Tutorial 7: Information Retrieval

Informatics 1 Data & Analysis — Tutorial Notes

Week 9, Semester 2, 2016/17

This worksheet has three parts: tutorial *Questions*, followed by some *Examples* and their *Solutions*.

- Before your tutorial, work through and attempt all of the Questions in the first section. If you get stuck or need help then ask a question on *Piazza*.
- The Examples are there for additional preparation, practice, and revision.
- Use the Solutions to check your answers, and read about possible alternatives.

You must bring your answers to the main questions along to your tutorial. You will need to be able to show these to your tutor, and may be exchanging them with other students, so it is best to have them printed out on paper.

If you cannot do some questions, write down what it is that you find challenging and use this to ask your tutor in the meeting.

Tutorials will not usually cover the Examples, but if you have any questions about those then write them down and ask your tutor, or post a question on *Piazza*.

It's important both for your learning and other students in the group that you come to tutorials properly prepared. Students who have not attempted the main tutorial questions will be sent away from the tutorial to do them elsewhere and return later.

Some exercise sheets contain material marked with a star ⋆. These are optional extensions.

Data & Analysis tutorials are not formally assessed, but the content is examinable and they are an important part of the course. If you do not do the exercises then you are unlikely to pass the exam.

Attendance at tutorials is obligatory: if you are ill or otherwise unable to attend one week then email your tutor, and if possible attend another tutorial group in the same week.

Please send any corrections and suggestions to Ian.Stark@ed.ac.uk

Introduction

This tutorial is about *Information Retrieval* (IR). It deals with two aspects of the information retrieval task discussed in lectures: evaluating performance of IR systems, and methods for document ranking. Note that these exercises are running concurrently with the Inf1-DA assignment. Your tutorial meeting will be to help with that, too: please come prepared to discuss your progress on the assignment.

Question 1: Evaluating an Information Retrieval System

Consider the following hypothetical information retrieval scenario. Suppose it has been found at Edinburgh Royal Infirmary that due to equipment malfunction, the results of blood tests taken on 2016-12-01 are unreliable for diabetic patients. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

(a) Calculate the *precision* and *recall* for this system, showing the details of your calculations.

(Tutor Notes) Starting from the scenario description, we can present the relevant information as a table:

	Relevant	Not relevant	Total
Retrieved	**125** TP	125 FP	**250**
Not retrieved	25 FN	9725 TN	9750
Total	**150**	9850	**10000**

Here the **bold** values indicate those specifically mentioned in the scenario, and the others are then computed from these. The calculations are:

$$\begin{array}{llll} \text{True Positives} & TP = 125 & \text{False Negatives} & FN = 150 - 125 = 25 \\ \text{False Positives} & FP = 250 - 125 = 125 & \text{True Negatives} & TN = (10000 - 250) - (150 - 125) \\ & & & \quad = 9725 \end{array}$$

From these we can calculate precision and recall.

$$\text{Precision } P = \frac{TP}{TP + FP} = \frac{125}{125 + 125} = 0.5 \quad \text{Recall } R = \frac{TP}{TP + FN} = \frac{125}{125 + 25} = 0.83$$

(b) Based on your results from (a), explain what the two measures mean for this scenario. How well would you say that the hospital's information retrieval system works?

(Tutor Notes) Precision of 0.5 means that 50% of the patients whose records were retrieved by the system did indeed take a blood test on 2016-12-01 and are diabetic. Recall of 0.83 means that 83% of records for diabetic patients who took a blood test on that date were retrieved by the system.

The hospital's IR system seems to be performing moderately at recall, but not very well at precision.

(c) According to the precision-recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall?

(Tutor Notes) Unless the underlying performance of the IR system is uniformly improved, tuning to achieve recall of 100% will likely reduce precision towards zero. Although we may retrieve almost all relevant records, these will be among a much larger number of irrelevant records also retrieved.

(d) For the given scenario, which measure do you think is more important, precision or recall? Why? Given your answer, what value would you give to the weighting factor α when calculating the F-score measure for the hospital's IR system?

(Tutor Notes) In the given scenario, recall is more important than precision. This means that we would prefer to retrieve all records for diabetic patients who took a blood test on 2016-12-01, even if this had the cost of also retrieving records for others that did not take a blood test on that day, or are not diabetic. This is because we wish to avoid missing a patient whose serious illness was not diagnosed at the time.

Since recall is more important for this scenario, the weighting factor α should be given a low value, say 0.3. (Any value under 0.5 would be acceptable here.)

⋆ **(e)** Last semester, in *Informatics 1: Computation and Logic*, you encountered the properties of *soundness* and *completeness* for a logic. Can you relate them to precision and recall of an IR system?

