Beuth Hochschule Berlin
Datenbanksysteme Übung, Medieninformatik (Bachelor), 2. Semester
Sommersemester 2018

Übungsaufgaben
DBS1 Bachelor 2. Semester

Prof. Dr. habil. Alexander Löser
Aljoscha Marcel Everding

Version: 1.0 (9. April 2018)

Inhaltsverzeichnis

1 Einleitung

Ihre „große“, übergreifende Aufgabe für die nächsten 16 Wochen:

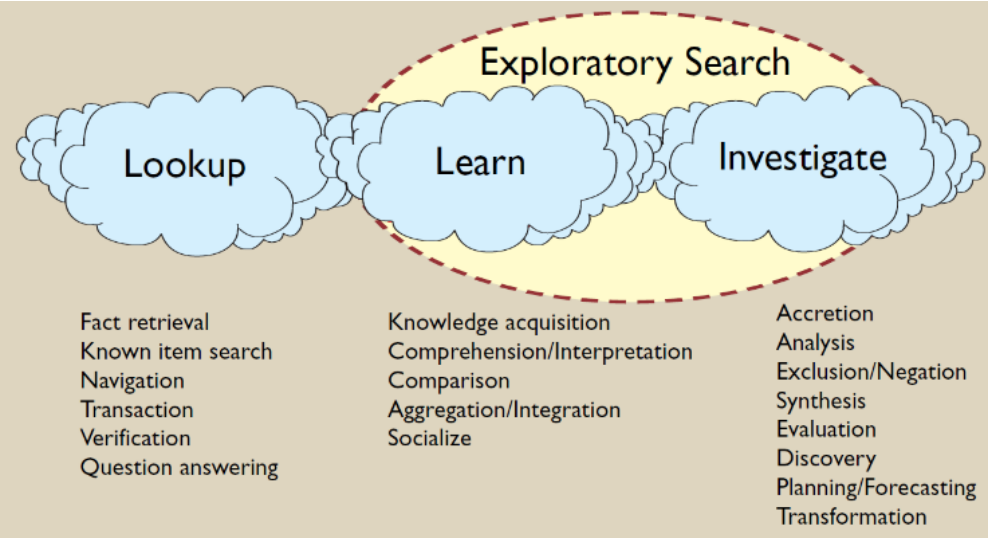

Abbildung 1.1: Menschen folgen bei der Verarbeitung von Informationen, z.B. mit Datenbanken oder Suchmaschinen einem iterativen Prozess aus drei Teilschritten „Lookup, Learn und Investigate“. Sie werden im Laufe des Semesters in den Anfragen Muster für diesen Prozess wiederfinden. Quelle [6, 7]

Stellen Sie sich vor, Sie würden bei *Google*, in der *Bayer AG*, bei *Zalando* oder bei *Springer* an der nächsten Generation von Websuchmaschinen bzw. Suchmaschinen für das Intranet arbeiten. Eine Grundlage für das Verstehen der Frage: *„Warum suchen wir und wie sind typische Muster abgebildet“* stellen die Klick-Logs im AOL-Korpus dar, die in Abbildung 1.1 gezeigt werden.

Dieser Korpus stammt aus dem Jahre 2006 (März - Juni) und erlaubt uns besser zu verstehen, wie Menschen zu bestimmten Ereignissen in 2006 gesucht und Informationen verarbeitet haben.

Wir stellen Ihnen eine Datenbank zum „Spielen“ zur Verfügung. Die Datenbank beinhaltet

- Anfragen der AOL-Nutzer (USA und in Englisch),

2028	294499	alaska's average temperature in the winter	42B	2006-05-26 17:59:21	1	http://www.farmersalmanac.com
2029	294499	alasks average winter temperature	33B	2006-05-26 18:00:31	0	
2030	294499	alaska average winter temperature	33B	2006-05-26 18:00:35	1	http://www.farmersalmanac.com
2031	294499	alaska average winter temperature	33B	2006-05-26 18:00:35	4	http://www.alaska.com
2034	294499	what is the average tempture of alaska in the w...	52B	2006-05-26 18:03:54	0	
2035	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	1	http://www.alaska.com
2036	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	6	http://instaar.colorado.edu
2037	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	10	http://www.worldviewofglobalwarming.org
6588	393765	wal mart stock average price per share	38B	2006-04-01 08:57:13	0	
6589	393765	walmart stock average price per share	37B	2006-04-01 08:58:48	0	
0474	672368	what is the average merit raise	31B	2006-05-31 13:31:41	4	http://www.salary.com
0987	881500	what is the average cost of beating a dui wrap ...	72B	2006-03-02 15:15:36	0	
1235	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	2	http://www.city-data.com
1236	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	10	http://mattsonworks.com
1237	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	7	http://www.epodunk.com
1238	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	10	http://mattsonworks.com

