TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 6

Information Retrieval

9. Evaluation and Result Summaries

Warm up

Dresden Database Systems Group

ok

Exercise 9.1

- Are the following statements true or false? Give reasons for your answer.

a) The key measure for a search engine is user happiness.

b) The F_β measure combines both, precision and recall, into one number.

c) The 11-point interpolated average precision projects the precision-recall curve to a single number.

d) The Mean Average Precision (MAP) is yet another measure for evaluating the result of one query.

not only one

e) Pooling means experts manually judge the relevance of each document in the collection.

f) The kappa value is 1 if two judges always agree and 0 if they never agree.

g) Dynamic result summaries can be constructed efficiently from the positional inverted index.

Unranked and Ranked Evaluation

Dresden Database Systems Group

Exercise 9.2

- Consider the following ranked query result retrieved from a collection of 10,000 documents
- Each document in this list is judged as either **R**elevant or **N**on-relevant
- Assume there are 8 relevant documents in total in the collection

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
R	R	N	N	N	N	N	N	R	N	R	N	N	N	R	N	N	N	N	R

a) What is the precision of the system on the top-20?
b) What is the F_1 on the top-20?
c) What is the uninterpolated precision of the system at 25% recall? 根据题目，已经可以写出不同p@k
d) What is the interpolated precision at 33% recall? ！本身没有数值，使用差值技术。但是这里取后面的最大数值
e) Assume that these 20 documents are the complete result set of the system. What is the MAP just for this query? 不够补零 NEXT

Unranked and Ranked Evaluation

Dresden Database Systems Group

Exercise 9.2

- Consider the following ranked query result retrieved from a collection of 10,000 documents
- Each document in this list is judged as either **R**elevant or **N**on-relevant
- Assume there are 8 relevant documents in total in the collection

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
R	R	N	N	N	N	N	N	R	N	R	N	N	N	R	N	N	N	N	R

- **Assume now, instead, that the system returned the entire 10,000 documents in a ranked list and the ranking above shows just the first 20 documents returned.**

f) What is the largest possible MAP that this system could have?

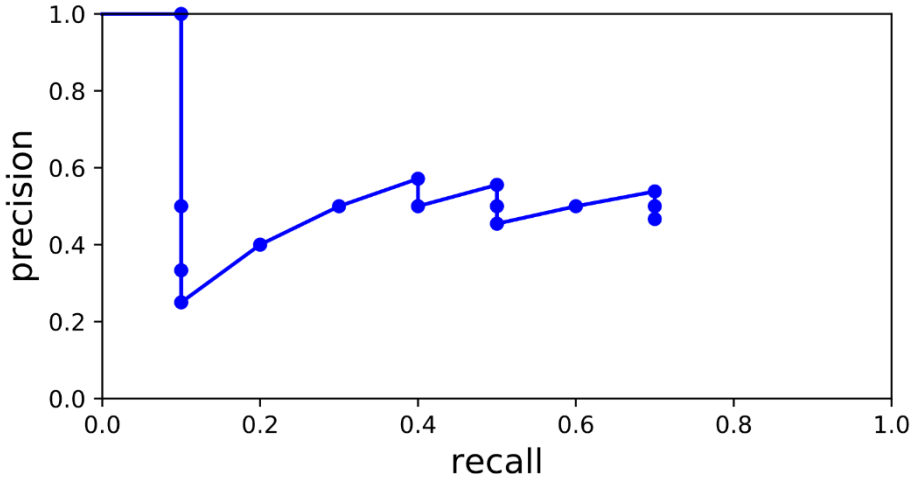
g) What is the smallest possible MAP that this system could have?

Precision-Recall Curve

Dresden Database Systems Group

Exercise 9.3

- Imagine we have a collection of 20 documents
- The precision-recall curve below visualizes the top-15 documents returned for some query (each dot represents one document in the result)



斜率大于等于0

a) How many relevant documents does the result contain and at which ranks are they?

b) How many relevant documents does the collection contain in total? 找一个点就可以

c) Sketch the precision-recall curve for the best possible top-20 result

d) Sketch the precision-recall curve for the worst possible top-20 result

最好最坏，都是在全部检索得到的条件下？主要是最坏，为啥不是一直零？ ？？ NEXT

Kappa Measure

Dresden Database Systems Group

Exercise 9.4

- On the right there is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (Relevant or Non-relevant)
- Assume an IR system returns documents {4, 5, 6, 7, 8} for a query suiting this information need

a) Calculate the kappa measure between the two judges

-1/3 一开始表写错

- When creating a gold standard, we could consider a document relevant, if
  - Both judges agree that it is relevant, or
  - At least one judge thinks it is relevant

b) How do these options affect recall and precision for the considered query?

liu答案没看懂，找标准答案？ NEXT

docID	Judge 1	Judge 2
1	N	N
2	N	N
3	R	R
4	R	R
5	R	N
6	R	N
7	R	N
8	R	N
9	N	R
10	N	R
11	N	R
12	N	R