(Tutor Notes) A system of logic is *sound* if it can prove only true statements. Similarly, an IR system with 100% precision returns no irrelevant documents. A system of logic is *complete* if it can prove all true statements. Similarly, an IR system with 100% recall returns all relevant documents.

We can loosely relate soundness to precision, and completeness to recall.

Question 2: Ranking Documents

You are looking for information on the **Economic Growth in Scotland** in a large document collection. You decide to search using the terms: **economy**, **growth**, **Scotland**, **banks** and **business** using an information retrieval system and this recommends three possible documents. You are given the frequency of each of the terms in each document, shown in the table below:

Terms	economy	Scotland	growth	banks	business
Document 1	10	8	0	2	1
Document 2	0	0	9	9	8
Document 3	2	2	4	4	6
Query	1	1	1	1	1

You have no additional information about the documents; and to actually retrieve any one document will cost money.

(a) One possible measure for determining which of the 3 documents is the *cosine similarity measure*, which measures the cosine of the angle between the query vector and that of each document. Compute this measure for each of the three documents.

(Tutor Notes) Here are the required cosine calculations:

$$\cos(\vec{query}, \vec{Document1}) = \frac{10+8+0+2+1}{\sqrt{5}\sqrt{100+64+0+4+1}} = \frac{21}{29.07} = 0.72$$

$$\cos(\vec{query}, \vec{Document2}) = \frac{0+0+9+9+8}{\sqrt{5}\sqrt{0+0+81+81+64}} = \frac{26}{33.62} = 0.77$$

$$\cos(\vec{query}, \vec{Document3}) = \frac{2+2+4+4+6}{\sqrt{5}\sqrt{4+4+16+16+36}} = \frac{18}{19.49} = 0.92$$

(b) Based on your results of (a), which document is the best match for this query? Why?

(Tutor Notes) The one with the the largest cosine (indicating the smallest angle between the vectors, and closest alignment between query and document) would be taken as the best match. In our case, this is Document 3, followed by Document 2 and then Document 1.

(c) Do you agree with the results of this analysis? What are the strengths and weaknesses of cosine measure?

(Tutor Notes) In this case, the cosine measure does yield a reasonable choice of best match. Document 1 may not be about growth at all. Document 2 seems unlikely to be about Scotland.

In general, however, the cosine measure is very crude. The only information it takes into account is relative proportions of word frequencies. Moreover, as used here, all query words are weighted equally. You might be able to suggest other means of measuring relevance that are more useful for such searches.

Another — slightly technical — point is that the method used in this question is not quite a correct implementation of the cosine measure as introduced in the lectures. In the lecture, the vectors were indexed by all words occurring in the document collection, not just by words in the query. This does affects the results of the calculation (it boosts the relevance score for documents in which the query words occur relatively often compared with other words).

Examples

This section contains further exercises on information retrieval. All are based on parts of past exam papers. Following these there is a section presenting solutions and notes on all the examples.

Example 1

(a) What is the *information retrieval task*? Give an example of such a task, indicating how it matches your description.

(b) The performance of an information retrieval system can be evaluated in terms of its *precision*, P, and *recall*, R. Give an English-language definition of these two terms.

(c) Precision and recall are computed as follows:

$$P = \frac{TP}{TP + FP} \qquad\qquad R = \frac{TP}{TP + FN}$$

Name and define the three values TP, FP, FN appearing here.

(d) Two retrieval systems, X and Y, are being compared. Both are given the same query, applied to a collection of 1500 documents. System X returns 400 documents, of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents relevant to the query.

Tabulate the results for each system, and compute the precision and recall for both X and Y. Show your working.

(e) Both precision and recall need to be taken into account when evaluating retrieval systems. Why is it not sufficient to pick one and use only that?

(f) The *F-score* is a measure which combines both measures using a *weighting factor α*, where high α means that precision is more important. Give the formula defining the F-score for weighting α.

(g) How is F-score related to the *harmonic mean*?

(h) For the example task you gave in part (a), suggest an appropriate weighting factor α. Justify your choice.

Example 2

Suppose you wish to find economic reports regarding the impact of oil extraction in the North Sea on the Scottish economy. A commercial document retrieval service offers the following suggested matches: the table shows how often some key phrases appear in each report.

	North Sea	oil	Scotland	economy
Report A	12	0	3	24
Report B	10	5	20	10
Report C	0	12	9	8
Query	1	1	1	1

Actually obtaining the reports will cost real money, so you would like to select the one most likely to be relevant. Your task now is to assess this using the cosine similarity measure.

(a) Write out the general formula for calculating the cosine of the angle between two 4-dimensional vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

(b) Use this formula to rank the three documents in order of relevance to the query according to the cosine similarity measure. What do you think of the results?

