

Information Retrieval [IR 4]

Übungen und Wiederholungsfragen zur Prüfungsvorbereitung

Winfried Gödert / Klaus Lepsky

5. Januar 2015

Institut für Informationswissenschaft
Fachhochschule Köln
Claudiusstraße 1, 50678 Köln
klaus@lepsy.de

1 Suchen in Suchmaschinen

1. Geben Sie eine Erklärung für die Ergebnisse der folgenden Suchanfragen mit Google:
 - information retrieval
 - retrieval information
 - „information retrieval“
 - „retrieval information“
 - information und retrieval
 - information and retrieval
 - „information und retrieval“
 - „information and retrieval“
2. Welchen Rückschluss auf die Verwendung von Stoppwörtern bei *Google* lassen die Ergebnisse auf die folgenden Suchanfragen zu? Testen Sie Ihre Schlussfolgerungen in anderen Suchumgebungen!
 - und
 - and
3. Macht es für eine Suche nach „und“ einen Unterschied, ob die Suche über ein sequenzielles Pattern Matching oder über einen Suchindex durchgeführt wird?
4. Welche Beziehung zwischen *Albert Einstein* und *S. B. Preuss* lässt sich angeben?
5. Welche Eigenschaft soll durch die informatrische Kenngröße *Hirsch-Index* zum Ausdruck gebracht werden?

2 Index, Invertierte Liste

[Vgl. Abschnitt 5.1.1 - 5.1.3 in: Gödert, W., K. Lepsky, and M. Nagelschmidt (2012). Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. X.media.press. Berlin u.a.: Springer.]

1. Erstellen Sie für den nachfolgenden Text eine wortinvertierte Liste:

„Ein Information-Retrieval-System (kurz: IR-System) verwendet für die schnelle Suche einen sog. Index, häufig auch invertierte Liste genannt.“

Erläutern Sie Ihre Kriterien dafür, was ein „Wort“ ist.

Skizzieren Sie mögliche Probleme einer Wortinvertierung.

2. Erstellen Sie für die nachstehenden Einzeldokumente jeweils eine invertierte Liste nach der Methode der Wortinvertierung und nach der Methode der Phraseninvertierung.

Dok.-Nr.	Dokumenttitel
056	Grundlagen der praktischen Information und Dokumentation
197	CD-ROM Netze in Bibliotheken
735	Leben und Werk von Hanns W. Eppelsheimer
811	Grundlagen des Bibliothekswesens

Benutzen Sie die Wörter Ihrer wortinvertierten Liste sowie die Boole'schen Operatoren zur Formulierung einer Fragestellung, die nur die drei Dokumente *056*, *735* und *811* als Treffer findet.

Welche Dokumente werden bei Phraseninvertierung und der Eingabe von

Grundlagen des*

gefunden?

3. Welche Methode — Wortinvertierung oder Phraseninvertierung — ist jeweils geeigneter für die Herstellung einer invertierten Liste für die Kategorien:

1. Verfasser
2. Titel
3. Abstract

Geben Sie jeweils eine Begründung!

4. Stellen Sie Vor- und Nachteile einer sequenziellen Suche der Suche in einer invertierten Liste gegenüber.
5. Ist es zweckmäßig, sog. „Stoppwörter“ aus dem Index auszuschließen? Begründen Sie die Antwort.
6. Ist die Sprache eines Dokuments für den Indexaufbau von Belang?
7. Was versteht man unter einer inkrementellen Suche im Rahmen einer Index-Suche?
8. Erläutern Sie die Erstellung einer Invertierten Liste für die Kategorien einer Dokumentrepräsentation unter Benutzung der Dokumentnummer und den Inhalten der Kategorien. Erstellen Sie ein Beispiel aus 3 Dokumenten mit je mindestens 8 Kategorien.

9. Geben Sie unter Verwendung der Konzepte *Sequenzielle Suche* bzw. *Suche in Binärbäumen* eine textliche Beschreibung des maschinellen Suchvorgangs in Invertierten Listen.

