

Diplomarbeitsvortrag

Colleague-2-Colleague Information Retrieval

Nutzung von Kompetenzeinschätzungen für die
Dokumentrecherche in P2P-Netzwerken

Stefan Weisenberger

Kaiserslautern, 28.09.2005



Gliederung des Vortrags

- Ziel der Diplomarbeit
- Vorüberlegungen
 - Nutzung der P2P-Technologie
 - Colleague-2-Colleague Szenario
- Konzeptionelle Herangehensweise
 - Konzeptsuche
 - Persönliche Sicht auf das Netzwerk
 - Kompetenz
 - Konfidenz
- Anforderungen für die Suche (Use Cases)
- Konzeptionelle Umsetzung der Algorithmen
- Demonstration



Ziel der Arbeit

- Entwurf eines Systems für den Austausch von Dokumenten in Netzwerken mit bekannter Topologie
 - Dezentrales Information Retrieval
 - P2P-Netzwerke mit bekannten Teilnehmern
 - Colleague-2-Colleague
 - Personalisierte Suche
 - Berücksichtigung von persönlichen Einschätzungen
 - Berücksichtigung von persönlichen Strukturen
- Implementierung des Systems



Nutzung eines P2P-Netzwerkes

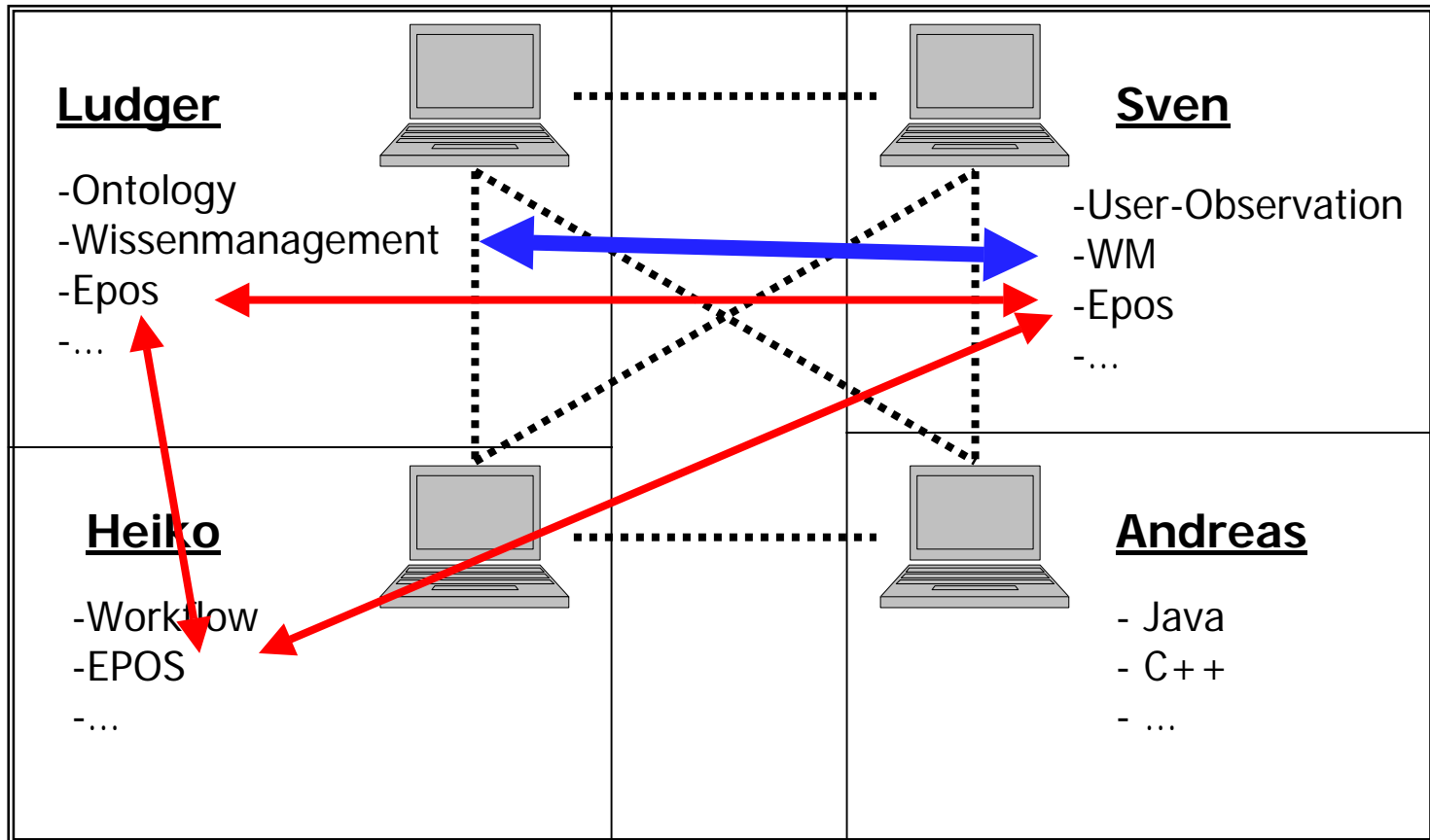
- Traditionelle P2P-Netzwerke
 - Redundanz
 - Fluktuation
 - Skalierbarkeit
 - Anonymität
- C2CIR-System
 - Natürliche Vernetzung zwischen Personen
 - Wissensaustausch von Person zu Person
 - Verbindungen nur zu Teilbereich der Personen
 - Keine zentrale Datenhaltung
 - Verbindungen variieren themenabhängig
 - Herkunft der Daten wichtiger Faktor