Abbildung 1.2: Beispiel von Suchanfragen die das Wort „average“ enthalten, den angeklickten Seiten, u.a. für den Nutzer 294499, dem Zeitpunkt des Klicks auf das Resultat und dessen Ranking in der Ergebnismenge durch Google. Der Wert 0 in der Spalte ItemRank bedeutet, dass der Nutzer keine Seite des Google Ergebnisses angeklickt hat.

- Kategorien zu den angeklickten Webseiten von DMOZ[1]
- Zip Codes und Orte in den USA,
- Wirbelstürme
- und Oscar Gewinner.

Die Aufgaben stellen 25 von 100 Punkten (Klausur = 75 Punkte) dar. Sie dürfen diese Aufgabe in Teams in der Größe von 2 bis 4 Personen lösen. Die Aufgaben bauen aufeinander auf.

[1] `http://www.dmoz.org`

2 Aufgabe 1: Diskursbereich, Anfragen und Modellierung (6 Punkte)

2.1 Aufgabe 1a: Anfragen in natürlicher Sprache

1. Überlegen Sie bitte was im Jahre 2006 in den USA zwischen März und Juni geschah, z. B. Ereignisse wie Erdbeben oder das Herausbringen einer neuen Spielkonsole. Eventuell haben die Suchenden aber auch ganz persönliche Ereignisse bearbeitet, wie die Auswahl der Hochschule für das naheliegende Studium.

2. Bitte formulieren Sie **zehn analytische Anfragen in natürlicher Sprache**, die Sie sich mit Hilfe der vorliegenden Datenbank zu beantworten erhoffen. Beispiele wären:

   - *Wie häufig sind . . . ?*
   - *Wann waren . . . .?*
   - *Gruppiere Anfragen nach x, y?*
   - *Welche Webseite . . . ?*
   - *Mit welcher Anfrage . . . ?*
   - *Welche Orte . . . mit . . . ?*
   - *Welche Muster/ Wörter . . . mit . . . ?*
   - *Welche Nutzer haben . . . was . . . wann . . . wo . . . am meisten . . . ?*

3. Bitte schreiben Sie **pro Anfrage auch das erhoffte Ergebnis auf. Das kann eine Tabelle, ein Diagramm oder eine andere angemessene Darstellung sein.**

4. Prüfen Sie bitte, ob die Tabellen in der Ihnen vorliegenden Datenbank Ihnen die Bearbeitung der Anfragen ermöglicht oder ob Sie weitere Datenquellen benötigen. Falls das letztere der Fall ist:

   - Variante 1: Nennen Sie mögliche Datenquellen, die Sie mit den Daten verknüpfen müssten. Denken Sie daran, dass Sie später diese Datenquellen auch wirklich verknüpfen müssen! Die Quelle sollte möglichst Ihre Daten als Tabelle „preisgeben“, falls die Datenquelle nur Text bereitstellt, erhalten Sie von uns dafür die benötigten Tools.
     Beispiele für Datenquellen sind Wikidata[1], Freebase[2] (Read-Only; in wie weit der Dienst in Zukunft noch verfügbar sein wird, ist derzeit leider unklar), ProgrammableWeb[3], Excel-Sheets, etc.
   - Variante 2: Sie suchen sich besser passende Anfragen.

5. Bitte stellen Sie pro Anfrage dar, warum die Antwort hilfreich für die Suchenden ist.

2.2 Aufgabe 1b: ER-Schema (→ VL ER-Modelle)

Entwickeln Sie für die Beantwortung der Anfragen aus Aufgabe 1 ein logisches ER-Schema. Das Schema sollte **mindestens 10 Elemente** (Entity- bzw. Relationshiptypen) enthalten. Bitte ergänzen Sie auch die notwendigen Integritätsbedingungen.