3 Boole'sche Operatoren



1. Welche Ergebnisse erzielen jeweils die folgenden Suchanfragen:

- Eins AND Drei
- Eins AND Drei AND Fünf
- Zwei OR Vier
- Eins OR Zwei OR Vier
- Zwei NOT Drei
- Zwei AND (Drei OR Vier)
- (Eins NOT Zwei) OR (Drei AND Vier)
- (Eins OR Drei OR Fünf) NOT (Zwei AND Vier)
- ((Zwei OR Drei) NOT (Vier AND Fünf)) OR (Zwei NOT Vier)
- Eins AND Zwei AND Drei AND Vier AND Fünf
- Eins OR Zwei OR Drei OR Vier OR Fünf

2. Welche Suchen führen zu diesen Ergebnissen?

- A E S
- 3 7 ß
- Ö Z 3 7
- S
- A H Z 3 7

4 Recall und Precision

[Vgl. Abschnitt 6.2 in: Gödert, W., K. Lepsky, and M. Nagelschmidt (2012). Informationser-schließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. X.media.press. Berlin u.a.: Springer.]

1. Ein Information Retrieval System findet bei einer Suche 20 relevante Dokumente. Die Treffermenge enthält insgesamt 60 Dokumente. In der Dokumentkollektion befinden sich für diese Suche insgesamt 40 relevante Dokumente. Berechnen Sie *Recall* und *Precision* für die Suche.
2. In einem Information Retrieval System führen Sie eine Suche mit einem Term zunächst als Volltextsuche aus, anschließend -- da der Term auch als Deskriptor für die Erschließung verwendet wurde -- suchen Sie mit diesem Term als Deskriptor. Welche Unterschiede in Bezug auf *Recall* und *Precision* erwarten Sie bei einem Vergleich der Suchergebnisse?
3. Wie wirkt sich der Einsatz einer Rechtstrunkierung auf *Recall* und *Precision* aus?
4. Wie wirkt sich der Einsatz eines Stemming-Verfahrens auf *Recall* und *Precision* aus? Ist es dabei von Bedeutung, ob vom Stemmer Grundformen oder Wortstämme produziert werden?
5. Welche Auswirkungen auf *Recall* und *Precision* erwarten Sie, wenn für eine Dokumentkollektion eine linguistische automatische Indexierung durchgeführt wird, bei der Komposita zerlegt werden? Unterscheiden Sie dabei zwischen einer Suche nach einem Kompositum und der Suche nach einem Teilwort.
6. Lässt sich eine sehr schlechte *Precision* durch ein *Relevance Ranking* kompensieren? Begründen Sie Ihre Antwort.

5 Statistik im Information Retrieval

[Vgl. Abschnitt 5.1.5 u. 5.3.5 in: Gödert, W., K. Lepsky, and M. Nagelschmidt (2012). Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. X.media.press. Berlin u.a.: Springer.]

1. Erläutern Sie die Begriffe TF , WDF und IDF und geben Sie Beispiele für den Nutzen von Termgewichtungsverfahren.
2. Wie entsteht eine nach TF sortierte Trefferliste?
3. Erhalten Hochfrequenzterme bei der Anwendung des IDF ein besonders hohes oder ein besonders niedriges Termgewicht? Warum?
4. Sind Hochfrequenzterme für den Einsatz einer Termgewichtung problematisch?
5. Sollte man bei der Anwendung eines Stemming-Verfahrens und einer Termgewichtung zunächst die Termgewichtung durchführen und danach das Stemming-Verfahren, oder eher umgekehrt? Begründen Sie die jeweiligen Vor- und Nachteile.

5.1 Zipf und Kant — Übung zur Termhäufigkeit



Zur Ermittlung von Termhäufigkeiten und für Übungen zu *Zipf's Law* wird der Volltext von Immanuel Kants „Kritik der reinen Vernunft“ verwendet.