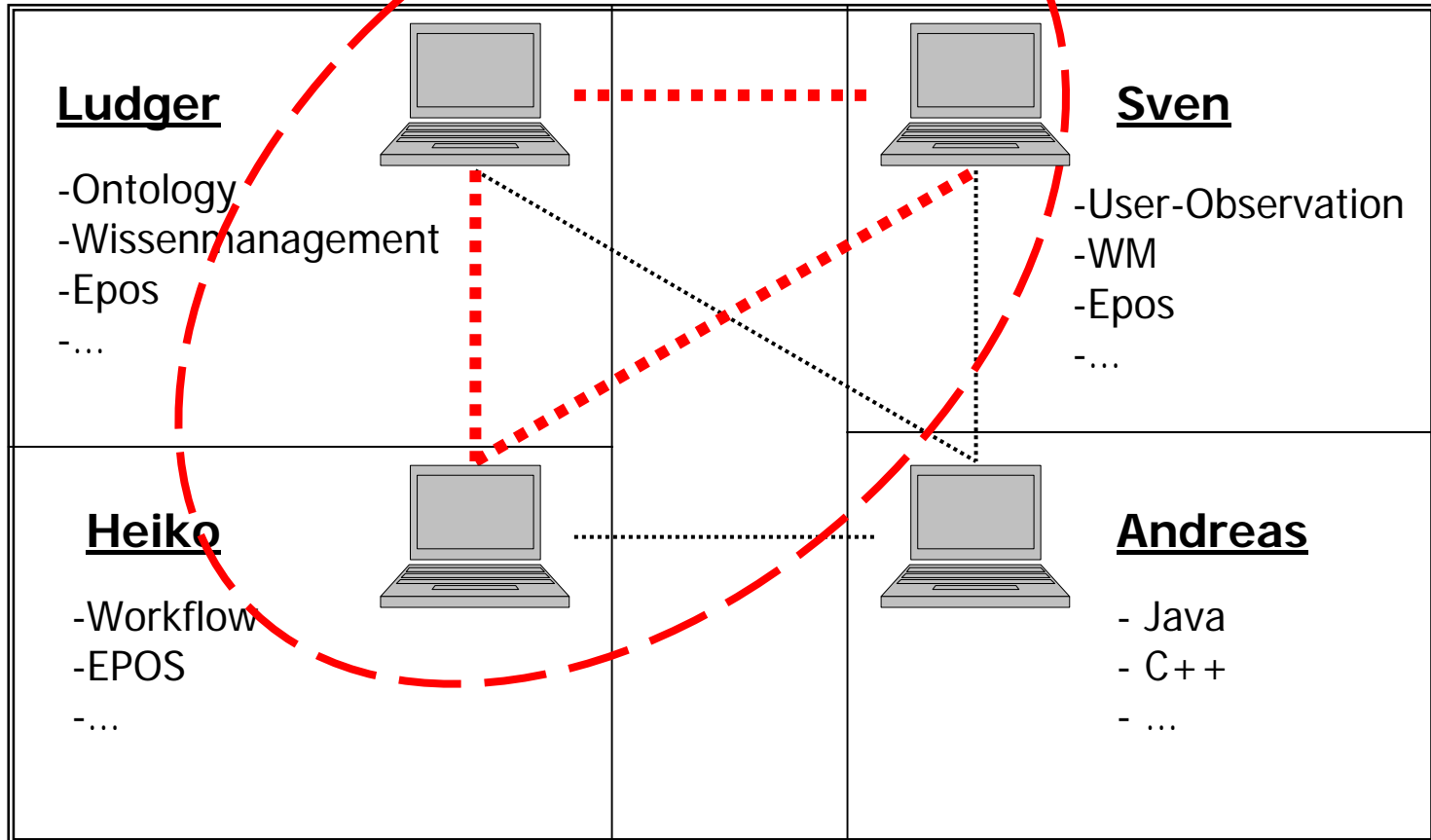
Colleague-2-Colleague Szenario

- Abteilung von Wissenschaftlern
 - Produzieren Wissen in Textform (schreiben, lesen, publizieren)
 - Ordnen ihre Dokumente selbständig
 - Arbeiten in verschiedenen Teams zusammen
 - Arbeiten an verschiedenen Projekten
- Feste Anzahl von Netzwerkteilnehmern
 - Keine Fluktuation
 - Bekannte/feste Topologie
- Rechner und Benutzer sind bekannt
 - Keine Anonymität
 - Nutzer kennen Kompetenzen der anderen Benutzer
 - Nutzer kennen Spezialisten im Kollegenkreis

