Modulprüfung Information Retrieval [MP 2200]

Übungen und Wiederholungsfragen zur Prüfungsvorbereitung

Winfried Gödert / Klaus Lepsky

1. Vorbemerkung

Die folgenden Wiederholungsfragen dienen der Prüfungsvorbereitung für die Modulprüfung MP 2200 (Information Retrieval) im Studiengang BA Informationswirtschaft. Die Modulprüfung wird ab dem Sommersemester 2014 als Klausur mit zwei Teilen in folgender Gewichtung durchgeführt:

- 50% Information Retrieval,
- 50% Automatisches Indexieren.

Die Inhalte des Klausurteils „Information Retrieval“ beziehen sich auf die Kapitel 5 und 6 des Lehrbuchs „Informationserschließung und Automatisches Indexieren“. Die Inhalte des Klausurteils „Automatisches Indexieren“ beziehen sich auf das Laborpraktikum „Automatisches Indexieren“ [IB 23] und das Kapitel 5 des Lehrbuchs „Informationserschließung und Automatisches Indexieren“.

2. Wiederholungsübungen und -fragen Information Retrieval

2.1 Index, Invertierte Liste

a) 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.

b) 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 Booleschen 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?

c) 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!

d) Welche Vorteile bietet ein **Index** gegenüber der **sequenziellen Suche**?

e) Ist es zweckmäßig, sog. „Stoppwörter“ aus dem Index auszuschließen? Begründen Sie die Antwort.

f) Ist die Sprache eines Dokuments für den Indexaufbau von Belang?

2.2 Recall und Precision

g) 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.

h) 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?

i) Wie wirkt sich der Einsatz einer Rechtstrunkierung auf Recall und Precision aus?

k) Wir 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?

l) 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.

m) Lässt sich eine sehr schlechte Precision durch ein Relevance Ranking kompensieren? Begründen Sie Ihre Antwort.

2.3 Statistik im Information Retrieval

n) Erläutern Sie die Begriffe TF, WDF und IDF und geben Sie Beispiele für den Nutzen von Termgewichtungsverfahren.

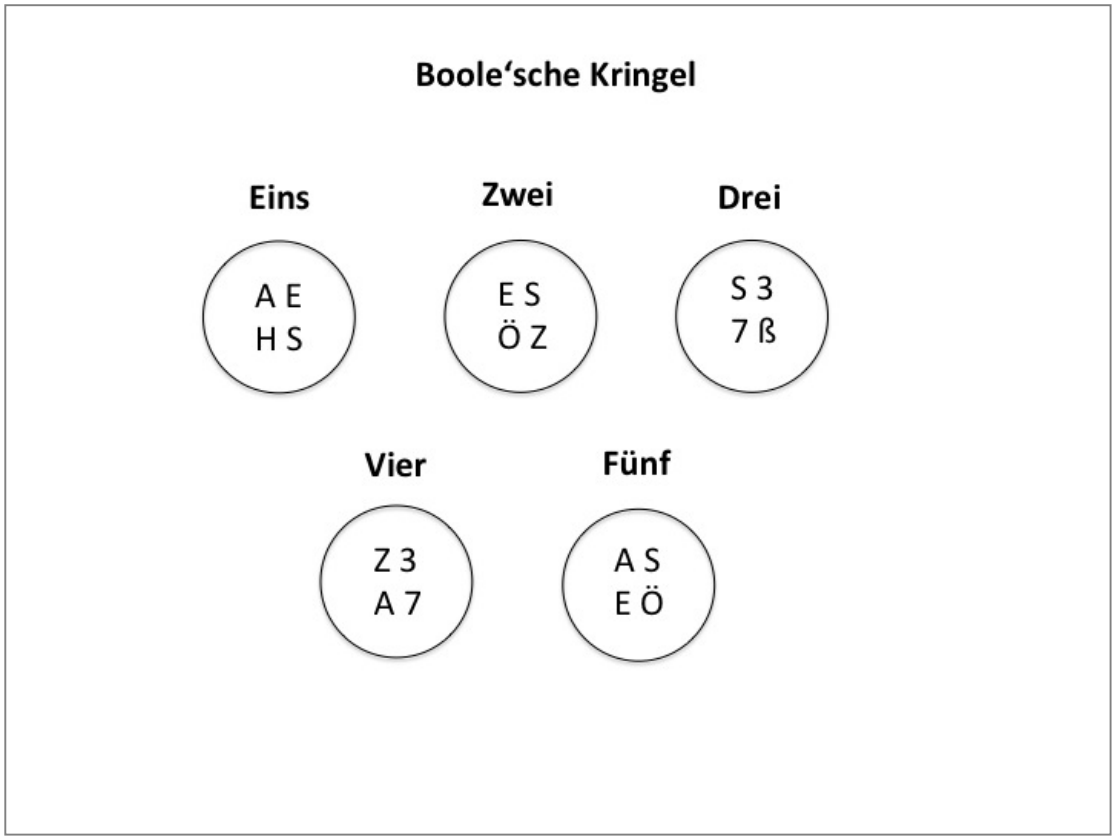
o) Wie entsteht eine nach TF sortierte Trefferliste?

