Text Based Information Retrieval
H02C8a
H02C8b

Academic year 2010-2011

Instructors

Marie-Francine Moens
Katholieke Universiteit Leuven
Department of Computer Science
Room 04.25
Celestijnenlaan 200A
B-3001 Heverlee
tel: 016-325383
e-mail: sien.moens@cs.kuleuven.be
http://www.cs.kuleuven.be/~liir

Karl Gyllstrom
Katholieke Universiteit Leuven
Department of Computer Science
Room 04.49
Celestijnenlaan 200A
B-3001 Heverlee
tel: 016-325383
e-mail: karl.gyllstrom@cs.kuleuven.be
http://www.cs.kuleuven.be/~liir

The **aim of the course** to study the current techniques and algorithms commonly used in text based information retrieval and the challenges of this field. The theoretical insights are the basis for discussions of commercial systems and ongoing research projects.

After the study of this course the student should be able to:

1) describe and understand fundamental concepts and algorithms in information retrieval and text mining
2) design, partially implement, and evaluate a text based information retrieval system.

- **4 study points:**
  - Lectures every Tuesday 17:30 - 19:00
  - 5 exercise sessions: 2 groups:
    - some Wednesdays 14:00-16:00
    - ?
  - Small project
- **6 study points:**
  - Exercises and lectures: see above
  - 3 extra lectures/seminars on Thursdays 10:30-12:00: 17/3, 24/3 and 31/3
  - 2 extra exercise sessions
  - Larger project

Chapter 1: Introduction
Text Based Information Retrieval

Information retrieval

- **Information retrieval** (IR) =
  - representation, storage and organization of information items in databases or repositories and their retrieval according to an information need
- **Information items**:
  - format of text, image, video, audio, ...
    - e.g., news stories, e-mails, web pages, photographs, music, statistical data, biomedical data, ...
- **Information need**:
  - format of text, image, video, audio, ...
    - e.g., search terms, natural language question or statement, photo, melody, ...

Information retrieval

- Text-based information retrieval:
  - **information item**: **document** (or document element) in textual format (written or spoken) or has textual description
  - **information need**: **query** in textual format

Is IR needed? Yes

- Large document repositories (archives):
  - of companies: e.g., technical documentation, news archives
  - of governments: e.g., documentation, regulations, laws
  - of schools, museums: e.g., learning material
  - of scientific information: e.g., biomedical articles
  - on hard disk: e.g., e-mails, files
  - of police and intelligence information: e.g., reports, e-mails, taped conversations
  - accessible via P2P networks on the Internet
  - accessible via the **World Wide Web**
  - ...

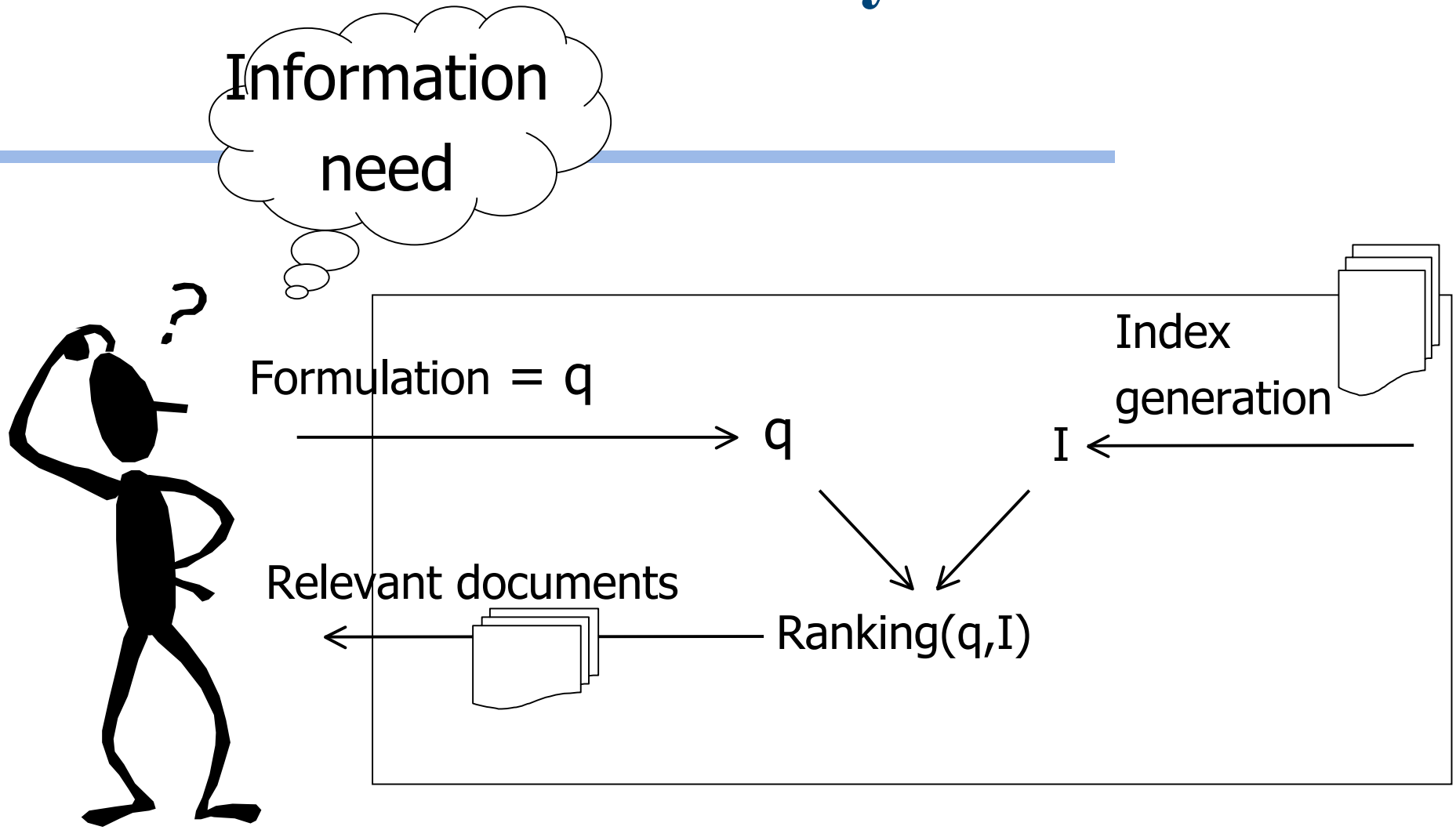
Information retrieval process

- Classical information retrieval system: 3 steps:
  1. generation of a representation of the content of each information item
  2. generation of a representation of the content of the information need of user
  3. the two representations are compared in order to select items that best suit the need

step 1: usually performed before the actual querying

steps 2 and 3 : performed at query time

Classical retrieval system



Data retrieval versus IR

- **Data retrieval** (e.g., from a relational database):
  - data retrieval language (e.g., SQL) with well-defined structure and semantics
  - data have well-defined structure and semantics
  - **deterministic matching** of query and data
- **Information retrieval**:
  - query and documents lack well-defined structure and semantics
  - often **non-deterministic matching** of query and documents
  - result of the retrieval: **ranking** of documents according to relevance

Data retrieval versus IR

- Current retrieval often deals with document bases that are mixture of structured and unstructured information: **bridging structured and unstructured data**
  - e.g., XML retrieval
  - e.g., retrieval of extracted entities and relationships combined with full text
  - e.g., multimedia retrieval

Information retrieval process

- Current retrieval systems
  - information need expressed as:
    - keywords
    - query by example
    - question in natural language
  - Results expressed as:
    - list of documents
    - clusters of documents and visualization of topics
    - short answer to natural language question
- Variant: navigation via linked content
- Future: exploration, synthesis, multimodal queries, ...

Information retrieval problem

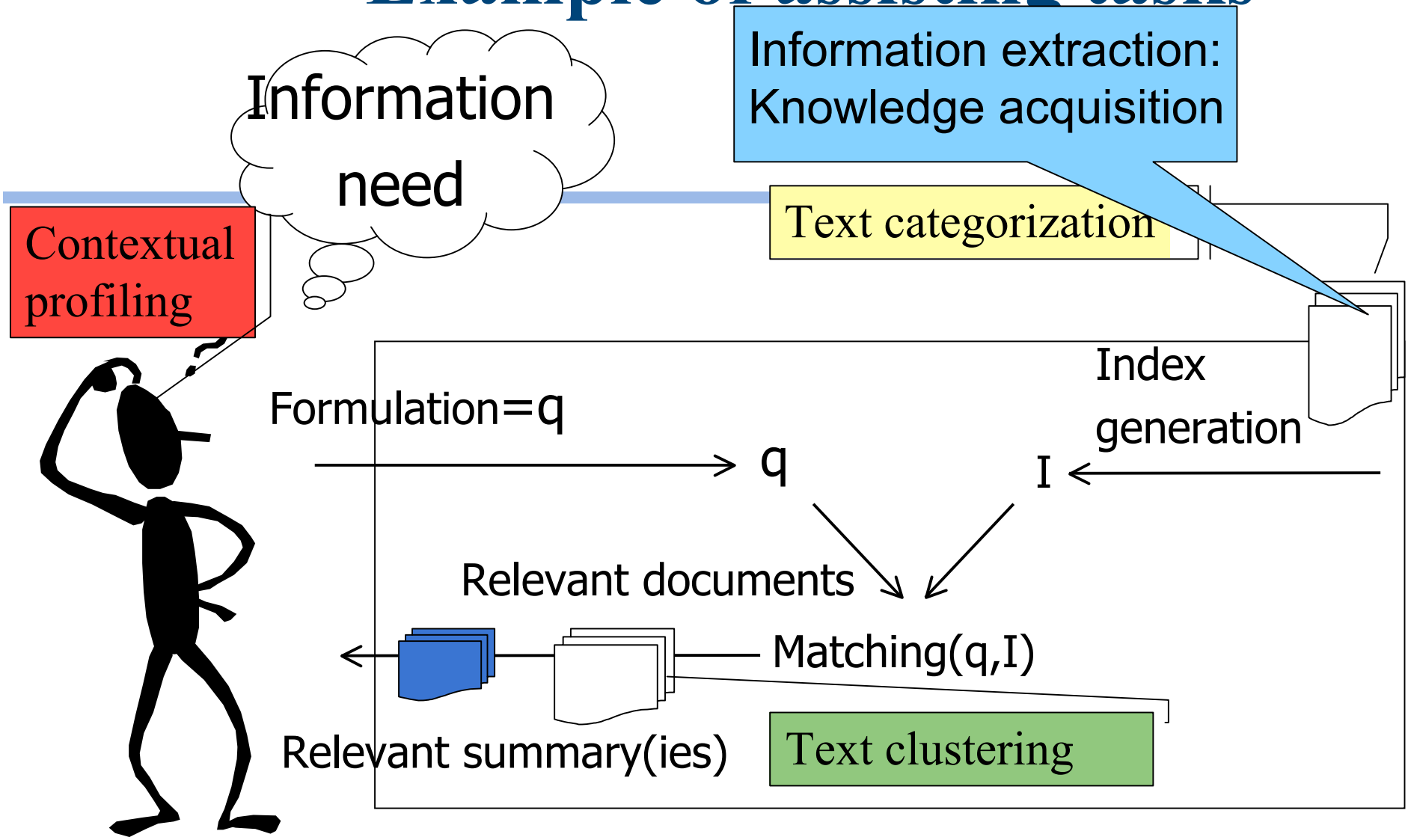
- Natural language = powerful and natural means of communication
- Classical text retrieval:
  - does not yield all information relevant to the information need and/or does supply information that is not or is only marginally relevant to the information need
  - when searching large databases (e.g., World Wide Web): returns too many hits which are impossible to consult

=>information retrieval problem

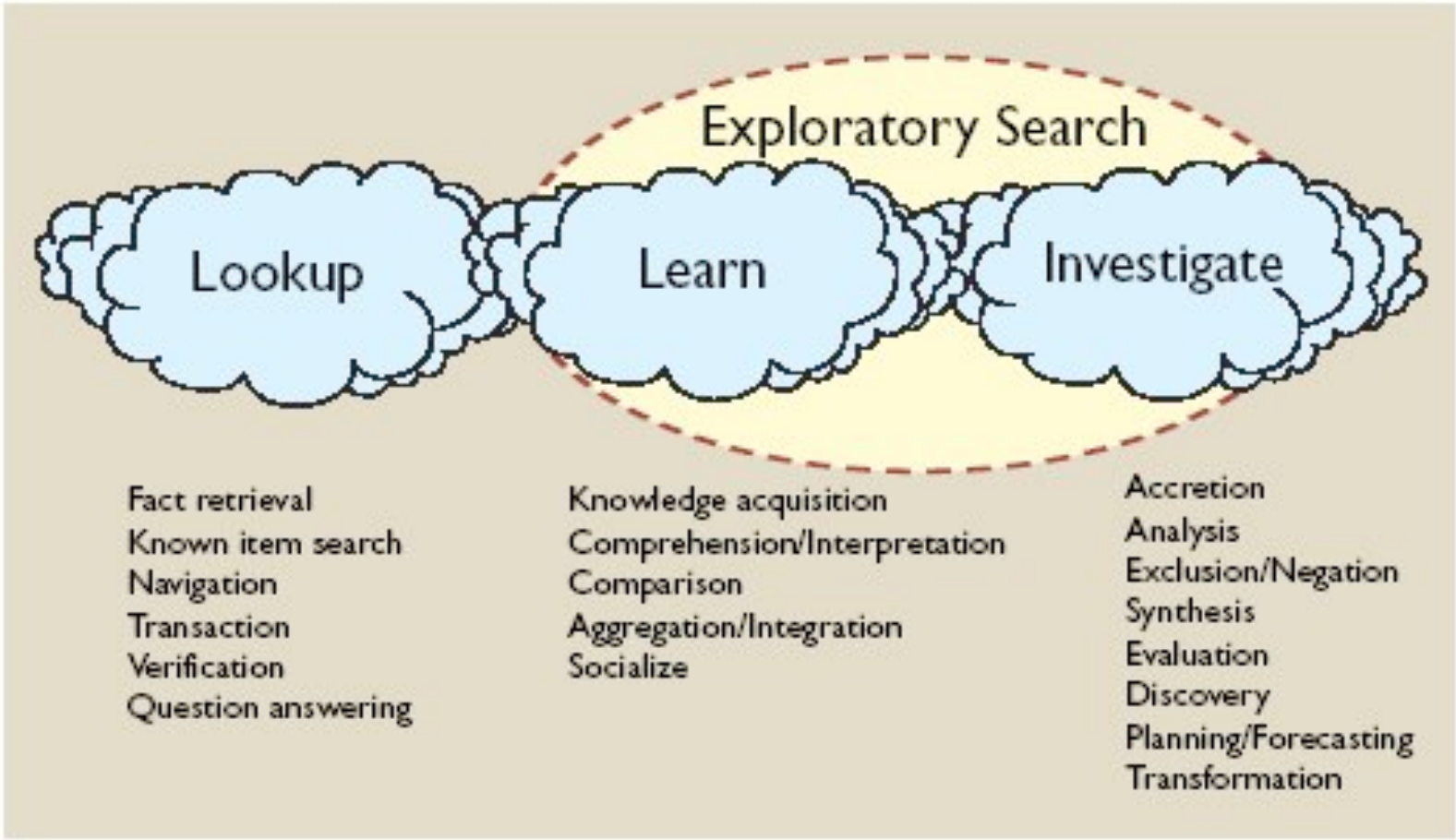
Information retrieval problem

- Solution:
  - **context based** and **personalized** retrieval: attempts to capture the user's goals, intentions, profile, background knowledge, location, ...
  - **information extraction** technologies that provide a rough understanding, ...
  - **advanced retrieval models** that capture probabilistic evidence from a variety of sources: e.g. language models, inference network models, ...
  - **information visualization** and presentation

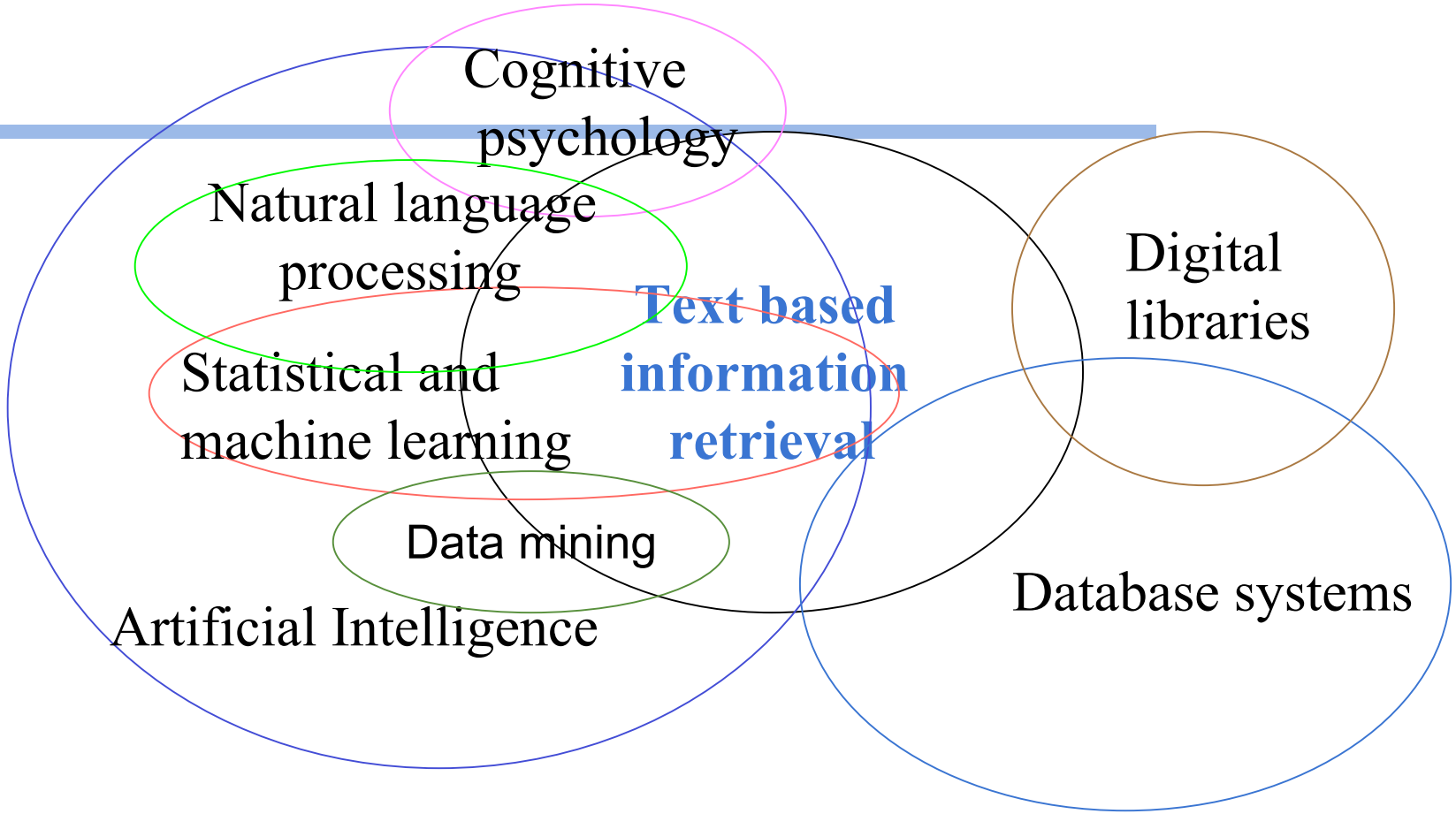
Example of assisting tasks



Information retrieval evolves towards exploration



[Marchionini CACM 2006]



Important application: **Web search and mining**

Methods

- Text based information retrieval uses methods of:
  - probability theory
  - linear algebra
  - information theory
  - machine learning
  - natural language processing

Overview of the course

1. Introduction
2. Background
   - Text and its characteristics: the features
   - Link and graph based algorithms
   - Statistical and machine learning techniques
3. Classical representation of documents and information needs
4. Retrieval models including link-based models
5. Advanced representations
6. Indexing structures and search techniques
7. Interactive retrieval
8. Evaluation measures

Fundamental techniques

9. Text categorization
10. Text clustering
11. Information extraction and
12. question answering
13. Result presentation and summarization

Assisting text mining tools

14. Cross-language information retrieval
15. Multimedia information retrieval

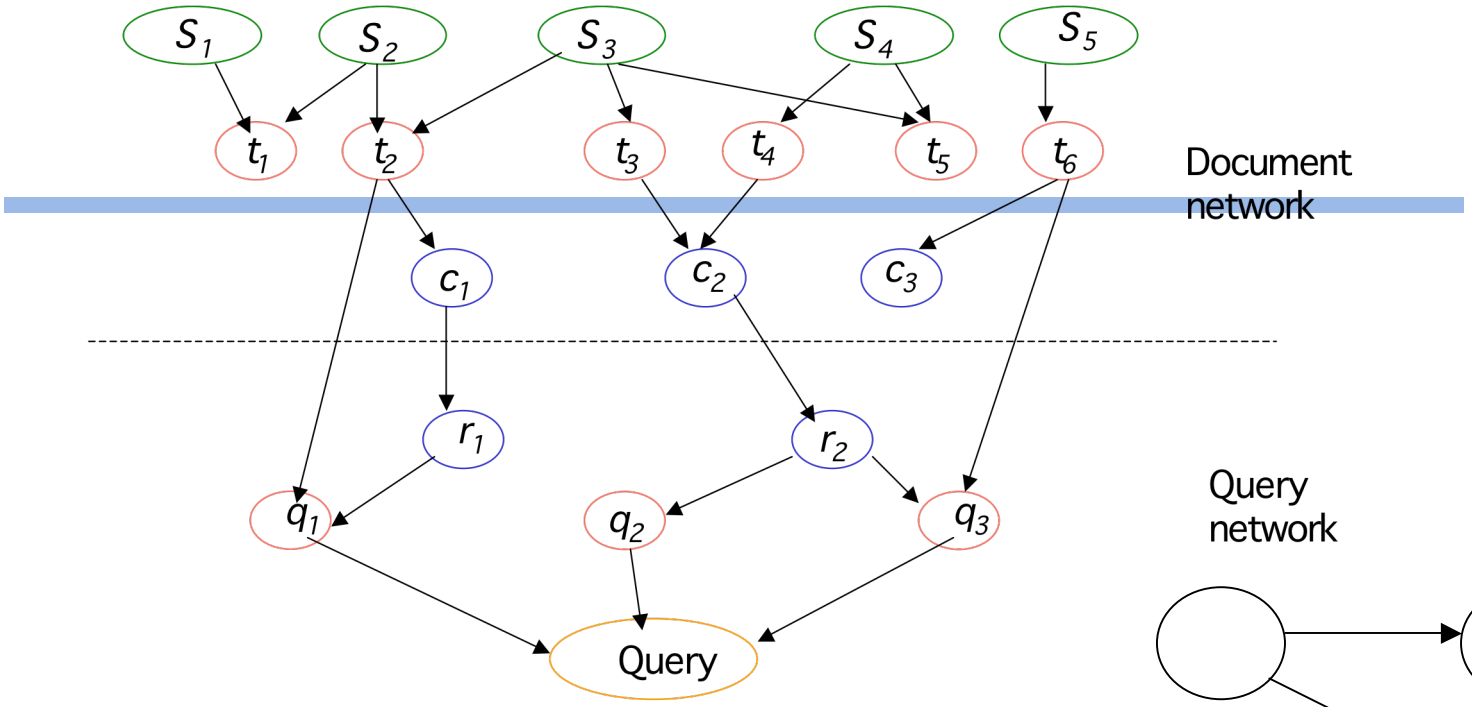
Broader framework

Representation of documents and information needs

- Traditional representations
  - natural language index terms
  - controlled language index terms
  - term weighting
- Current interest
  - **adding structural and semantic information** to the document representation
  - **adding contextual information** to the information need

Retrieval models

- Retrieval model defined by:
  - document and information need representation format
  - by matching algorithms
- Types of models:
  - **set theoretic**: e.g., Boolean, extended Boolean
  - **algebraic**: e.g., vector space
  - **probabilistic**: e.g., language, inference network
  - **link based**: e.g., PageRank
  - ...



$$P(cq_1,\ldots,cq_m|D_j) = \prod_{i=1}^{m}(\alpha\sum_{l=1}^{l} P(cq_i|w_l)P(w_l|D_j) + \beta P(cq_i|D_j) + (1-\alpha-\beta)P(cq_i|C))$$

Advanced representations

- Latent semantic indexing (LSI)
- Probabilistic latent semantic indexing (pLSI)
- Latent Dirichlet Allocation
- Incorporation in retrieval models

Web information retrieval

- Issues such as:
  - Scalability
  - Heterogeneous content
  - Multimedia
  - User generated content
  - Computational advertisements
  - ...

  - And the consequences for information retrieval

Indexing structures and search techniques

- Sequential searching
  - finite state automaton
- Indexing structures
  - inverted files
- Crawling-indexing architecture of search engines on the World Wide Web
- MapReduce architectures
- Compression and searching

- Limited for 4 study points, in detail for 6 study points

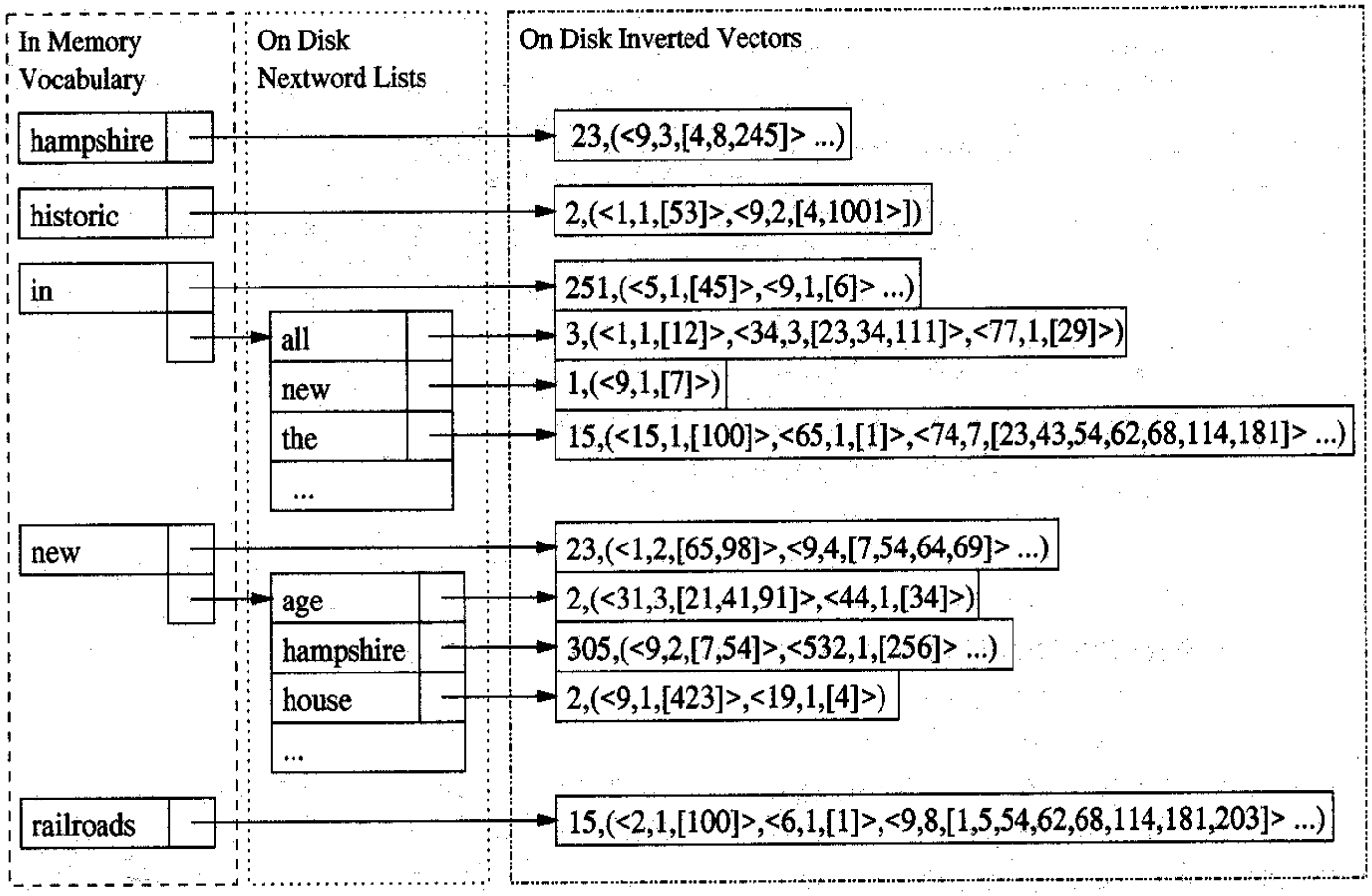


Figure 3: A combined inverted file and nextword index.

[Bahle et al., SIGIR 2006]

Evaluation measures

- The concept of relevance
- Classical measures:
  - Recall
  - Precision
  - F-measure
- Other measures: MAP, area under ROC, confusion matrix, ...

Text categorization

- = **semantic labeling** of whole documents for filtering purposes:
  - assignment of subject descriptors
- Techniques:
  - focus on supervised learning of classification patterns: e.g., learning of rules and trees, naive Bayes, support vector machines
  - focus on hierarchical classification

AP

Business Schools Sprout Greener MBAs

Sunday September 23, 1:57 pm ET

By Michelle Locke, Associated Press Writer

Green MBAs Seek to Balance Profit and Planet

Ecology

Education

...

OAKLAND, Calif. (AP) -- Business professor John Stayton remembers when eyes would start rolling at the idea of a "green MBA."

These days, business schools across the country are incorporating the environmental and social costs of doing business into their curricula, and a few, like the program Stayton directs at Dominican University of California, aim for an all-green program.

The goal? How to succeed in business without really frying the planet.

"Essentially we've got to change the way we're doing everything and making everything," said Stayton.

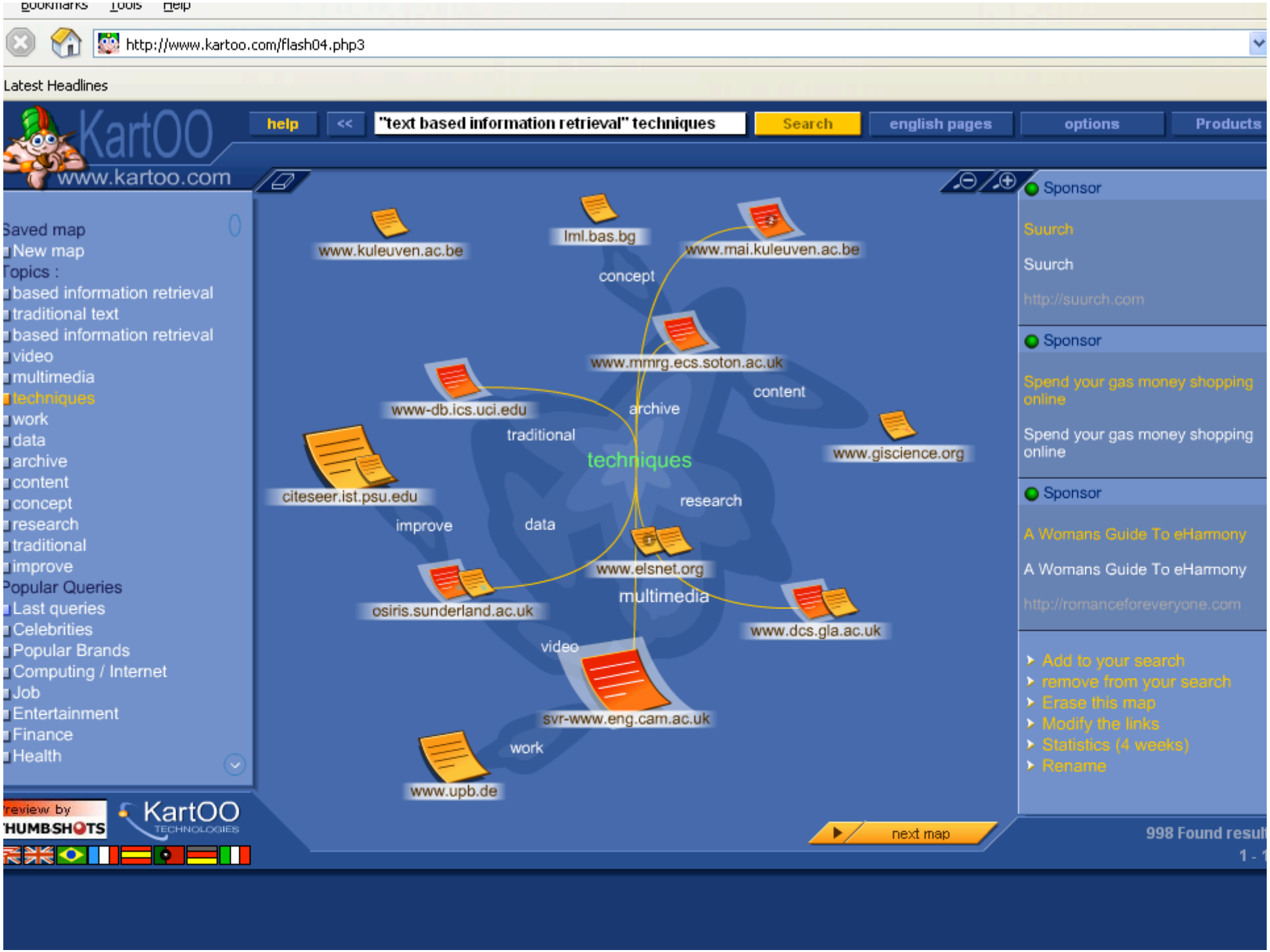
The program Stayton directs was launched at Santa Rosa's New College of California North Bay in 2000 as a Master of Arts in the humanities department and transferred to Dominican last spring. It's one of a handful of such degrees; others include MBAs offered at the Presidio School of Management in San Francisco and the Bainbridge Graduate Institute in Washington state.

The move to balance economy and ecology is showing up all over, said Rich Leimsider, director of the Center for Business Education at The Aspen Institute, a leadership think tank which reports on how Master of Business Administration programs are adding social and environmental issues to their courses in its biennial "Beyond Grey Pinstripes" report.

"It matters what the senior executives of companies do, say and think," said Leimsider. "If you can change business education to include an appreciation for the social and environmental context you wind up with leaders who are really good at creating

Text clustering

- Important unsupervised learning technique for text based retrieval
- Clustering of:
  - **terms** based on their co-occurrence in documents, paragraphs or sentences
    - e.g., used in query expansion, automatic thesaurus construction
  - **documents, sentences** or **paragraphs** based on shared terms or other textual features
    - e.g., used in the detection of similar content, classification, visualization


Bookmarks Tools Help
http://www.kartoo.com/flash04.php3
Latest Headlines
KartOO
www.kartoo.com
help
<<
"text based information retrieval" techniques
Search
english pages
options
Products
Saved map
New map
Topics :
based information retrieval
traditional text
based information retrieval
video
multimedia
techniques
work
data
archive
content
concept
research
traditional
improve
Popular Queries
Last queries
Celebrities
Popular Brands
Computing / Internet
Job
Entertainment
Finance
Health
www.kuleuven.ac.be
lml.bas.bg
www.mai.kuleuven.ac.be
concept
www.mmrg.ecs.soton.ac.uk
content
www-db.ics.uci.edu
archive
traditional
techniques
www.giscience.org
citeseer.ist.psu.edu
research
improve
data
www.elsnet.org
osiris.sunderland.ac.uk
multimedia
www.dcs.gla.ac.uk
video
svr-www.eng.cam.ac.uk
work
www.upb.de
Sponsor
Suurch
Suurch
http://suurch.com
Sponsor
Spend your gas money shopping online
Spend your gas money shopping online
Sponsor
A Womans Guide To eHarmony
A Womans Guide To eHarmony
http://romanceforeveryone.com
Add to your search
remove from your search
Erase this map
Modify the links
Statistics (4 weeks)
Rename
Preview by
THUMBSHOTS
KartOO
TECHNOLOGIES
next map
998 Found result
1 - 1

Information extraction

- = **semantic labeling** of content in documents,
  - e.g., named entities, coreferent resolution, relation detection, scenario detection, ...
- Techniques:
  - focus on supervised learning of classification patterns: e.g., maximum entropy models, hidden Markov models, conditional random fields
  - weak supervision
- Special application: **question answering** systems

Autonomy and Verity Join Forces in Enterprise Search Market *by Paula J. Hane*

November 14, 2005 Two competitors in the enterprise search space have joined forces. U.K.-based Autonomy Corp. PLC announced it will acquire the larger U.S.-based Verity, Inc. for approximately $500 million. The companies said that their products are complementary and that the combination will provide customers with a broader and more powerful set of solutions that address the increasingly demanding requirements of information access. The combined entity will be branded Autonomy and will maintain global headquarters in Cambridge, England, while Verity will become the base for U.S. operations. Mike Lynch, Autonomy's group CEO and co-founder, will continue as CEO of the expanded group. Anthony J. Bettencourt, CEO of Verity, will assume the role of CEO, Autonomy, Inc., the company's U.S. unit. He …

…

Result presentation and summarization

- = generating the content of text in a condensed form
  - single-document summarization
  - multi-document summarization: aggregating content
  - text compaction for display on very small screens
- Techniques:
  - compression of content
  - alignment and fusion of content
  - use of supervised and unsupervised learning

[Otterbacher et al. IP & M 2008]

Interactive information retrieval

- How do ***real users*** search for, identify, and collect information?
- How does this affect how we ***build*** and ***evaluate*** systems? What assumptions change?
- How are new search paradigms (e.g., twitter) changing the landscape?

Cross language information retrieval (CLIR)

- = query in one language, but retrieval of documents in multiple languages
- Techniques: focus on:
  - **query translation**
  - **query expansion** to resolve ambiguity in query term translation
  - acquiring knowledge from parallel and comparable corpora
  - retrieval models

搜尋

設定雅虎中

藝術與人文
文學, 繪畫, 表演藝術...

新聞與媒體
網路廣播, 雜誌, 報紙, 電視...

商業與經濟
公司, 金融與投資, 就業...

休閒與生活
體育運動, 生活資訊, 遊戲, 旅遊...

電腦與網際網路
網際網路, 聊天室, 軟體...

參考資料
圖書館, 字典與辭典, 電話號碼...

教育
大專院校, 中小學, 留學...

區域
中國大陸, 台灣, 香港, 美國...

娛樂
酷站, 音樂, 電影, 明星照片...

科學
生物學, 工程學, 另類科學...

政府與政治
各地政府, 法律, 軍事...

社會科學
人類學, 社會學, 經濟學...

健康與醫藥
醫學, 疾病與症狀, 中國醫藥...

社會與文化
個人網頁, 飲食, 宗教信仰...

MADRID (Reuters) - La tormenta tropical Gordon se alejó el jueves de Galicia tras azotar por la mañana la región con vientos que llegaron a superar los 130 kilómetros por hora que provocaron cortes de luz a miles de personas y la caída de decenas de árboles.
La Xunta de Galicia había suspendido las clases y reforzado los servicios de emergencia ante la llegada de Gordon, que tocó tierra a las siete de la mañana en la zona atlántica y se fue desplazando hacia el interior hasta empezar a disminuir hacia el mediodía.
Los incidentes consistieron principalmente en la caída de decenas de árboles así como postes de luz, de teléfono, marquesinas y farolas, según datos de Protección Civil.
En total, los servicios de emergencias gallegos realizaron 332 intervenciones. Una persona resultó herida al caer un árbol sobre un coche en la provincia de La Coruña.
Unión Fenosa calculó en alrededor de 29.000 los clientes afectados por cortes de luz en toda Galicia,

Multimedia information retrieval

- = retrieval of images, video and audio
- Focus upon=
  - retrieval models
  - automatic content recognition taking into account accompanying text
  - spoken document retrieval

Cleveland Cavaliers' LeBron James (23) shoots between Detroit Pistons' Richard Hamilton, left, and Chauncey Billups late in the fourth quarter of the Pistons' 84-82 win in a second-round NBA playoff basketball game Friday, May 19, 2006, in Cleveland. The series is tied at three games each.

LeBron James: 0.75 Chauncey Billups: 0.464 Richard Hamilton: 0.333

Examples of commercial text retrieval systems

- General **Web search engines**
- Legal domain:
  - Westlaw, Lexis-Nexis
- Medical domain:
  - Medline
- Enterprise settings

=> Invited lecture

Discussion of interesting research projects

- Illustration of topics of the course by current research projects

Course material

- Course slides and exercise solutions can be downloaded from the Toledo platform of the K.U.Leuven
  - http://toledo.kuleuven.be (**H02C8a** and **H00G9a**)
- Further reading at the end of each chapter and available at the Toledo platform: if you want to know more ...

Exercises

- **H00G9a**
- Give an in-depth understanding of important algorithms
- Demonstrations
- 4 study points: 5 sessions of 2 hours
- 6 study points: 2 sessions of 2 hours

Exam

- **Project / assignment** (grading: 33.3%):
  - 4 study points: short paper (ca. 3-4 pages): formulation and justification of solution to the problem OR small programming assignment: **due April 28**
  - 6 study points: large programming assignment:
    - 1st part: **due April 28**; second part: **due May 26**
- **Oral, closed book theoretical exam** (grading: 33.3 %):
  - broad overview questions, high level design of system that solves a particular problem, comparison of technologies, ...
- **Written, open book exercise exam** (grading: 33.3 %):
  - focus on algorithms, evaluation metrics, writing of small high level program, ...

Practical information

- Instructors:
  - Marie-Francine Moens (sien.moens@cs.kuleuven.be)
  - Karl Gyllstrom (karl.gyllstrom@cs.kuleuven.be): lesson on interactive retrieval

- Teaching assistant:
  - Raquel Mochales Palau (raquel.mochales@cs.kuleuven.be)