A Statistical View of Binned Retrieval Models

Donald Metzler [Yahoo! Research], Trevor Strohman [Google]
and W. Bruce Croft [U. of Massachusetts]

Language Modeling

- Language modeling in speech recognition
  - Goal: model generation process of text
  - P(information retrieval) > P(information zebra)
  - Bigram multinomial models (1st order HMMs) often used
- Language modeling for IR [Ponte and Croft '98]
  - Goal: model 'topical' nature of text
  - Queries and documents are represented as statistical models that encode topicality
  - Unigram multinomial models typically assumed

Copyright Donald Metzler

Language Modeling for IR

- Document model estimation
  - Smoothed estimates
    - Dirichlet
    - Jelinek-Mercer
- Query model estimation
  - Empirical estimate
  - Pseudo-relevance feedback
    - Relevance-based language models [Lavrenko and Croft '01]
    - Model-based feedback [Zhai and Lafferty '01]

Copyright Donald Metzler

Language Modeling for IR

- Ranking function (negative KL divergence)

$$\begin{aligned} -KL(\theta_Q||\theta_D) &= H(\theta_Q) - CE(\theta_Q, \theta_D) \\ &\stackrel{rank}{=} \sum_{w \in \mathcal{V}} \theta_{w,Q} \log \theta_{w,D} \end{aligned}$$

- When query model is estimated using empirical estimates, this is equivalent to the 'query likelihood' ranking strategy
- Dot product between query model and component-wise log of document model

Copyright Donald Metzler

Language Modeling for IR

- Pros
  - Formally motivated
  - Highly effective
  - Relatively easy to implement and understand
- Cons
  - Models term occurrences, not relevance
  - Assumes queries and documents are generated from the same or ‘comparable’ models
  - Large model complexity
    - |Vocabulary| - 1 parameters per document

Copyright Donald Metzler

Binned Retrieval Models

- Anh and Moffat explored a number of so-called 'binned retrieval' models [Anh and Moffat '05]
- Models are designed with efficiency in mind
- Assumes that there are very few very important terms in a document and many unimportant terms (Zipfian)
- Rather than assign floating point tf.idf-like values to each term, terms of equal importance are binned together and all terms within the same bin are assigned the same weight