p) Erhalten Hochfrequenzterme bei der Anwendung des IDF ein besonders hohes oder ein besonders niedriges Termgewicht? Warum?

q) Sind Hochfrequenzterme für den Einsatz einer Termgewichtung problematisch?

r) 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.

2.4 Boole'sche Operatoren



s) 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

t) Welche Suchen führen zu diesen Ergebnisse?

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

3. Wiederholungsübungen und -fragen Automatisches Indexieren

3.1 Wiederholungsfragen zum Laborpraktikum Automatisches Indexieren [IB 23]

Die Wiederholungsfragen dienen der Vertiefung der im Laborpraktikum behandelten Materie. Sie ergänzen die Übungsaufgaben, die sich im Buch am Ende jedes Kapitels finden.

3.1.1 Automatische Schlagwortvergabe mit Midos (Kapitel 5.2)

1. Warum ist es nicht zweckmäßig, die Automatische Schlagwortvergabe für alle Kategorien eines Datensatzes durchzuführen?

2. Welchen Verweisungstyp enthält die Datei „auto-sw.wtx“ (bzw. „synonym.wtx“) und wie erkennt man die Richtung der Verweisung?

3. In welcher Form enthält die Datei „synonym.txt“ diese Verweisungen?

4. Geben Sie zwei Datensätze an, an denen man die Wirkung der Verweisungen auf die Zuteilung der Auto-Schlagwörter sehen kann?

5. Wäre es sinnvoll, mit der Datei „synonym.wtx“ zu einem Unterbegriff auch Oberbegriffe zuteilen zu lassen?

6. Was müsste man tun, um zu Unterbegriffen auch Oberbegriffe zu erzeugen?

7. Nennen Sie drei Beispiele, für die das Erzeugen von Oberbegriffen sinnvoll wäre.

3.1.2 Das Indexierungssystem Lingo (Kapitel 5.3)

a) tokenizer und Konfigurationsdatei (Kapitel 5.3.2)

8. Wie wird in einer Konfigurationsdatei „*.cfg“ festgelegt, dass ein Attendee mitarbeitet oder nicht?

9. Wie wird in einer Konfigurationsdatei „*.cfg“ festgelegt, ob eine Ergebnisdatei geschrieben wird?

10. Welche Attendees müssen bei einem Verarbeitungslauf mindestens eingeschaltet sein, damit Lingo sinnvolle Ergebnisse produzieren kann?