Colleague-2-Colleague Szenario(2)



Colleague-2-Colleague Szenario(2)





Konzepte

- Konzepte = thematische Zusammenfassung von Dokumenten
 - Beziehungen zwischen Dokumente
- Konzepte eines Peers repräsentieren dessen Wissen
- Freie Wahl der Konzepte
 - Nutzer benennt Konzepte selbst
 - Nutzer wählt passende Dokumente selbst aus
 - Positiv: Konzepte auf Bedürfnisse des Nutzers angepasst
Nutzer kann Ähnlichkeit von Dokumenten beeinflussen
 - Negativ: Fremder Zugriff schwierig

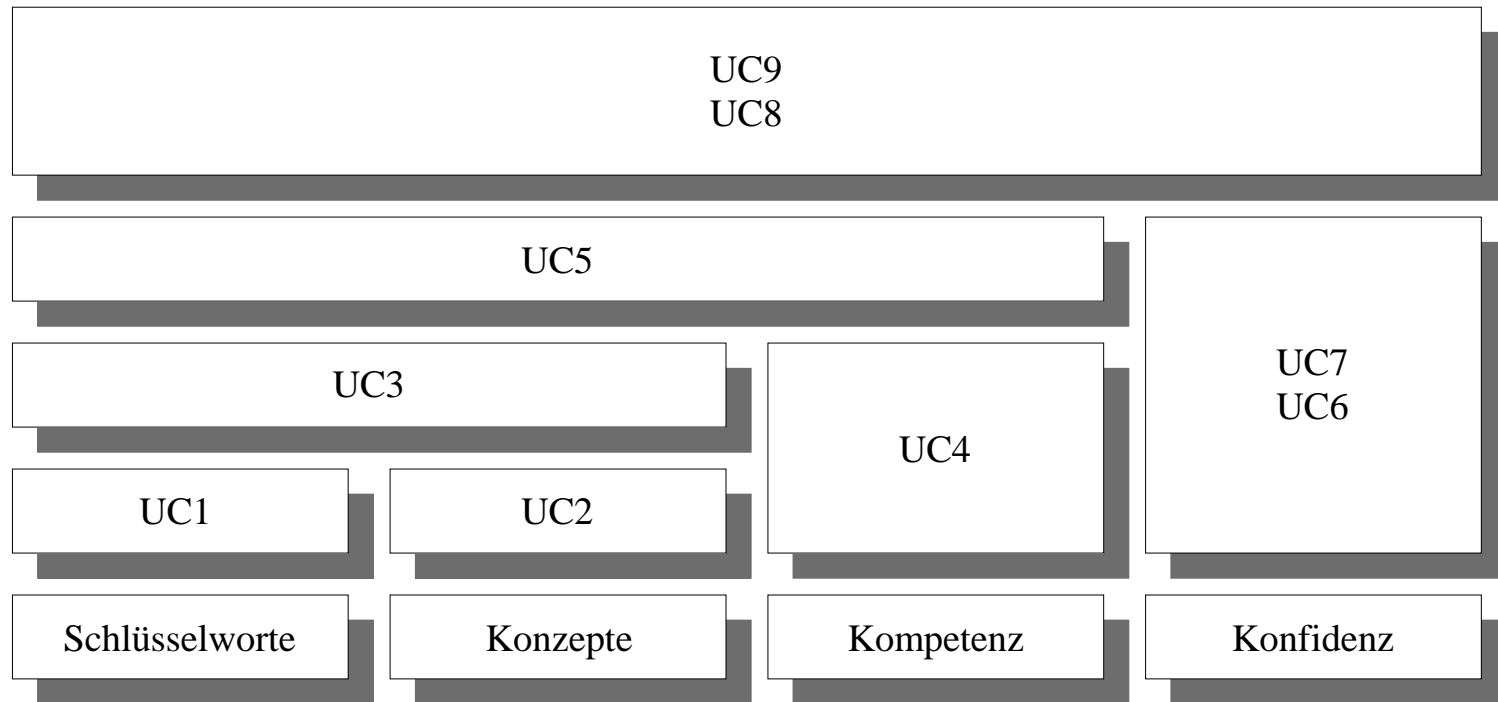


Persönliche Einschätzungen

- Relevanz eines Dokuments/Themas für eine Anfrage ist subjektiv
→ Persönliche Einschätzungen wichtig bei der Suche
- Unterscheidung nach Blickwinkel
 - Bewertung der eigenen Konzepte = Kompetenz
 - Bewertung von fremden Konzepten = Konfidenz

