Meaning-Based Computing Retrieving Information from text, video and audio

Universität Salzburg
Computer Science
Seminar aus Informatik SS2009
Sabine Klausner
Simon Kranzer

Agenda

- Meaning-Based Computing (MBC)
- Knowledge Management
- Information Retrieval
  - Text-Mining
  - Video-Mining

MBC

- Structured data
  - Database, data-centric Applications
- Unstructured data
  - emails, blogs, micro-blogging, pictures, videos and other material

MBC (cont.)

- Meaning-Based Computing is the technology to capture tacit information from unstructured data

- 85% of data generated in an unstructured format as being probably more interesting than the data collected in a structured format

MBC (cont.)

- Methods:
  - speech-, video-recognition
  - text-comprehension (Text-Mining)
  - implistic query
    - Instead of seeking the help of a search engine with a query, implistic technology would not only read what is on the screen at any time, be it an email or a web page, but understand what is on the screen and summon up related information.

MBC (cont.)

- MBC is
  - Knowledge Management and
  - Information Retrieval
  - in
    - structured and
    - unstructured
  - data.

Knowledge Management

- Knowledge Management refers to the processes and/or tools an organization uses to collect, analyze, store, and disseminate its intellectual capital.

- Knowledge Management refers to making optimum use of experience and understanding of organizational knowledge, in general.

- Knowledge Management includes all sources of Information, structured, unstructured, spoken, writen and any other availiable.

Information Retrieval

- Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Information retrieval is fast becoming the dominant form of information access, overtaking traditional databasestyle searching

Information Retrieval (cont.)

- Boolean Retrieval:
- Search for A and B where NOT C in a large text
  - The simplest form for a computer to do this is a linear scan through the text (grepping) but we need more:
    - The amount of online data has grown at least as quickly as the speed of computers, and we would now like to be able to search collections that total in the order of billions to trillions of words.
    - Allow more flexible matching operations. For example, a search "within 5 words" or "within the same sentence".
    - Allow ranked retrieval: in many cases you want the best answer to an information need among many documents that contain certain words.

Information Retrieval (cont.)

- To avoid linearly scanning the texts for each query is to *index* the documents in advance.

- Boolean Retrieval takes every term (word) of a text and builds up an index marking the appearance of it in every lexical unit.

- Searching the index means just a bitwise AND.

- An example should demonstrate that!

IR (Example)

Search for the term „Simon" NOT „Tom" in boolean index documents:

Incidence matrix

Incident	Text1	Text2	Text3	Text4
Simon	1	1	0	0
Tom	1	0	1	0

Simon -> 1 1 0 0

Tom -> 1 0 1 0   NOT Tom -> 0 1 0 1

Simon AND NOT Tom -> 0 1 0 0

Information Retrieval

- But the world is not always 0 (false) or 1 (true)?!
- The Boolean approach therefore needs a fuzzy improvement:
- Tolerant Retrieval
  - Introduces techniques that are robust to typographical errors in the query, as well as alternative spellings
  - Will be discussed in more detail in the Seminar-Paper

Text Mining is…

….how to generate and prepare information in eg.: databases

It is a very new disziplin.

