CS 674/INFO 630: Advanced Language Technologies Fall 2007

Lecture 15 — October 23, 2007

Prof. Lillian Lee *Scribes: Nam Nguyen*
Myle Ott

1 Recall

In our current setting, we are using "single-term" distributions,

$$P_{\vec{\theta}} : V \mapsto [0, 1],$$

where $\vec{\theta}$ is an element of the m-dimensional probability simplex. Hence the probability assigned to a single term v_j is defined as:

$$P_{\vec{\theta}}(v_j) \overset{def}{=} \theta[j].$$

Also recall from the previous lecture that the Kullback–Leibler (KL) divergence between two probability distributions $P_{\vec{\theta}}$ and $P_{\vec{\theta}'}$, i.e. the expected log-likelihood ratio with respect to $P_{\vec{\theta}}$, is defined as:

$$D(P_{\vec{\theta}} \| P_{\vec{\theta}'}) = \sum_{j=1}^{m} \theta[j] \log \frac{\theta[j]}{\theta'[j]}.$$

2 Ranking by KL Divergence

Lafferty and Zhai [2] propose a ranking function based on KL divergence. Given a language model induced from the query, $P_{\vec{\theta}_q}$, and a language model induced from the document, $P_{\vec{\theta}_d}$, the score assigned to the document is given as: $D(P_{\vec{\theta}_q} \| P_{\vec{\theta}_d})$.

Question

Recall our previous "query-likelihood" ranking function:

$$QL(d, q) = \prod_{j=1}^{m} (\theta_d[j])^{q[j]}.$$

LZ claim that the "query-likelihood" ranking function can be derived from the KL divergence ranking function. Give a derivation, making sure to explain any assumptions made.

Hint: because the KL divergence measures dissimilarity and query-likelihood measures similarity, it will be helpful to use the negative KL divergence instead, namely $-D(P_{\vec{\theta}_q} \| P_{\vec{\theta}_d})$.

Answer

Starting with the negative KL divergence, we have:

$$\begin{aligned}
& -\sum_{j=1}^{m} \theta_q[j] \log \frac{\theta_q[j]}{\theta_d[j]} \\
= \quad & \underbrace{-\sum_{j=1}^{m} \theta_q[j] \log \theta_q[j]}_{\text{document independent}} + \sum_{j=1}^{m} \theta_q[j] \log \theta_d[j] \\
\overset{rank}{=} \quad & \sum_{j=1}^{m} \theta_q[j] \log \theta_d[j].
\end{aligned}$$

If we assume that $\theta_q[j]$ can be estimated by $\frac{q[j]}{q[\cdot]}$, then we have:

$$\begin{aligned}
& \frac{1}{q[\cdot]} \sum_{j=1}^{m} q[j] \log \theta_d[j] \\
\overset{rank}{=} & \prod_{j=1}^{m} (\theta_d[j])^{q[j]}.
\end{aligned}$$

3 Incorporating Implicit Feedback

Shen, Tan and Zhai [1] propose incorporating implicit feedback into the query language model, P_{θ_q}, using the following framework. First, let us group queries into sessions, i.e. queries based on a single information need.[1] Next, for each session, let us maintain both the query history and any relevant clickthrough data. Thus, supposing the user has just issued the k^{th} query, q_k, in a given session, we have:

- user query history, $QH_k = q_1, ..., q_{k-1}$
- clickthough data, $CT_k = ct_1, ..., ct_{k-1}$ where ct_i corresponds to the concatenation of all summaries clicked in response to query q_i.

We also wish to distinguish the language model induced from $q_k + CT_k + QH_k$ from the language model induced from q_k alone. We will refer to these language models as $\vec{\theta}_{IF,k}$ and $\vec{\theta}_{q_k}$, respectively. Using this notation, documents will be ranked by $D(P_{\vec{\theta}_{IF,k}} \| P_{\vec{\theta}_d})$, where $P_{\vec{\theta}_d}$ can be estimated by any applicable smoothing methods discussed in previous lectures.

STZ propose several different estimates for $P_{\vec{\theta}_{IF,k}}$, namely fixed-coefficient interpolation, length-adaptive interpolation, and round-dependent interpolation.

