Informatics 1

School of Informatics, University of Edinburgh

# Data and Analysis

## Part III

## Unstructured Data

Ian Stark

February 2011

# Part III — Unstructured Data

Data Retrieval:

**III.1 Unstructured data and data retrieval**

Statistical Analysis of Data:

**III.2** Data scales and summary statistics

**III.3** Hypothesis testing and correlation

**III.4** $\chi^2$ and collocations

## Staff-Student Liaison Meeting

- Today 1pm
- Informatics 1 teaching staff and student reps
- Send mail to the reps at inf1reps@lists.inf.ed.ac.uk if there with any comments you would like them to make at the meeting

## Coursework Assignment

- Three sample exam questions, download from course web page
- Due 4pm Friday 11 March, to box outside ITO
- Marked by tutors and returned for discussion in week 11 tutorial
- Not for credit; you can discuss and ask for help (do!)

## Examples of Unstructured Data

- Plain text.

  There is structure, the sequence of characters, but this is *intrinsic* to the data, not imposed.
  We may wish to impose structure by, e.g., annotating (as in Part II).

- Bitmaps for graphics or pictures, digitized sound, digitized movies, etc.

  These again have *intrinsic* structure (e.g., picture dimensions).
  We may wish to impose structure by, e.g., recognising objects, isolating single instruments from music, etc.

- Experimental results.

  Here there may be structure in how represented (e.g., collection of points in $n$-dimensional space).
  But an important objective is to uncover implicit structure (e.g., confirm or refute an experimental hypothesis).

## Topics

We consider two topics in dealing with unstructured data.

1. *Information retrieval*

   How to find data of interest in within a collection of unstructured data documents.

2. *Statistical analysis of data*

   How to use statistics to identify and extract properties from unstructured data (e.g., general trends, correlations between different components, etc.)

## Information Retrieval

The *Information retrieval (IR) task*: given a query, find the documents in a given collection that are relevant to it.

Assumptions:

1. There is a large document collection being searched.
2. The user has a need for particular information, formulated in terms of a query (typically keywords).
3. The task is to find all and only the documents relevant to the query.

Example: Searching a library catalogue. Document collection to be searched: books and journals in library collection. Information needed: user specifies query giving details about author, title, subject or similar. Search program returns a list of (potentially) relevant matches.

## Key issues for IR

Specification issues:

- Evaluation: How to measure the performance of an IR system.
- Query type: How to formulate queries to an IR system.
- Retrieval model: How to find the best-matching document, and how to *rank* them in order of relevance.

Implementation issues:

- Indexing: how to represent the documents searched by the system so that the search can be done efficiently.

The goal of this lecture is to look at the three *specification issues* in more detail.

## Evaluation of IR

The performance of an IR system is naturally evaluated in terms of two measures:

- *Precision:* What proportion of the documents returned by the system match the original objectives of the search.
- *Recall:* What proportion of the documents matching the objectives of the search are returned by the system.

We call documents matching the objectives of the search *relevant documents*.

## True/false positives/negatives

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | true positives | false positives |
| Not retrieved | false negatives | true negatives |

- *True positives (TP):* number of relevant documents that the system retrieved.
- *False positives (FP):* number of non-relevant documents that the system retrieved.
- *True negatives (TN):* number of non-relevant documents that the system did not retrieve.
- *False negatives (FN):* number of relevant documents that the system did not retrieve.

# Defining precision and recall

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | true positives | false positives |
| Not retrieved | false negatives | true negatives |

*Precision*

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

*Recall*

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

# Comparing 2 IR systems — example

Document collection with 130 documents.
28 documents relevant for a given theory.

System 1: retrieves 25 documents, 16 of which are relevant

$\mathrm{TP}_1 = 16,\ \ \mathrm{FP}_1 = 25 - 16 = 9,\ \ \mathrm{FN}_1 = 28 - 16 = 12$

$$P_1 = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FP}_1} = \frac{16}{25} = 0.64 \qquad R_1 = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FN}_1} = \frac{16}{28} = 0.57$$

System 2: retrieves 15 documents, 12 of which are relevant

$\mathrm{TP}_2 = 12,\ \ \mathrm{FP}_2 = 15 - 12 = 3,\ \ \mathrm{FN}_2 = 28 - 12 = 16$

$$P_2 = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FP}_2} = \frac{12}{15} = 0.80 \qquad R_2 = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FN}_2} = \frac{12}{28} = 0.43$$

N.B. System 2 has higher precision. System 1 has higher recall.

## Precision versus Recall

A system has to achieve both high precision and recall to perform well. It doesn't make sense to look at only one of the figures:

- If system returns all documents in the collection: 100% recall, but low precision.
- If system returns only one document, which is relevant: 100% precision, but low recall.

*Precision-recall tradeoff:* System can optimize precision at the cost of recall, or increase recall at the cost of precision.

Whether precision or recall is more important depends on the application of the system.

## F-score

The *F-score* is an evaluation measure that combines precision and recall.

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

Here $\alpha$ is a *weighting factor* with $0 \leq \alpha \leq 1$.

High $\alpha$ means precision more important. Low $\alpha$ means recall is more important.

Often $\alpha = 0.5$ is used, giving the *harmonic mean* of $P$ and $R$:

$$F_{0.5} = \frac{2PR}{P+R}$$

## Using F-score to compare — example

We compare the examples on slide III: 11 using the F-score (with $\alpha = 0.5$).

$$F_{0.5}(\text{System}_1) = \frac{2P_1R_1}{P_1 + R_1} = \frac{2 \times 0.64 \times 0.57}{0.64 + 0.57} = 0.60$$

$$F_{0.5}(\text{System}_2) = \frac{2P_2R_2}{P_2 + R_2} = \frac{2 \times 0.80 \times 0.43}{0.80 + 0.43} = 0.56$$

The F-score (with this weighting) rates System 1 as better than System 2.