Notes 6

Recap of inverted index

- input user query -> query parsing -> throw that into inverted index -> find associated documents and return to user
- same approach for mapping stemming to actual queries
- query expansion: term A -> synonym of A

Judging Criteria

Brainstorm - What is important when evaluating a search engine?

- % of Relevant articles
- Customer satisfaction
- Speed
- How "Smart" the engine is -> if it can learn

What is actualy used in industry

1. Dwell time on ranked result
2. Amount of scrolling -> more common for mobile searching

'Correct' metric

- The ideal goal is to satisfy users' information needs
- We try to approximate this.

Metric approximation

1. Information need = Reflected my query

- Categorize information need into : Navigational, Informational , Transactional
- **Navigation**: User doesn't know infromation need much. Kind of like using search engine as bookmark
- **Informational**: User knows query. Prefer more long clicks
- **Transactional**: You get led to results

2. Satisfaction

- We approximate satisfaction as less effort = satisfaction
- Quality of search result . Higher quality = satisfaction

Classic IR Evaluation

1. Define collection

2. Fix set of queries
3. Set of relevance judgements. Check to see if you satisfy this metric. Not how you rank, etc.

- Revelevance is with respect to information need. NOT the key words of the query
- Two ways: unranked retrieval sets vs ranked retrieval

Unranked

- Boolean retrieval:
- Precision: fraction of retrieved documents are relevant p(rel| retr). Return less , be more conservative
- Recall: fraction of relevant docs retrieved p(retr | rel). Return more.
- Choosing one over the other. Unless you have NO results or PERFECT ranking

Summarize precision and recall to single value.

- In order to compare different systems
- Computer F-measure: weighted harmonic mean of precision and recall. . Alpha balances trade off. F1 score is more sensitive to lower value than arithmetic average. The F1 score / harmonic mean tells you worst case .

Ranked

- Calculate precision and recall with respect to rank. At every precision, calculate precision and recall
- Decide which curve is better. Area under curve => effort user has to spend

Factoring in Ranking

TODO *clean this up*

- Relevant Docs = { A , B, C, D}
- Ranking Algo returns => A, E, B, F , G, C, H , D
- positions 1, 2, 3, 4, 5 ,6 , 7, 8
- @ pos 1 precision= 1/1 recal 1/4
- @ pos 2 precision= 1/2 recal 1/4
- why? (precision- returned 2 docs thusfar, only 1 rel recal : only 1 from 4 still returned)
- @ pos 3 precision = 2/3 recall (2/4)
- ETC.

Plot this with respect to recall (X axis = recall, y axis = precision when you have perfect ranking recall never drops Changes when we have a rel doc

Decide which curve is better. Area under curve => effort user has to spend. Largest area is closest to PERFECT ranking.

How to calculate area: Compute area using series of rectangles

Other approximations

1) Elevent-point interpolated average precision

- At 11 recall levels [0,0.1,0.2,. . . ,1.0] compute arithmetic mean of interpolated precision over all queries

2) Precision@K

- Assume User cares about top k results. Ignore docs ranked lower than K. Compute precision in top k retrieved docs.

3) Mean Average Precision - need to know relevant docs

- K defined as ranking position of every relevant doc
- In ranking look at all positions with relevant docs returned. Sum up all precisions, divide by count of relevant docs = instead of rewarding your system you penalize it
- Emphasizes recall
- Average within query. Mean is between miltiple queries
- MAP = Requires us to annonate a lot. If rel doc never gets retrieved corresponding precision =0. Each query counts equally. Also assumes users are interesting in finding many relevant docs per query

4) Mean reciprocal Rank

- Measures effectivness of ranked results
- Suppose users are only looking for one relevant doc
- Use rank of answer (where is first rel doc take reciprocal of it 1/k = penalizing it if it's lower than 1)

Problem with Binary relevance

- Kind of naive and won't find difference between something perfect , something good

Discounted Cumulative Gain

—- Lecture going a bit fast write notes later lol—-

we need significance test

. . .