

HPHA 5330

Health Informatics and Data Analytics

Information Retrieval

Definitions -

Database – A collection of items

Field – a single, specific, piece of information in a database

Record (document) – an individual collection of information in a database. A record is made up of fields that are related in some way

File – a collection of similar **records** in a database

Search Engine – a computer application that is used to search through a database

Database management program – an application that is used to organize and retrieve items in a database, based on desired criteria

Indexing – the act of assigning a term or terms to each record in a database to make it easier to search through the items in the database

Index term (Search field) – The name used to refer to a characteristic of a record that is used for indexing the items in a database

Query – a search request using a term or terms from the indexing vocabulary, or general terms. A query usually includes Boolean terms

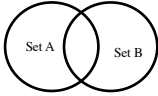
Output – The result that a computer program returns

Boolean algebra –

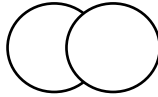
The (an) algebra of logic.
Used in discussing the philosophy of logic, designing computer logic circuits, in the mathematics of set theory, and for creating search queries

Three basic Boolean operators:
OR, AND, and NOT

The Boolean operator '**OR**'



$A \text{ OR } B = C$

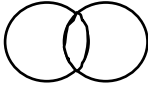


The set $C = A \text{ OR } B$

The Boolean operator "AND"

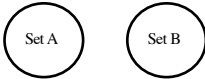


$A \text{ AND } B = C$

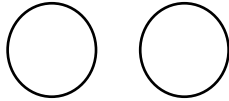


The set $C = A \text{ AND } B$

If the set A and the set B do not intersect there is no overlap, that is, there are no elements that are **common** to **both** sets,



The set $C = A \text{ OR } B$
will still be the total of all of
the elements in **the two** sets

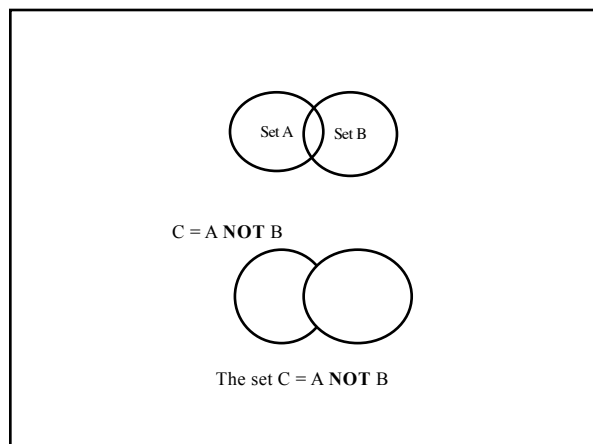


The set $C = A \text{ OR } B$

However, the set $C = A \text{ AND } B$
will be the “null” set, a set
containing no elements, since
there are no elements that belong
to **both** sets, Set A **AND** Set B

The set $C = A \text{ AND } B$

The Boolean operator **NOT**
defines all of the elements that
are contained in one set but **NOT**
in the other



Retrieval -
Finding records in a database

How 'good' is the output?

Measures of success -

Recall
Precision

Recall (R) – the proportion of relevant documents in the database that are retrieved

Recall (R) = Number of relevant documents that are retrieved /
Number of relevant documents in the entire database

Relative Recall (RR) = Number of relevant documents that are retrieved / Number of relevant documents that are retrieved by a number of searches on the topic using different query terms

Precision (P) – the proportion of relevant documents in the output

Precision (P) = Number of relevant documents that are retrieved / Total number of documents retrieved (relevant + non-relevant)

Example

Total number of documents in the database = 10,436
Total number of relevant documents in the database = 297
Number of documents retrieved by search A = 853
Number of relevant documents retrieved by search A = 255
Number of relevant documents retrieved by several searches (search A, search B, and search C) on the same topic, using different search queries = 274

Calculate the Recall

Recall (R) = Number of relevant documents that are retrieved / Number of relevant documents in the entire database

$$\text{Recall} = 255 / 297$$
$$\text{Recall} = 0.859, \text{ or } 85.9\%$$

Calculate the Relative Recall

Relative Recall (RR) = Number of relevant documents that are retrieved / Number of relevant documents that are retrieved by a number of searches on the topic using different query terms

$$\text{Relative Recall} = 255 / 274$$
$$= 0.930, \text{ or } 93\%$$

Calculate the Precision

Precision (P) = Number of relevant documents that are retrieved / Total number of documents retrieved (relevant + non-relevant)

$$\text{Precision} = 255 / 853$$
$$= 0.299, \text{ or } 29.9\%$$

What makes searching and retrieval difficult?

Polysemy – the situation
in which a word has
more than one meaning

e.g., “fair”

Fair – reasonable, equitable:
“I thought that our trade was fair.”

Fair – a light complexion:
“She has fair skin.”

Fair – an event:
“We went to the county fair.”

Synonymy – two or more words
mean the same thing

Cancer, neoplasm, tumor

Cut, laceration

Fracture, “broken bone”

Good search strategy:

Simple, Succinct, Specific

References

Hersh, WR
Information Retrieval, A Health and Biomedical Perspective
- Springer, 2008.

PubMed
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

End of the lesson

Information Retrieval