2.3 Aufgabe 1c: Relationales Schema, DDL und Beispiele (→ VL Relationales Datenmodell)

- Transformieren Sie bitte das ER Schema aus Aufgabe 1b in ein relationales Schema.
- Erstellen Sie die Tabellen für die 1:1-, 1:N- und M:N-Beziehungen.
- Fügen Sie pro Tabelle per SQL drei Beispieldatensätze ein.
- Beachten Sie Schlüssel, Kardinalitäten, Totalitäten und andere Integritätsbedingungen.

[1] `https://www.wikidata.org/`
[2] `http://www.freebase.com/`
[3] `http://www.programmableweb.com/`

Abgabe

Bitte stellen Sie Ihre Übersicht in einer Präsentation zusammen. Eine Slide sollte eine Anfrage, die erhofften Resultate, die Liste der zu verknüpfenden Datenquellen und den Wert enthalten. Erzeugen Sie von der Präsentation eine **PDF-Datei**![4]

Bitte erstellen Sie eine ZIP-Datei unter dem Namen `aufgabe1_<Gruppenname>`[5]) und laden Sie bitte die Datei in das Moodle. Den Abgabetermin für diese Aufgabe entnehmen Sie bitte dem Moodle-Kurs. Die Datei enthält in geeigneter Form Ihre Anfragen, das dazu passende ERM und relationale Schema, sowie Screenshots (vom SQL-Developer) der Tabellen mit den drei Beispieldatensätzen pro Tabelle aus der Oracle-Datenbank als „Beweis“.

[4] Andere Dateiformate (Keynote, Powerpoint, LibreOffice etc.) werden nicht akzeptiert!

[5] Ihr Gruppenname setzt sich aus Ihren Nachnamen in alphabetischer Reihenfolge zusammen. Bei gleichen Nachnamen ist an den Nachnamen der erste Buchstabe des Vornamens zu setzen z. B. Max Mütze → MützeM

3 Aufgabe 2: SQL, Anfrageausführung und Optimierung (5 Punkte)

3.1 Aufgabe 2a: SQL-Anfragen

Bitte formulieren Sie für Ihre zehn Anfragen aus Aufgabe 1a die deklarativen Anfragen in der Sprache SQL. Bitte verwenden Sie in geeigneter Weise Joins, Funktionen zum Konvertieren von Zeitangaben, Aggregationsfunktionen, Group Bys sowie **mindestens** eine View und Sub-Queries. Bitte verknüpfen Sie in mindestens acht Ihrer Anfragen die Tabellen aus der Oracle-Datenbank aus dem Schema AOL mit den von Ihnen zusätzlich angelegten Tabellen.

3.2 Aufgabe 2b: Anfrageplanung und Optimierung

- Bitte formulieren Sie zwei angemessen (!) komplexe Anfragen aus Aufgabe 2a als Anfrageplan in der Notation der relationalen Algebra.
- Bitte diskutieren Sie zwei mögliche heuristische oder statistische Optimierungen für diese Pläne.

Abgabe

Bitte erstellen Sie eine ZIP-Datei unter dem Namen `aufgabe2_<Gruppenname>`[1] und laden Sie bitte die Datei in das Moodle. Den Abgabetermin für diese Aufgabe entnehmen Sie bitte dem Moodle-Kurs. Die Datei enthält in geeigneter Form

[1] Ihr Gruppenname setzt sich aus Ihren Nachnamen in alphabetischer Reihenfolge zusammen. Bei gleichen Nachnamen ist an den Nachnamen der erste Buchstabe des Vornamens zu setzen, z. B. Max Mütze → MützeM

in einer Präsentation (als PDF-Datei![2]), Ihre natürlich-sprachigen Anfragen, die SQL-Anfragen die Resultate und die Querypläne sowie Optimierungen. **Bitte geben Sie zusätzlich als „Beweis", dass Ihre Anfragen wirklich die Resultate liefern, Screenshots der Ausgaben im Oracle SQL Developer ab.**

[2] Andere Dateiformate (Keynote, Powerpoint, LibreOffice etc.) werden nicht akzeptiert!

4 Aufgabe 3: Datenintegration und Anwendungsentwicklung (10 Punkte)

Für diese Aufgabe dürfen Sie aus einer der drei Optionen wählen. In jedem Fall ist ein Programm (z. B. in Java oder Python) zu schreiben, dass mit der Datenbank interagiert!

4.1 Option 1: Eigene Datenquellen

Füllen Sie jetzt Ihre Dummy-Daten aus Aufgabe 2 mit realen Daten auf. Benutzen Sie dazu die in Aufgabe 1 erwähnten Datenquellen bzw. ergänzen Sie die Datenquellen. Testen Sie dann Ihre Anfragen aus den vorigen Aufgaben mit den neuen Daten. Bitte erstellen Sie wiederum Screenshots als Beweis.

4.2 Option 2: Textdaten aus dem Internet Archive

Bitte ergänzen Sie die AOL-Query Log Datenbank mit den tatsächlichen Dokumenten aus dem Jahre 2006. Die Anfragen enthalten die Ziel-URL und das Datum des Besuches der Webseite.

- Bitte nutzen Sie diese Information, um aus dem Internetarchive um die HTML Quelle der Seite in der Datenbank zu ergänzen. Heuristik: Bitte wählen Sie die Seite aus dem Archiv mit dem nächstmöglichen Zeitstempel nach dem Besuchsdatum durch den AOL Nutzer. Sie können auf das Archiv mit dem Memento Protokoll zugreifen [9].
- Bitte ergänzen Sie Ihr logisches und relationales Schema so, dass es nun auch die HTML Seiten aufnehmen kann. Bitte ergänzen Sie Ihr Schema zur Aufnahme folgender weitere Daten:

  - der Quellcode der Seite
  - der bereinigte Quellcode der Seite, nutzen Sie z. B. die Software Boilerpipe[1].
  - das Datum aus dem Internet-Archive
  - die Sprache der bereinigten Seite nach Boilerpipe. Nutzen Sie eventuell dafür die JTCL - Java Text Categorizing Library[2].
  - die Domäne der Seite
  - die Top Level Domäne der Seite
  - der Titel der Seite und eventuell weitere DC Attribute z. B. mit dem Jericho HTML-Parser[3].

- Bitte extrahieren Sie die obigen Daten für 2000 HTML-Seiten in Ihre Datenbank.

4.3 Option 3: Visualisierung

Bitte visualisieren Sie die Daten aus der Oracle Datenbank. Sie sollten zeigen, dass Sie die Daten anfragen und in eine Datenvisualisierungsumgebung überführen können bzw. dort die Anfragen an die Oracle Datenbank weitergeleitet werden. *Beispiele* für Visualisierungen sind:

- `http://d3js.org`
- JFreeChart[4]
- `https://www.google.com/publicdata/directory`
- `https://google-developers.appspot.com/chart/interactive/docs/gallery`
- `https://groups.google.com/forum/#!topic/google-visualization-api/UdOFybnvFo0`

[1] `https://github.com/kohlschutter/boilerpipe/`
[2] `http://textcat.sourceforge.net`
[3] `http://jericho.htmlparser.net/docs/index.html`
[4] `http://www.jfree.org/jfreechart/`

Abgabe

Bitte erstellen Sie ein Zip-Archiv unter dem Namen `aufgabe3_<Gruppenname>`[5] und laden Sie bitte die Datei in das Moodle. Den Abgabetermin für diese Aufgabe entnehmen Sie bitte dem Moodle-Kurs.

Die Datei enthält:

- ein Präsentationsfoliensatz als PDF-Datei[6] mit Ihrem Datenmodell und eventuell erfolgten Erweiterungen,
- Screenhots von 3 Anfragen (bitte mindestens 2x Joins) aus den vorherigen Aufgaben mit Resultaten auf Ihrem Datenbestand, die nun möglich werden (nur Option 1)
- Screenshots Ihrer Visualisierungen (nur Option 3)
- den Quelltext ihres Programms
- Ihre Überlegungen für eine Fortsetzung Ihrer Arbeit.

[5] Ihr Gruppenname setzt sich aus Ihren Nachnamen in alphabetischer Reihenfolge zusammen. Bei gleichen Nachnamen ist an den Nachnamen der erste Buchstabe des Vornamens zu setzen z. B. Max Mütze → MützeM

[6] Andere Dateiformate (Keynote, Powerpoint, LibreOffice etc.) werden nicht akzeptiert!

5 Aufgabe 4: Analyse und Verwertung der Erkenntnisse (4 Punkte)

Herzlichen Glückwunsch, Sie haben jetzt den AOL-Datensatz erweitert und können auf den Daten neue Erkenntnisse gewinnen.

