

ADAM – Current research status

Research Day 09.03.2006

Ralph Weires

Supervision: Prof. Christoph Schommer

MINE Group – Management of Information and Net-Centric Computing

Overview

Introduction

Information Retrieval

Basic Concept

Research News

Next Steps

Overview

Introduction

Information Retrieval

Basic Concept

Research News

Next Steps

Field of Interest

- **Information retrieval out of the Web**
- **Search engines**
 - often poor results
 - only small amount of information really relevant to a user
- **Enhanced information retrieval**
 - Bring more intelligence into the search process
 - Take information about the users into account

Overview

Introduction

Information Retrieval

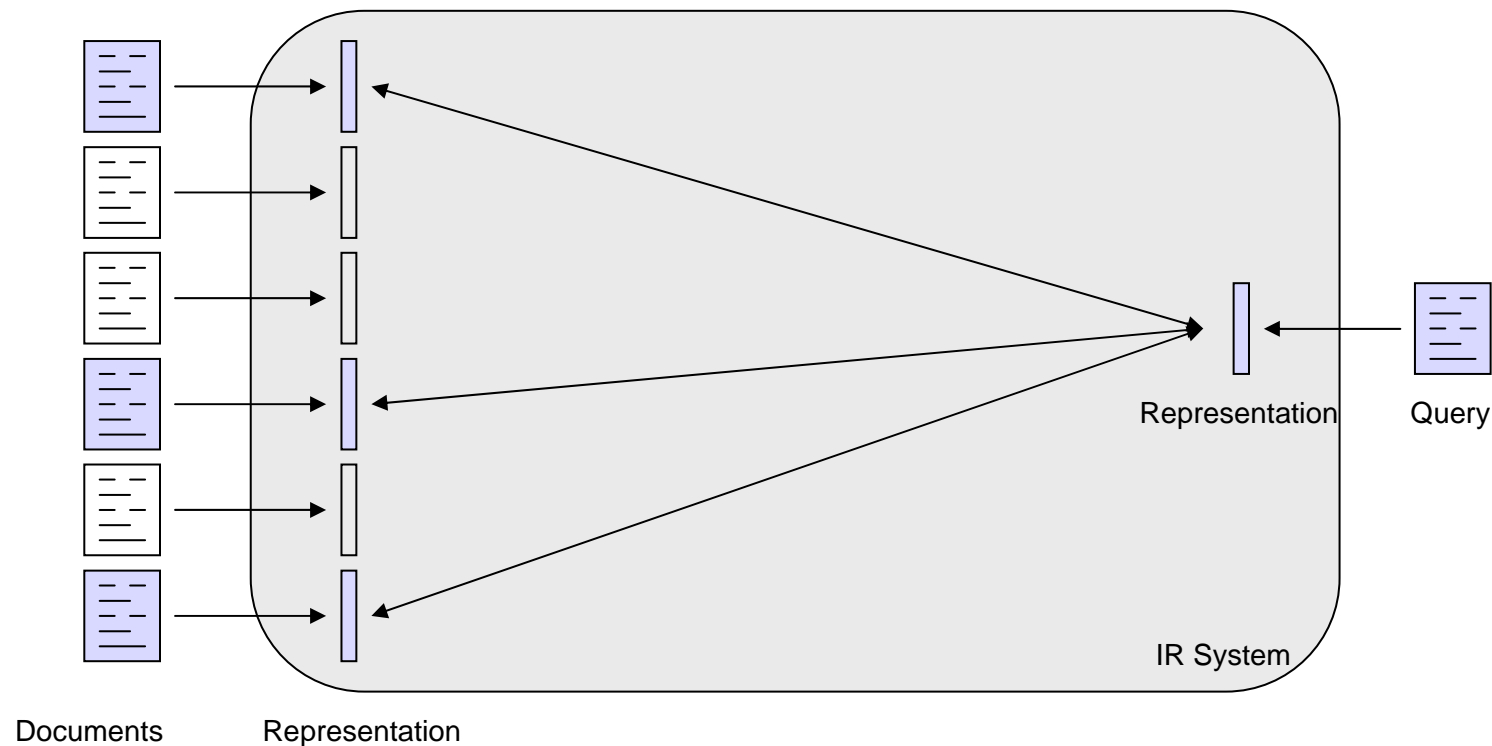
Basic Concept

Research News

Next Steps

Classical Information Retrieval

- How to find the right documents according to a given query



Boolean Retrieval

- **Representation of documents and queries by a list of binary attributes (each either 0 or 1)**
 - (0, 1, 1, 0, 0, 0, 1)
- **Result of query q**
 - All documents d which fulfill the condition $(q \text{ AND } d) = q$
 - Set of documents (\rightarrow unordered)

