Text Analysis and Topic Models

Wray Buntine
National ICT Australia (NICTA)
Australian National University (ANU)

MLSS, Oct., 2010

Part III

Information Retrieval

Outline

We motivate standard information retrieval.

Probability ranking principle

If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

see C.J. van Rijsbergen, *Information Retrieval*, 1999.

On probabilistic versions of IR ranking

- Ranking is really a ubiquitious problem in so-called intelligent systems.
  - e.g. case-based failure diagnosis, and entity matching in NLP can be recast as ranking problems.
- Useful to have a fully probabilistic framework for
  - personalised search,
  - entity ranking,
  - semantic-based search, and
  - search in customised domains.

On Learning to Rank

- Learning to rank is a big growth area in machine learning research.
- Big search companies wont currently allow machine learning to control ranking: too many subtleties, complexities and editorial decisions.
- Smaller organisations don't have terabytes of users' search behavior. They cannot use learning easily.
- So they need to be smart about using learning:
  - They need to be smart about "learning to rank" methods.
  - Most importantly, they need to get the basic problem set up right, so they need to understand information retrieval properly.

Outline

We describe three interesting branches of the main theory.

1. Motivation
2. Probabilistic theories of IR
3. Introducing latent variables

Probabilistic versions of IR ranking

Theory overview from Roelleke and Wang, *SIGIR 2006*:

- t is term, d is document, q is query, r_d is relevance of a document,
- two-Poisson model (PM), basis for BM25 and friends,
- binary independent retrieval (BIR) model similar to PM in formula,
- language modelling (LM) works well in combining evidence, vanilla version sometimes related to PM

Large growing literature of variations, alternatives, ...

e.g. Bayesian sets by Ghahramani and Heller, 2006.

Eliteness

- Of those terms in a document, which is the document really *about*? These are called *elite* terms, and would be listed in some index.
- Notion of *eliteness* has its origins in library and information science in the 70's (Harter, 1975).
- Crudely, and moving things around, we can ask the question:

  For each query term $t \in q$, is the document d elite for t, or does t just happen to occur in d by some random chance.

- What about a probabilistic model of this?

Two-Poisson model (PM)

$$\text{Score} = \text{odds}(r_d|q,d) \;=\; \text{odds}(r_d|q)\frac{p(d|q,r_d)}{p(d|q,\overline{r_d})}$$

- $\text{odds}(r_d|q) = \text{odds}(r_d)$ is the document (relevance) odds, assumed constant in IR, but very important in web search,
  e.g. consider page-rank.
- Ranking score for query term in document, $\log \frac{p(t|q,r_d)}{p(t|q,\overline{r_d})}$, based on different rates for relevant versus non-relevant (*i.e.*, elite vs. non-elite terms)

$$n_{t,d} \log \frac{\lambda(t,r_d)}{\lambda(t,\overline{r_d})}$$

  where $n_{t,d}$ is count of terms t in document d.
- For non-query terms t in document, $p(t|q,r_d) = p(t|q,\overline{r_d})$, hence they drop out of log odds.

Two-Poisson model (PM), cont.

$$\text{Score} = \log \text{odds}(r_d|q,d) = \text{odds}(r_d) + \sum_{t \in q} n_{t,d} \log \frac{\lambda(t, r_d)}{\lambda(t, \overline{r_d})}$$

- Document frequency df_t is a robust estimate of rate $\lambda(t, \overline{r_d})$
  - $df_t = \frac{\#\text{docs with term}}{\#\text{docs}}$;
  - since occurrences greater than 1 in a document probably due to eliteness, and eliteness is assumed to be quite rare.
- Then $\lambda(t, r_d)$ estimated as is constant, as 1.
- Thus we get variant of standard TF-IDF formula.

$$\log \text{odds}(r_d|q,d) = \text{odds}(r_d) + \sum_{t \in q} n_{t,d} \log \frac{1}{df_t}$$

Note: the best known version, BM25, is more complex.

Language model (LM)

$$\text{Score} = p(r_d|d,q) \;=\; p(q|r_d,d)\frac{p(r_d|d)}{p(q|d)}$$

- Assumes $p(q|d) = p(q)$ which as a constant. *i.e.*, user query generated independently of document.
- *Document (relevance) prior* $p(r_d|d)$ usually assumed constant.
- Ranking score for query terms in document based on smoothed document frequencies, so

$$p(t|r_d,d) \;\approx\; (\delta\hat{p}_{MLE}(t|\text{collection}) + (1-\delta)\hat{p}_{MLE}(t|d))$$

- Probabilities $\hat{p}_{MLE}(t|\text{collection})$ and $\hat{p}_{MLE}(t|d)$ are observed proportions in the collection and the document, respectively.

Bayesian Sets (BS)

$$\text{Score} = p(r_d|d,q) = \frac{p(d|r_d,q)p(r_d)p(q|r_d)}{p(d)p(q|d)}$$

- $p(q|r_d)$ and $p(q|d)$ are constants, we ignore.
- $p(r_d)$ is document (relevance) prior.
- Published model has explicit notion of relevance. Above is "interpretation."
- Score based on

$$\log p(r_d) + \log \frac{p(d|r_d,q)}{p(d)}$$

i.e., while virtually the same as PM, PM seems more sophisticated, and works better.

Bayesian Sets (BS), cont.

$$\text{Score} = \log p(r_d) + \log \frac{p(d|r_d, q)}{p(d)}$$

- Assume documents d generated by multinomials with unknown probabilities.
- Knowing r_d, q makes q behave as "prior data" for the multinomial for d.
- Simple implementation by Ghahramani and Heller using standard Bayesian tricks (Dirichlet smoothing).

Latent Variable Models.

- Basic TF-IDF and BM25 usually given as increasingly better versions of two-Poisson models.
- The two-Poisson model here is really a latent variable model where parameters are approximated using simple tricks.
- Why don't we do full latent variable modelling using MCMC?
- We really *need* improved theoretical versions of information retrieval to properly incorporate semantics, personalisation, *etc.*, into scores.

See "Recommended Reading for IR Research Students," Alistair Moffat, Justin Zobel and David Hawking, *SIGIR Forum,* December, 2005.