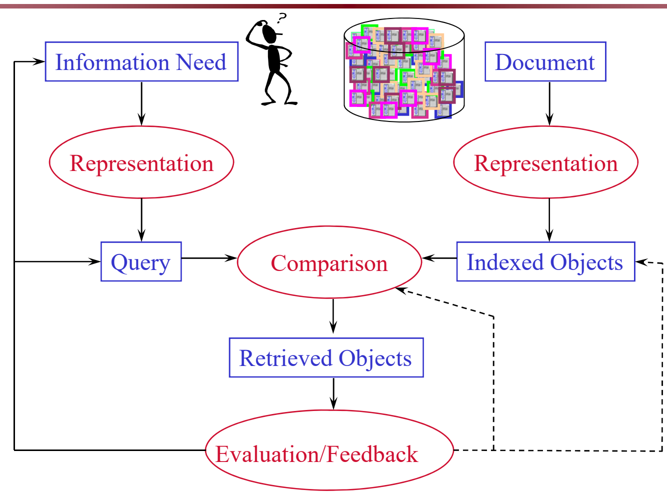
11-642:
Search Engines

Relevance and Pseudo Relevance Feedback

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Overview of Information Retrieval Processes



Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

Introduction to Relevance Feedback

A query only approximates an information need

- Users often start with short queries (poor approximations)
- People can improve queries after seeing relevant and non-relevant documents
  - by adding and removing terms
  - by reweighing terms

Question: Can a better query be created automatically?

- Machine learning

Introduction to Relevance Feedback: Initial Query and Top 10 Results

Original query: New space satellite applications

1. Soviets May Adapt Parts of SS-20 Missile For Commercial… ✓, +
2. NASA Hasn't Scrapped Imaging Spectrometer ✓, +
3. When the Pentagon Launches a Secret Satellite, Space …
4. NASA Uses 'Warm' Superconductors For Fast Circuit
5. NASA Scratches Environment Gear From Satellite Plan ✓, +
6. Pentagon Lags in Race To Match the Soviets In Rocket Launchers
7. Rescue of Satellite By Space Agency To Cost $90M
8. Telecommunications Tale of Two Companies ✓, +

✓: Judged by the user **+: Relevant document**

1987-1992 Wall Street Journal, (173,252 documents, 533.2 MB)

Introduction to Relevance Feedback: A Learned Query

#weight (

2.074942	new	15.106679	space
30.816116	satellite	5.660316	application
5.991961	nasa	5.196587	eos
4.196558	launch	3.972533	aster
3.516046	instrument	3.446570	arianespace
3.004332	bundespost	2.806131	ss
2.790090	rocket	2.053300	scientist
2.003333	broadcast	1.172533	earth
0.836515	oil	0.646711	measure)

Introduction to Relevance Feedback: Initial Query and Top 10 Results

Original query: New space satellite applications

1. NASA Hasn't Scrapped Imaging Spectrometer ✓, +
2. NASA Scratches Environment Gear From Satellite Plan ✓, +
3. Science Panel Backs NASA Satellite Plan, But … +
4. A NASA Satellite Project Accomplishes Incredible Feat …
5. Scientist … Proposes Satellites for Climate Research +
6. Report Provides Support for the Critics Of Using Big Satellites to Study Climate +
7. Arianespace Receives Satellite Launch Pact From Telesat … +
8. Telecommunications Tale of Two Companies ✓, +

✓: Judged by the user **+: Relevant document**

1987-1992 Wall Street Journal, (173,252 documents, 533.2 MB)

Introduction to Relevance Feedback

Relevance feedback is a machine learning problem

- **Ideally:** Learn f (document) → {relevant, not relevant}
- **Typically:** Learn f (document) → score

Use your favorite machine learning algorithm

- Perceptron (Rocchio)
- Naïve Bayes
- …

Introduction to Relevance Feedback

Key issue: How much training data?

- In the previous example, 4 positive examples, 0 negative examples

How much training data would you expect to be reasonable?

