

Informatics 1B, 2008  
School of Informatics, University of Edinburgh

## **Data and Analysis**

### **Note 11**

### **Unstructured Data and Information Retrieval**

**Alex Simpson**

## So far in course

- Part I: Structured data:
  - Relational databases
- Part II: Semistructured data
  - XML documents
  - Marked-up text and corpora

## Part III — Unstructured Data

**Note 11 Unstructured data and information retrieval**

**Note 12** Statistical analysis of data I

**Note 13** Statistical analysis of data II

## Unstructured data — examples

- Plain (unannotated) 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 note 9).

- 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).

## Contents of Part III

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.

The topic of this lecture,

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.)

The topic of notes 12 and 13.

## 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: Internet search engine.** Document collection to be searched: web pages. Information need: find pages on a particular topic. Query: user specifies keywords. Search engine returns a ranked list of relevant web pages.

## Key issues for IR

### Specification issues:

- **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.
- **Evaluation:** How to measure the performance of the system.

### Implementation issues:

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

## Simple queries

We shall only consider *simple queries* of the form:

- Find documents containing *word1, word2, ..., wordn*

More specific tasks are:

- Find documents containing all the words *word1, word2 ... wordn*;
- or find documents containing as many of the words *word1, word2 ... wordn* as possible.

In real-world applications, queries can be more complex than this (e.g., they can be combined using boolean operations, one can search for substrings of words or whole phrases, etc.).

## Recognising words

The task, as we have specified it, relies on recognising *words* in documents.

Nevertheless, this task is typically performed in contexts in which words are not specifically annotated as such in documents.

For example, on a typical webpage, the text appearing on the webpage is written in HTML as contiguous text; words are not marked up (in contrast to a corpus, such as the BNC, where **w** tags mark words).

So one task of the IR system is to identify word tokens.

This is typically done using simple separation criteria for words, e.g., using white space and punctuation as word separators.



## A retrieval model

If all documents containing all words of the query are returned this might result in a large number of varying relevance (at least if the document collection is large and the query general).

IR systems need to rank documents according to likely relevance.

There are many such ranking methods.

We focus on one, which uses the *vector space model*.

This model is the basis of many IR applications.

In this course, we shall only use it in one particularly simple way.

## The vector space model

### Core ideas:

- Treat documents as points in a high-dimensional vector space, based on words in the document collection.
- The query is treated in the same way.
- The documents are ranked according to document-query similarity.

**N.B.** You do not need to know much about vector spaces to understand this.

## The vector associated to a document

Suppose  $\text{Term}_1, \text{Term}_2, \dots, \text{Term}_n$  are all the different words occurring in the entire collection of documents  $\text{Doc}_1, \text{Doc}_2, \dots, \text{Doc}_N$ .

Each document,  $\text{Doc}_i$ , is assigned an  $n$ -valued vector:

$$(m_{i1}, m_{i2}, \dots, m_{in})$$

where  $m_{ij}$  is the number of times word  $\text{Term}_j$  occurs in document  $\text{Doc}_i$ .

Similarly, the query is assigned an  $n$ -valued vector by considering it as a document itself.

## Example

Consider the document

*Sun, sun, sun, here it comes*

and suppose the only words in the document collection are: *comes, here, it, sun*.

The vector for the document is  $(1, 1, 1, 3)$

comes	here	it	sun
1	1	1	3

Similarly, the vector for the query *sun comes* is  $(1, 0, 0, 1)$

## Document matrix

The frequency information for words in the document collection is normally precompiled in a *document matrix*.

This has:

- Columns represent the words appearing the document collection
- Rows represent each document in the collection.
- each entry in the matrix represent the frequency of the word in the document.

## Document matrix — example

	Term <sub>1</sub>	Term <sub>2</sub>	Term <sub>3</sub>	...	Term <sub>n</sub>
Doc <sub>1</sub>	14	6	1	...	0
Doc <sub>2</sub>	0	1	3	...	1
Doc <sub>3</sub>	0	1	0	...	2
...	...	...	...	...	...
Doc <sub>N</sub>	4	7	0	...	5

**N.B.** Each row gives the vector for the associated document.

## Vector similarity

We want to rank documents according to relevance to the query.

We implement this by defining a measure of *similarity* between vectors.

The idea is that the most relevant documents are those whose vectors are most similar to the query vector.