Die Archiv-Datei <http://ixtrieve.fh-koeln.de/lehre/kant-kdrv.zip> zur Übung enthält folgende Dateien:

- kant-kdrv.txt – der Volltext von Kants „Kritik der reinen Vernunft“
 - ling-zipf-kdrv.xls – die Auswertungen einer Zählung der reinen Zeichenketten und der Grundformen
 - kant-kdrv.abs – die Ergebnisse für eine *Lingo*-Indexierung mit Häufigkeit der Token
 - kant-kdrv.ven – die Ergebnisse für eine *Lingo*-Indexierung mit Häufigkeit der Grundformen
 - zipf.cfg – die *Lingo*-Konfiguration zur Indexierung von *token*
 - vector_filter.rb – ein modifizierter *Lingo*-Vectorfilter, der die Ausgabe von *token* erlaubt
1. Erstellen Sie aus der Datei „kant-kdrv.txt“ eine nach Häufigkeiten sortierte Liste der enthaltenen Zeichenketten. Verwenden Sie notepad++, um eine Liste der Einzelwörter zu erstellen und *Excel* (Pivot-Tabellen), um die Worthäufigkeit zu berechnen. Analysieren Sie das Ergebnis.
 2. Überprüfen Sie, ob in der Liste nur Einträge enthalten sind, die Ihrer Vorstellung von einem „Wort“ entsprechen. Wiederholen Sie ggf. Ihre Ersetzung, bis Sie mit dem Ergebnis zufrieden sind. Halten Sie die Gesamtmenge der zu ersetzenden Zeichen in Form eines regulären Ausdrucks fest.
 3. Überprüfen Sie die Liste auf das Vorkommen von Wortvarianten und auf die Gesamtanzahl der enthaltenen Einträge. Erstellen Sie mit *Lingo* eine nach Häufigkeiten sortierte Liste der Zeichenketten (*token*). Benutzen Sie für diesen Indexierungslauf die Konfiguration „zipf.cfg“ und ersetzen Sie die Originaldatei „vector_filter.rb“ durch die modifizierte Datei im Archiv. Speichern Sie die Ergebnisse für eine spätere Verwendung in geeignet benannte Dateien ab.
 4. Erstellen Sie mit *Lingo* eine nach Häufigkeiten sortierte Liste der Grundformen. Verwenden Sie dafür die Standardkonfiguration „lingo.cfg“. Speichern Sie die Ergebnisse für eine spätere Verwendung in geeignet benannten Dateien ab.
 5. Führen Sie für die in 3. und 4. erstellten Ergebnisse eine vergleichende Analyse durch:

- Übereinstimmung bzw. Unterschiede der *Excel*-Pivot-Liste mit der durch *Lingo* ermittelten Liste der *token*. Erklären Sie ggf. vorhandene Unterschiede.
- Vergleich der Anzahl der Einträge in der durch *Lingo* ermittelten Liste der *token* und der durch *Lingo* ermittelten Liste der Häufigkeit der Grundformen. Erklären Sie ggf. vorhandene Unterschiede.
- Vergleich der Häufigkeit der Einträge für ausgewählte Wörter in der Liste der *token* und der Liste der Grundformen. Muss die Häufigkeit in einer Liste immer größer (oder kleiner) sein als in der anderen?
- Welcher Effekt für die Häufigkeiten ergibt sich durch Zu- oder Abschalten einer Zerlegung von Komposita.

5.2 Übung zur Termgewichtung

Gegeben seien 5 Dokumente aus einer Kollektion von insg. 10.000 Dokumenten und eine Tabelle der Dokumenthäufigkeiten (Tab. 1).