Vectorspace Model

- **Similar to boolean retrieval, but with real attribute values (often only in range [0, 1])**
 - (0.0, 0.8, 0.7, 1.0, 1.0, 0.2)
 - values e.g. resulting out of the number of occurrences of terms
 - In queries: weighted terms possible
- **Result of a document d for a query q**
 - e.g. defined by the scalar product $d * q$
 - $d * q = (d_1 * q_1 + d_2 * q_2 + \dots + d_n * q_n)$
 - Each document results in a certain value for the query
 - Ranking of documents possible, in order of relevance
 - Thresholds for defining relevance

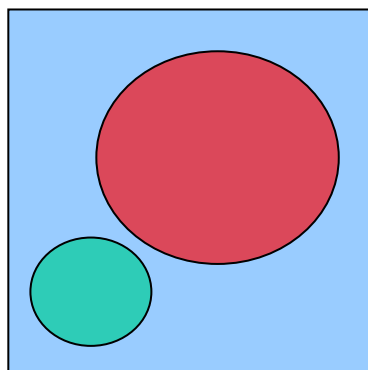
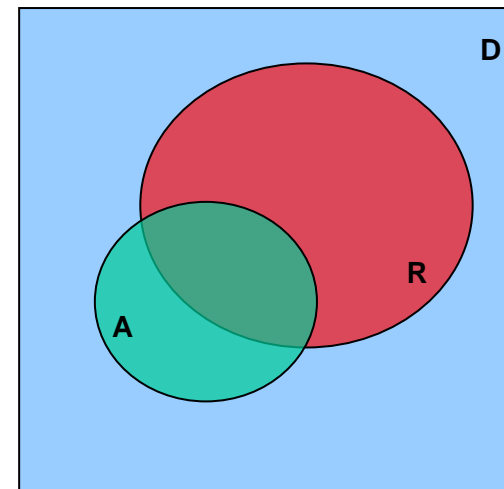
Measurement of IR Systems

- **Precision / Recall**

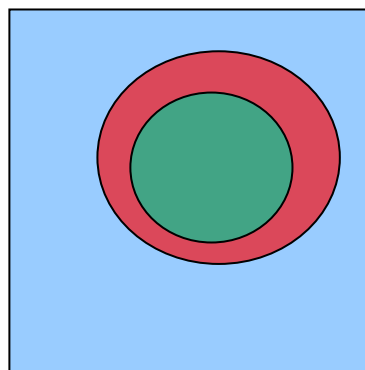
- D: All documents
- R: Relevant documents (for some query q)
- A: Documents returned by the IR system that is to be measured (for query q)

$$\text{Precision} = \frac{|A \cap R|}{|A|}$$

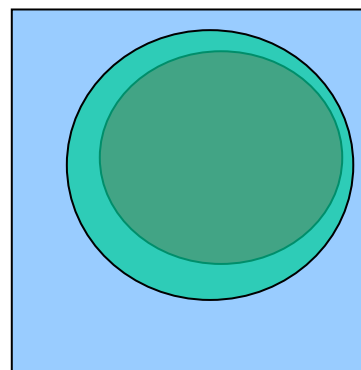
$$\text{Recall} = \frac{|A \cap R|}{|R|}$$



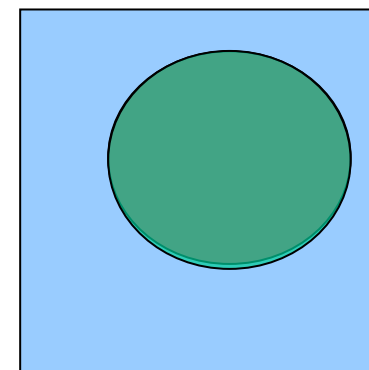
P=R=0



P=1, R<1



P<1, R=1



P=R=1

Some Additional Approaches

- **Document / Query preprocessing**
 - Elimination of stopwords (the, a, of, and, or, ...)
 - Stemming (Reduce words to their lexical basis)
 - Synonyms, Homonyms (e.g. Query Expansion)
- **Web-based IR**
 - Considering dependencies between documents
 - link structure, e.g. like PageRank in Google
 - Making use of the document structure
 - Document parts can be rated differently (title, headlines, ...)
 - Returning only parts of documents to the user
 - But: difficult to rely on certain information parts, for documents can contain anything