3.1 Fixed-Coefficient Interpolation

For fixed-coefficient interpolation, we interpolate between the current query q_k and the history. Within history, we interpolate between the query history and the clickthrough data. Thus,

$$\theta_{IF,k}[j] \overset{def}{=} \alpha\, \theta_{q_k}[j] + (1-\alpha) \underbrace{(\beta\, \theta_{QH,k}[j] + (1-\beta)\, \theta_{CT,k}[j])}_{\text{History}},$$

[1] STZ point out that grouping queries into sessions is not a trivial task and, in fact, they cite several papers that discuss the topic. We assume for the sake of brevity, however, that sessions have already been detected.

where,

$$\begin{aligned}
\vec{\theta}_{QH,k} &\overset{def}{=} \frac{1}{k-1}\sum_{i=1}^{k-1}\vec{\theta}_{q_i} \\
\theta_{q_i}[j] &\overset{def}{=} \frac{q_i[j]}{q_i[\cdot]} \\
\vec{\theta}_{CT,k} &\overset{def}{=} \frac{1}{k-1}\sum_{i=1}^{k-1}\vec{\theta}_{ct_i} \\
\theta_{ct_i}[j] &\overset{def}{=} \frac{ct_i[j]}{ct_i[\cdot]} \\
\alpha,\beta &\in [0,1].
\end{aligned}$$

3.2 Length-Adaptive Interpolation

One possible objection to fixed-coefficient interpolation is that it ignores any properties of the current query, q_k. However, as STZ point out, "intuitively, if our current query $[q_k]$ is very long, we should trust the current query more, whereas if $[q_k]$ has just one word, it may be beneficial to put more weight on the history." To satisfy this intuition, we use length-adaptive interpolation, where the coefficients are weighted proportionally to the query length.

Question

Applying Dirichlet smoothing to the query, with the query history and clickthrough data as priors, we have:

$$\theta_{IF,k}[j] \overset{def}{=} \frac{q_k[j] + \mu\,\theta_{QH,k}[j] + \lambda\,\theta_{CT,k}[j]}{q_k[\cdot] + \mu + \lambda},$$

where μ, λ are unobserved sample counts from prior distributions.

Given the above, show that $\theta_{IF,k}[j]$ is an interpolation with a length-dependent coefficient.

Answer

$$\begin{aligned}
& \frac{q_k[j] + \mu\,\theta_{QH,k}[j] + \lambda\,\theta_{CT,k}[j]}{q_k[\cdot] + \mu + \lambda} \\
= \; & \frac{\theta_{q_k}[j] \times q_k[\cdot] + \mu\,\theta_{QH,k}[j] + \lambda\,\theta_{CT,k}[j]}{q_k[\cdot] + \mu + \lambda} \\
= \; & \underbrace{\frac{q_k[\cdot]}{q_k[\cdot] + \mu + \lambda}}_{\alpha}\theta_{q_k}[j] + \underbrace{\frac{\mu + \lambda}{q_k[\cdot] + \mu + \lambda}}_{1-\alpha}\left(\underbrace{\frac{\mu}{\mu + \lambda}}_{\beta}\theta_{QH,k}[j] + \underbrace{\frac{\lambda}{\mu + \lambda}}_{1-\beta}\theta_{CT,k}[j]\right).
\end{aligned}$$

Clearly α depends on the query length, $q_k[\cdot]$, and longer queries are given greater weight.

3.3 Round-Dependent Interpolation

While length-adaptive interpolation gives us length-dependent coefficients, historical data is still weighted uniformly among different rounds in the session. However, STZ reason that "as the user interacts with the system and acquires more knowledge about the information in the collection, presumably, the reformulated queries will become better and better." Thus, we should ideally give greater weight to history data from later rounds and lesser weight to history data from earlier rounds, i.e. a round-dependent interpolation.

The obvious way to obtain a round-dependent interpolation is to consider $\theta_{IF,k-1}$ as a prior,

$$\theta_{IF,k}[j] \overset{def}{=} \frac{q_k[j] + \nu\,\theta_{IF,k-1}[j]}{q_k[\cdot] + \nu}.$$

Unfortunately, this obvious solution does not make use of available clickthrough data. A solution is to include the clickthrough data into our prior in the following manner:

$$\theta_{IF,k}[j] \overset{def}{=} \frac{q_k[j] + \nu_1 \left(\dfrac{ct_{k-1}[j] + \nu_2\,\theta_{IF,k-1}[j]}{ct_{k-1}[\cdot] + \nu_2} \right)}{q_k[\cdot] + \nu_1}.$$

Question

Consider the following (simple) round-dependent interpolation:

$$\theta'_{IF,k}[j] \overset{def}{=} \sum_{i=0}^{k-1} \alpha(i) \cdot \theta_{q_{k-i}}[j] + \sum_{i=1}^{k-1} \beta(i) \cdot \theta_{ct_{k-i}}[j],$$

where α, β are both decreasing functions, i.e. less weight is given to earlier (larger i) rounds. Show that $\theta_{IF,k}$ is a round-dependent interpolation by reducing it to to the form of $\theta'_{IF,k}$. How does this compare to the case where $\alpha(i) = \alpha^i$ and $\beta(i) = \beta^i$?