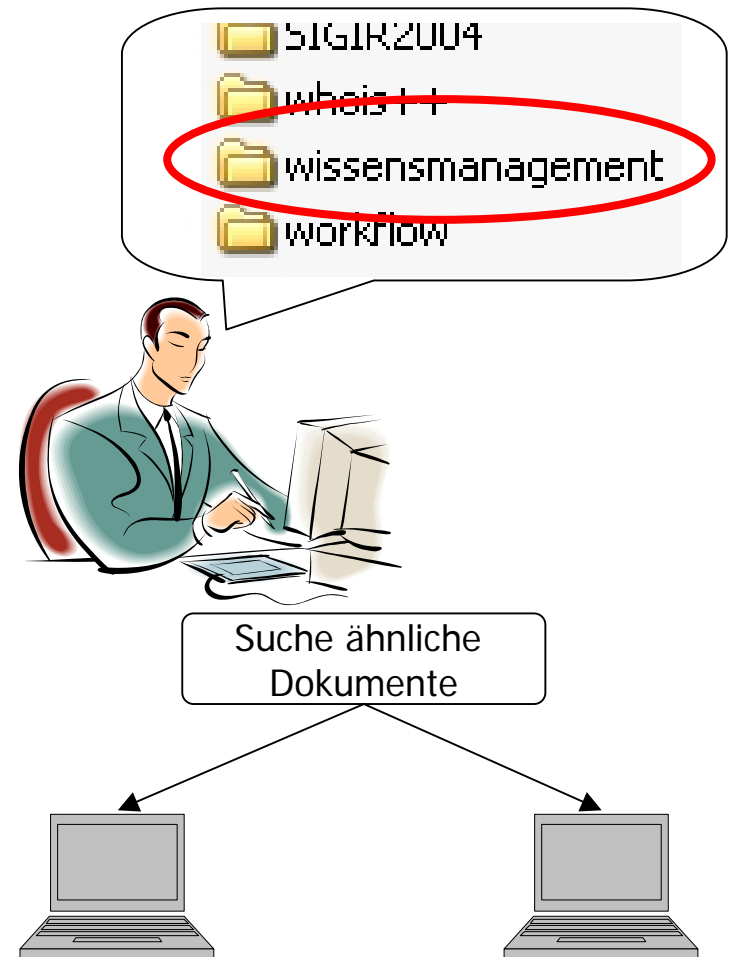
Anforderungen an das System

- Anforderungen werden durch neun Use Cases formuliert



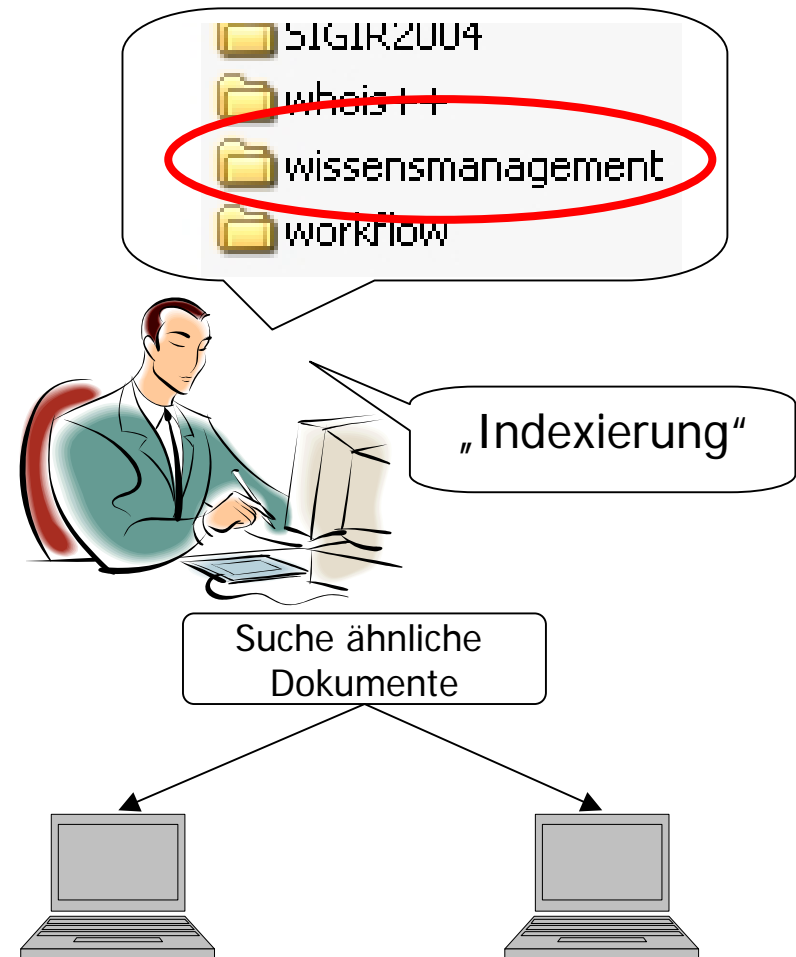
Suche mit Konzepten

- Konzept = thematische Zusammenfassung von Dokumenten
- Suche nach passenden Dokumenten zu Konzepten
- Vorteil:
 - Konzepte können Suchthema besser beschreiben als Schlüsselworte
 - Konzepte können Sicht des Nutzers darstellen
 - Beziehungen zwischen Konzepten