- (a) Anwendung des Prinzips Thesaurus für das Information Retrieval
- (b) Zusammenhang zwischen Thesaurus und Klassifikation
- (c) Klassifikation und OPAC: verbesserter Sucherfolg durch Einsatz einer Klassifikation im Retrieval
- (d) Thesaurus für die Physik und Thesaurus für die physikalische Chemie
- (e) Klassifikation für die Chemie

Tabelle 1: Dokumenthäufigkeiten für Substantive

Anwendung	2000
Chemie	200
Einsatz	100
Klassifikation	100
OPAC	600
Physik	300
Prinzip	1500
Retrieval	400
Sucherfolg	50
Thesaurus	200
Zusammenhang	3000

1. Berechnen Sie für alle Substantive in den Dokumenten die relative Termhäufigkeit WDF und die inverse Dokumenthäufigkeit $TF*IDF$.
2. Berechnen Sie die Retrievalergebnisse für folgende Suchanfragen:

Thesaurus im Retrieval
Klassifikation in der Chemie

6 Vektorraummodell

(I) Gegeben seien die folgenden Dokumente mit Titeln und Deskriptoren:

- (1) TI: Entwicklung eines akustischen Feedbacksystems als ruderspezifisches Trainingsgerät
DE: Akustische Lernhilfe; Bewegungsablauf; Bewegungslehre; Entwicklung; Feedback; Kinematik; Rudern; Sportinformatik; Sportliches Training; Sporttechnologie; Sportwissenschaft; Trainingsgerät
 - (2) TI: Entwicklung eines Mess- und Analysesystems zur Optimierung der Bootsbewegung im Wassertraining und Ruderrennen
DE: Belastungsintensität; Belastungsumfang; Bewegungsanalyse; Leistungsdiagnostik; Messgerät; Rudern; Rudersport; Sportinformatik; Sportliche Technik; Sportliches Training; Sporttechnologie; Sportwissenschaft; Trainingssteuerung; Trainingswissenschaft
 - (3) TI: Aktuelle Probleme der Renngestaltung mit trainingsmethodischen Schlussfolgerungen für die olympische Regatta im Rudern
DE: Olympische Spiele; Rudern; Rudersport; Sportliches Training; Trainingsmethode; Trainingswissenschaft; Wassersport; Wettkampfgestaltung; Wettkampfvorbereitung
 - (4) TI: Übung macht den Meister : gilt das auch für das Rudern?
DE: Leistungssteigerung; Rudern; Sportliches Training; Superkompensation; Trainingslehre; Übertraining
 - (5) TI: Techniktraining nach Gehör : Steigerung der Bootsgeschwindigkeit durch akustische Wahrnehmung
DE: Akustik; Auditive Wahrnehmung; Beschleunigung; Bewegung; Bewegungslehre; Boot; Feedback; Mentales Training; Ruderboot; Rudern; Rudersport; Techniktraining; Trainingsmittel; Visuelle Wahrnehmung; Wassertraining
 - (6) TI: Trainingsbegleitende Leistungsdiagnostik im Weitsprung
DE: Bewegungsanalyse; Bewegungslehre; Biomechanik; Biomechanische Analyse; Leistungsdiagnostik; Leistungsmessung; Leistungssport; Messplatz; Messverfahren; Sportartspezifische Leistungsfähigkeit; Sportliches Training; Techniktraining; Trainingswissenschaft; Weitsprung
1. Stellen Sie einen Dokument-/Deskriptorvektorraum auf und berechnen Sie die Ähnlichkeit der Dokumente durch das vereinfachte Skalarprodukt:

$$\dot{A}(\mathbf{d}_i, \mathbf{d}_j) = \frac{1}{n} \sum_{k=1}^n d_{ik} d_{jk}$$

2. Erstellen Sie für die Suchanfrage

Belastungsintensität; Belastungsumfang; Bewegungsanalyse;
Rudern; Sportliches Training; Techniktraining; Trainingssteuerung;
Wettkampfvorbereitung

den Anfragevektor und berechnen Sie das ähnlichste Dokument auf der Basis des vorstehenden vereinfachten Skalarprodukts.