11. Welche Zeichen benutzt der Attendee *tokenizer* zur Bestimmung von Wörtern?

12. Ist die Zeichenkette „3-D-Simulation“ für den *tokenizer* ein Wort?

13. Arbeitet der *tokenizer* regelbasiert oder gestützt auf Wortlisten?

14. Welche Endungen haben die Ergebnisdateien, die jeweils von den folgenden Attendees produziert werden?

- *sequencer*
- *decomposer*
- *multiworder*

15. Welche Wortklassenkennung besitzen die Wörter der „*.non“-Datei in der „*.log“-Datei ?

16. An welcher Stelle im Abschnitt „Inhalte verarbeiten“ der „*.cfg“-Datei muss die Angabe „out: datei“ erfolgen, damit die erzeugten Ergebnisse vom *vector_filter* als „in: datei“ verarbeitet werden können?

b) Grundformerzeugung und Wortklassenerkennung (Kapitel 5.3.2)

17. Welche Bedeutung hat das „?“ hinter Wörtern in der „*.log“-Datei?

18. Welchen Nutzen kann es haben, die Wörter der „*.non“-Datei einer Analyse zu unterziehen?

19. Woher bezieht Lingo die Wortklassenkennungen, die in der „*.log“-Datei angegeben werden?

20. Welcher Attendee erzeugt Grundformen und wie macht er das?

21. Kann Lingo das Wort „Menschern“ verarbeiten und eine korrekte Grundform generieren?

22. Bilden Sie aus der Grundform „Mensch“ durch Anhängen eines Suffixes ein Wort (kein Kompositum!), das vom *word_searcher* nicht erkannt wird.

23. Ist die Reihenfolge der Attendees beliebig oder nicht? Gilt diese Aussage für alle Attendees gleichermaßen?

24. Besitzt der *word_searcher* eine Kenntnis darüber, ob ein erkanntes Wort ein korrektes Wort der deutschen Sprache ist?

25. Warum gibt es keine Wortklassenkennung für Homonyme?

26. Wie muss der *vector_filter* für die Erzeugung der „*.vec“-Datei eingestellt werden, damit nur

- Adjektive
- Verben
- Komposita

ausgegeben werden?

c) Kompositumerkennung, Longest Matching, Wörterbücher (Kapitel 5.3.3)

27. Welche Wörter werden nach dem *word_searcher* vom *decomposer* weiterverarbeitet?

28. Werden durch den *decomposer* immer nur zwei Zerlegungsbestandteile eines Kompositums ermittelt? In welcher Datei kann man darauf durch welche Einstellung Einfluss nehmen?

29. Welchen Einfluss hat es auf das Zerlegungsergebnis von Komposita, ob das Verfahren des Longest Matching von links oder von rechts durchgeführt wird? Welche Variante wird von Lingo verwendet? Erklären Sie das folgende Ergebnis:

```
lex:)  <Wirkungsorte|KOM = [(wirkungsorte/k), (sorte/s+), (wir-
kung/s+)]>
```

30. Wie kann man erreichen, dass „Wirkungsorte“ korrekt zerlegt wird?

31. Welche Wortklassenkennungen können die durch den *decomposer* erkannten Zerlegungsbestandteile haben?

32. Warum kennt Lingo kein Wörterbuch, das nur Komposita enthält?

33. Erläutern Sie den Zusammenhang zwischen „usr-dic“ und „user-dic.txt“ in der Datei „de.lang“.

34. Was bedeutet der Eintrag source: in einer „*.cfg“-Datei?

35. Was muss man tun, damit der *decomposer* ein Benutzerwörterbuch verwendet?

36. Was bedeutet der Eintrag „mode: first“ in einer „*.cfg“-Datei?

37. Ist es möglich, ein Wörterbuch zu gestalten, das nur Adjektive enthält und in die Konfiguration einer Lingo-Verarbeitung einzubinden?

d) Semantische Analyse: multiworder, sequencer, synonymer (Kapitel 5.3.4)

38. Welcher Eintrag muss im Benutzerwörterbuch für Mehrwortgruppen vorhanden sein, damit die Phrase „Menschen für Menschen“ vom Attendee *multi_worder* erkannt und ausgegeben wird?