Suche mit Konzepten und Text

- Ergebnisse der Konzeptsuche werden durch Schlüsselworte eingeschränkt
 - Suche nach Konzepten
 - Einschränkung der Ergebnisse durch Schlüsselworte



Publikation eigener Kompetenzen

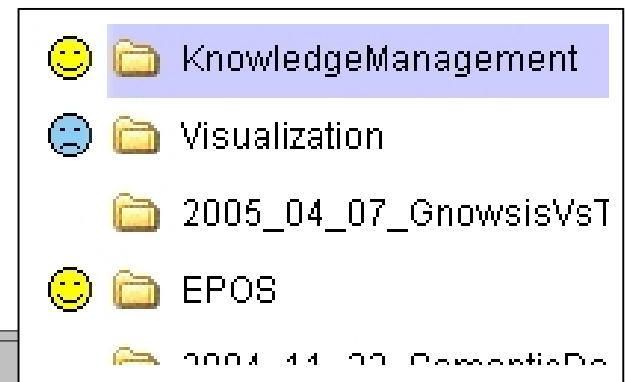
Suche mit Kompetenzen

- Einschätzung der eigenen Konzepte
 - **Kompetenz:** Themen zu denen auf dem Peer gute Dokumente zu finden sind
 - **Keine Kompetenz:** alle anderen Dokumente
- Andere Anwender können Kompetenzen eines Peers leichter erfassen und werden dadurch bei der Auswahl des richtigen Peers unterstützt
- Bei der Suche werden Dokumente aus Kompetenzkategorien bevorzugt



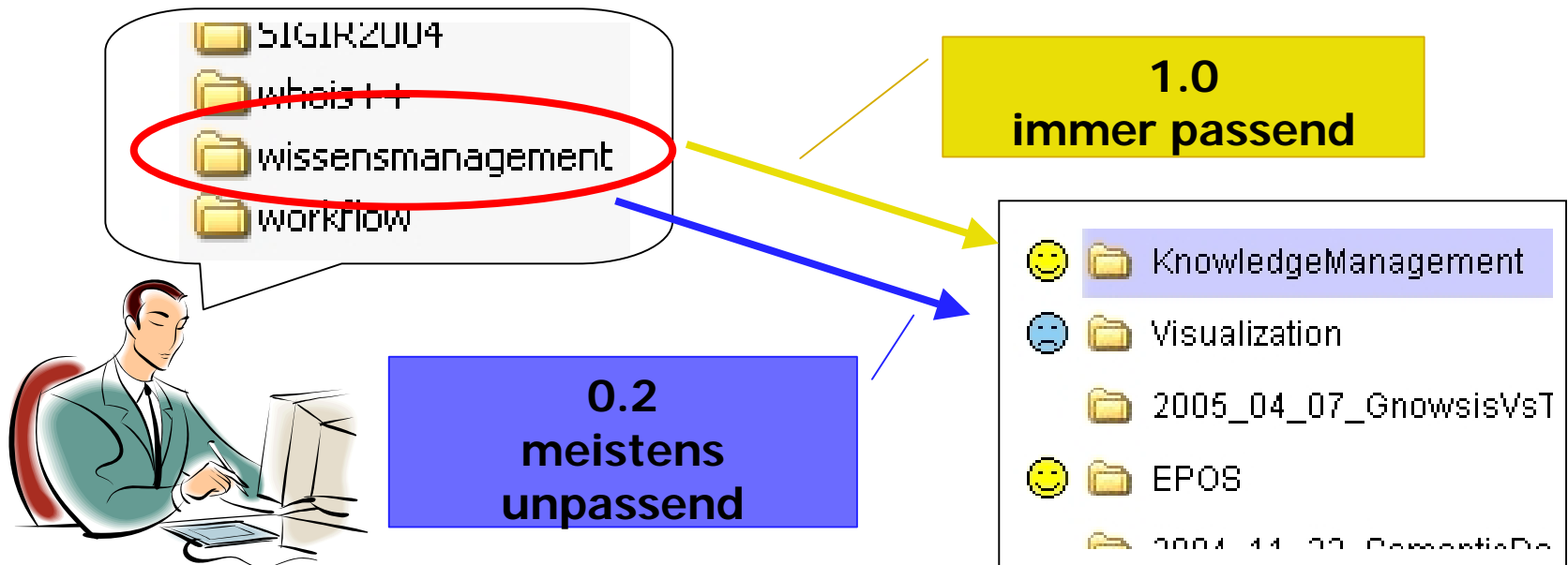
Anfrageunabhängige Konfidenz

- Bewertung von fremden Kompetenzen
 - Nach meinen Interessen
 - Nach meinem Wissen
- 2 Zustände
 - **Gut** - Für mich generell Interessant
 - **Schlecht** – Für mich generell nicht interessant



