**Adama Science and Technology University**

**School of Electrical Engineering and Computing**

**Department of Computer Science and**

**Engineering**

**Information Storage and Retrieval Assignments**

**Submitted by** Nitsuh Demissew

A/ur146889/10

Section 1

2012 E.c

***Chapter One:***

1. **Compare and contrast *Information* *Retrieval* and *Data* *Retrieval*.**

Information retrieval is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. **Data retrieval** means obtaining **data** from a database management system such as ODBMS.

**Both** are means of acquiring knowledge. Retrieve something to know.

**But**

– IR deals with unstructured/semi-structured data while a data retrieval (a database management system or DBMS) deals with structured data with well-defined semantics

– Querying a DBMS system produces exact/precise results or no results if no exact match is found

– Querying an IR system produces multiple results with ranking. Partial match is allowed

|  |  |  |
| --- | --- | --- |
|  | Data Retrieval | Info Retrieval |
| Data organization | Structured(Clear Semantics:Name,age…) | Unstructured(No fields (other than text)) |
| Query Language | Artificial (defined,SQL) | Free text(“natural Language”),Boolean |
| Query Specification | Complete | Incomplete |
| Items wanted | Exact Matching | Partial Best matching |
| Accuracy | 100% (results are always “correct”) | <50% |
| Error response | Sensitive | Insensitive |

1. Explain in detail the two *subsystem* of an IR system.

There are two subsystems

* **Indexing** –it is an offline process of organizing documents using keywords extracted from the collection. It is used to speed up access to desired information from document collection as per users query.

In indexing there are certain steps to follow. First the documents to be indexed should be *tokenized* and then to modified tokens. Finally the index file would be built. An ***Index*** file of a document is a file consisting of a list of index terms and a link to one or another that has the index term.

* **Searching** – it is an online process that scans document corpus to find relevant documents that matches users query.

The process of building an index file in searching is different from indexing. Parse query ---> stop list ---🡪stemming and normalize-🡪Term weighting

The query terms and index terms from the index file go through similarity measure and then there will be relevant document set. After ranking the ranked document set.

1. What are the three types of *IR* *models* that have been used for years?
2. The Boolean model
3. The vector space
4. Probabilistic model
5. Write about the two types of *evaluation* *strategies* of IR systems?

**System centered Studies** - It makes use of experimental collections consisting of: sets of documents representing domains of interest; topics, which simulate and abstract actual user information needs; and ground-truth, i.e. the “correct”' answers, where relevant documents for each topic are pre-determined.

**User centered Studies** - It focus on users’ affective and cognitive experiences with IIR systems, and the behaviors they exhibit during use or as a result of interacting with information. Thus, the goal in user-centered evaluation is to understand users’ motivations, cognitive involvement and processes, and emotional responses to systems and/or search tasks, and how this relates to system performance and other outcomes

1. What are the *challenges* for researchers and practitioners of IR?

***Technical challenge*** – They must answer what tools should IR systems provide to allow effective and efficient manipulation of information within such diverse media of text, image video and audio?

***Interaction challenge*** – They must answer what features should IR systems provide in order to support a wide variety of users in their search for relevant information.

***Evaluation challenge*** – They must answer how can we measure effectiveness of retrieval? Which features are effective and usable, given the increasing diversity of end-users and information seeking situations?

***Chapter Two:***

1. What are the statistical properties of a given text?

* How is the frequency of different words distributed?
* How fast does vocabulary size grow with the size of a corpus
* A few words are very common
* Most words are very rare

1. Explain the main operations required for selecting the index terms in text operations.

* **Lexical analysis/tokenization of the text**- digits, hyphens, punctuation marks, and the case of letters
* **Stemming words** – remove affixes (prefixes and suffixes)
* **Construction of term categorization structures -** such as thesaurus, to capture relationship for allowing the expansion of the original query with related terms.

1. What is thesaurus and what is the aim of using thesaurus?

In the context of information retrieval, a thesaurus (plural: "thesauri") is a form of controlled vocabulary that seeks to dictate semantic manifestations of metadata in the indexing of content objects. A thesaurus serves to minimize semantic ambiguity by ensuring uniformity and consistency in the storage and retrieval of the manifestations of content objects.

The prime function of a thesaurus is to support information retrieval by guiding the choice of terms for indexing and searching

***Chapter Three:***

1. Write and explain three index file *evaluation* *metrics*.

* Running time
* Indexing time
* Access/search time
* Update time (Insertion time , deletion time ,modification time…)
* Space overhead

Computer storage space consumed

* Access types supported efficiently
* Is the indexing structure allows to access:
  + - * Records with a specified term, or
      * Records with terms falling in a specified range of values

1. What is the difference between *sequential* *file* and *inverted* *file*?

|  |  |
| --- | --- |
| **Sequential Files** | **Indexed Files** |
| Sequential files are called as QSAM files. | Indexed files are also called as VSAM files. |
| Data is entered in entry sequential order. | Data is entered in key sequential order. |
| Duplicate data is allowed. | Duplicate data is not allowed. |
| Data is not in sorted order. | Data is in sorted order based on key. |
| Delete is not applicable. | Delete is applicable. |
| Access is slow. | Access is faster. |
| Key is not available. | Key is available. Key is user defined. Key is part of a record. |
| Data is stored on tape/disk. | Data is stored on disk only. |
| File modes. INPUT. OUTPUT. I-O. EXTEND. | File modes. INPUT. OUTPUT. I-O. |
| Frequently used files. | Rarely used files. |

1. What is text compression and what is its advantage?

Text compression is about finding ways to represent the text in fewer bits or bytes

**Advantages**:

* save storage space requirement
* Speed up document transmission time
* Takes less time to search the compressed text

***Chapter Four:***

1. How we can evaluate IR models?

We need to investigate what procedures they follow and what techniques they used for:

* Are they using binary or non-binary weighting for measuring importance of terms in documents?
* Are they using similarity measurements?
* Are they applying partial matching?
* Are they performing exact matching or best matching for document retrieval?
* Any Ranking mechanism?

1. What are the advantages and the disadvantages of the Boolean Model?

**Advantages**

* Clean formalism
* Easy to implement
* Intuitive concept

**Disadvantages**

* Exact matching may retrieve too few or too many documents
* Hard to translate a query into a Boolean expression
* All terms are equally weighted
* More like data retrieval than information retrieval
* Frequently returns either too few or too many documents in response to a user query

1. What are the advantages and disadvantages of the Vector space model?

**Advantage**

1. Simple model based on linear algebra
2. Term weights not binary
3. Allows computing a continuous degree of similarity between queries and documents
4. Allows ranking documents according to their possible relevance
5. Allows partial matching documents.

**Disadvantage**

1. Long documents are poorly represented because they have poor similarity values (a small scalar product and a large dimensionality)
2. Search keywords must precisely match document terms; word substrings might result in a "false positive match"
3. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".
4. The order in which the terms appear in the document is lost in the vector space representation.
5. Theoretically assumes terms are statistically independent.
6. Weighting is intuitive but not very formal.
7. Suffers from synonym and polysemy