TDT4117_2

paaedwl - Pål-Edward Larsen, johannkv - Johannes Kvamme

October 2018

1 Relevance Feedback

1.1 Explain the difference between automatic local analysis and automatic global analysis

The differnce is that global analysis uses the collection of all documents to build a thesaurus, while the local analysis retrieves documents from the initial query and builds a thesaurus based on those documents.

1.2 What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

The purpose behind relevance feedback is to retrieve the results given from the user's initial query, the user indicates which are relevant, and use the information about if the results are relevant to perform a new query. In a hope to give a more succesfull retrieval.

Query Expansion:
Query expansion is to reformulate a query to improve the effectiveness of the search and hopefully retrieve a 'better' result for the user. Retrieving more relevant and less not relevant documents.

Term Re-weighting:
Re-weighting is to use the statistics found in the retrieved documents. In non relevant documents it will reduce the weight of the terms in the documents while it will increase the weight of terms in relevant documents. It does not use query expansion.

The difference between the two is that query expansion takes user input to possibly retrieve a new set of (hopefully) more relevant documents while term re-weighting takes the set of retrieved documents and will sort them based on the relevant and non relevant documents.

2 Language model

2.1 Explain the language model, what are the weaknesses and strengths of this model?

A language model is a probability distribution model of regularities in a language. The language model's purpose is to separate words from a sentence by their relevance.

The strengths of the language model are:

1. The concept of the language model is simple and easy to explain
2. Has a formal mathematical model
3. Uses mostly statistics and not heuristics

The weaknesses of the language model are:

1. Depends on that the language model are accurate representations of data
2. Difficult to know what the user wants
3. Difficult to improve the relevance

2.2 Given the following documents and queries, build the language model according to the document collection

d1 = An apple a day keeps the doctor away.
d2 = The best doctor is the one you run to and can't find.
d3 = One rotten apple spoils the whole barrel.
q1 = doctor
q2 = apple orange
q3 = doctor apple

Terms	Doc1	Doc2	Doc3
an	1/9	0	0
apple	1/9	0	1/8
a	1/9	0	0
day	1/9	0	0
keeps	1/9	0	0
the	1/9	2/13	1/8
doctor	1/9	1/13	0
away	1/9	0	0
.	1/9	1/13	1/8
best	0	1/13	0
is	0	1/13	0
one	0	1/13	1/8
you	0	1/13	0
run	0	1/13	0
to	0	1/13	0
and	0	1/13	0
can't	0	1/13	0
find	0	1/13	0
rotten	0	0	1/8
spoils	0	0	1/8
whole	0	0	1/8
barrel	0	0	1/8

Table 1: Unigram language model

Applying $P(t|Md) = (1-\lambda)\hat{p}mle(t|Md) + \lambda\hat{p}mle(t|C), \lambda = 0.5$

Doc/Query	doctor
doc1	$(1-0.5)*(\frac{1}{9}+\frac{2}{30}) = 0.0889$
doc2	$(1-0.5)*(\frac{1}{13}+\frac{2}{30}) = 0.0718$
doc3	$(1-0.5)*(\frac{0}{8}+\frac{2}{30}) = 0.0333$

Table 2: Query 1

Doc/Query	apple orange
doc1	$((1-0.5)*(\frac{1}{9}+\frac{2}{30}))*((1-0.5)*(\frac{0}{9}+\frac{0}{30})) = 0$
doc2	$((1-0.5)*(\frac{0}{13}+\frac{2}{30}))*((1-0.5)*(\frac{0}{13}+\frac{0}{30})) = 0$
doc3	$((1-0.5)*(\frac{1}{8}+\frac{2}{30}))*((1-0.5)*(\frac{0}{8}+\frac{0}{30})) = 0$

Table 3: Query 2

Doc/Query	doctor apple
doc1	$((1-0.5)*(\frac{1}{9}+\frac{2}{30}))*((1-0.5)*(\frac{1}{9}+\frac{2}{30}))=0.0079$
doc3	$((1-0.5)*(\frac{0}{8}+\frac{2}{30}))*((1-0.5)*(\frac{1}{8}+\frac{2}{30}))=0.0032$
doc2	$((1-0.5)*(\frac{1}{13}+\frac{2}{30}))*((1-0.5)*(\frac{0}{13}+\frac{2}{30}))=0.0024$

Table 4: Query 3

2.3 Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.

Smoothing is to fine tune the ranking function of a language model. This is done by avoiding zero probability to document terms by using statistics from the entire document collection. This will smooth out the model.
A popular way of smoothing is to move some mass probability from the terms in the document to the terms not in the document.
Jelinek-Mercer method uses λ (set empirically) between 0 and 1. When λ is closer to 1 the influence of the term document frequency is higher and when λ is closer to 0 the influence of the term collection frequency is higher.
Bayesion method uses Dirichlet distribution. The λ parameter has the same function as with Jelinek-Mercer, but when it is larger the effect of the smoothing is higher.
For example, with the query "doctor apple", we used Jelinek-Mercer with λ at 0.5 for a medium influence, and used statistics on both terms, "doctor" and "apple" from the entire document collection.

3 Evaluation of IR Systems

3.1 Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system

R is the set of relevant documents and —R— is the number of documents in the set. A is the set of answers generated by a retrieval algorithm and —A— is the number of documents in the set. $|R \cap A|$ is the number of documents in the intersection.
Precision is the fraction of the retrieved documents (of set A) which is relevant. It describes the accuracy of the returned results. With 100% precision it would say that all the documents retrieved by a query are relevant, but not if all relevant documents were retrieved. Precision = p = $\dfrac{|R \cap A|}{|A|}$
Recall is the fraction of the relevant document (of set R) which has been retrieved. It describes the amount documents retrieved that are relevant to the query. With 100% recall it would retrieve all of the relevant documents, but not how many aren't. Recall = r = $\dfrac{|R \cap A|}{|R|}$

Precision and recall are metrics that descirbe a document's relevancy.

3.2 Given the following set of relevant documents, *rel*, and the set of retrieved documents, *ret*, provide a table with the calculated precision and recall at each level

rel = 23, 10, 33, 500, 70, 59, 82, 47, 72, 9
ret = 55, 500, 2, 23, 72, 79, 82, 215
$\|R\| = 10$
$\|A\| = 8$

Retrieved	Relevant	Precision	Recall
d55	No	$0/1 = 0$	$0/10 = 0$
d500	Yes	$1/2 = 0.5$	$1/10 = 0.1$
d2	No	$1/3 = 0.33$	$1/10 = 0.1$
d23	Yes	$2/4 = 0.5$	$2/10 = 0.2$
d72	Yes	$3/5 = 0.6$	$3/10 = 0.3$
d79	No	$3/6 = 0.5$	$3/10 = 0.3$
d82	Yes	$4/7 = 0.57$	$4/10 = 0.4$
d215	No	$4/8 = 0.5$	$4/10 = 0.4$

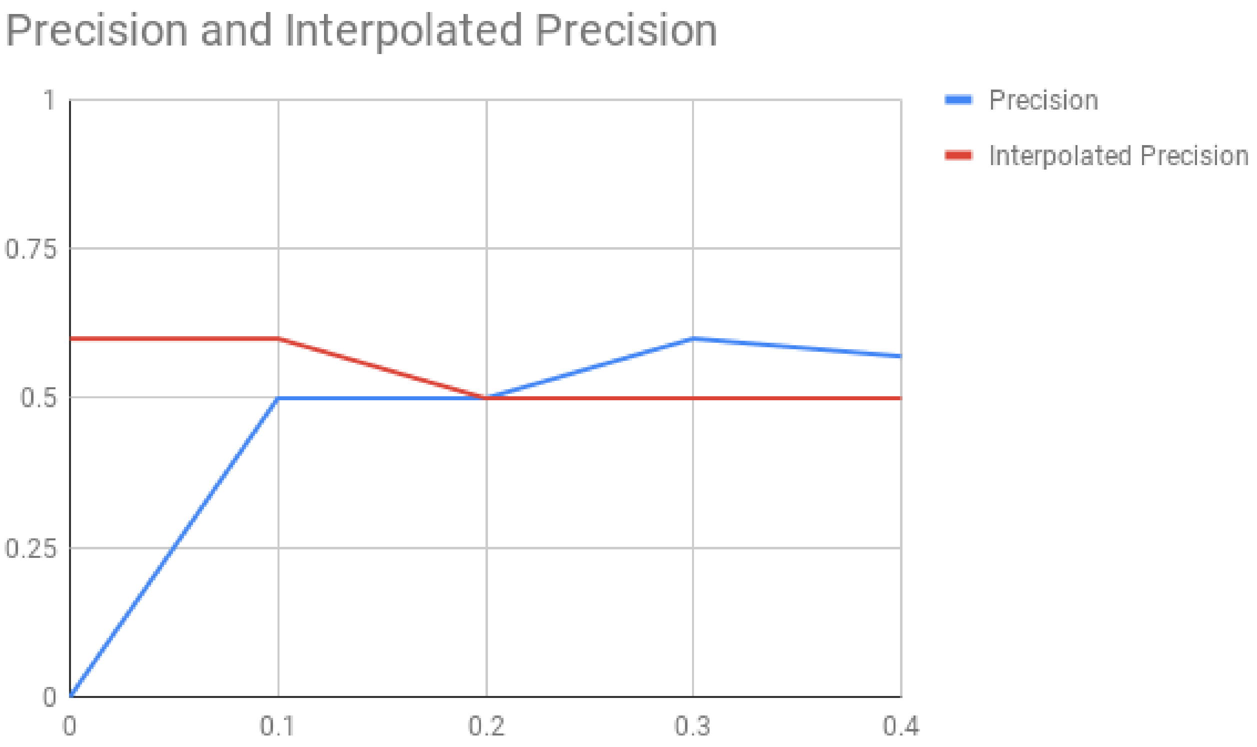
Table 5: Prciesion and recall table

4 Interpolated Precision

4.1 What is interpolated precision?

Precision-Recall curves have very jagged, pointy lines. This happens because when the document is irrelevant the recall is the same and precision has dropped, but when the document is relevant the preicsion and recall has increased which makes the jagged lines. Interpolated precision is used to smooth out those jagged lines.

4.2 Given the example in Task 3.2, find the interpolated precision and make a graph.


Precision and Interpolated Precision
Precision
Interpolated Precision