Many different similarity measures are used. A simple one that is conceptually appealing and enjoys some good properties is the *cosine* of the angle between two vectors.

## Cosines (from school trigonometry)

Recall that the *cosine* of an angle  $\theta$  is:

$$\frac{\text{adjacent}}{\text{hypotenuse}}$$

in a right-angled triangle with angle  $\theta$ .

Crucial properties:

$$\cos(0) = 1 \quad \cos(90^\circ) = 0 \quad \cos(180^\circ) = -1$$

More generally, two  $n$ -dimensional vectors will have cosine: **1** if they are identical, **0** if they are orthogonal, and **-1** if they point in opposite directions.



## Vector cosines

Suppose  $\vec{x}$  and  $\vec{y}$  are  $n$ -value vectors:

$$\vec{x} = (x_1, \dots, x_n) \qquad \vec{y} = (y_1, \dots, y_n)$$

Their *cosine* (that is, the cosine of the angle between them) is calculated by:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

## Vector cosines — example

Continuing the example from slide 11.12, suppose:

$$\vec{x} = (1, 1, 1, 3) \qquad \vec{y} = (1, 0, 0, 1)$$

Then:

$$\vec{x} \cdot \vec{y} = 1 + 0 + 0 + 3 = 4$$

$$|\vec{x}| = \sqrt{1 + 1 + 1 + 9} = \sqrt{12}$$

$$|\vec{y}| = \sqrt{1 + 0 + 0 + 1} = \sqrt{2}$$

So

$$\cos(\vec{x}, \vec{y}) = \frac{4}{\sqrt{12} \times \sqrt{2}} = \frac{2}{\sqrt{6}} = 0.82$$

to two decimal places.

## Ranking documents

Suppose  $\vec{y}$  is the query vector, and  $\vec{x}_1, \dots, \vec{x}_N$  are the  $N$  document vectors.

We calculate the  $N$  values:

$$\cos(\vec{x}_1, \vec{y}), \dots, \cos(\vec{x}_N, \vec{y})$$

The documents are then ordered so that those with the highest cosine values are counted as most suitable, and those with the lowest cosine values are counted as least suitable.

**N.B.** On this slide  $\vec{x}_1, \dots, \vec{x}_N$  are  $N$  (potentially) different vectors, each with  $n$  values.

## Discussion of cosine measure

The cosine similarity measure, as discussed here, is very crude.

- It only takes word frequency into account
- It takes all words in the document collection into account (whether very common “stop” words which are useless for IR, or very uncommon words unrelated to the search)
- All words in the document collection are weighted equally
- It ignores document size (just the angles between vectors not their magnitude are considered)

Nevertheless, the cosine method can be refined in various ways to avoid these problems. (This is beyond the scope of this course.)

## 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.

**N.B.** These measures are relevant to all IR systems, not just to ones based on the vector space model and cosine test.

## 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*

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

### *Recall*

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

## 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

$$TP_1 = 16, \quad FP_1 = 25 - 16 = 9, \quad FN_1 = 28 - 16 = 12$$

$$P_1 = \frac{TP_1}{TP_1 + FP_1} = \frac{16}{25} = \mathbf{0.64} \quad R_1 = \frac{TP_1}{TP_1 + FN_1} = \frac{16}{28} = \mathbf{0.57}$$

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

$$TP_2 = 12, \quad FP_2 = 15 - 12 = 3, \quad FN_2 = 28 - 12 = 16$$

$$P_2 = \frac{TP_2}{TP_2 + FP_2} = \frac{12}{15} = \mathbf{0.80} \quad R_2 = \frac{TP_2}{TP_2 + FN_2} = \frac{12}{28} = \mathbf{0.43}$$

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



## Precision-Recall Tradeoff

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 11.24 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.

## Other issues

- Precision and recall, as defined, only evaluate the set of documents returned, they do not take *ranking* into account. Other more complex evaluation measures can be introduced to deal with ranking (e.g., *precision at a cutoff*).
- We have not considered the efficient implementation of the search for documents matching a query. This is often addressed using a purpose-built index such as a *inverted index* which indexes all documents using the words in the document collection as keys.
- Often useful ranking methods make use of information extraneous to the document itself. E.g., Google's *pagerank* method evaluates documents according to their degree of *connectivity* with the rest of the web (e.g., number of links to page from other pages).

These are important issues, but are beyond the scope of Data and Analysis.