Stellen Sie in 12 Minuten die wichtigsten Erkenntnisse aus den Daten Ihren Kommiliton*innen vor. Bewerten Sie in Ihrem Vortrag den Erkenntnisgewinn, z. B. gegenüber Ihren Kommiliton*innen oder der Literatur. Welche Erkenntnisse hätten einen kommerziellen Wert?

Ihre Lösung soll vor allen Gruppen präsentiert werden.

Inhalte:

- Führen Sie in Ihr Projekt ein (u.a.: Ziele, Gegenstand, Inhalt, Ergebnisse, Beteiligte, Arbeitsteilung).
- Zeigen Sie Ihr Datenmodell, mit dem Sie gearbeitet haben, und erläutern Sie dieses kurz.
- Die wichtigsten Anfragen die sie lösen wollten und die Anfragen die sie lösen konnten.
- Machen Sie den Entwicklungsweg Ihres Projektes deutlich und stellen Sie Entwurfs-/Implementierungsalternativen vor.
- Zeigen Sie Ihren Erkenntnisgewinn auf.

Abgabe

Bitte erstellen Sie von ihrer Präsentation eine als `aufgabe4_<Gruppenname>`[1] benannte PDF-Datei[2] und laden Sie diese in das Moodle. Den Abgabetermin für diese Aufgabe entnehmen Sie bitte dem Moodle-Kurs.

Tag der Präsentation

Die Präsentationen finden in den Vorlesungsblöcken und Übungen statt. Sie bekommen von den Dozenten einen Slot zugewiesen, dieser ist einzuhalten. Es ist Ihre Aufgabe, am Tag der Präsentation Ihrer Ergebnisse dafür zu sorgen, dass Sie präsentieren können. Sorgen Sie dafür, dass ein Rechner im Veranstaltungsraum vorhanden ist, der Ihre Präsentationsdatei abspielt und der eine Verbindung zum Beamer hat. Klären Sie im Vorfeld, ob etwaige Adapter (HDMI, VGA) von Nöten sind. Die Dozenten stellen weder Laptops, noch Adapter! Aus Fairness den präsentierenden Gruppen gegenüber wird Anwesenheit erwartet.

[1] Ihr Gruppenname setzt sich aus Ihren Nachnamen in alphabetischer Reihenfolge zusammen. Bei gleichen Nachnamen ist an den Nachnamen der erste Buchstabe des Vornamens zu setzen z. B. Max Mütze → MützeM

[2] Andere Dateiformate (Keynote, Powerpoint, LibreOffice etc.) werden nicht akzeptiert!

Weiterführende Literatur

1. Greg Pass, Abdur Chowdhury, Cayley Torgeson: A picture of search. Infoscale 2006: 1

2. `http://en.wikipedia.org/wiki/AOL_search_data_leak` (letzter Zugriff 26.9.2013)

3. Daniel E. Rose, Danny Levinson: Understanding user goals in web search. WWW 2004: 13-19

4. David J. Brenes, Daniel Gayo-Avello: Stratified analysis of AOL query log. Inf. Sci. 179(12): 1844-1858 (2009)

5. Clarissa Mayer: Kategorisierung von Suchanfragen und deren Unterstützung durch Indexstrukturen. Diplomarbeit TU Berlin 2010.

6. Benjamin Bloom: Taxonomy of educational objectives: Handbook I: Cognitive Domain New York, Longmans, Green, 1956

7. G. Marchionini. Exploratory search: from searching to understanding. Communications of the ACM, 49:41-46, Apr. 2006

8. Alexander Löser, Sebastian Arnold, Tillmann Fiehn: The GoOLAP Fact Retrieval Framework. Lecture Notes in Business Information Processing Volume 96, 2012, pp 84-97

9. Memento Protokoll Informationen für Zugriff auf Wayback-Machine. `http://mementoweb.org/depot/native/ia/` (letzter Zugriff 26.9.2013)

10. `http://www.freebase.com/queries` (letzter Zugriff 26.9.2013)

11. `http://archive.org/web/web.php`
Internet Archive (letzter Zugriff 26.9.2013).

12. Sebastian Arnold: GoOLAP User Interaktion. Bachelorarbeit Technische Universität Berlin 2011.

13. Jeff Huang, Efthimis N. Efthimiadis: Analyzing and evaluating query reformulation strategies in web search logs. CIKM 2009: 77-86

14. `http://www.dmoz.org` (letzter Zugriff 30.9.2013).