=> Problem
to choose or to know how to choose appropriate informationfragments from a text (in eg. information engines, mails, digitaly stored documents)
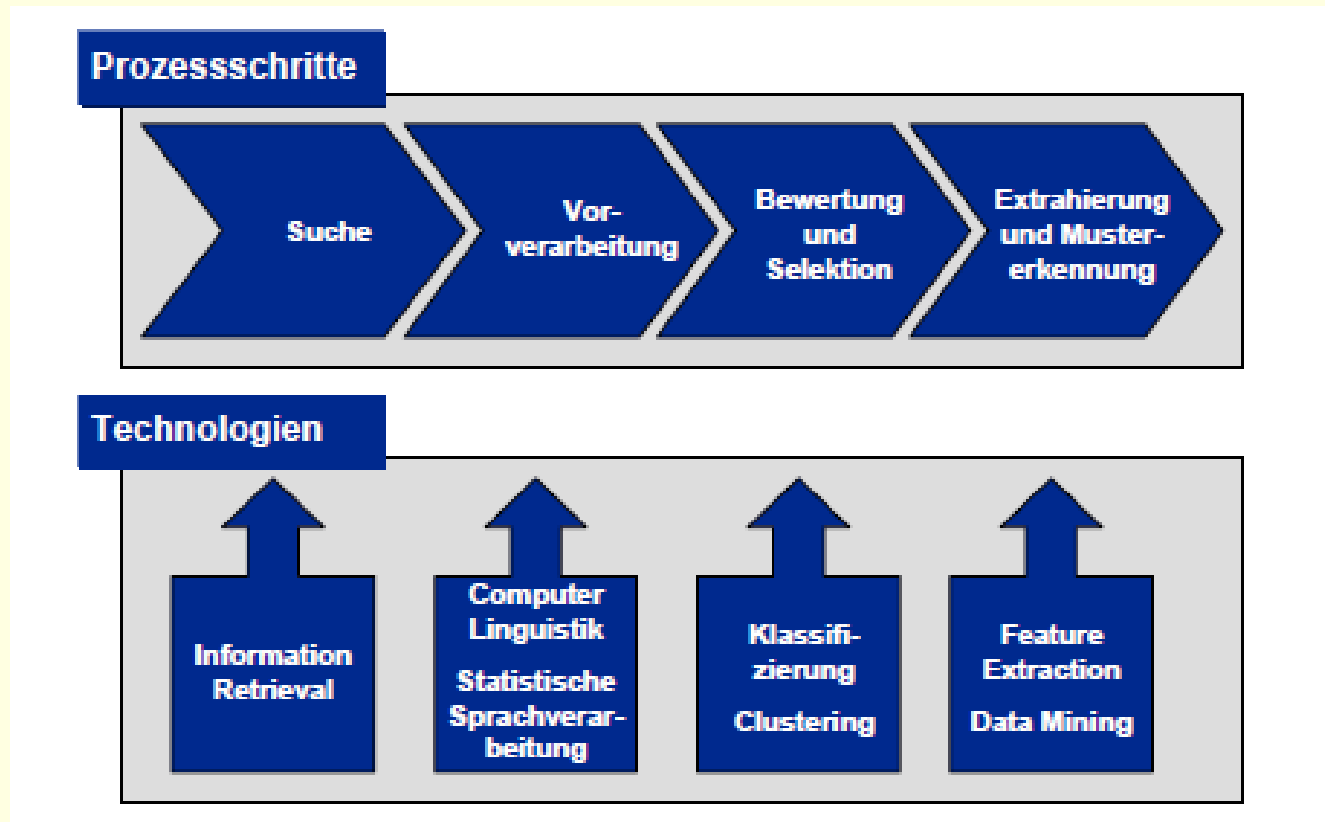
Web Farming (used for internet staff)

Text Mining

Abbildung: Text Mining Prozess

Usage

- Analysation of elections behaviour
- Optimisation of information (Spam filtering)
- Communication monitoring – crime fighting

- Commercial usage in companies

Video Mining… What is it?

- Digital Videos are stored in Databases

- Analysis followed by indexing

- Rich Medium…It is usually accompanied by other information sources such as speech, music and closed captions

How to do Video Mining

Exploring the concepts and meanings of the videos
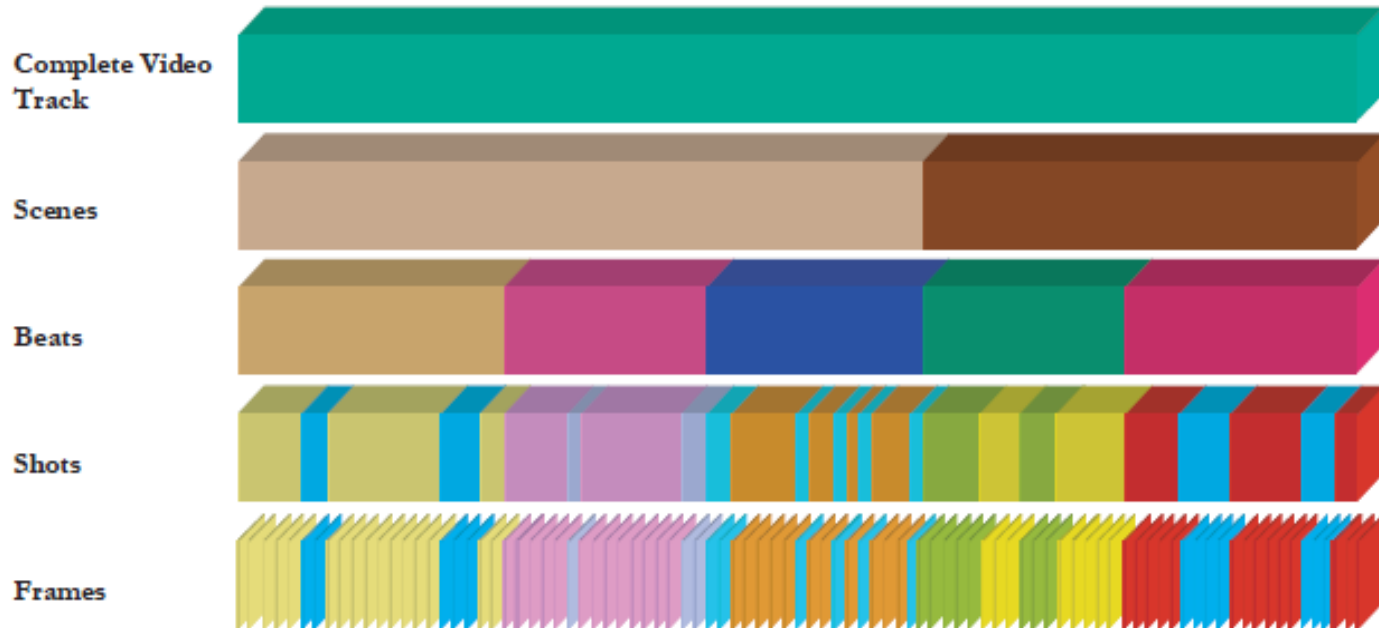
By learning rules of the video (Film Grammar):

• Learn the theories and practices of film aesthetics, such as the effect of color on the mood, the effect of music on the scene situation and the effect of post processing of the audio and video on human perception.

• Develop a model to integrate this information to explore concepts.
.

Idea of Video Mining
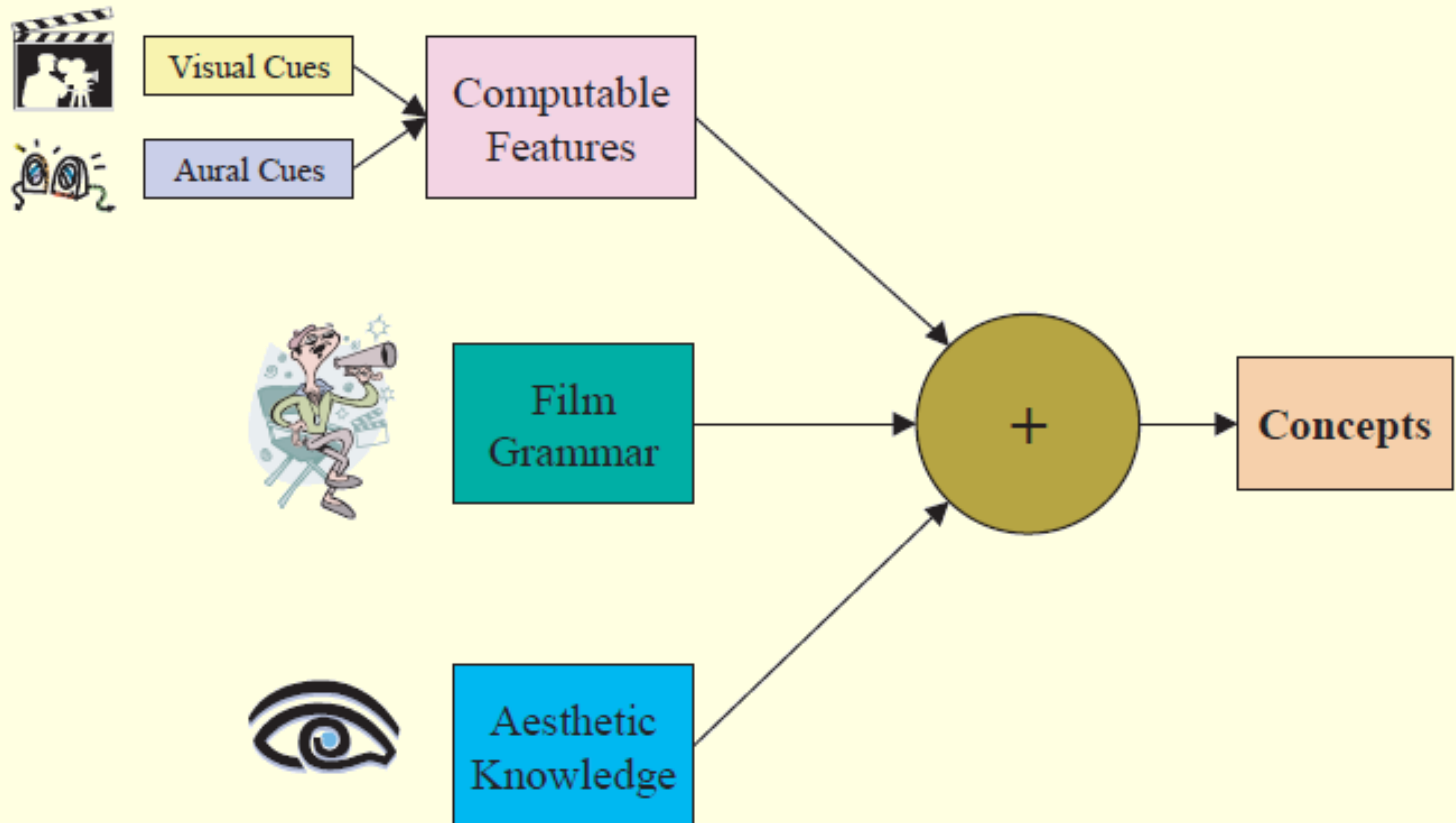
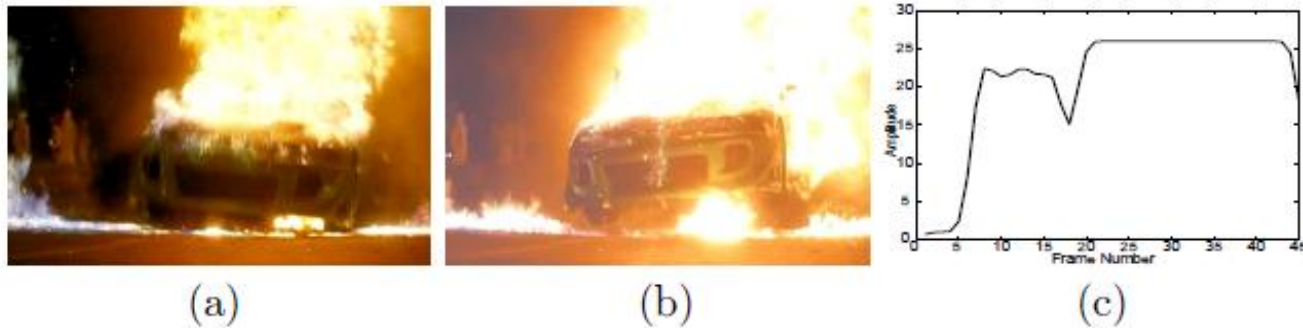# How a video is devided

Example

Figure 7.14. Detection of *fire/explosion* in a shot of the movie "The Fast And The Furious". (a) and (b) are two frames of the shot. (c) the plot of the index of the histogram peak against time. The shot was successfully identified as *fire/explosion* (images courtesy of "Original Film".)

References

- Autonomy, MBC, http://www.autonomy.com/

- Manning, Prabhakar, Schütze, „An Introduction to Information Retrieval" (Draft), Cambridge University Press, Cambridge, April 2009;

- Dr. René Witte und Jutta Mülle: Text Mining: Wissensgewinnung aus natürlichsprachigen

- AZRIEL ROSENFELD, DAVID DOERMANN, DANIEL DEMENTHON: Video Mining