Anfrageabhängige Konfidenz

- Subjektive Bewertung wie gut zwei Konzepte zusammen passen
 - **Gut** – Konzept passt zu Anfrage
 - **Schlecht** – Konzept passt nicht zu Anfrage
- Anzahl guter und schlechter Bewertungen → Beziehungen zwischen Konzepten





Konzeptionelle Umsetzung

- System arbeitet mit Vektorraummodell
 - Dokumente werden durch Termvektor repräsentiert
 - Termvektor der Konzepte wird dynamisch aus den repräsentierten Dokumenten berechnet
- *Similarity* berechnet ähnliche Dokumente auf Grundlage der Termvektoren
 - Ausgangsfunktion (von System vorgegeben)
 - Textsuche
 - Konzeptsuche
- *Similarity* für mehrere Konzepte
 - Berechne Dokumente für jedes einzelne Konzept
 - Zusammenfügen der Ergebnisliste

Dokumentrelevanz

- Relevanz $rel(doc, query)$ eines Dokuments doc zu einer Anfrage $query$ von 3 Faktoren abhängig:
 - Similarity sim
 - Kompetenz $comp$
 - Konfidenz
 - Anfrageunabhängig $gconf$
 - Anfrageabhängig $qconf$
- Berechnung des Dokumentrangs

$$rel(doc, query) = \frac{\alpha \cdot sim(doc, query) + \beta \cdot comp(doc) + \gamma \cdot gconf(doc) + \delta \cdot qconf(doc, query)}{\alpha + \beta + \gamma + \delta}$$

Similarity

Competency

gConfidence

qConfidence



Kompetenz

- Dokumente gehört zur Kompetenz eines Peers, wenn es zu einer Kompetenzkategorie gehört
 - $\text{comp}(\text{doc}) = 1$, wenn Dokument in einer Kompetenzkategorie
 - $\text{comp}(\text{doc}) = 0$, sonst



Berechnung der Konfidenz

- Anfrageunabhängige Konfidenz kennt 3 Zustände:
 - Gut → Wert 1
 - Schlecht → Wert 0
 - Neutral → Wert 0.5
- Dokumentkonfidenz abhängig von den Konzepten
- Mittelwert der Konfidenzen, wenn Dokument in mehreren Konzepten ist.



Berechnung der Konfidenz(2)

- Berechnung der anfrageabhängige Konfidenz aus den Konzeptbeziehungen $qconf(\text{concept1}, \text{concept2})$

$$qconf = \frac{\text{positive Bewertungen}}{\text{pos. Bewertungen} + \text{neg. Bewertungen}}$$

- Berechnung der anfrageabhängige Konfidenz für ein Dokument
 - Mittelwert der Beziehungen aller Anfragekonzepte mit den Dokumentkonzepten

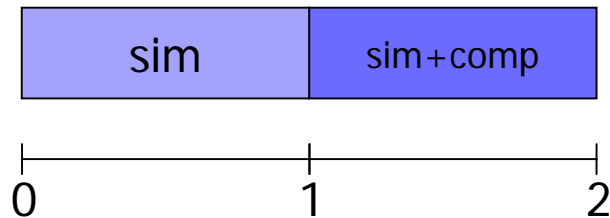
Bsp.: Anfrage $q(c1, c2)$; Dokumentekategorien $(c3, c4)$

$$qconf(\text{doc}, \text{query}) = \frac{qconf(c1, c3) + qconf(c1, c4) + qconf(c2, c3) + qconf(c2, c4)}{4}$$

Anfrageberechnung

- Ähnlichkeit und Kompetenz
 - Wertebereich der Funktionen
 - similarity: [0;1]
 - comp= 0 oder 1
 - Bsp.: Gewichtungsfaktoren auf 1

$$rel(doc, query) = \frac{\alpha \cdot sim(doc, query) + \beta \cdot comp(doc)}{\alpha + \beta}$$



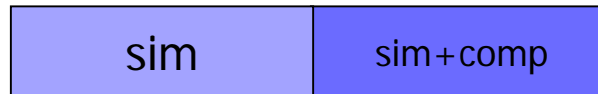
- 0-1 → keine Kompetenz
- 1-2 → Kompetenz

Anfrageberechnung(3)

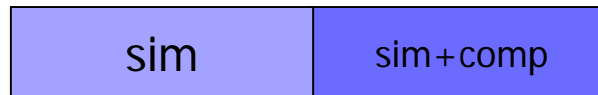
- **Generelle Konfidenz gconf**
 - Gut -> 1
 - Neutral -> 0.5
 - Schlecht -> 0
 - Bsp. Gewicht 2

$$rel(doc, query) = \frac{\alpha \cdot sim(doc, query) + \beta \cdot comp(doc) + \gamma \cdot gconf(doc)}{\alpha + \beta + \gamma}$$