Introduction to Relevance Feedback: What We Know

Machine learning is effective if given enough training data

- 10-20 judge documents is <u>good</u>, 100-200 is <u>great</u>

But, people do not enjoy judging documents

- It is boring, and there is no immediate reward
- It is faster to reformulate the query

Typically, relevance feedback is only used in situations where it is practical to expect <u>many</u> judged documents

- E.g., review of legal documents

Relevance Feedback: State of the Art

Relevance feedback works

- Improved queries can be learned from judged documents

Relevance feedback is not used in many deployed systems

- People don't like giving relevance judgments
- Search providers don't like the risk of doing something stupid
  - If many documents are judged, results are very reliable
  - If few documents are judged, results are highly variable

Major open problems

- Stability and consistency (e.g., don't ever be stupid)
- Inferring relevance from implicit feedback

11

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

12

Pseudo-Relevance Feedback (Automatic Relevance Feedback)

Relevance feedback is supervised machine learning

Pseudo relevance feedback is unsupervised machine learning

- Treat the initial query as a classifier
- Use it to label some data
  - i.e., rank the documents
- Use the labeled data to generate a better classifier
  - Noisy training data

Typically there is just one iteration of this cycle

- Additional iterations increase risk but do not increase reward

13

Pseudo-Relevance Feedback (Automatic Relevance Feedback)

Typically…

- Use the original, unexpanded query to retrieve documents
- Assume that the top N documents are relevant, e.g., N=50
  - This is the positive training data
  - Some documents won't be relevant, but the goal is to learn vocabulary patterns
- Apply a relevance feedback algorithm
  - Term weighting and term selection
- Use the modified query to retrieve documents

14

Relevance Feedback in the Vector Space: The Rocchio Algorithm

Goal: Make the query more similar to relevant documents

New Query: A weighted average of original query vector, the relevant document vectors, and non-relevant document vectors

$$Q_{expanded} = Q_{original} + \alpha \boxed{\frac{1}{|R|} \sum_{\vec{d} \in R} \vec{d}} - \beta \boxed{\frac{1}{|NR|} \sum_{\vec{d} \in NR} \vec{d}}$$

Average of Rel docs

Average of Non-rel docs

Notation

- **R and NR:** Judged Relevant and Non-Relevant documents
- **d:** A document vector (e.g., (log ($tf_{t,d}$)+1) × idf_t
- **α and β:** Weights on Relevant and Non-Relevant judgments

15

Relevance Feedback in the Vector Space: The Rocchio Algorithm

Goal: Make the query more similar to relevant documents

New Query: A weighted average of original query vector, the relevant document vectors, and non-relevant document vectors

$$Q_{expanded} = Q_{original} + \alpha \boxed{\frac{1}{|R|} \sum_{\vec{d} \in R} \vec{d}} - \beta \boxed{\frac{1}{|NR|} \sum_{\vec{d} \in NR} \vec{d}}$$

Average of Rel docs

Average of Non-rel docs

Variations:

- Different values of α and β
- Vector length (number of terms added to the query)
- Which documents are used for training (all, best, uncertain, etc)

16

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

Relevance Feedback in Okapi

Features: Any term in any relevant document

Term weight

$$
\begin{aligned}
w_{\text{expansion}}(t) &= P(t \mid R)\, w_t && \textbf{w_t: RSJ weight} \\
&\approx \frac{rdf_t}{|R|} w_t && \textbf{MLE estimate of P(t|R)} \\
&\propto rdf_t\ w_t && \textbf{Drop the constant |R|} \\
&= rdf_t \left(\log \frac{N - df_t + 0.5}{df_t + 0.5} \right) && \textbf{Showing the RSJ weight}
\end{aligned}
$$

rdf_t: # of relevant docs containing t

R: Set of relevant docs

(Robertson and Zaragoza, 2009)

Relevance Feedback in Okapi

Okapi uses a typical pseudo relevance feedback architecture

1. The initial query $Q_{original}$ retrieves the top-ranked *n* documents
2. Extract potential expansion terms from top *n* documents
3. Calculate a score for each potential expansion term
4. Use the top *m* terms to create a new query $Q_{learned}$
5. $Q_{learned}$ retrieves a new (better) set of documents

Relevance Feedback in Okapi

Select the top *n* documents

- E.g., n=10-30

Select the top *m* terms

- E.g., m=10-30

Treat $w_{expansion}(t)$ as a user query term weight (qtf_t)

- k_3=7

$$\sum_{t \in q \cap d} \left(\log \frac{N - df_t + 0.5}{df_t + 0.5} \right) \frac{tf_{t,d}}{tf_{t,d} + k_1 \left((1-b) + b \frac{doclen_d}{avg\_doclen} \right)} \frac{(k_3 + 1)\, qtf_t}{k_3 + qtf_t}$$

(Walker et al, 1997; Robertson and Zaragoza, 2009)