(II) Gegeben seien die folgenden Dokumente mit Titeln und gewichteten Deskriptoren:

- (1) TI: Entwicklung eines akustischen Feedbacksystems als ruderspezifisches Trainingsgerät
DE: Akustische Lernhilfe (5); Bewegungsablauf; Bewegungslehre (3); Entwicklung; Feedback; Kinematik; Rudern (3); Sportinformatik; Sportliches Training; Sporttechnologie; Sportwissenschaft; Trainingsgerät (4)
 - (2) TI: Entwicklung eines Mess- und Analysesystems zur Optimierung der Bootsbewegung im Wassertraining und Ruderrennen
DE: Belastungsintensität; Belastungsumfang; Bewegungsanalyse; Leistungsdiagnostik (3); Messgerät; Rudern; Rudersport (4); Sportinformatik (2); Sportliche Technik; Sportliches Training; Sporttechnologie; Sportwissenschaft (2); Trainingssteuerung; Trainingswissenschaft (4)
 - (3) TI: Aktuelle Probleme der Renngestaltung mit trainingsmethodischen Schlussfolgerungen für die olympische Regatta im Rudern
DE: Olympische Spiele (3); Rudern (5); Rudersport; Sportliches Training (2); Trainingsmethode (2); Trainingswissenschaft (4); Wassersport; Wettkampfgestaltung (2); Wettkampfvorbereitung
 - (4) TI: Übung macht den Meister : gilt das auch für das Rudern?
DE: Leistungssteigerung (3); Rudern (4); Sportliches Training; Superkompensation; Trainingslehre (2); Übertraining
 - (5) TI: Techniktraining nach Gehör : Steigerung der Bootsgeschwindigkeit durch akustische Wahrnehmung
DE: Akustik (4); Auditive Wahrnehmung (4); Beschleunigung; Bewegung; Bewegungslehre; Boot; Feedback (2); Mentales Training (3); Ruderboot (3); Rudern; Rudersport (2); Techniktraining; Trainingsmittel (4); Visuelle Wahrnehmung; Wassertraining
 - (6) TI: Trainingsbegleitende Leistungsdiagnostik im Weitsprung
DE: Bewegungsanalyse (2); Bewegungslehre; Biomechanik (3); Biomechanische Analyse; Leistungsdiagnostik (5); Leistungsmessung; Leistungssport; Messplatz; Messverfahren; Sportartspezifische Leistungsfähigkeit (2); Sportliches Training; Techniktraining (2); Trainingswissenschaft (2); Weitsprung (5)
1. Stellen Sie einen Dokument-/Deskriptorvektorraum auf und berechnen Sie die Ähnlichkeit der Dokumente durch das vereinfachte Skalarprodukt:

$$\ddot{A}(\mathbf{Ad}_i) = \frac{1}{M+n} \sum_{k=1}^n a_k d_{jk}, M = \max_{1 \leq i \leq m} \left(\sum_{j=1}^n d_{ij}, A \right), i = 1, \dots, m$$

2. Erstellen Sie für die Suchanfrage

Belastungsintensität (4); Belastungsumfang; Bewegungsanalyse (5);
Rudern (3); Sportliches Training; Techniktraining (2);
Trainingssteuerung; Wettkampfvorbereitung (2)

den Anfragevektor und berechnen Sie das ähnlichste Dokument auf der Basis des vorstehenden vereinfachten Skalarprodukts.

7 PageRank

1. Berechnen Sie den PageRank für die Knoten der nachstehend abgebildeten Netzstrukturen als Vektor, indem Sie eine Iteration in 5 Schritten für die Matrix H der Übergangswahrscheinlichkeiten durchführen, wie es in der Vorlesung dargestellt wurde:

$$r_{i+1} = r_i H, \quad i=0, \dots, 4;$$

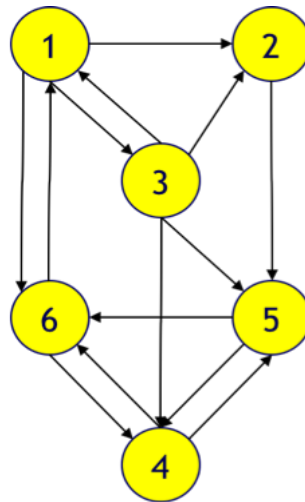
$H = H_{kl}$; H_{kl} steht für die Übergangswahrscheinlichkeit des Knoten P_k zum Knoten P_l ;

k, l stehen für die Zahl der Zeilen und Spalten der Matrix H.