Overview

Introduction

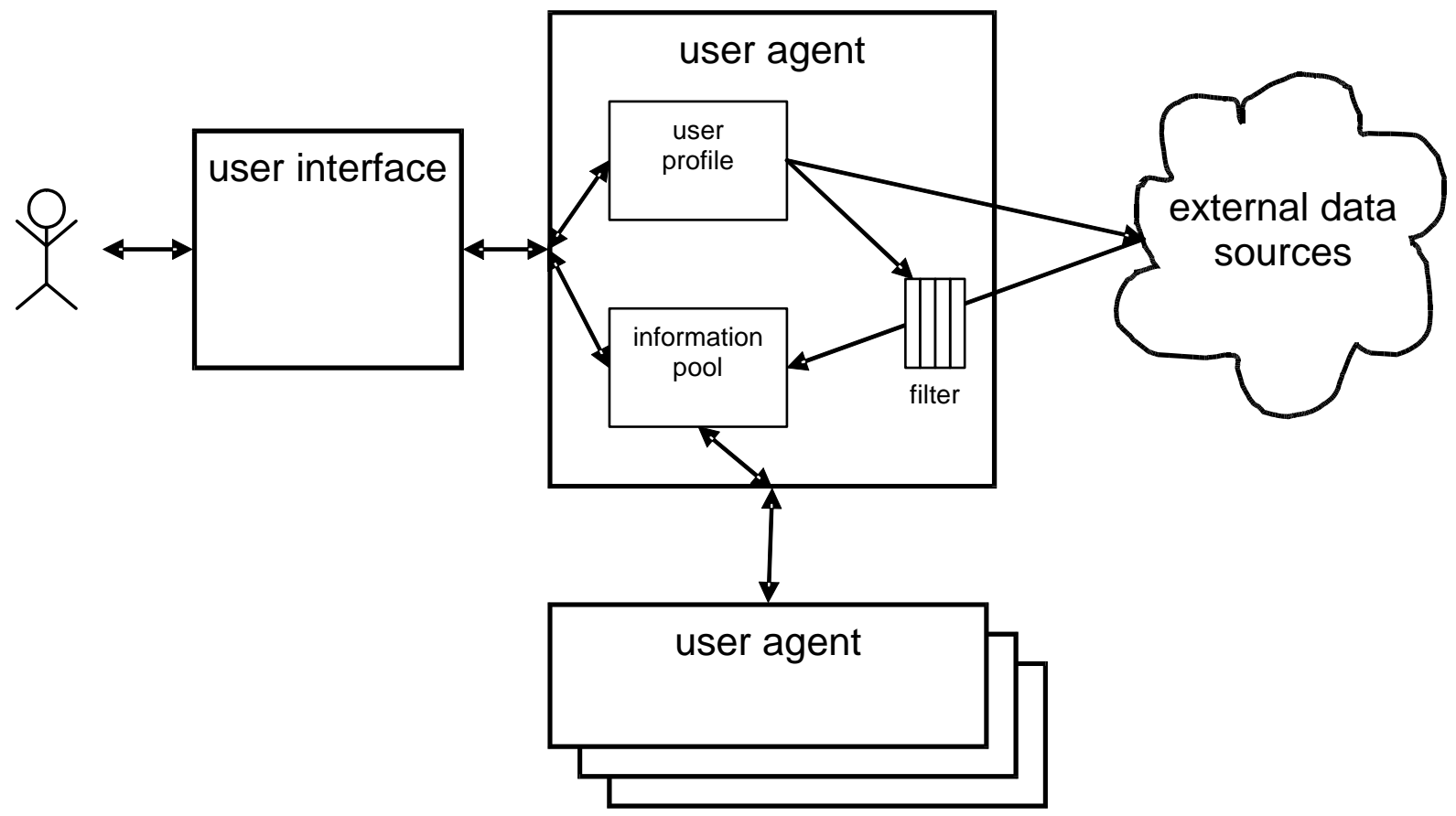
Information Retrieval

Basic Concept

Research News

Next Steps

Architecture



Example Application: FAQ Generation

- **Frequently Asked Questions (FAQs)**
 - e.g. for certain computer programs
 - many common problems with solutions can e.g. be found around forums in the web
 - additional reference for a user in case of problems
- **Main focus here on:**
 - find new and relevant information according to the profile

