

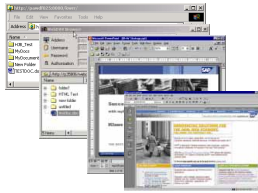
Retrieval & Classification with the mySAP Workplace

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

**Publisher/
Poster**



- Authoring tools
- XML-based forms
- Push capabilities

- Feedback
- Rating

Seeker



- 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

**T
o
d
a
y**

- **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

**n
e
x
t**

- **Document classification:**
assign a document to one or more predefined categories
- **Person search:**
find knowledgeable experts

**y
e
a
r**

- **Term search:**
find better search terms; discover interesting relationships
- **Document clustering:**
discover sets of related documents

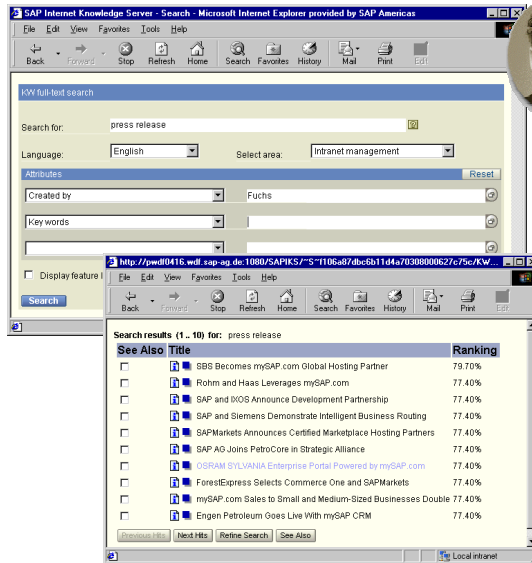
Full-Text Search

- **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)



“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)



Search results (1..10) for: sap

See Also	Title
<input type="checkbox"/>	KM News 10/99
<input type="checkbox"/>	KM News 12/99
<input type="checkbox"/>	SAP & Abaco To Set Standards
<input type="checkbox"/>	KM News 03/00
<input type="checkbox"/>	PDF: FactSheet of Competency
<input checked="" type="checkbox"/>	SAPMarkets Provides Supplier On-Ramps to e-Marketplaces
<input type="checkbox"/>	SAP Launches Major Promotions and Advertising
<input type="checkbox"/>	mySAP.com Sales to Small and Medium-Sized Business
<input type="checkbox"/>	Owens Corning Builds Its Internet Future With mySAP.com
<input type="checkbox"/>	SITA and SAP Join Forces to Create an ASP Company

95.20%

Previous Hits Next Hits Refine Search See Also

Search results (1..10) for: http://pwd0416.wdf.sap-ag.de:1080/SAPIKS/~S~f106a951bc6b11d4a7030...

See Also	Title
<input type="checkbox"/>	SAPMarkets Provides Supplier On-Ramps to e-Marketplaces
<input type="checkbox"/>	SAPMarkets, Commerce One to Power Utilities E-Marketplace
<input type="checkbox"/>	ForestExpress Selects Commerce One and SAPMarkets
<input type="checkbox"/>	SAPMarkets and Commerce One Deliver Joint Solutions
<input type="checkbox"/>	Commerce One and SAPMarkets Announce Joint Solutions
<input type="checkbox"/>	SAPMarkets Announces Certified Marketplace Hosting Partners
<input type="checkbox"/>	SAPMarkets Introduces Services for e-Market-Makers
<input type="checkbox"/>	Babcock Borsig, SAPMarkets Create E-Marketplace for Engineer
<input type="checkbox"/>	Owens Corning Builds Its Internet Future With mySAP.com
<input type="checkbox"/>	"cc-markets" Prepared for Launch

Previous Hits Next Hits Refine Search See Also

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

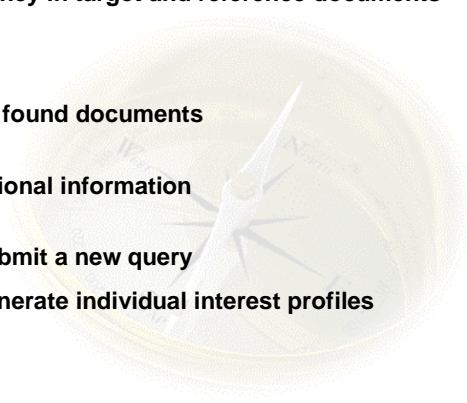
The system thus also enhances access to implicit knowledge available only in people's heads.

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

- **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

- **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

- **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



- **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



The Portal - Microsoft Internet Explorer provided by SAP AG

Address: http://pawd025.wdi.sap-ag.de:8080/n2/ins/portaldemo2/default.html

T-Rep

TREX: Text Mining (Jürgen Kreuziger)

Search: Go! Settings

Tree Gurus What is

What is ?

Word: Harry Potter

Search

CNN News FAZ SAP Docu Reuters Spiegel KREUZIGER

CNN News -> Entertainment

Arts (cuba, art, museum, castro, guevara, ...)

Food (wine, food, recipe, chef, cooking, ...)

Music (music, song, album, napster, ...)

Travel (travel, hotel, tour, city, room, ...)

Books (book, writer, harry, life, novel, ...)

Movies (film, movie, buck, actor, grace, ...)

Performing Arts (festival, edinburgh, theatre, ...)

TV (show, cbs, nbc, drama, tv, ...)

Display Documents Knowledge Finder Request Knowledge Finder Results

Checkin New Document Show my Favorites

News Document Java Google

Single Terms

Term	Ranking
harry	990
potter	847
harry+potter	829
goblet	808
goblet+fire	800
rowling	798
potter+goblet+fire	793
potter+goblet	793
harry+potter+goblet+fire	793
harry+potter+goblet	793
hogwarts	740
scholastic	708
think+new+book	625
think+new	625
board+what+do	546
message+board+what+do	546
childrens+book	486
wwand	430
prejudice	405
wizard	396

Local intranet

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

A Any questions ?



Copyright

- No part of this presentation may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.
- Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.
- Microsoft®, WINDOWS®, NT®, EXCEL®, Word® and SQL Server® are registered trademarks of Microsoft Corporation.
- IBM®, DB2®, OS/2®, DB2/6000®, Parallel Sysplex®, MVS/ESA®, RS/6000®, AIX®, S/390®, AS/400®, OS/390®, and OS/400® are registered trademarks of IBM Corporation.
- ORACLE® is a registered trademark of ORACLE Corporation, California, USA.
- INFORMIX®-OnLine for SAP is a registered trademark of Informix Software Incorporated.
- UNIX®, X/Open®, OSF/1®, and Motif® are registered trademarks of The Open Group.
- HTML, DHTML, XML, XHTML are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Laboratory for Computer Science NE43-358, Massachusetts Institute of Technology, 545 Technology Square, Cambridge, MA 02139.
- JAVA® is a registered trademark of Sun Microsystems, Inc. , 901 San Antonio Road, Palo Alto, CA 94303 USA.
- JAVASCRIPT® is a registered trademark of Sun Microsystems, Inc., used under license for technology invented and implemented by Netscape.
- SAP, SAP Logo, mySAP.com, mySAP.com Marketplace, mySAP.com Workplace, mySAP.com Business Scenarios, mySAP.com Application Hosting, WebFlow, R/2, R/3, RIVA, ABAP, SAP Business Workflow, SAP EarlyWatch, SAP ArchiveLink, BAPI, SAPHIRE, Management Cockpit, SEM, are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world. All other products mentioned are trademarks or registered trademarks of their respective companies.