or intranets)
- Search engines are an example of IR systems
- Result of IR is a set of relevant documents
 - filtering huge collections based on a query
 - no fine-grained information, just whole documents
 - users would need to read and analyse these documents on their own

continued

Information retrieval (IR)

- IR is based on **indexing** by **keywords**
 - pre-calculated and stored for easier retrieval
 - each document represented by an index vector
- Simple techniques are used
 - index all words and "phrases"
 - use meta-information (cross-links, titles, etc.)
- Calculate "similarity" between documents & query
 - vector space model & distances (Euclidian, cosine)
 - ranked** list of documents, based on similarity

IR challenges

- Number of documents (e.g. on the Web) is growing much faster than any present IR technology can index
 - also, many (web) pages are updated frequently, which forces the IR engines to revisit them periodically (refresh)
- Dynamically generated sites/documents may be difficult to index, or may result in excessive results from a single site
- Queries one can make are currently limited to searching by/for **keywords**, which may result in many
 - false positives* (wrong hits), because of ambiguity
 - false negatives* (non-returns), because of variability

Information extraction (IE)

- Extract information i.e. facts from text
- Identify instances of pre-defined *entities* (dates, names of people, locations, etc.) and relations between them
- Fill in database-like tables with "facts"

Slot	Information
Date	7/10/96 (today)
Location	San Salvador
Victim injured	policeman
Victim attacked	guards
Perpetrator	urban guerrillas

San Salvador, 7/10/96
 It has been officially reported that a policeman was wounded today when urban guerrillas attacked the guards at a power substation located downtown San Salvador.



continued

Information extraction (IE)

- Identification of *entities* and *terms* of interest
 - e.g. persons, jobs, companies, dates, locations, ...
- Extract specific relations and events
 - use of templates (regular expressions)
<PERSON> is appointed as a <JOB> of <COMPANY>
 - typically designed around important verbs –
attacked, bought, appointed, merged, acquired, etc.

Question answering (QA)

- Produce factual (short) answers on a user query that is formulated as a question
 - *"When was the takeover of AstraZenica?"*
 - *"Who is the CEO of Software Ltd.?"*
- Combine IR, IE and heuristics
 - if the question starts with "When ..." then look for date and time expressions
 - if the question starts with "Who ..." then look for persons

Text mining

- **Text mining** is now widely used as an umbrella for large variety of natural language processing techniques to denote all approaches to retrieve, extract and analyse textual information
 - e.g. find and analyse specific news reports (e.g. related to a certain company)
 - e.g. extract and analyse user complaints mails
 - e.g. find causal links between symptoms or diseases and drugs or chemicals

example

Personalised movie "matcher"

- Match movies to individuals based on their preference profiles
- Information sources
 - written reviews of movies
 - users' lists of favorite movies





recommendation portal

- A recommendation portal for movies and tv shows
 - provides recommendations, answering a free given search.
 - ~10,000 movie, TV and video titles
- Based on semantic technologies, Jinni uses text mining on plot, mood, style, setting, soundtrack and more in combination with an ontology, created by film professionals
- You don't need to know about exact title, actor, director, place or year of production to get an result, you can enter simply a phrase describing the mood, genre or place the movie is about, and you will guided through a facilitated search to narrow your search and get at the end what you want.
- Or alternative, if you search for a movie and you have only a vague idea of the plot, you can formulate a plot's description in free phrasing.
- As it also offers APIs for Internet and TV content providers you can make your way direct to an online store to download or purchase the movie.



Restaurant Reputation Report

- A service targeting restaurant owners to provide them reports of positive and negative reviews of food, service and ambiance at their restaurants.
- For that the service monitors negative and positive trends across hundreds of online review sites.
 - Now restaurant owners can subscribe to receive a PDF of their monthly reports. This PDFs came with charts, trends, rankings, summaries and some quotes from users, month by month. The reports may enable those restaurant owners to react and improve their services in the specific field.
 - A simple but straight forward way to using semantic technologies in business.

Challenges

- Text mining is possible but difficult
 - language variability and ambiguity
 - currently – only approximation is used, but needs interpretation, context, background knowledge (not everything is explicit)

Nicholas Andrews was succeeded by Gina Torretta as chair-person of BNC Holdings. She was appointed 2 days ago.



Summary

- XML can be used to represent semi- and unstructured data
- Differences between retrieving information from structured and unstructured data
 - computers do not understand
- Text mining techniques: IR, IE, QA, etc.
 - problems with understanding text documents (language variability and ambiguity, growing number and size of documents, etc.)