Overview

Introduction

Information Retrieval

Basic Concept

Research News

Next Steps

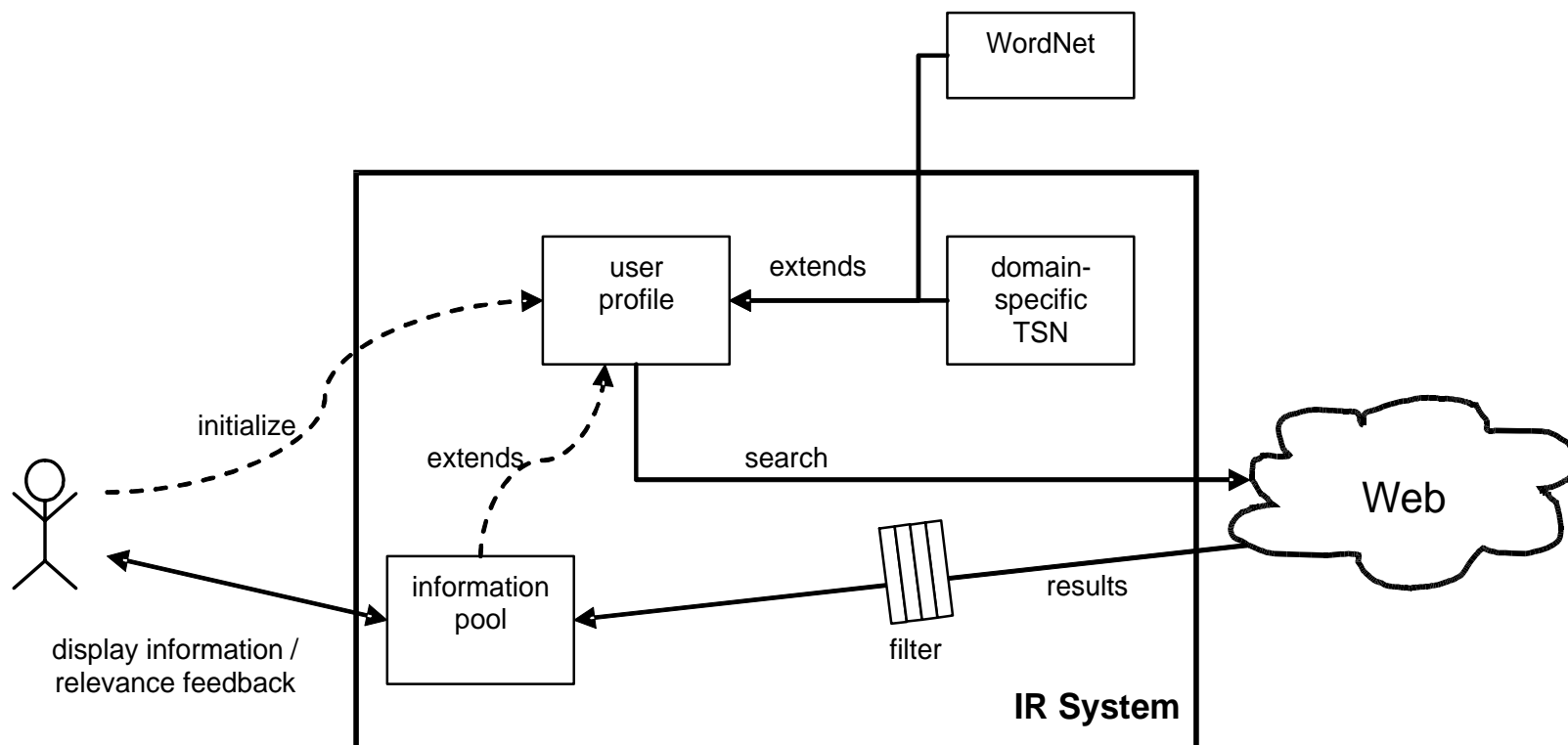
Related Work (1)

- R. Arezki, P. Poncelet, G. Dray, and D. W. Pearson. **Information retrieval model based on user profile**. In C. Bussler and D. Fensel, editors, AIMSA, volume 3192 of Lecture Notes in Computer Science, pages 490–499. Springer, 2004.
- Z. Gong, C. W. Cheang, and L. H. U. **Web query expansion by wordnet**. In K. V. Andersen, J. K. Debenham, and R. Wagner, editors, DEXA, volume 3588 of Lecture Notes in Computer Science, pages 166–175. Springer, 2005.
- L. Khan, D. McLeod, and E. H. Hovy. **Retrieval effectiveness of an ontology-based model for information selection**. VLDB J., 13(1):71–85, 2004.

Related Work (2)

- Y. Chiaramella. **Information retrieval and structured documents**. In M. Agosti, F. Crestani, and G. Pasi, editors, ESSIR, volume 1980 of Lecture Notes in Computer Science, pages 286–309. Springer, 2000.
- N. Fuhr and G. Weikum. **Classification and intelligent search on information in xml**. IEEE Data Eng. Bull., 25(1):51–58, 2002.
- G. Salton and C. Buckley. **Improving retrieval performance by relevance feedback**. JASIS, 41(4):288–297, 1990
- C. G. von Wangenheim, A. Bortolon, and A. Von Wangenheim. **A hybrid approach for the management of FAQ documents in latin languages**. In D. W. Aha and I. Watson, editors, ICCBR, volume 2080 of Lecture Notes in Computer Science, pages 204–218. Springer, 2001.

Adapted Architecture



Overview

Introduction

Information Retrieval

Basic Concept

Research News

Next Steps

What now?

- **Try to present out ideas on e.g.**
 - SIGIR, August 2006 in Seattle
 - 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval
 - SPIRE 2006, October 2006, Glasgow
 - 13th Symposium on String Processing and Information Retrieval.
- **Development of a prototype for our example application**
- **Measurement of our system**
- **How to create the domain-specific TSN?**
 - ISS-Presentation