39. Ist es empfehlenswert, für den Attendee *multi_worder* in der „*.cfg“-Datei neben dem Wörterbuch „sys-mul“ auch das Wörterbuch „sys-dic“ anzugeben? In welcher Reihenfolge?

40. Mit welcher Einstellung des Attendee *sequencer* lässt sich die Phrase „Menschen für Menschen“ identifizieren? Wird bei der Ausgabe das Wort „Menschen“ oder das Wort „Mensch“ ausgegeben?

41. Welche Einstellung für den Attendee *sequencer* erzeugt aus der Wortfolge „Kommunale öffentliche Bibliothek“ die Ausgaben:

- Bibliothek, kommunal öffentlich
- öffentlich Bibliothek, kommunal
- kommunal öffentlich Bibliothek

42. Welche der Varianten empfiehlt sich für den Aufbau eines Suchindex in einer Retrievalumgebung? Welche für den Aufbau eines Mehrwortgruppen-Wörterbuchs, das vom Attendee *multi_worder* benutzt werden soll?

43. Wie müsste ein Eintrag im Benutzerwörterbuch für den Attendee *synonymer* aussehen, damit aus „Bücherei“ im Text das Ergebnis „Bibliothek“ erzeugt wird?

44. Wie müsste ein Eintrag im Benutzerwörterbuch für den Attendee *synonymer* aussehen, damit aus „Schlagwort“ im Text die Ergebnisse „Deskriptor“ und „Vorzugsbenennung“ erzeugt werden? Geht das auch wechselweise, z. B. aus „Deskriptor“ im Text sollen „Schlagwort“ und „Vorzugsbenennung“ generiert werden?

45. Ist „deskriptoren=deskriptor“ ein sinnvoller Eintrag in einem Synonym-Wörterbuch?

46. Dürfen Synonyme auch Komposita sein?

e) LIR-Konfiguration (Kapitel 5.3.6)

47. Warum ist es nicht zweckmäßig, die aus einer Datenbank exportierten Datensätze mit der „lemma.cfg“ zu indexieren, obwohl sie alle in einer Datei stehen?

48. Wie zweckmäßig wäre es, die exportierten Datensätze einer Datenbank jeweils in getrennte Dateien zu schreiben, um sie mit der „lemma.cfg“ indexieren zu können?

49. Warum werden für die Zwecke der Indexierung nicht alle Felder der Datenbank exportiert? Unterscheidet sich die Antwort von der Antwort auf Frage 47?

50. Was muss beim Export der Datensätze aus einer Datenbank beachtet werden, damit durch Lingo erzeugte Indexierungsdaten dem zutreffenden Datensatz zugeordnet werden können?

51. Welche Unterschiede gibt es zwischen der automatischen Schlagwortvergabe mit MIDOS und der Indexierung mit Lingo:

( ) MIDOS-AutoSW kann keine Adjektiv-Substantiv-Verbindungen erzeugen

( ) MIDOS-AutoSW kann alle Singular-/Plural-Wendungen erkennen

( ) MIDOS-AutoSW kann Synonyme auf ihre Vorzugsbenennung abbilden

( ) Lingo kann alle synonymen Wortformen als Indexterme erzeugen

( ) Lingo kann Flektionsformen von Substantiven im Plural erkennen

( ) MIDOS-AutoSW kann Flektionsformen von Kompositabestandteilen erkennen

( ) Lingo kann Flektionsformen von Kompositabestandteilen erkennen

52. Ist es sinnvoll, die Ergebnisse einer automatischen Schlagwortvergabe mit MIDOS mit Indexierungsergebnissen von Lingo für einen gemeinsamen Suchindex einer Retrievalanwendung zusammenzufassen?