Copyright Donald Metzler

Binned Retrieval Models

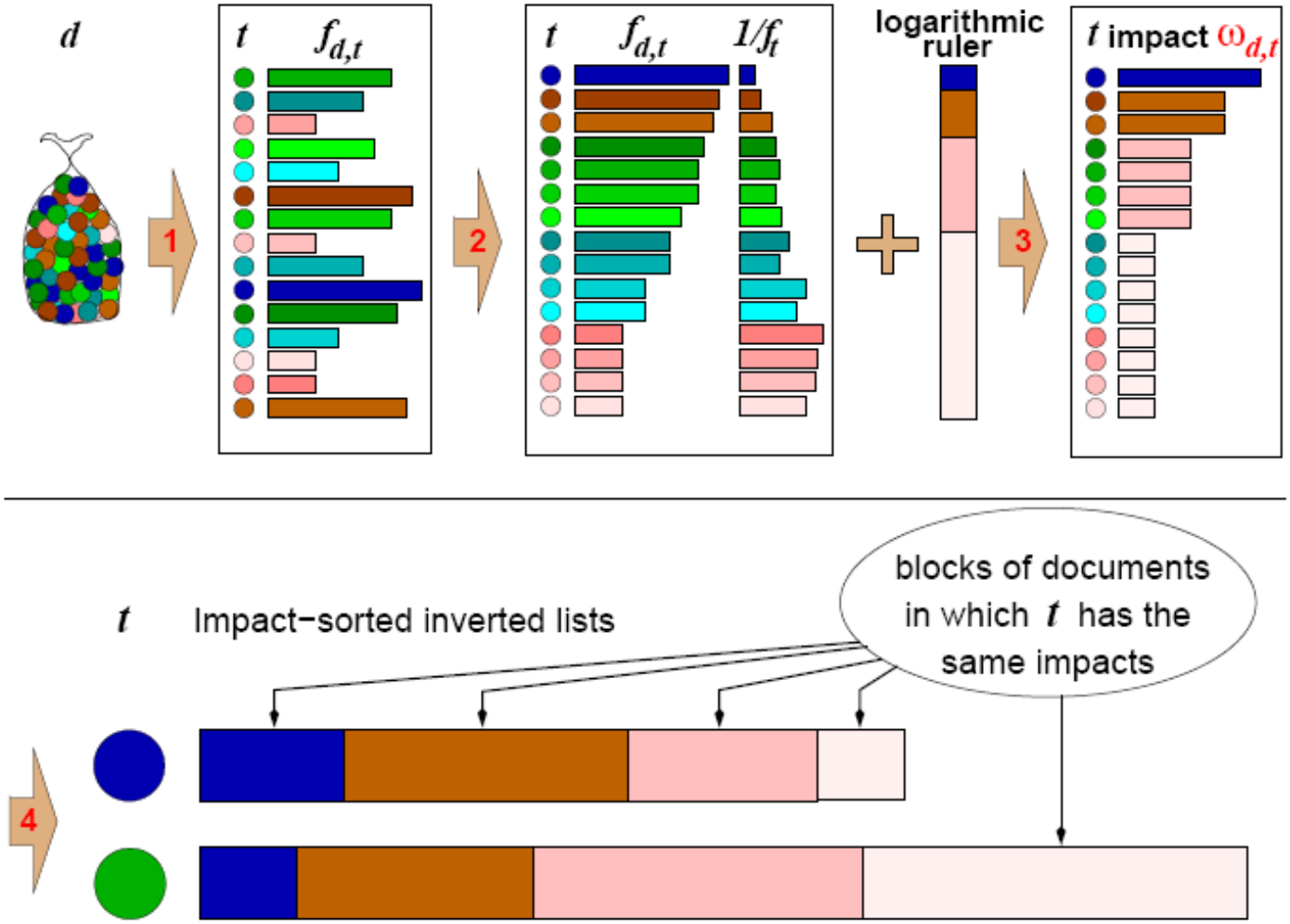

Figure borrowed from “Collection-Independent Document-Centric Impacts” (Anh and Moffat, ADCS 2004)

Copyright Donald Metzler

Document-Centric Impact Model

- Specific instantiation of binned retrieval models that assign integer weights to bins
- Pros
  - Efficient (space; indexing)
  - Efficient (time; query processing)
- Cons
  - Effectiveness typically not as good as language modeling or BM25
  - No formal motivation for choosing integral bin weights

Copyright Donald Metzler

A Statistical View of Binned Retrieval Models

- Language modeling for IR
  - Treats term importance (term weight) as a quantitative variable
    - Term A has weight 0.3, Term B has weight 0.1, …
  - Smoothing and estimation techniques dictate how weights are assigned
- Document-centric impact model
  - Treats term importance as an ordinal variable
    - Term A > Term B
  - Assigns terms to bins according to the ordering
  - Assigns quantitative importance value to each bin
    - Bin 1 has weight 10, Bin 2 has weight 5, …

Copyright Donald Metzler

A Statistical View of Binned Retrieval Models

- We propose to interpret binned retrieval models in a statistical, language modeling-like manner
- General idea
  - Treat term importance as ordinal variables
  - Bin terms according to term ordering
  - Impose probability distribution over bins

Copyright Donald Metzler

Basics of Proposed Model

- Let $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ be a set of bins
  - Ordinal variables
  - B_1 is "most important" bin
  - B_K is the "least important" bin
- Binning function
  - $b_D : \mathcal{V} \rightarrow \mathcal{B}$
  - Maps terms to bins
  - Depends on document D
- Bin probabilities
  - Multinomial over bins
  - $\theta_{B,D} \equiv P(\mathcal{B} = B|\theta_D)$

Copyright Donald Metzler

Basics of Proposed Model

- Ranking function

$$P(Q|D) = \prod_{w \in \mathcal{V}} \theta_{b_D(w),D}^{wt_{w,Q}}$$
$$\stackrel{rank}{=} \sum_{w \in \mathcal{V}} wt_{w,Q} \log \theta_{b_D(w),D}$$

- Weighted likelihood
  - Weighted geometric mean of bin probabilities
  - Note similarity with –KL from language modeling and RSV from binned models

Copyright Donald Metzler

Query Term Binning and Weighting

- Use Anh and Moffat's IDF-weighted binning
- Assign each query term a weight as follows:

$$v_w = (1 + \log tf_{w,Q}) \log \left(1 + \frac{maxtf_w}{cf_w}\right)$$

- Bin terms linearly by their weight
  - Term with highest weight in most important bin, term with lowest weight in least important bin
- Assign integral weights to terms, based on their importance
  - Least important bin gets weight 1, most important gets weight |Q|

