Retrieval & Classification with the mySAP Workplace

Otto Kühn
SAP AG

mySAP Workplace (Definition)

- **mySAP Workplace provides**
  - **personalized**
  - **convenient access to**
  - **everything needed**
  - **for all users**
  - **anytime, anywhere**
  - **via the Internet**
  - **to get their tasks/jobs done**
- **This requires in particular**
  - **easy access to relevant documents**
  - **direct access to information contained in document collections**
  - **support in building well-structured document repositories**

From Poster to Seeker and Back

- Authoring tools
- XML-based forms
- Push capabilites

- Feedback
- Rating

- Retrieval
- Subscription

mySAP Workplace

The Web
Diverse file systems

Crawler

- Document classification
- Publishing pipeline
- Authorizations

Text Retrieval and Information Extraction

Powerful retrieval and text mining engines

- based on “State of the Art” technology in
  - computer linguistics
  - machine learning
  - statistical natural language processing
- efficient implementation for
  - handling millions of documents
  - in different languages
  - from arbitrary sources
  - with fast response times
- multiple functionalities
  - based on common document index
  - provide optimal support for knowledge management

Overview of Functionalities

Today

- **Full-Text search (with Verity search engine):**
  find matching documents
- **"See Also" search:**
  get more documents like this
- **Feature extraction from documents:**
  find characteristic keywords

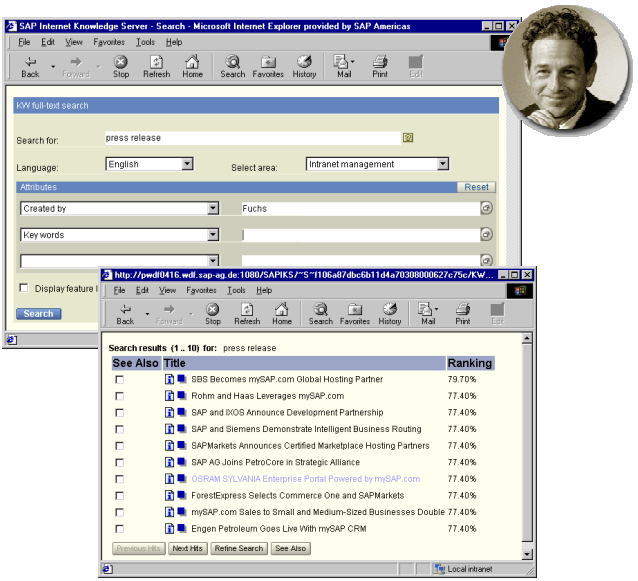
next year

- **Document classification:**
  assign a document to one or more predefined categories
- **Person search:**
  find knowledgeable experts
- **Term search:**
  find better search terms; discover interesting relationships
- **Document clustering:**
  discover sets of related documents

Full-Text Search

- **Standard functionalities**
  - keyword search (e.g. Clinton)
  - phrase search (e.g. „Hillary Clinton")
  - boolean search (e.g. „Hillary Clinton" & „New York")
  - attribute search (e.g. Clinton & category~=USA)
  - area-specific search (e.g. Clinton <in> title)
- **Advanced functionalities**
  - linguistic search (e.g. want will also find wants or wanted)
  - fuzzy search (e.g. Hillary will also find Hilary)
  - Search refinement
    (expand query by related terms suggested by system)

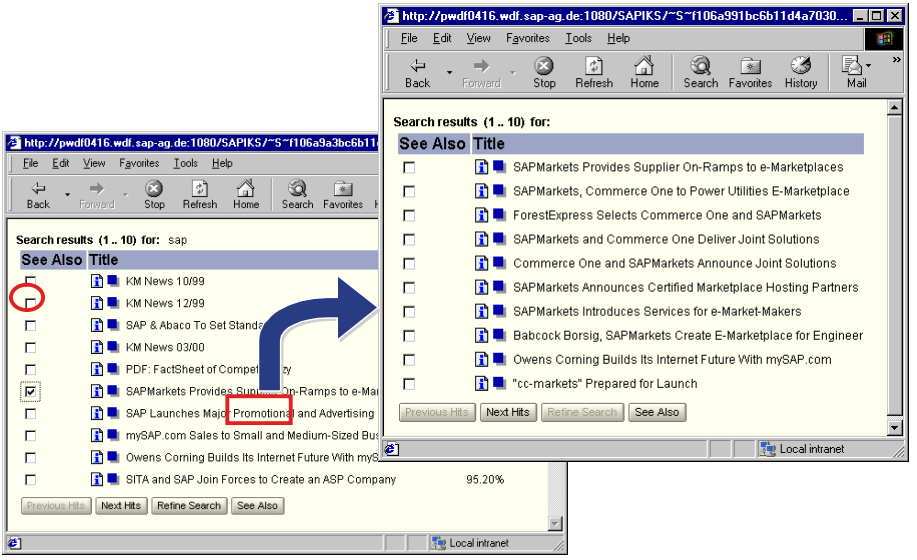
Find Documents



"See Also" Search

- **Query consists of**
  - **one entire document**
  - **several selected documents**
  - **a class of documents (defined by an attribute value)**
- **Target documents are**
  - **other documents in current collection**
    - **similarities between the query documents(s) and all other documents are computed**
    - **the most similar documents are returned**
  - **external documents**
    - **the most important terms of the query document(s) are extracted**
    - **a keyword search including these terms is submitted**
    - **returned documents are ranked wrt. similarity to query document(s)**