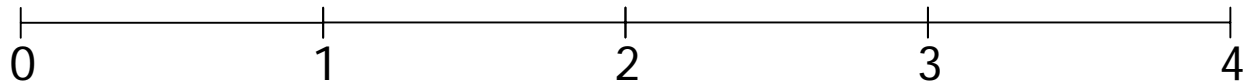
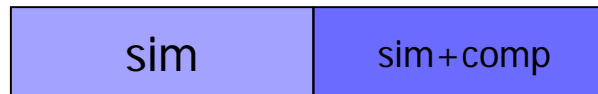
gut



neutral



schlecht



Diplomarbeit von Stefan
Weisenberger

Anfrageberechnung(4)

Anfrageabhängige Konfidenz q_{conf}

$$0 \leq q_{conf} \leq 1$$

$q_{conf} = 0$ = schlecht

$$q_{conf} = 0.5$$

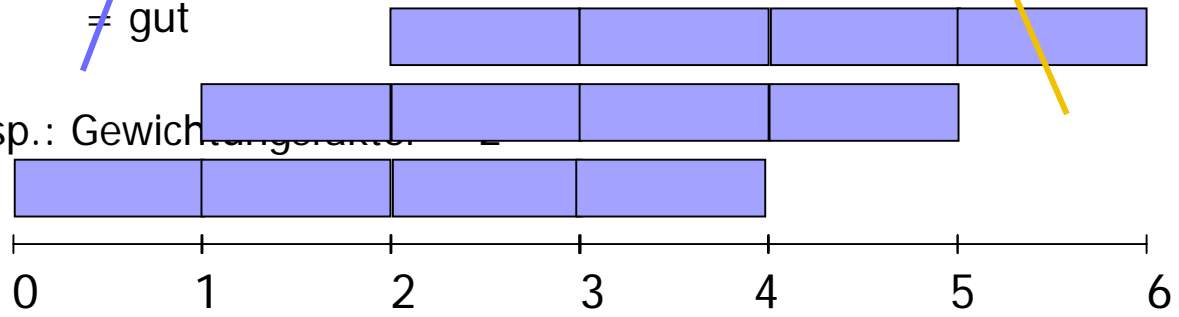
Nur negative
Bewertungen

Nur positive
Bewertungen

$$q_{conf} = 1$$

sp.: Gewichtungsfaktor = 1

gut
neutral
schlecht



C2CIR-GUI -Browser

The screenshot shows the C2CIR weissenbe browser interface. A red box highlights the 'Categories' and 'Documents' sections. On the left, four red labels with arrows point to specific items in the 'Categories' list: 'Peerauswahl' points to 'Epos', 'Konzepte' points to 'Konzeptsuche', 'Dokumente' points to 'allgemein', and 'Kompetenzen' points to 'COMPETENCY'. The 'Documents' list shows several files with globe icons, including 'vision_slides_get', 'EPOS_M1.pdf', 'T23VBLUA738D1', and 'EPOS_GuidingEx'. The interface also includes a search bar, an 'Execute Feedback' button, and checkboxes for 'Competency' and 'Confidence'.

