Informatics 1B: Data and Analysis

Lecture 11: Semi-structured Data: Information Retrieval

Frank Keller

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

February 7, 2005

Reading: lecture notes; Manning and Schütze (1999: ch. 15)

Information Retrieval

Last lecture:

- information retrieval (IR): based on a user query, return relevant documents from a document collection;
- need to index documents: extract terms that can be matched against the query;
- simplest case: inverted index: list of all words in the document collection.

This lecture:

- vector space model of IR: rank documents by relevance;
- evaluation: verify if you're returning the right documents.

Vector Space Model

An IR system using an inverted index could simply return all documents that contain all the words in the query.

Problem: this often returns a large number of documents (if document collection is large and query is general).

IR system needs to perform *document ranking* by relevance.

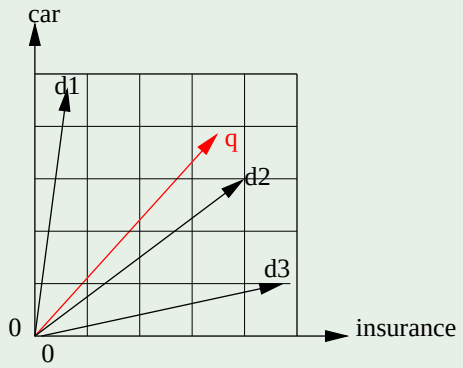
There are lots of ranking methods; we will focus on the *vector space model of document relevance.* Core idea:

- treat documents as points in high-dimensional vector space, based on words they contain;
- also represent queries in vector space;
- return documents with the highest document-query similarity.

Example

Example: document-query similarity

Query: *car insurance*: two dimensional vector space with terms *car* and *insurance* as dimensions. Vector q represents query; vectors d_1, d_2, and d_3 represent documents: d_2 closest to q.



Data Structures

Data structure for the vector space model:

- document vectors created by tabulating the frequencies of the terms in a document;
- matrix as data structure: columns represent terms, rows represent documents;
- each cell of the matrix specifies how often the term occurs in the document;
- query is represented in the same way
- vectors for computing similarity are the rows in the matrix.

Example

Example: document matrix

Matrix containing the vectors for a document collection, with $\text{Doc}_1 \ldots \text{Doc}_N$ as rows, and terms $\text{Term}_1 \ldots \text{Term}_N$ as columns; Q is the query vector.

	Term_1	Term_2	Term_3	...	Term_n
Doc_1	14	6	1	...	0
Doc_2	0	1	3	...	1
Doc_3	0	1	0	...	2
...	...	...	...	...	...
Doc_N	4	7	0	...	5
Q	0	1	0	...	1

Vector Similarity

Formally: a document in the collection is represented as a vector of n values, the term frequencies:

(1) $\vec{x} = (x_1, x_2, \ldots, x_n)$

Now use a *vector similarity measure* to determine the similarity between document vector $\vec{x}$ and query vector $\vec{y}$.

A number of similarity measures are available. *Cosine* (angle between $\vec{x}$ and $\vec{y}$) is successful for IR:

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

Other Issues in IR

Things we can do to improve the performance of the basic vector space model:

- *term weighting:* implicit in the model: vectors contain term frequencies: high frequency terms more important for retrieval;
- better weighting schemes use combination of term frequency, document frequency, collection frequency;
- *normalization:* factor the effect of document length: divide term frequency by number of terms in this document;
- *term manipulation:* modify the terms that end up in the vectors: stop word removal, stemming.

See Manning and Schütze (1999: Ch. 15) for details.

Evaluation

Evaluation: measuring the performance of a system (e.g., IR system) by comparing its output against pre-defined criteria:

- demonstrates systematically that a system achieves the task it is designed for;
- compares objectively the performance of several systems on the same task.

Discuss evaluation techniques using IR as example; these techniques are applicable in many areas of informatics.

Precision and Recall

Precision: how many of the documents the system retrieved are correct (i.e., relevant to the query).

Recall: how many of the relevant documents in the collection the system managed to find.

Precision and recall can be defined in terms of:

- *True positives* (TP): number of relevant documents that the system retrieved.
- *False positives* (FP): number of non-relevant document 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.

Confusion Matrix

Confusion matrix tabulates these quantities:

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

We can now formally define precision and recall as:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

Example

Example: comparing two IR systems

Document collection with 130 documents, 28 of which are relevant for a given query. System 1 retrieves 25 documents, 16 of which are relevant: TN $= 16$; FP $= 25 - 16 = 9$; FN $= 28 - 16 = 12$.

$$P_1 = \frac{TP_1}{TP_1 + FP_1} = \frac{16}{16+9} = .64 \quad R_1 = \frac{TP_1}{TP_1 + FN_1} = \frac{16}{16+12} = .57$$

System 2 retrieves 15 documents, 12 of which are relevant for the query: TP $= 12$; FP $= 3$; FN $= 16$.

$$P_2 = \frac{TP_2}{TP_2 + FP_2} = \frac{12}{12+3} = .80 \quad R_2 = \frac{TP_2}{TP_2 + FN_2} = \frac{12}{12+16} = .43$$

Sys. 2 beats sys. 1 for precision; sys. 1 beats sys. 2 for recall.

Precision-Recall Tradeoff

A system has to achieve both high precision and recall to perform well; 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.

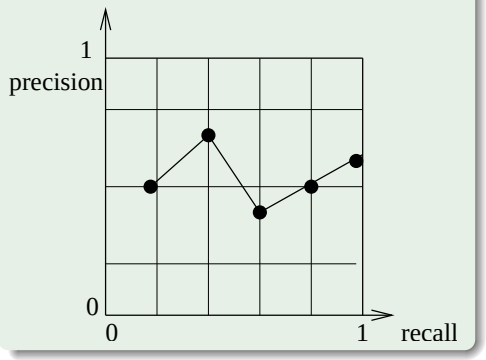
Precision-recall tradeoff: system can optimize precision at the cost of recall, or increase recall at the cost of precision.

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

Precision-Recall Curves

Precision-recall curves are a way of visualizing the precision-recall tradeoff for a particular system.

Example



F-Score

F-score: evaluation measure that combines precision and recall:

$$(5) \quad F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

α: weighting factor: high α means recall is more important; low α means precision is more important.

Often $\alpha = .5$ is used: harmonic mean of P and R:

$$(6) \quad F_{.5} = \frac{2PR}{P+R}$$

Example

Example: comparing two IR systems using F-score

Return to the two systems of the previous example. Now compare them in terms of F-score:

$$F_{.5}(\text{System}_1) = \frac{2P_1R_1}{P_1 + R_1} = \frac{2 \cdot .64 \cdot .57}{.64 + .57} = .60$$

$$F_{.5}(\text{System}_2) = \frac{2P_2R_2}{P_2 + R_2} = \frac{2 \cdot .80 \cdot .43}{.80 + .43} = .56$$

System 1 beats system 2 in terms of F-score.

Precision at a Cutoff

We introduced the vector space model as a way of *ranking* the results output by an IR system.

Standard precision and recall measures only talk about relevant documents, don't take ranking into account.

New evaluation measure to address this: *precision at a cutoff*. See lecture notes for details.

Summary

- vector space model for IR: ranks documents by relevance to the query;
- idea: treat documents and queries as vectors: term frequencies as elements of the vectors;
- the compare the query and documents using vector similarity measures; relevant documents more similar to query;
- evaluation: systematic measurement of system performance;
- precision: number of documents retrieved that are relevant; recall: number of documents retrieved out of all relevant ones;
- precision-recall tradeoff; precision-recall curve; F-score.

References

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.