Copyright Donald Metzler

Query Term Binning and Weighting

- Example
  - Query terms => IDF weights
    - v(A,Q) = 12.6, v(B,Q) = 5.2, v(C,Q) = 19.9
  - IDF weights => Bins
    - bin(A) = B_2, bin(B) = B_3, and bin(C) = B_1
    - $B_1 > B_2 > B_3$
  - Bins => Integral weights
    - $wt_{A,Q} = 2, wt_{B,Q} = 1, wt_{C,Q} = 3$

Copyright Donald Metzler

Document Binning

- Use Anh and Moffat's (TF, IDF) binning
- Order terms in ascending order with respect to term frequency (primary key) and inverse document frequency (secondary key)
- Geometrically assign terms to the *k* bins as follows:

$$\begin{aligned} x_i &= (|D|+1)^{1/k} x_{i+1} \\ x_{|\mathcal{B}|} &= (|D|+1)^{1/k} - 1 \end{aligned}$$

where x_i is the number of terms assigned to bin *i* and $|D|$ is the total number of unique terms in D.

Copyright Donald Metzler

Document Bin Model Estimation

- How do we estimate $\theta_{B_1,D}, \ldots, \theta_{B_K,D}$?
- Language modeling-like estimates not appropriate for bins
- Integral impacts not well motivated or understood
- Our proposed approach
  - Find bin probabilities that directly maximizes the retrieval metric of interest
  - Most metrics non-differentiable, so we have to use a local search technique

Copyright Donald Metzler

Document Bin Model Estimation

- For simplicity, we assume, for all *i*, *j*, *k*:

$$P(B_i|\theta_{D_j}) = P(B_i|\theta_{D_k})$$

- That is, bins of equal importance have the same probability for all documents
- This drastically reduces the model complexity
  - |B| - 1 parameters to estimate
- Perform greedy hill climbing over retrieval metric space to find best parameter setting

Copyright Donald Metzler

Model Equivalences

- Model is equivalent to language modeling if…
  - Each term has its own bin
  - $wt_{w,Q} = tf_{w,Q}$
  - Bin probabilities estimated using standard LM techniques
- Model is equivalent to document-centric impact model if…
  - Terms are binned geometrically
  - Query weights computed using IDF method
  - Bin probabilities are ‘estimated’ as follows:

$$\hat{\theta}_{B,D} = \frac{\exp\left[I_{B,D}\right]}{\sum_{B' \in \mathcal{B}} \exp\left[I_{B',D}\right]}$$

Copyright Donald Metzler

Evaluation

- Experiments run using Galago
  - Next generation Indri
  - http://www.galagosearch.org/
- Effectiveness experiments on 3 TREC data sets

Collection	# Docs	Train Topics	Test Topics
TREC Disks 1,2	741,856	51-150	151-200
TREC Disks 4,5	528,155	301-450	601-700
WT10g	1,692,096	451-500	501-550

Copyright Donald Metzler

Data	θ_D Estimation	$wt_{w,Q}$ Estimation	2 bins	4 bins	8 bins	16 bins
TREC12	Integral	IDF	0.2067	0.2241	0.2273	0.2273
	Discriminative	IDF	0.2105	0.2269	0.2315	$0.2336^{\dagger}$
	Language Modeling		0.2633			
TREC45	Integral	IDF	0.2325	0.2417	0.2427	0.2459
	Discriminative	IDF	$0.2430^{\dagger}$	$0.2494^{\dagger}$	$0.2577^{\dagger}$	$0.2567^{\dagger}$
	Language Modeling		0.2920			
WT10g	Integral	IDF	0.1522	0.1598	0.1863	0.1886
	Discriminative	IDF	0.1570	$0.1692^{\dagger}$	$0.1879^{\dagger}$	0.1887
	Language Modeling		0.1861			

Comparison of Techniques

A superscript indicates statistically significant improvements in effectiveness over the cell immediately above it using a one-tailed t-test with $p < 0.05$.

Copyright Donald Metzler

Review of Our Proposed Model

- Pros
  - Directly models term importance, not term occurrences
  - Can be used for translation, cross-language, relevance modeling, etc, since it is probabilistic
  - Efficient
- Cons
  - Theoretical underpinnings less well founded
  - Slightly less effective

Copyright Donald Metzler

Conclusions

- Described how binned retrieval models can be interpreted within a statistical framework
  - Can be used for translation, cross-lingual, and relevance modeling
- Proposed novel parameter estimation technique
- Resulting model is just as efficient as, and consistently more effective than traditional binned retrieval models
  - Less effective than LM on small collections
  - Equally as effective as LM on large collections

Copyright Donald Metzler