Solutions to Examples

These are not entirely "model" answers; instead, they indicate a possible solution. Remember that not all of questions necessarily have a single "right" answer. If you have difficulties with a particular example, or have trouble following through the solution, please raise this as a question in your tutorial.

Solution 1

(a) The *information retrieval task* is to find those documents relevant to a user query from among some large collection of documents.

For example, searching for previous legal rulings relevant to a certain topic from a judicial archive. The judicial archive is the document collection; the query is some words related to the topic; and the previous rulings are the relevant documents to be retrieved.

Other examples are possible, of course; but you would still need to identify the document collection, the query, and which documents are relevant.

(b) Precision records what proportion of the documents retrieved do in fact match the query; recall is the proportion of relevant documents in the collection which are successfully retrieved.

This kind of question is often referred to as "bookwork" — however, even though the required information can indeed be found in books, it's still important to be able to explain it clearly in any given context.

(c) Here are the full names and definitions for the three terms.

- *TP* is *True Positives*, the number of relevant documents correctly returned.
- *FP* is *False Positives*, the number of irrelevant documents returned.
- *FN* is *False Negatives*, the number of relevant documents incorrectly rejected.

Note that this question asks you to both "name" and "define" the values, so it wouldn't be enough to say just "True Positives": you need the definition as well.

(d) "Tabulate" means to exhibit in a table, so this question requires a table showing the results for each system.

X	Relevant	Not relevant	Total
Retrieved	40	360	400
Not retrieved	10	1090	1100
Total	50	1450	1500

Y	Relevant	Not relevant	Total
Retrieved	15	15	30
Not retrieved	35	1435	1470
Total	50	1450	1500

$$\text{System } X \text{ precision } P = \frac{40}{400} = 0.1 \qquad \text{System } Y \text{ precision } P = \frac{15}{30} = 0.5$$

$$\text{System } X \text{ recall } R = \frac{40}{50} = 0.8 \qquad \text{System } Y \text{ recall } R = \frac{15}{50} = 0.3$$

(e) Depending on just one out of precision and recall can lead to extreme but unhelpful solutions. A system that returns every document indiscriminately has 100% recall; while one that returns only a single correct document is 100% precise. As information retrieval systems, the first is no help at all, and the second is not much better.

(f) Here is the formula for F-score in terms of α.

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

(g) For $\alpha = 0.5$ the $F_{0.5}$-score, or *balanced* F-score is the harmonic mean of precision and recall.

$$F_{0.5} = \frac{1}{\frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R}} = \frac{2PR}{P+R}$$

(h) For the retrieval of legal judgements, recall is of particular importance (you really don't want to miss anything), so value of α below 0.5, say 0.2, might be appropriate.

For other examples, either recall or precision might be more important, depending on the exact choice of example.

Solution 2

(a) The cosine formula for 4-vectors is:

$$\cos(\vec{x},\vec{y}) = \frac{x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4}{\sqrt{x_1^2+x_2^2+x_3^2+x_4^2}\sqrt{y_1^2+y_2^2+y_3^2+y_4^2}}$$

It's also possible to give a more compact presentation using vector notation:

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{|\vec{x}||\vec{y}|}$$

although that's only useful if you are confident in how to then calculate the dot product and modulus of 4-dimensional vectors.

(b) For the three reports listed, the appropriate calculation is the cosine between each report and the original query.

$$\begin{aligned}
\cos(\text{Report A}, \text{Query}) &= \frac{12+3+24}{\sqrt{4}\sqrt{12^2+3^2+24^2}} &= \frac{39}{54} = 0.72 \\
\cos(\text{Report B}, \text{Query}) &= \frac{10+5+20+10}{\sqrt{4}\sqrt{10^2+5^2+20^2+10^2}} &= \frac{45}{50} = 0.90 \\
\cos(\text{Report C}, \text{Query}) &= \frac{12+9+8}{\sqrt{4}\sqrt{12^2+9^2+8^2}} &= \frac{29}{34} = 0.85
\end{aligned}$$

The best fit is where the cosine is largest, closest to 1. This ranks the three documents in order of similarity to the query as:

- Report B
- Report C
- Report A

These results seem reasonable: Report B is the only document which contains all the keywords; while Report C does mention oil it doesn't mention the North Sea specifically; and Report A doesn't mention oil at all. The superiority of C over A seems clear in the cosine measure, but I don't think it is altogether obvious from simply inspecting the numbers.

Notice that it's *not* necessary to take the inverse cosine and compute the actual angle between the vectors. The questions doesn't ask for this. However if you did, then the best match would be the smallest angle, closest to 0.