University of Regina

Department of Computer Science

Winter 2019

CS 824 - Information Retrieval

Assignment 4

Submitted to Dr. Yiyu Yao

By

Chao Zhang

Regina, April 6, 2019

1. Derive the results of probabilistic models of IR.

   a) Probabilistic retrieval model

   b) Probabilistic indexing model

   c) A unified model.

(a) In the probabilistic retrieval model, a document is represented by a min-term m_d defined by $d = m_d = t_1^{\alpha 1} \cap \cdots \cap t_n^{\alpha n}$. By substituting it into the log-odds transformation of the precision-oriented measure, we obtain:

$$\log \frac{\psi(d \to q)}{1-\psi(d \to q)} = \log \frac{P(d \mid q)}{P(d \mid \overline{q})} + \log \frac{P(q)}{P(\overline{q})} = \log \frac{P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid q)}{P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid \overline{q})} + \log \frac{P(q)}{P(\overline{q})}.$$

Suppose we use the following approximations:

$$\begin{cases} P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid q) = \prod_{l=1}^{n} P(t_i^{\alpha_l} \mid q) \\ P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid \overline{q}) = \prod_{l=1}^{n} P(t_i^{\alpha_l} \mid \overline{q}) \end{cases}$$

By substituting these values into the above formula, we can obtain:

$$\log \frac{\psi(d \to q)}{1-\psi(d \to q)} = \log \frac{P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid d)}{P(t_1^{\alpha_1} \cap \cdots \cap t_n^{\alpha_n} \mid \overline{d})} + \log \frac{P(d)}{P(\overline{d})} = \sum_{l=1}^{n} \log \frac{P(t_l^{\alpha_l} \mid q)}{P(t_l^{\alpha_l} \mid \overline{q})} + \log \frac{P(q)}{P(\overline{q})}.$$

Let $u_l = P(t_l \mid q), v_l = P(t_l \mid \overline{q})$. Then the probabilities can be written as:

$P(t_l^{\alpha_l} \mid q) = u_l^{\alpha_l} (1-u_l)^{1-\alpha_l}$, $P(t_l^{\alpha_l} \mid \overline{q}) = v_l^{\alpha_l} (1-v_l)^{1-\alpha_l}$.

By combing all of these value, we can obtain the probability retrieval model:

$$\log \frac{\psi(d \to q)}{1-\psi(d \to q)} = \sum_{l=1}^{n} \log \frac{P(t_l^{\alpha_l} \mid q)}{P(t_l^{\alpha_l} \mid \overline{q})} + \log \frac{P(q)}{P(\overline{q})} = \sum_{l=1}^{n} \alpha_l \log \frac{u_l(1-v_l)}{(1-u_l)v_l} + \sum_{l=1}^{n} \log \frac{(1-u_l)}{(1-v_l)} + \log \frac{P(q)}{P(\overline{q})}.$$

(b) In the conventional probabilistic indexing model, a query is represented by a single atomic concept m_q, namely:

$q = m_1 = t_i^{\beta_1} \cap \cdots \cap t_n^{\beta n}$, where $\beta_i \in \{0,1\}$ and $t_l^{\beta_1} = \begin{cases} t_l & if \beta_l = 1, \\ \bar{t}_l & if \beta_l = 0. \end{cases}$

That is, an index term is not negated if it appears in the text of the query otherwise, it is negated. For the recall-oriented measure, we obtain:

$$\log it\,\psi(q \to d) = \log \frac{\psi(q \to d)}{1-\psi(q \to d)} = \log \frac{P(d \mid q)}{P(\bar{d} \mid q)} = \log \frac{P(q \mid d)}{P(q \mid \bar{d})} \frac{P(d)}{P(\bar{d})} = \log \lambda(d,q) + \log O(d).$$

By substituting the query into the recall-oriented measure we can obtain:

$$\log \frac{\psi(q \to d)}{1-\psi(q \to d)} = \log \frac{P(q \mid d)}{P(q \mid \bar{d})} + \log \frac{P(d)}{P(\bar{d})} = \log \frac{P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid d)}{P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid \bar{d})} + \log \frac{P(d)}{P(\bar{d})}.$$

From the probabilistic-like definition of intersection, we know:

$$\begin{cases} P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid d) = \prod_{l=1}^{n} P(t_i^{\beta_l} \mid d) \\ P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid \bar{d}) = \prod_{l=1}^{n} P(t_i^{\beta_l} \mid \bar{d}) \end{cases}$$

By substituting the probabilistic-like definition of intersection into what we obtain above, then we can get:

$$\log \frac{\psi(q \to d)}{1-\psi(q \to d)} = \log \frac{P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid d)}{P(t_1^{\beta_1} \cap \cdots \cap t_n^{\beta_n} \mid \bar{d})} + \log \frac{P(d)}{P(\bar{d})}$$

$$= \sum_{l=1}^{n} \log \frac{P(t_l^{\beta_l} \mid d)}{P(t_l^{\beta_l} \mid \bar{d})} + \log \frac{P(d)}{P(\bar{d})} = \sum_{l=1}^{n} \log \lambda(d, t_l^{\beta_l}) + \log O(d).$$

Thus, the indexer can either provide $\lambda(d,t)$ and $\lambda(d,\bar{t})$, or the probabilities $P(t \mid d)$ and $P(t \mid \bar{d})$. Let $r_l = P(t_i \mid d), s_l = P(t_i \mid \bar{d})$. Then the probability can be written as: $P(t_l^{\beta_l} \mid d) = r_l^{\beta_l}(1-r_l)^{1-\beta_l}$, $P(t_l^{\beta_l} \mid \bar{d}) = s_l^{\beta_l}(1-s_l)^{1-\beta_l}$.

By combing all of these value, we can obtain the probability indexing model:

$$\log \frac{\psi(q \to d)}{1-\psi(q \to d)} = \sum_{l=1}^{n} \log \frac{P(t_l^{\beta_l} \mid d)}{P(t_l^{\beta_l} \mid \bar{d})} + \log \frac{P(d)}{P(\bar{d})} = \sum_{l=1}^{n} \beta_l \log \frac{r_l(1-s_l)}{(1-r_l)s_l} + \sum_{l=1}^{n} \beta_l \log \frac{(1-r_l)}{(1-s_l)} + \log \frac{P(d)}{P(\bar{d})}.$$

(c) According to the Bayes decision procedure, a document described by x is judged to be relevant to a query described by y if $P(R \mid x, y) > P(\overline{R} \mid x, y)$. then we can construct a discriminant function: $g(x, y) = \log \frac{P(R \mid x, y)}{P(\overline{R} \mid x, y)}$ and the function can be rewritten as $g(x, y) = \log \frac{P(x, y \mid R)}{P(x, y \mid \overline{R})} + \log \frac{P(R)}{P(\overline{R})}$ where $P(R)$ and $P(\overline{R})$ are the priori probabilities. In the proposed model, we make the following independence assumptions: $\begin{cases} P(x, y \mid R) = P(x_1, y_1 \mid R)P(x_2, y_2 \mid R)\cdots P(x_n, y_n \mid R) \\ P(x, y \mid \overline{R}) = P(x_1, y_1 \mid \overline{R})P(x_2, y_2 \mid \overline{R})\cdots P(x_n, y_n \mid \overline{R}) \end{cases}$. It means that the cooccurrence of each index term in the document-query pairs with respect to R and $\overline{R}$ is assumed to be independent of other terms. These assumptions can be considered as the generalization of the independence approximations. Then they can be $\begin{cases} P(x, y \mid R) = \prod_{i=1}^{n} p_{i0}^{(1-xi)(1-yi)} p_{i1}^{(1-xi)yi} p_{i2}^{(1-yi)xi} p_{i3}^{xiyi} \\ P(x, y \mid \overline{R}) = \prod_{i=1}^{n} q_{i0}^{(1-xi)(1-yi)} q_{i1}^{(1-xi)yi} q_{i2}^{(1-yi)xi} q_{i3}^{xiyi} \end{cases}$. By substituting $P(x, y \mid R)$ and $P(x, y \mid \overline{R})$, we can obtain:

$$g(x, y) = \sum_{i=1}^{n} x_i \log \frac{p_{i2}q_{i0}}{p_{i0}q_{i2}} + y_i \log \frac{p_{i1}q_{i0}}{p_{i0}q_{i1}} + x_i y_i \log \frac{p_{i0}p_{i3}q_{i1}q_{i2}}{p_{i1}p_{i2}q_{i0}q_{i3}} + \sum_{i=1}^{n} \log \frac{p_{i0}}{q_{i0}} + \log \frac{P(R)}{P(\overline{R})}$$

$$= \sum_{i=1}^{n} [a_i x_i + b_i y_i + c_i x_i y_i] + C.$$

where $a_i = \log \frac{p_{i2}q_{i0}}{p_{i0}q_{i2}}, b_i = \log \frac{p_{i1}q_{i0}}{p_{i0}q_{i1}}, ci = \log \frac{p_{i0}p_{i3}q_{i1}q_{i2}}{p_{i1}p_{i2}q_{i0}q_{i3}}$ and $C = \sum_{i=1}^{n} \log \frac{p_{i0}}{q_{i0}} + \log \frac{P(R)}{P(\overline{R})}$.

Considering the problem of estimating the probabilities, the parameters can be computed from the formulas:

$$p_{ik} = \frac{n_{ik}}{\sum_{j=0}^{3} n_{ij}}, q_{ik} = \frac{m_{ik}}{\sum_{j=0}^{3} m_{ij}} \qquad (k = 0,1,2,3).$$

2. Discuss the main ideas of probabilistic inference models.

Based on probabilistic inference models, we firstly discuss the Bayesian inference which is a method of statistical inference for updating the probability for a hypothesis as more evidence or information becomes available. The computation for Bayesian inference is $P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$ where H stands for hypothesis, P(H) is prior probability, E is the evidence, P(H|E) is the posterior probability, P(E|H) is the probability of H given E and P(E) is the marginal likelihood. For different values of H, only the factors P(H) and P(E|H) affect the value of P(H|E). So Baye's rule can be written as follows: $P(H \mid E) = \frac{P(E \mid H)}{P(E)} \cdot P(H)$ where the factor $\frac{P(E \mid H)}{P(E)}$ can be interpreted as the impact of E on the probability of H. Pr(H) is belief before seeing evidence and Pr(H|E) is belief after seeing the evidence. In probabilities inference models, binary vector space can be classified into three cases: 1. precision-based: $\Pr(d \to q) = \Pr(q \mid d)$; 2. recall-based: $\Pr(q \to d) = \Pr(d \mid q)$; 3. balanced: $\Pr(q \leftrightarrow d) = \frac{\Pr(d \cap q)}{\Pr(d \cup q)}$. What we want to know is Q-T-D and what we know is query and indexing. For vector representation, it is $Q - (t_1, t_2 \cdots t_m) - D$. With respect to Bayesian inference, what we want to know is $\Pr(d \to q)$. Based on document representation $\Pr(d \to t_i)$, for t_j in T, query representation is $\Pr(t_i \to q)$. We view each t_j as one piece of evidence and T is a pool of many pieces of evidence. Based on t_j, we can get $\Pr(d \to q \mid t_i) \cong \Pr(d \to t_i)\Pr(t_i \to q)$. Assume that t_j are independent or non-overlap which is $ti \cap tj \neq 0$. Based on T, $\Pr(d \to q) = \sum_{t_j} \Pr(d \to t_j)\Pr(t_j \to q)$ where $\Pr(d \to t_j)$ is document and $\Pr(t_j \to q)$ is query.