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

21

Indri's Pseudo-Relevance Feedback: Overview

Indri uses a typical pseudo relevance feedback architecture

1. The initial query $Q_{original}$ retrieves the top-ranked *n* documents
2. Extract potential expansion terms from top *n* documents
3. Calculate a score for each potential expansion term
4. Use the top *m* terms to create an expansion query $Q_{learned}$
5. Combine $Q_{original}$ and $Q_{learned}$ to create Q_{expanded}
6. Q_{expanded} retrieves a new (better) set of documents

(http://ciir.cs.umass.edu/~metzler/indriretmodel.html#prf)

22

Indri Relevance Feedback: Potential Expansion Terms

Any document feature can be a potential expansion term

Unigrams (terms) are the most common choice

- Bigrams and trigrams ("phrases") are also possible
  - Higher computational cost, because there are more of them
  - Usually only a small additional value

Indri Relevance Feedback: Scoring Potential Expansion Terms

For each candidate expansion term t, calculate *p(t|I)*

$$p(t \mid I) = \sum_d p(t \mid d) p(d \mid I)$$ **A relevance model**

$$= \frac{\sum_d p(t \mid d) p(I \mid d) p(d)}{p(I)}$$ **Apply Bayes Rule to *p(d|I)***

$$\propto \sum_d p(t \mid d) p(I \mid d) p(d)$$ ***p(I)* is constant, so drop it**

$$\propto \sum_d \underbrace{p(t \mid d)}_{\text{the usual } p(t|d) \text{ calculation}} \underbrace{p(I \mid d)}_{\text{Indri score for } Q_{original}}$$ **Assume *p(d)* is uniform**

Indri score for $Q_{original}$

$$p(t \mid d) = \frac{tf_{t,d} + \mu \, p_{MLE}(t \mid C)}{length(d) + \mu}$$

The usual *p(t|d)* calculation

Indri default for PRF is $\mu = 0$

(Metzler, 2007)

Indri Relevance Feedback: Scoring Potential Expansion Terms

The original Indri score for expansion terms does not include a penalty for frequent terms

- E.g., an 'idf-like' weight
- It can select words that are 'almost stopwords'

A later version corrects this problem

$$p(t \mid I) \propto \sum_d p(t \mid d) p(I \mid d) \log \frac{1}{p(t \mid C)}$$

$$\propto \sum_d p(t \mid d) p(I \mid d) \boxed{\log \frac{length_{terms}(C)}{ctf_t}}$$ **A form of idf**

Use this in HW3 (Metzler, 2007)

Indri Relevance Feedback: Create an Expanded Query

The original query is $Q_{original}$

The learned query is $Q_{learned}$

#wand ($p(t_1 \mid I)$ t_1
$p(t_2 \mid I)$ t_2
$p(t_3 \mid I)$ t_3
…)

The expanded query is

$Q_{expanded}$ = #wand (w $Q_{original}$ (1-w) $Q_{learned}$)

http://ciir.cs.umass.edu/~metzler/indriretmodel.html#prf

Indri Relevance Feedback: Parameters

Parameters are set heuristically

- **fbdocs:** The number of judged documents
- **fbterms:** The number of terms to add to the query
  - Indri's default is 10
  - It is not unusual to use many more if fbdocs is high
- **μ:** The smoothing weight to use for new terms [0 - ∞]
  - Indri's default is 0
- **w:** The amount of weight to place on the original query [0 - 1]
  - Indri's default is 0.5

27

Query Expansion in Indri

Topic 523: facts about the five main clouds
Original query: facts five main clouds
Expansion terms:

0.321 clouds	0.116 cloud	0.070 weather
0.050 main	0.049 earth	0.046 facts
0.045 space	0.038 water	0.033 atmosphere
0.032 ice	0.029 radiation	0.029 jupiter
0.029 rain	0.027 planet	0.023 jupiters
0.017 atmospheric	0.016 sky	0.011 wavelength
0.011 hydrogen	0.010 infrared	

Effect: MAP 0.06 → 0.29 (**+350%**)

28

Query Expansion in Indri

Topic 509: steroids; what does it do to your body
Original query: steroids your body
Expansion terms:

0.152 steroids	0.124 body	0.079 effects
0.078 drug	0.071 drugs	0.062 treatment
0.051 steroid	0.050 side	0.046 anabolic
0.037 muscle	0.037 skin	0.035 doctor
0.034 disease	0.030 blood	0.023 taking
0.022 hormones	0.021 hormone	0.019 symptoms
0.019 testosterone	0.012 inhaled	

Effect: MAP 0.20 → 0.46 **(+126%)**

29

Query Expansion in Indri

Topic 522: how is water supplied to the mojave desert region?
Original query: water supplied mojave desert region
Expansion terms:

0.197 desert	0.181 water	0.115 valley
0.077 california	0.063 san	0.056 mojave
0.042 619	0.033 area	0.032 park
0.030 hesperia	0.028 victorville	0.026 bernardino
0.024 basin	0.016 natural	0.016 adelanto
0.016 land	0.015 canyon	0.014 sierra
0.010 geological	0.009 nevada	

Effect: MAP 0.27 → 0.11 **(–61%)**

30

Query Expansion in Indri

Topic 538: fha
Original query: fha
Expansion terms:

0.163 mortgage	0.159 fha	0.120 hud
0.104 home	0.098 loan	0.058 loans
0.050 housing	0.048 mortgages	0.033 insurance
0.031 financing	0.027 payment	0.020 conventional
0.016 lender	0.016 purchase	0.015 insured
0.012 lenders	0.011 borrower	0.007 borrowers
0.007 refinance	0.006 adjustable	

Effect: MAP 0.31 → 0.22 **(–30%)**

Outline

Relevance feedback

Pseudo relevance feedback
- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

Parameters: How Many Documents?

There is no good theory about how many documents to use for pseudo relevance feedback

The most common solution is the top n documents

- n = 10, 50, 100, …
- n is based on a guess about the quality of the initial retrieval
  - The number of relevant documents
  - The quality of the documents

Treat it as a collection-dependent parameter to be tuned

Parameters: How Many Terms is Enough?

The 'right' number of expansion terms is related to the number of documents used for query expansion

- More documents → More evidence for selecting terms
- More documents → A larger candidate vocabulary

5-50 terms is common

Parameters: How Many Terms is Enough?

Most systems expand by a static number of terms

- i.e., use the same number of terms for all queries
- Different systems use different numbers of expansion terms

Research shows that these systems are well-tuned

- No other static number of terms would provide better results

The best number of terms varies by query

- Picking the right number yields significant improvements

An interesting research topic

(Billerbeck and Zobel, 2003; Ogilvie, et al., 2009)

35

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

36

The Corpus

Usually the initial and final query are run on the same corpus

- Learn the vocabulary patterns in this corpus

The initial and final query can be run on different corpora

- Done if the initial retrieval is likely to produce a noisy result
- E.g., web search, Twitter search, …

Example

- Run the initial query on wikipedia
- Generate high-quality expansion terms
- Run the expanded query on the web corpus

37

Does Query Expansion Improve Accuracy? Wikipedia-Based Query Expansion

ClueWeb09 (500 million documents)

Method	MAP	P@10
Indri	0.0751	0.3120
Indri + wikipedia PRF (initial exact match)	0.1399	0.4520
Indri + wikipedia PRF (initial best match)	0.1169	0.3980

ClueWeb09 Category B (50 million document subset)

Method	MAP	P@5
Indri + spam filter + SDM	0.1135	0.3250
Indri + spam filter + SDM + wikipedia PRF	0.1482	0.4875

(Nguyen and Callan, 2011)
(Bendersky, Fisher, and Croft, 2011)

38

The Corpus

Wikipedia-based query expansion is widely done

…but you can use any high-quality corpus

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

Does Query Expansion Improve Accuracy?

Two query expansion methods on four TREC datasets

Data	Method	MAP	P@10	Data	Method	MAP	P@10
T-1,2	Indri	0.1818	0.4443	T-7	Indri	0.1890	0.4200
	PRF_1	0.2419	0.4913		PRF_1	0.2175	0.4320
	PRF_2	0.2406	0.5363		PRF_2	0.2169	0.4480
T-8	Indri	0.2013	0.3960	wt10g	Indri	0.1741	0.2760
	PRF_1	0.2361	0.4160		PRF_1	0.1829	0.2630
	PRF_2	0.2268	0.4340		PRF_2	0.1946	0.2960

(Collins-Thompson and Callan, 2007)

41

Does Query Expansion Improve Accuracy? Wikipedia-Based Query Expansion

ClueWeb09 (500 million documents)

Method	MAP	P@10
Indri	0.0751	0.3120
Indri + wikipedia PRF (initial exact match)	0.1399	0.4520
Indri + wikipedia PRF (initial best match)	0.1169	0.3980

ClueWeb09 Category B (50 million document subset)

Method	MAP	P@5
Indri + spam filter + SDM	0.1135	0.3250
Indri + spam filter + SDM + wikipedia PRF	0.1482	0.4875

(Nguyen and Callan, 2011)
(Bendersky, Fisher, and Croft, 2011)

42

Does Query Expansion Improve Accuracy?

PRF is viewed as 'recall enhancing'

- Adding more query terms allows more documents to match

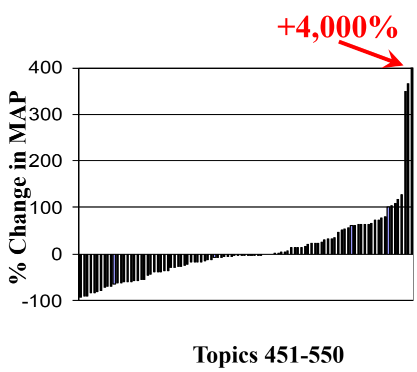
PRF improves MAP consistently

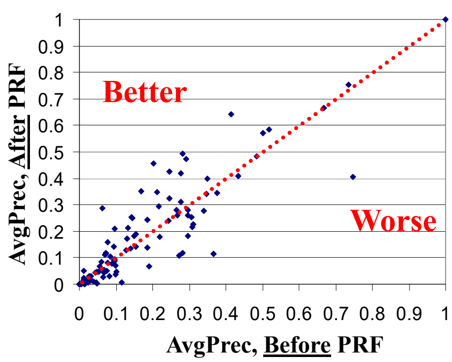
- 15-20% improvement is typical (over many test collections)

PRF may or may not improve P@10 or NDCG@10

- Very sensitive to the quality of the initial retrieval

Does Query Expansion Improve Accuracy?





TREC Topics 451-500

Does Query Expansion Improve Accuracy?

Average effectiveness improves

- But, many queries are harmed

Query expansion is used when Recall is important, or when average performance matters

- E.g., legal retrieval, TREC, research papers, …

Used less often or very carefully for interactive systems

- Too many queries are hurt
- People remember search engine mistakes more than successes
- Most needed on 'hard' queries, but most effective on 'easy' queries

Outline

Relevance feedback

Pseudo relevance feedback

- Vector space (Rocchio)
- Okapi BM25
- Inference networks (Indri)
- Parameter values
- Corpus
- Effect on retrieval accuracy

Pseudo Relevance Feedback: Summary

We covered three pseudo relevance feedback methods

- Vector space (Rocchio)
- Okapi BM25
- Indri

All of them look similar

- Select terms from the n top-ranked documents
- Weight terms based on something that looks like tf or tf.idf
- Select the best m terms
- Form an expansion query that uses terms and weights
- Use the expansion query to retrieve documents

Pseudo Relevance Feedback: Summary

Pseudo relevance feedback is unsupervised machine learning

- Treat the initial query as a classifier
- Use it to label some data
  – The top-ranked documents
- Use the labeled data to generate a better classifier

Typically there is just one iteration of this cycle

- Additional iterations increase risk but do not increase reward

Pseudo Relevance Feedback: Summary

Pseudo relevance feedback is widely used by researchers

- It is rare to see a TREC system that does not use PRF

Why?

- Typically researchers focus on average-case analysis
- PRF reliably improves MAP
  - MAP is the most common evaluation metric due to its reliability for evaluating which system is best

49

Why Isn't Pseudo-Relevance Feedback More Common?

The main unsolved problem is how to reduce variance

- Query expansion improves MAP by about 20% on average
- But, you may really annoy 1/3 of your users

Typical research directions

- Only expand if the initial retrieval results seem high quality
- Only expand if the top N documents appear homogeneous
  - Thus, they agree about which terms are important
- Generate several candidate expansion queries
  - Use only terms that appear in most of the candidate expansions
- …

50

For Additional Information

- M. Bendersky, D. Fisher, and W. B. Croft. UMass at TREC 2010 Web Track: Term dependence, spam filtering, and quality bias. In *TREC 2010 Conference Proceedings (TREC 2010).* 2011.
- K. Collins-Thompson and J. Callan. "Estimation and use of uncertainty in pseudo-relevance feedback." In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam. 2007.
- D. A. Metzler, Beyond bags of words: Effectively modeling dependence and features in information retrieval. PhD dissertation, University of Massachusetts. 2007.
- D. Nguyen and J. Callan. Combination of evidence for effective web search. In *TREC 2010 Conference Proceedings (TREC 2010).* 2011.
- S. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4). 2009.
- G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley. 1989.
- S. Walker, S.E. Robertson, M. Boughanem, G.J.F. Jones, K. Sparck Jones. "Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering, and QSDR. TREC-6 Proceedings. 1997.