Beginnen Sie die Iteration mit einem Startvektor, der als Startwert für jeden der n Knoten $1/n$ enthält.

Benutzen Sie für die Multiplikation eines Vektors mit einer Matrix die Rechenregel (für andere Zeilen- und Spaltenzahlen entsprechend anpassen):

$$\begin{array}{c} (a_1, a_2, a_3) * \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \\ \hline = (a_1 b_{11} + a_2 b_{21} + a_3 b_{31}, a_1 b_{12} + a_2 b_{22} + a_3 b_{32}, a_1 b_{13} + a_2 b_{23} + a_3 b_{33}) \end{array}$$



Vergleichen Sie das Ergebnis mit dem Rechner unter:

http://www.webworkshop.net/pagerank_calculator.php

2. Benutzen Sie einen PageRank-Rechner im Web, z. B.

http://www.prchecker.info/check_page_rank.php

<http://checkpagerank.net/>

<http://may.be/pagerank>

um die PageRank-Werte für die Startseite einer Web-Site mit den Werten untergeordneter Seiten zu vergleichen.

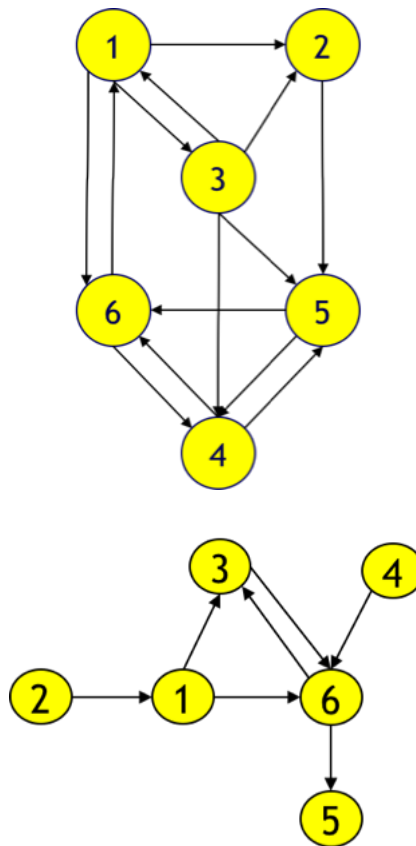
- Bestimmen Sie für die nachfolgend abgebildeten Netze unter Benutzung der Rechner auf den Seiten

http://www.webworkshop.net/pagerank_calculator.php

<http://www.cs.sjsu.edu/faculty/pollett/masters/Semesters/Spring10/amith/HITS.html>

die jeweiligen Werte für den PageRank bzw. die Werte für die Autoritätsknoten (authorities) sowie für die Netzknoten (hubs). Benutzen Sie jeweils unterschiedliche Anzahlen von Iterationsschritten.

Vergleichen Sie die erzielten Ergebnisse hinsichtlich der jeweiligen Rangfolge der Knoten.



8 Retrievaltests

8.1 Hinweis

Die folgende Übung zur Durchführung von Retrievaltests setzt voraus:

1. Kenntnisse über die Berechnung von *Recall* und *Precision* [Vgl. Abschnitt 6.2 in: Gödert, W., K. Lepsky, and M. Nagelschmidt (2012). Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. X.media.press. Berlin u.a.: Springer.];
2. Kenntnisse über die Anwendung des sog. *Pooling-Modells* für Retrievaltests [Vgl. Sachse, Elisabeth, Martina Liebig und Winfried Gödert: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse eines Retrievaltests zum MILOS-II-Projekt, Köln: Fachhochschule Köln 1998 (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft 14). (<http://ixtrieve.fh-koeln.de/lehre/sachse-1998-a.pdf>)]
3. Kenntnisse über die suchtechnischen Möglichkeiten der einzusetzenden Suchumgebungen.

