Novelty & Diversity

CISC489/689-010, Lecture #25

Monday, May 18th

Ben Carterette

IR Tasks

- Standard task: ad hoc retrieval
  - User submits query, receives ranked list of top-scoring documents
- Cross-language retrieval
  - User submits query in language E, receives ranked list of top-scoring documents in languages F, G, …
- Question answering
  - User submits natural language question and receives natural language answer
- Common thread: documents are scored independently of one another

Independent Document Scoring

- Scoring documents independently means the score of a document is computed without considering other documents that might be relevant to the query
  - Example: 10 documents that are identical to each other will all receive the same score
  - These 10 documents would then be ranked consecutively
- Does a user really want to see 10 copies of the same document?

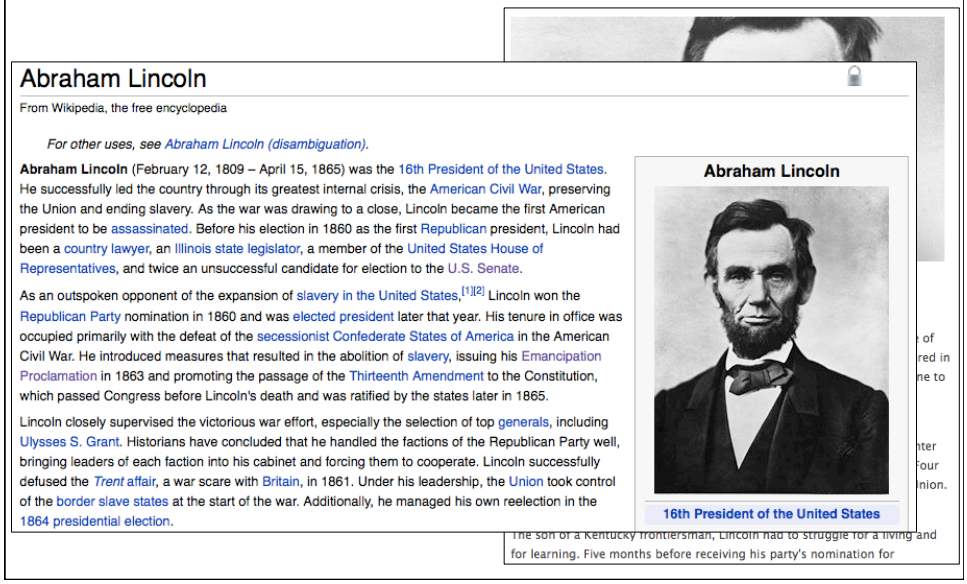
Duplicate Removal

- Duplicate removal (or *de-duping*) is a simple way to reduce redundancy in the ranked list
- Identify documents that have the same content and remove all but one
- Simple approaches:
  - *Fingerprinting*: break documents down into blocks and measure similarity between blocks
  - If there are many blocks with high similarity, documents are probably duplicates

Redundancy and Novelty

- Simple de-duping is not necessarily enough
  - Picture 10 documents that contain the same information but are written in very different styles
  - A user probably doesn't need all 10
    - Though 2 might be OK
  - De-duping will not reduce the redundancy
- We would like ways to identify documents that contain *novel* information
  - Information that is not present in the documents that have already been ranked

Example: Two Biographies of Lincoln



Abraham Lincoln

From Wikipedia, the free encyclopedia

For other uses, see Abraham Lincoln (disambiguation).

Abraham Lincoln (February 12, 1809 – April 15, 1865) was the 16th President of the United States. He successfully led the country through its greatest internal crisis, the American Civil War, preserving the Union and ending slavery. As the war was drawing to a close, Lincoln became the first American president to be assassinated. Before his election in 1860 as the first Republican president, Lincoln had been a country lawyer, an Illinois state legislator, a member of the United States House of Representatives, and twice an unsuccessful candidate for election to the U.S. Senate.

As an outspoken opponent of the expansion of slavery in the United States,[1][2] Lincoln won the Republican Party nomination in 1860 and was elected president later that year. His tenure in office was occupied primarily with the defeat of the secessionist Confederate States of America in the American Civil War. He introduced measures that resulted in the abolition of slavery, issuing his Emancipation Proclamation in 1863 and promoting the passage of the Thirteenth Amendment to the Constitution, which passed Congress before Lincoln's death and was ratified by the states later in 1865.

Lincoln closely supervised the victorious war effort, especially the selection of top generals, including Ulysses S. Grant. Historians have concluded that he handled the factions of the Republican Party well, bringing leaders of each faction into his cabinet and forcing them to cooperate. Lincoln successfully defused the *Trent* affair, a war scare with Britain, in 1861. Under his leadership, the Union took control of the border slave states at the start of the war. Additionally, he managed his own reelection in the 1864 presidential election.

Abraham Lincoln

16th President of the United States

The son of a Kentucky frontiersman, Lincoln had to struggle for a living and for learning. Five months before receiving his party's nomination for

Novelty Ranking

- *Maximum Marginal Relevance* (MMR) – Carbonell & Goldstein, SIGIR 1998
- Combine a query-document score S(Q, D) with a similarity score based on the similarity between D and the (k-1) documents that have already been ranked
  - If D has a low score give it low marginal relevance
  - If D has a high score but is very similar to the documents already ranked, give it low marginal relevance
  - If D has a high score and is different from other documents, give it high marginal relevance
- The k^{th} ranked document is the one with maximum marginal relevance

MMR

$$MMR(Q, D) = \lambda S(Q, D) - (1 - \lambda) \max_i sim(D, D_i)$$

Top-ranked document = $D_1 = \max_D MMR(Q, D) = \max_D S(Q, D)$
Second-ranked document = $D_2 = \max_D MMR(Q, D) = \max_D \lambda S(Q, D) - (1 - \lambda) sim(D, D_1)$
Third-ranked document = $D_3 = \max_D MMR(Q, D) = \max_D \lambda S(Q, D) -$
$(1 - \lambda) \max\{sim(D, D_1), sim(D, D_2)\}$
…

When λ = 1, MMR ranking is identical to normal ranked retrieval

A Probabilistic Approach

- "Beyond Independent Relevance", Zhai et al., SIGIR 2003
- Calculate four probabilities for a document D:
  - P(Rel, New | D) = P(Rel | D)P(New | D)
  - P(Rel, ~New | D) = P(Rel | D)P(~New | D)
  - P(~Rel, New | D) = P(~Rel | D)P(New | D)
  - P(~Rel, ~New | D) = P(~Rel | D)P(~New | D)
  - Four probabilities reduce to two: P(Rel | D), P(New | D)

A Probabilistic Approach

- The document score is a cost function of the probabilities:

$$\begin{aligned} S(Q,D) = \quad & c_1 P(Rel|D)P(New|D) \\ + \quad & c_2 P(Rel|D)P(\neg New|D) \\ + \quad & c_3 P(\neg Rel|D)P(New|D) \\ + \quad & c_4 P(\neg Rel|D)P(\neg New|D) \end{aligned}$$

- c_1 = cost of new relevant document
- c_2 = cost of redundant relevant document
- c_3 = cost of new nonrelevant document
- c_4 = cost of redundant nonrelevant document

A Probabilistic Approach

- Assume the following:
  - $c_1 = 0$ – there is no cost for a new relevant document
  - $c_2 > 0$ – there is some cost for a redundant relevant document
  - $c_3 = c_4$ – the cost of a nonrelevant document is the same whether its new or not
- Scoring function reduces to

$$S(Q, D) = P(Rel|D)(1 - \frac{c_3}{c_2} - P(New|D))$$

A Probabilistic Approach

- Requires estimates of P(Rel | D) and P(New | D)
- P(Rel | D) = P(Q | D), the query-likelihood language model score
- P(New | D) is trickier
  - One possibility: KL-divergence between language model of document D and language model of ranked documents
  - Recall that KL-divergence is a sort of "similarity" between probability distributions/language models

Novelty Probability

- P(New | D)
- The smoothed language model for D is

$$P(w|D) = (1 - \alpha_D)\frac{tf_{w,D}}{|D|} + \alpha_D \frac{ctf_w}{|C|}$$

- If we let C be the set of documents ranked above D, then α_D can be thought of as a "novelty coefficient"
  - Higher α_D means the document is more like the ones ranked above it
  - Lower α_D means the document is less like the ones ranked above it

Novelty Probability

- Find the value of α_D that maximizes the likelihood of the document D

$$P(New|D) = \arg\max_{\alpha_D} \prod_{w \in D} (1 - \alpha_D)\frac{tf_{w,D}}{|D|} + \alpha_D \frac{ctf_w}{|C|}$$

- This is a novel use of the smoothing parameter: instead of giving small probability to terms that don't appear, use it to estimate how different the document is from the background

Probabilistic Model Summary

- Estimate P(Rel | D) using usual language model approaches
- Estimate P(New | D) using smoothing parameter
- Combine P(Rel | D) and P(New | D) using cost-based scoring function and rank documents accordingly

Evaluating Novelty

- Evaluation by precision, recall, average precision, etc, is also based on independent assessments of relevance
  - Example: if one of 10 duplicate documents is relevant, all 10 must be relevant
  - A system that ranks those 10 documents at ranks 1 to 10 gets a better precision than a system that finds 5 relevant documents that are very different
- The evaluation does not reflect the utility to the users

Subtopic Assessment

- Instead of judging documents for relevance to the query/information need, judge them with respect to subtopics of the information need
- Example:

Information need

```
Number:   392i
Title:    robotics
Description:
  What are the applications of robotics in the world
  today?

Instances:
  In the time alloted, please find as many DIFFERENT
  applications of the sort described above as you can.
  Please save at least one document for EACH such
  DIFFERENT application. If one document discusses
  several such applications, then you need not save
  other documents that repeat those, since your goal
  is to identify as many DIFFERENT applications of
  the sort described above as possible.
```

Subtopics

```
1  'clean room' applications in healthcare &
   precision engineering
2  spot-welding robotics
3  controlling inventory - storage devices
4  pipe-laying robots
5  talking robot
6  robots for loading & unloading memory tapes
... ...
```

Subtopics and Documents

- A document can be relevant to one or more subtopics
  - Or to none, in which case it is not relevant
- We want to evaluate the ability of the system to find non-duplicate subtopics
  - If document 1 is relevant to "spot-welding robots" and "pipe-laying robots" and document 2 is the same, document 2 does not give any extra benefit
  - If document 2 is relevant to "controlling inventory", it does give extra benefit

Evaluating Novelty

- We can evaluate novelty by evaluating the ability of the system to find *unique* subtopics
- Zhai et al. introduced *S-precision* and *S-recall*
- S-recall at rank k: number of unique subtopics in top-k ranked documents divided by total number of unique subtopics
- S-precision at rank k:
  - First calculate S-recall at rank k
  - Then S-precision at k is the minimum rank at which the same S-recall could be achieved divided by k

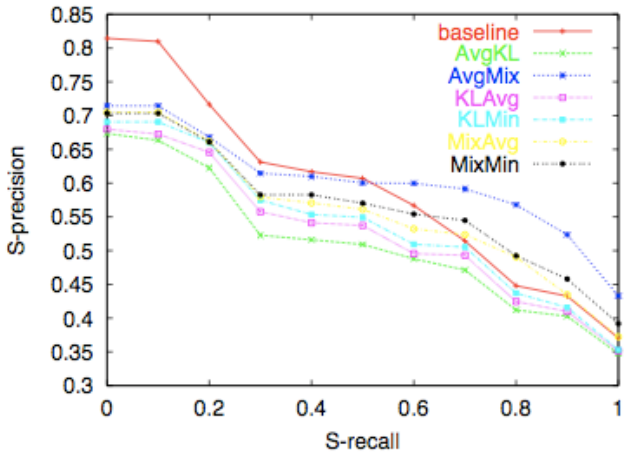
Novelty Evaluation

- One problem: S-precision is NP-complete
  - Specifically, calculating the minimum rank at which a given S-recall could be achieved is an instance of *minimum set cover*
  - Minimum set cover: given items U and a collection C of subsets of U, find the smallest subset of C that contains all items in U
  - U = subtopics, C = documents
- Some queries will be very difficult to evaluate

Novelty Data

- To do experiments, we need a collection of documents that have been judged w.r.t. subtopics of information needs
- There is not much data available
  - Only set used in literature: 20 information needs, 210,000 news articles judged for subtopic relevance

Some Experimental Results



From Zhai et al., "Beyond Independent Relevance", SIGIR 2003

Novelty & Diversity

- This is a growing area of interest in IR research
- A number of papers have been published in the last year
- Still not much data available for experiments
- TREC 2009 will have a diversity retrieval task
  - Slightly different from novelty: find documents that answer different interpretations of a query
- I am co-organizing a workshop on the subject at SIGIR this summer