“See Also” Search (Example)



Person Search

- Documents are indexed together with their authors
- the user specifies a search query as for a document search
- instead of returning a list of documents
  the system returns a ranked list of author names
- these authors are likely to be knowledgeable in the topics specified in the query

The system thus also enhances access to implicit knowledge available only in people‘s heads.

In contrast to a traditional skill database, no manual maintenance of skill profiles is required.

Document Classification

- Assign a new document to one or more predefined categories
  - The system analyzes the document and provides a ranked list of suggestions
  - User selects to which categories the document should be assigned
  - Automatic assignment to top category or to all categories with a rank value above some given threshold
- Important applications
  - Updating a structured document collection
    - Multi-user document repository
    - personal repository
    - online catalog
  - Automatic forwarding of incoming documents
    - Customer mail to appropriate department
    - News messages to people with corresponding interest profiles

Document Clustering

- Determine sets of documents with similar contents
  - Analyze all inter-document similarities
  - Build a hierarchical structure reflecting the document similarities
  - Compute an optimal assignment of all documents into a predefined number of categories
- Important applications
  - Identification of redundant (duplicate) documents
  - Development of an appropriate categorization scheme (category labels may be identified with feature extraction)
  - Facilitate browsing in a document collection

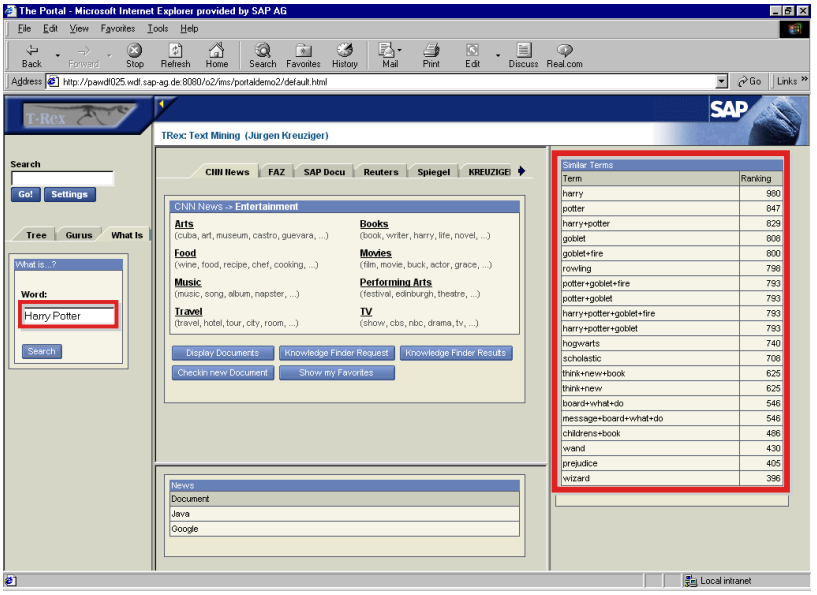
Feature Extraction

- Determine characteristic features (words and phrases) which distinguish a set of documents from other documents
  - Based on grammatical category (nouns and noun phrases)
  - Based on occurrence frequency in target and reference documents
- Important applications
  - Quickly assess relevance of found documents (alternative to abstracting)
  - Class features provide additional information besides the class label
  - Use extracted features to submit a new query
  - Use extracted features to generate individual interest profiles

Term Search

- Specify one or more search terms
- Find other terms which are related with these terms in the given document collection
- Important applications
  - Help to formulate more precise search queries
  - Support for domain-specific thesaurus generation
  - Automatic query expansion
  - Discovering interesting relationships (Text Mining) (e.g. In what context does the name of my company appear in the current press)
  - Finding the meaning of unknown terms
  - Answering specific queries

Analysis of term relations



Summary

Text retrieval and information extraction provide a rich set of functionalities for effective knowledge management

- Find relevant documents
- Extract knowledge from masses of documents
- Find knowledgeable experts
- Build structured document collections

mySAP Workplace provides

- Convenient user interface for the various functionalities
- Integration of Information, Applications and Services
- Role-based personalization
- Access via Internet from everywhere

Questions & Answers

Any questions ?

Copyright