8.2 Kollektionen und Suchumgebungen

Verwenden Sie für die Durchführung des vergleichenden Retrievaltests die folgenden drei Aufbereitungen der Datenbank *Literatur zur Informationserschließung*:

- Suchmaschine (Solr) *Birds*: <http://ixtrieve.fh-koeln.de/birds/litie/>
- Bibliografische Datenbank *Midos*: <http://ixtrieve.fh-koeln.de/mwr/litie/>
- Bibliografische Datenbank *allegro*: <http://ixtrieve.fh-koeln.de/a35/litie/>

8.3 Vorgehensweise

Gegenstand der Übung ist die Durchführung eines vergleichenden Retrievaltests für insg. 10 Themenstellungen. Verglichen werden sollen dabei die Suchergebnisse für die o.g. drei verschiedenen Suchumgebungen. Die vergleichende Bewertung soll auf der Basis selbst ermittelter Werte für *Recall* und *Precision* erfolgen. Zur Ermittlung der *Recall*-Werte ist das *Pooling-Verfahren* anzuwenden.

Gehen Sie für die Durchführung des Retrievaltests folgendermaßen vor.

- Suchen
 - Setzen Sie die 10 Themenstellungen in Suchformulierungen um, die die jeweiligen Eigenschaften der drei Suchumgebungen berücksichtigen.
 - Verwenden Sie für die Suche in den bibliografischen Datenbanken geeignete Suchindices.
 - Nutzen Sie Möglichkeiten zur Phrasensuche aus (Syntax).
- Relevanzbewertung
 - Ermitteln Sie für Ihre Suchformulierungen die jeweiligen Treffermengen in den verschiedenen Suchumgebungen.

- Bestimmen Sie für *jedes* Dokument in *jeder* Treffermenge die *Relevanz* im Hinblick auf die Suchanfrage. Als Relevanzkriterium soll gelten: „Im Sinne einer Frage ‚Ich suche etwas über ...‘ halte ich den Treffer für so interessant, dass ich mir das Dokument im Volltext anschauen würde“.
- Auswertung
 - Berechnen Sie für jede Treffermenge die *Precision*.
 - Berechnen Sie für jede Treffermenge den *Recall* mit dem *Pooling-Verfahren*.
- Analyse
 - Analysieren Sie die ermittelten *Recall*- und *Precision*-Werte im Hinblick auf eine Gesamtaussage zur Retrievaleffektivität.
 - Analysieren Sie Stärken und Schwächen einzelner Suchverfahren bzw. Suchumgebungen in Abhängigkeit vom Typus der Suchanfrage.
 - Analysieren Sie den (möglichen) Einfluss grammatikalischer Varianten auf das Suchergebnis. Führen Sie ggf. Testrecherchen mit alternativen Suchformulierungen durch.
 - Analysieren Sie den (möglichen) Einfluss der Dokumentsprache auf das Suchergebnis. Führen Sie ggf. Testrecherchen mit alternativen Suchformulierungen durch.
- Recherchethemen [in Klammern eine Typisierung]
 - (1) Serendipity [Einzelwort]
 - (2) Retrievaltest [Kompositum]
 - (3) h-index [Bindestrichwort, Verbindung mit Eigennamen, Verbindung mit Abkürzung]
 - (4) PRECIS [Akronym, Eigenname]
 - (5) Faceted classification [Phrase]
 - (6) Syntaktisches Indexieren [Phrase]
 - (7) Information-Retrieval-Verfahren [Bindestrichwort, Phrase]
 - (8) Linguistische Verfahren des automatischen Indexierens [Kombination von Phrasen]
 - (9) Visualisierung von Informationsräumen [Kombination Einzelwort und Kompositum]
 - (10) Effektivitäts- oder Effizienzmaße (z. B. von Rijsbergen) für Retrievaltests [komplexe Fragestellung]