Informatics 1: Data & Analysis

Lecture 15: Information Retrieval

Ian Stark

School of Informatics
The University of Edinburgh

Friday 15 March 2013
Semester 2 Week 8

http://www.inf.ed.ac.uk/teaching/courses/inf1/da

Unstructured Data

Data Retrieval

- The information retrieval problem
- The vector space model for retrieving and ranking

Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- χ^2 tests and collocations

also *chi-squared*, pronounced “kye-squared”

Unstructured Data

Data Retrieval

- The information retrieval problem
- The vector space model for retrieving and ranking

Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- χ^2 tests and collocations

also *chi-squared*, pronounced "kye-squared"

Examples of Unstructured Data

Almost all data in machine-readable form has at least *some* structure: bits, bytes, characters, files. By *unstructured* data we generally mean there is no additional annotation or data-specific structure. For example:

Plain text

No structure beyond a sequence of characters.

Graphics, photographs, digitized audio and video

A stream of values (bits, colours, sound waves) in one, two or more dimensions. File formats and compression techniques may be quite structured; but the data itself is not.

Sensor data, results from scientific experiments

Collections of points in n-dimensional space, one of which may be time, representing observations.

Analysis of Unstructured Data

Different kinds of unstructured data are open to different kinds of analysis; sometimes this then adds structure, but not always. For example:

Plain text

Analysis can add structure like POS annotations, syntax trees, etc.

Graphics, photographs, digitized audio and video

Within this unstructured data, we might annotate by recognising objects, or filtering certain sounds.

Sensor data, results from scientific experiments

The analysis task here is usually to apply statistical tests to confirm or refute an experimental hypothesis.

N.B. Crude but so far successful rule of thumb: everything is better in bits, eventually.

Topics

In this course we introduce two different areas of working with unstructured data. In each case this can only be a brief introduction, and later more specialist courses then build on these.

Information Retrieval

Finding items of interest in within a collection of unstructured documents.

Statistical Analysis

Using mathematics to identify and extract properties from unstructured data: summaries, trends, correlations, significant observations.

Information Retrieval

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

This makes some fixed assumptions about the task context:

1. There is a large document collection to be searched.
2. The user seeks some 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.

For example, when searching a library catalogue we have:

1. Books and journals in the library;
2. User supplies author, title, subject keywords, or similar information;
3. Retrieval system should return a list of potentially relevant matches.

Key Issues in Information Retrieval

Specification

- Evaluation: How to measure the performance of an IR system.
- Query type: How to formulate requests to an IR system.
- Retrieval model: Which are the most relevant documents, and how to rank them.

Implementation

- Indexing: What information to capture about the documents, and how to store it, so that the search can be done efficiently.

This course will only address the specification issues, starting with how to assess the effectiveness of an IR system, or compare two different ones.

Evaluation of Information Retrieval

In the information retrieval task, we assume that within the large document collection is a smaller set of *relevant documents* that meet the requirements of the search.

The standard performance assessment of an IR system is through two measures.

- Precision: What proportion of the documents returned by the system are relevant.
- Recall: What proportion of all the relevant documents are returned by the system.

These are appropriate measures regardless of the method used to retrieve or rank documents, its implementation, or which documents are deemed to be relevant.

Making Measures Precise

To make precise these measures for evaluation, we need some definitions.

True positives (TP): Number of relevant documents retrieved.

False positives (FP): Number of non-relevant documents retrieved.

True negatives (TN): Number of non-relevant documents not retrieved.

False negatives (FN): Number of relevant documents not retrieved.

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

Making Measures Precise

To make precise these measures for evaluation, we need some definitions.

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

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

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

Example: Comparing IR Systems

Suppose we have a collection containing 130 documents; and a query for which 28 of these are relevant.

System 1 retrieves 25 documents, of which 16 are relevant

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

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

System 2 retrieves 15 documents, of which 12 are relevant

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

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

System 2 has higher precision; System 1 has higher recall.

Precision versus Recall

In information retrieval it is not enough to consider only one performance measure alone. Suppose — as is typical — we have a collection with a large number of documents, of which some are relevant to our query.

- Consider a system which returns every document in the collection: this gives 100% recall, but very low precision.
- Consider a system which returns just one document, but it is relevant: 100% precision, but low recall.

The *precision-recall tradeoff* is that at a given level of performance, a system may be able to improve precision at the cost of recall, or increase recall at the cost of precision.

Which is more important depends on the intended application:

- A search for legal documents might need excellent recall;
- A search for second-hand cars for sale might favour precision.

The F-Score

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

$$\mathrm{F}_\alpha = \frac{1}{\alpha\frac{1}{\mathrm{P}} + (1-\alpha)\frac{1}{\mathrm{R}}}$$

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

Higher values of α, closer to 1, put more weight on precision. Lower values of α, closer to 0, put more weight on recall.

Taking $\alpha = 0.5$ gives a *balanced* F-score, the *harmonic mean* of P and R:

$$\mathrm{F}_{0.5} = \frac{1}{\frac{1}{2}\frac{1}{\mathrm{P}} + \frac{1}{2}\frac{1}{\mathrm{R}}} = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$$

Sorry, I don't know why "F"

Comparing IR Systems by F-Score

Here are the two examples from earlier, compared by balanced F-score.

System 1 had higher recall, but less precision

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

System 2 had lower recall, but better precision

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

The balanced F-score rates System 1 slightly above System 2.