Peerauswahl

Konzepte

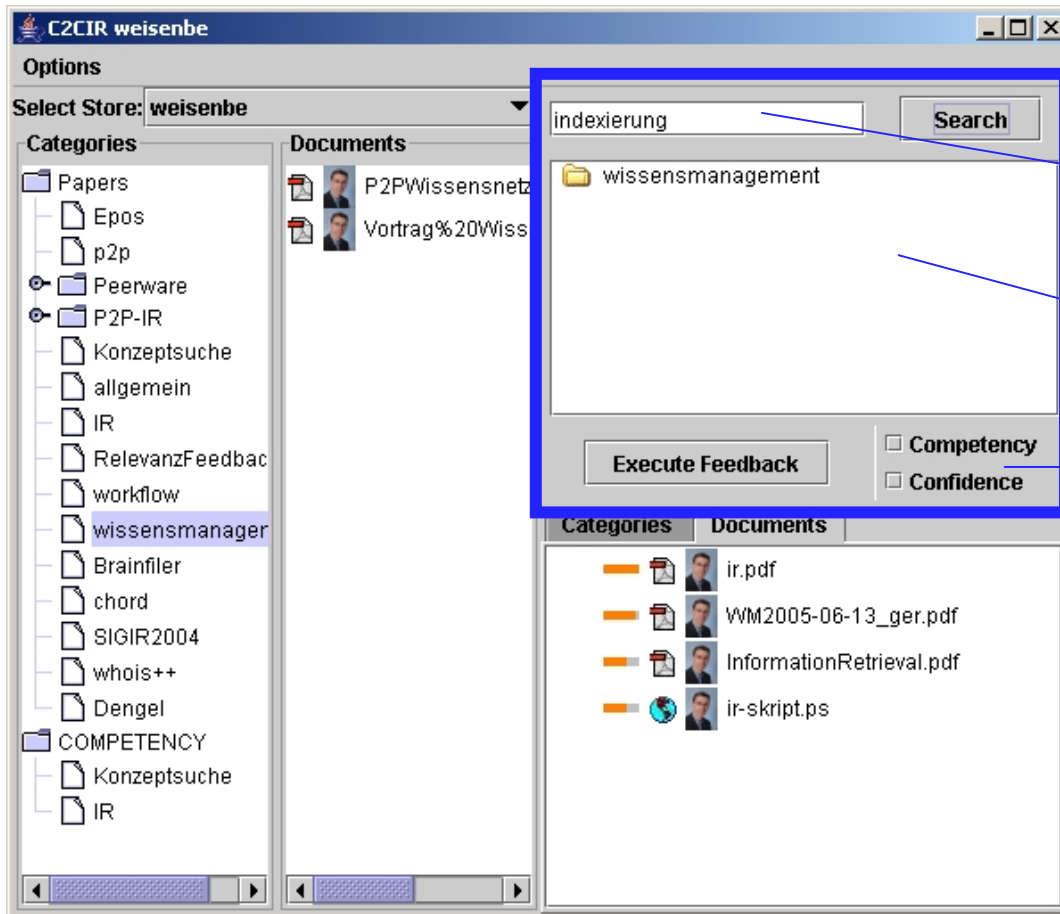
Dokumente

Kompetenzen

Diplomarbeit von Stefan
Weissenberger

28.09.2005

C2CIR-GUI -- Anfragebereich



Schlüsselworte

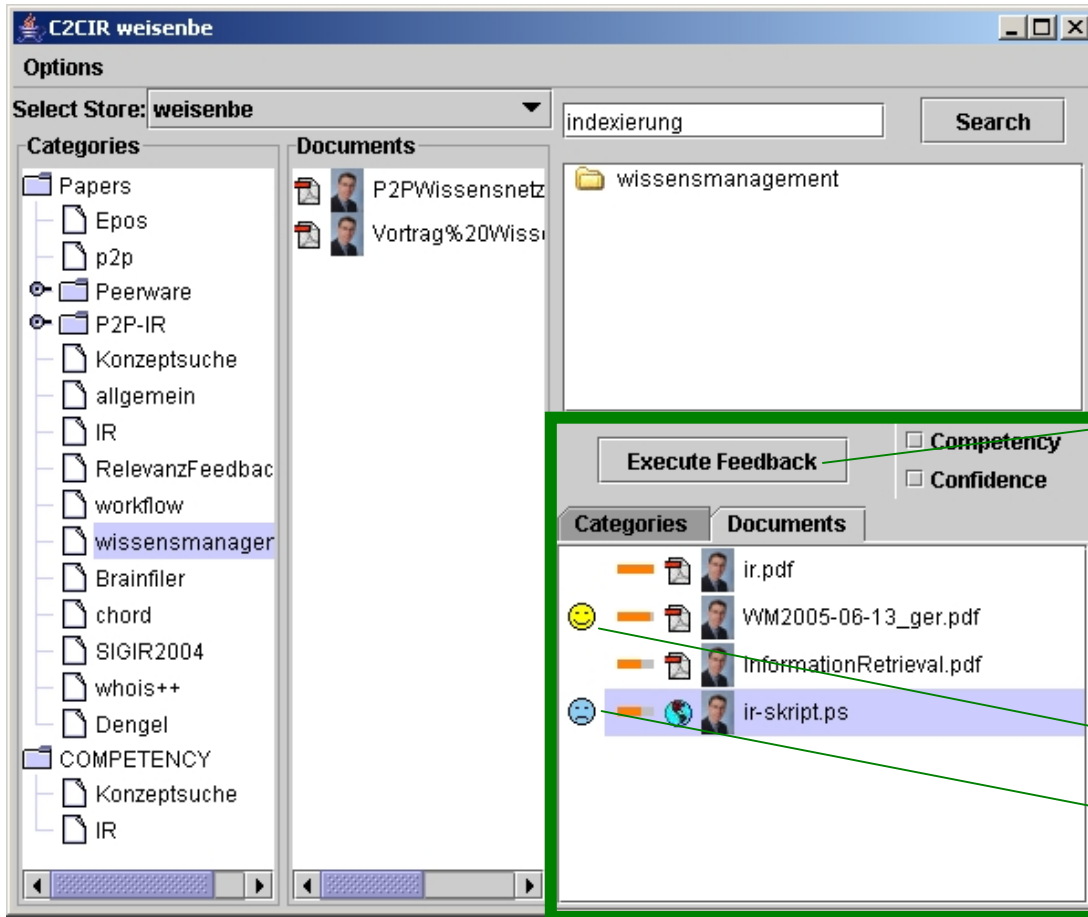
Konzepte

Suchmodus

Diplomarbeit von Stefan
Weissenberger

28.09.2005

C2CIR-GUI -- Ergebnisdarstellung



Feedback anwenden

Positives Feedback

Negatives Feedback

Diplomarbeit von Stefan
Weisenberger

28.09.2005