Anhui University of Science and Technology

2021~2022 Academic Year: Second Term International Postgraduate Exam Paper Course Title: Searching and Mining on Big Data Exam Mode: Close-book Paper Type: A

Major: COMPUTER SCIENCE Student ID: 2021250010 Name: MAHAD SAJJAD NELOY

Binding Line

	I	II	III	IV	V	VI	VII	VIII	IX	X	Total Score
Standard Score	40	15	15	30							
Score											

I. Short Answer Questions(5 points each question, 40 points in total)

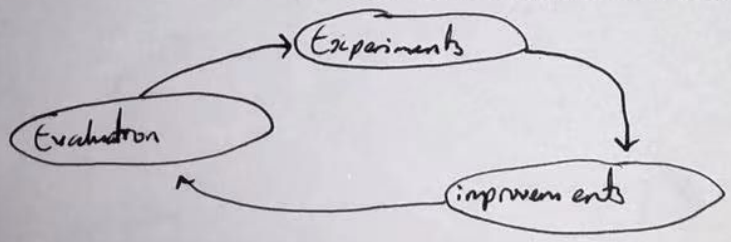
1. What is the meaning of search engine for human beings?

- Search engine means a system that brings information requirments in life and study.
- This kind of information requirements includes navigation, information and transaction.

2. What is the relationship between search engines and big data?

- In the context of big data era, search engine optimization plays a crucial role in the potential dissemination of personalized content that reflects quality.
- This quality is related to the curation of content and proper usability in the web-based systems in order to cover users information needs.

3. What is the common/universal model of performance evaluation?



4. How to build Query sample set?

→Build the query sample is based on three aspects, truth, typical and completeness.
→Assess search engine for example Baidu or google and type some words.

5. Please describe the Word segmentation method?

→ splitting a string of written language into component words
→ sgely divide the input string into substring.
→ Identify the best technique that would aid in the extraction of keyword from the text.
→ Chinese and English segmentation tools: Jieba and Spacy.

6. How to reduce the size of indexing keywords ?

→By either forward index or reverse index table.
→The data is stored in parallel storage units called buckets or fragments which share the task of index and retrieval to increase the overall throughput efficiency of the index
→ Precision recall, f-measure, average precision, string matching.

7. Please describe basis/main factors of sorting search results?

→The values differences
→ Navigation query: when querying "Exam" the link analysis results should account for the main weight.
→Information query: when querying the "graduate entrance examination", the weight of the content search results should be large.

8. Please describe there kinds of Information retrieval models.

→Boolean model → also known as Exact-Match retrieval.
• composed of keywords and logical operators that express the characteristics that the user wants the document to have.
→ Vector Space model
weights calculation in query word or document.
→ Probabilistic Model
retrieval weight to derive ranking functions.

Major: COMPUTER SCIENCE Student ID: 2021290010 Name: MAHADI SAJJAD NELOY

Binding Line

II. Comprehensive Analysis Question (15 points in total)

Please describe five steps for retrieve subsystem.

i) Receive user query words
ii) get candidate result documents from inverted index table,
iii) Calculate various parameters that affect the sorting of retrieval results.
iv) merge each parameter to get the final value used for something.
v) The sorted result document is returned to the user.

III. Comprehensive Designing Question (15 points in total)

Please describe the process from a given query to get final results.

The process from a given query to the final results is well illustrate as follows:

Given $G = (V, E)$,
if there is another graph $g' = (v', e')$,
if v' is included in V & e' is included in E,
then g' is a subgraph of G.
if v' is contained in V & e' contains all the edges between the node subset v',
g' is the derived subgraph of G.

Major: COMPUTER SCIENCE Student ID: 2021290010 Name: MAHADI SAJJAD NELOY

Binding Line

IV. Chinese literature reading and Translation(30 points in total)

数据挖掘是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示三个步骤。数据准备是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集；规律寻找是用某种方法将数据集所含的规律找出来；规律表示是尽可能以用户可理解的方式（如可视化）将找出的规律表示出来。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

Data mining is a technology that analyzes each piece of data and finds its rules from a large amount of data. It mainly includes three steps: data preparation, rule finding and rule presentation. Data preparation is to select necessary data from relevant data sources and integrate them into data sets for data mining. Law search is to find out the law contained in the dataset by some method; rule representation is to express the found rules in a way that users can understand (such as visualization) as much as possible. The task of data mining include association rule analysis, cluster analysis, classification analysis, anomaly analysis, special group analysis and evolution analysis.