53. Ist es sinnvoll, die Inhalte der „*.vec“-Datei und der „*.mul“-Datei einer Lingo-Indexierung zu einem gemeinsamen Suchindex zusammenzufassen?

54. Kann ich mir eine Liste von Indextermen ausgeben lassen, die für mindestens drei Datensätze erzeugt wurden?

3.2 Musteraufgabe für den Klausurteil Automatisches Indexieren

```
Zielvorstellung ist der Aufbau von Datenbanken zur formalen und
inhaltlichen Erschließung und die Gestaltung von Retrievalumge-
bungen, für bibliografische Daten auch die Erstellung von Biblio-
grafien. Als Methoden zur inhaltlichen Erschließung werden beson-
ders die semantische Strukturierung von Themenfeldern am Beispiel
des aspektorientierten Thesaurus-Konzepts und das Automatische
Indexieren bibliografischer Daten behandelt.
```

1. Welche Ergebnisse für das oben gezeigte Dokument erzielt der *wordsearcher* von Lingo bei einer Automatischen Indexierung?

2. Welche Ergebnisse erzielt der *decomposer* von Lingo?

3. Welche Ergebnisse erzielt der *multiworder* von Lingo?

4. Welche Ergebnisse erzielt der *sequencer* von Lingo?

5. Welche Ergebnisse erzielt der *synonymer* von Lingo?

!! Notieren Sie Ihre Ergebnisse als einfache Liste von Indextermen untereinander. !!!

!!! Verwenden Sie die Lingo-Wörterbücher „lingo-dic“, „lingo-mul“, „lingo-syn“ sowie die Suffixliste und die *sequences* auf den folgenden Seiten!!!

lingo-dic

```
aspekt=aspekt #s
automatisch=automatisch #a
behandeln=behandeln #v
behandelt=behandelt #a behandeln #v
beispiel=beispiel #s
besonders=besonders #a
bibliografie=bibliografie #s bibliographie #s
bibliografisch=bibliografisch #a bibliographisch #a
bibliographie=bibliographie #s bibliografie #s
bibliographisch=bibliographisch #a bibliografisch #a
daten=daten #s
datenbank=datenbank #s
erschließung=erschließung #s
erstellung=erstellung #s
feld=feld #s
gestaltung=gestaltung #s
indexieren=indexieren #v
inhaltlich=inhaltlich #a
ist=sein #v
konzept=konzept #s
methode=methode #s
orientiert=orientiert #s
retrieval=retrieval #s
semantisch=semantisch #a
strukturierung=strukturierung #s
thema=thema #s
themen=thema #s
thesaurus=thesaurus #s
umgebung=umgebung #s
vorstellung=vorstellung #s
ziel=ziel #s
```

lingo-syn

```
datenbank;dbms

erschließung;indexierung;indexing

bibliografie;bibliographie
```

lingo-mul

```
aspektorientiertes thesaurus-konzept

automatisches indexieren

formale erschließung

gestaltung von retrievalumgebungen

inhaltliche erschließung

semantische Strukturierung
```

Lingo-Suffixliste

suffix:

```
#    Suffixklasse:   s = Substantiv, a = Adjektiv, v = Verb, e =
Eigenwort, f = Fugung

#    Suffixe je Klasse:   „<Suffix>[,/'<Ersetzung>][ <suf-
fix>[,/'<Ersetzung>]]“

- [s, „e en er ern es n s se sen ses“]

- [a, „este ste ster sten stes ester estes esten e em en er ere
eren erer eres es erem“]

- [v, „e/en en/en est/en et/en st/en t/en te/en ten/en eten/en
ete/en etest/en s“]

- [e, „s“]

- [f, „s n e en es er ch/che /en“]
```

Lingo-sequences

```
   sequencer:

   sequences: [ [AS, „2, 1“], [AK, „2, 1“], [AAK, „3, 1 2“], [AAS,
„3, 1 2“] ]
```