Answer

Let $\mu_i = \frac{q_i[\cdot]}{q_i[\cdot]+\nu_1}$ and $\lambda_i = \frac{ct_i[\cdot]}{ct_i[\cdot]+\nu_2}$. Using simple algebraic manipulations, we can then rewrite $\theta_{IF,k}$ as follows:

$$\theta_{IF,k}[j] = \frac{q_k[j] + \nu_1 \left(\dfrac{ct_{k-1}[j] + \nu_2\,\theta_{IF,k-1}[j]}{ct_{k-1}[\cdot] + \nu_2} \right)}{q_k[\cdot] + \nu_1}$$

(Noting that $q_k[j] = q_k[\cdot] \times \theta_{q_k}[j]$ and $ct_{k-1}[j] = ct_{k-1}[\cdot] \times \theta_{ct_{k-1}}[j]$)

$$= \frac{q_k[\cdot] \times \theta_{q_k}[j] + \nu_1 \left(\dfrac{ct_{k-1}[\cdot] \times \theta_{ct_{k-1}}[j] + \nu_2\,\theta_{IF,k-1}[j]}{ct_{k-1}[\cdot] + \nu_2} \right)}{q_k[\cdot] + \nu_1}$$

$$= \underbrace{\left(\frac{q_k[\cdot]}{q_k[\cdot] + \nu_1} \right)}_{\mu_k} \theta_{q_k}[j] + \underbrace{\left(\frac{\nu_1}{q_k[\cdot] + \nu_1} \right)}_{1-\mu_k} \left(\underbrace{\left(\frac{ct_{k-1}[\cdot]}{ct_{k-1}[\cdot] + \nu_2} \right)}_{\lambda_{k-1}} \theta_{ct_{k-1}}[j] + \underbrace{\left(\frac{\nu_2}{ct_{k-1}[\cdot] + \nu_2} \right)}_{1-\lambda_{k-1}} \theta_{IF,k-1}[j] \right)$$

$$= \mu_k \times \theta_{q_k}[j] + (1 - \mu_k) \Big(\lambda_{k-1} \times \theta_{ct_{k-1}}[j] + (1 - \lambda_{k-1}) \times \theta_{IF,k-1}[j] \Big).$$

Recursively substituting $\theta_{IF,i}[j]$, we get the sum:

$$\begin{aligned} &\sum_{i=0}^{k-1} \underbrace{\mu_{k-i} \left(\prod_{j=0}^{i-1} (1-\mu_{k-j})(1-\lambda_{k-j-1}) \right)}_{\alpha(i)} \cdot \theta_{q_{k-i}}[j] \\ + &\sum_{i=1}^{k-1} \underbrace{\lambda_{k-i}\,(1-\mu_{k-i+1}) \left(\prod_{j=0}^{i-1} (1-\mu_{k-j})(1-\lambda_{k-j-1}) \right)}_{\beta(i)} \cdot \theta_{ct_{k-i}}[j]. \end{aligned}$$

Clearly for $i \in [0, k-1]$, both $\alpha(i)$ and $\beta(i)$ are decreasing functions, and thus $\theta_{IF,k}$ is a round-dependent interpolation.

Compared to the case where $\alpha(i) = \alpha^i$ and $\beta(i) = \beta^i$, notice that our $\alpha(i)$ and $\beta(i)$ depend both on query length and clickthrough length. Intuitively, having this length-adaptive weighting will provide better performance when compared to fixed weights.

Question

In the above formulation, we use the clickthrough data as a prior for the query. Suppose we instead use the query as a prior to the clickthrough data. In that case, we have:

$$\theta''_{IF,k}[j] \overset{def}{=} \frac{ct_{k-1}[j] + \nu_1 \left(\dfrac{q_{k-1}[j] + \nu_2\, \theta_{IF,k-1}[j]}{q_{k-1}[\cdot] + \nu_2} \right)}{ct_{k-1}[\cdot] + \nu_1}.$$

What is an advantage to using $\theta''_{IF,k}$ over $\theta_{IF,k}$?

Answer

STZ propose a scheme where $\theta''_{IF,k}$ can be used to re-rank documents before seeing the query q_k. Additionally, this re-ranking can be used to suggest a rewriting of the query by treating the top-ranked documents as pseudo-relevance feedback.

References

[1] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 43–50, New York, NY, USA, 2005. ACM.

[2] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM.