Processing of large document collections

Part 1 (Introduction)

Helena Ahonen-Myka

Spring 2006

1. Introduction

- course organization
- introduction to the topic
  - applications
  - methods
- learning goals
- schedule

Organization of the course

- lectures (Helena Ahonen-Myka)
  - Tue 12-14, Thu 10-12 B222
  - 14.3.-27.4. (no lectures 13.4. and 18.4.)
- exercise sessions (Miro Lehtonen)
  - Wed 10-12 B222
  - 20.3.-28.4. (no exercises 19.4.)
- exam: Thu 4.5. at 16-19
- points: exam 50 pts, exercises 10 pts
  - required: ~30 pts (= 1)

Course material

- slides on the course web page
- also other material available on the page
  - handouts used in the class (sample documents etc.)
  - original articles

Large document collections

- What is a document?
  - “a document records a message from people to people” (Wilkinson et al., Document Computing, 1998)
- each document has content, structure, and metadata (context)
  - in this course, we concentrate on content
  - particularly: textual content

Large document collections

- large?
  - some person may have written a document, but it is not possible later to process the document manually -> automatic processing is needed
  - large w.r.t to the capacity of a device (e.g. a mobile phone)
- collection?
  - documents somehow similar -> automatic processing is possible

Applications

- text categorization
- text summarization
- information extraction
- question answering
- ---------------------------------
- text compression
- text indexing and retrieval
- machine translation

...

Text categorization

- given a predefined set of categories and a set of documents
- label each document with one or more categories

Text summarization

- ”Process of distilling the most important information from a source to produce an abridged version for a particular user or task” (Mani & Maybury, Advances in Automatic Text Summarization, 1999)

Example

A Spanish priest was charged here today with attempting to murder the Pope. Juan Fernandez Krohn, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, Fernandez told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope ’looked furious’ on hearing the priest’s criticism of his handling of the church’s affairs. If found quilty, the Spaniard faces a prison sentence of 15-20 years.

Example

- summary could be, e.g.
  - “A Spanish priest is charged after an unsuccessful murder attempt on the Pope”
- or a set of phrases:
  - a Spanish priest was charged
  - attempting to murder the Pope
  - he trained for the assault
  - Pope furious on hearing priest´s criticisms

Information extraction

- ”Information extraction involves the creation of a structured representation (such as a database) of selected information drawn from the text” (Grishman, Information Extraction: Techniques and Challenges, 1997)

Example: terrorist events

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.

According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Example: terrorist events

Incident type	bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical target	destroyed
Effect on human target	no injury or death
Instrument	bomb

Example: terrorist events

- a document collection is given
- for each document, decide if the document is about terrorist event
- for each terrorist event, determine
  - type of attack
  - date
  - location, etc.
- = fill in a template (~database record)

Question answering systems

- the user asks a question in a natural language
- the question answering system finds answers from a document collection, e.g. from a collection of newspaper stories

Example

- question:
  - When did Chuck Yeager break the sonic barrier?
- a text fragment in the collection:
  - “For many, seeing Chuck Yeager – who made his historic supersonic flight Oct. 14, 1947 – was the highlight of this year’s show, in which...”
- answer: Oct. 14, 1947

Methods

- typically several methods (from several research fields) are combined in each application
  - statistics (or simply counting frequencies...)
  - machine learning
  - knowledge-based methods
  - linguistic methods
  - algorithmics

Learning goals

- learn to recognize components of applications/processes
- learn to recognize which (kind of) methods could be used in each component
- learn to implement some methods
- (meta)learn to control learning processes (What do I know? What should I know to solve this problem?)

Schedule

- 14.-23.3.
  - text representation, text categorization, term selection
- 28.3.-6.4.
  - text summarization
- 11.4.-20.4.
  - information extraction
- 25.4.
  - question answering systems
- 27.4.
  - closing