University of Regina

Department of Computer Science

Winter 2019

CS 824 - Information Retrieval

Assignment 3

Submitted to Dr. Yiyu Yao

By

Chao Zhang

Regina, March 27, 2019

1. Explain the ideas of a perfect ranking and an acceptable ranking. Why it is more meaningful to use an acceptable ranking?

Based on a weak order which is asymmetric and negative transitive and user performance relation $\succ_u$, we suppose u is a retrieval function used by an IR system, for any two documents d1 and d2 in set of documents D, we can say that u produce a perfect ranking: d1 $\succ_u$ d2 $\Leftrightarrow$ u(d1) $\succ$ u(d2). We also say that u produce an acceptable ranking: d1 $\succ$ d2 $\Rightarrow$ u(d1) $\succ$ u(d2). Acceptable ranking means the system rank preferred document to non-preferred document which is normally used in real information retrieval design. Different from perfect ranking which requires both documents and utility functions agree on each other, acceptable ranking is based on performance, it is easier to implement ranking and more practical in use, which means it is more meaningful than perfect ranking.

2. Derive the ndpm formula.

The distance-based performance measure(dpm) can be equivalently defined by $dpm(\succ_u,\succ_s)=\beta(\succ_a,\succ_s)$. For rankings $\succ_a$ and $\succ_s$, the number of agreeing, contradictory, and compatible pairs are: $\left|\succ_a\cap\succ_s\right|=C^+ + C^S$, $\left|\succ_a\cap\succ_s^c\right|=C^-$, $\left|\succ_a\cap\succ\sim_s\right|=C^u$. dpm provides an appropriate basis for comparing various retrieval systems with a fixed query. It may be considered as an absolute distance function but it can not evaluate the performance of every query equally. So we need a normalized distance-based performance measure in terms of distance relative to the maximum distance: $ndpm(\succ_u,\succ_s)=\frac{dpm(\succ_u,\succ_s)}{\max\succ_{\in\Gamma(D)}dpm(\succ_u,\succ)}$ where $\max\succ_{\in\Gamma(D)}dpm(\succ_u,\succ)$ is the maximum distance between $\succ_u$ and all rankings. Based on the definition of dpm, the converse ranking $\succ_u^c$ produces the maximum

dpm value: $\max_{\succ \in \Gamma(D)} dpm(\succ_u, \succ) = dpm(\succ_u, \succ_u^c) = 2\left|\succ_u^c\right| = 2C$. Then the formula of ndpm can be $ndpm(\succ_u, \succ_s) = \frac{dpm(\succ_u, \succ_s)}{dpm(\succ_u, \succ_u^c)} = \frac{2C^- + C^u}{2C}$.

3. Discuss the main ideas of vector space models.

*** document representation**

*** query representation**

*** matching**

Note: You need cover both binary and non-binary models. You need to discuss different retrieval function/similarity measures.

For vector space model, we have basic assumptions: 1. a document is represented as a m-dimensional vector; 2. a query is represented as a m-dimensional vector; 3. match is defined as distance or similarity between two vectors. For binary vector space model, TF matrix is changed to B matrix. Each row in B matrix is a document which is also a set. Query q is a vector. For similarity, we have 1. number of common terms; 2. number of common terms/~; 3.number of common terms which are smaller than number of terms in d: $\frac{|d \cap q|}{d}$ (precision based), number of terms in q: $\frac{|d \cap q|}{q}$ (recall based) and number of terms in d or q: $\frac{|d \cap q|}{|d \cup q|}$ (balanced). For weighted models which works for any weighting methods, we have document representation $\vec{d} = (d_1, d_2, d_3 ... d_m)$ and query representation $\vec{q} = (q_1, q_2, q_3 ... q_m)$. For matching, we have precision-based $\frac{\sum d_i q_i}{\sum d_i}$, recall-based $\frac{\sum d_i q_i}{\sum q_i}$ and balanced $\frac{\sum d_i q_i}{\sum d_i + \sum q_i}$.

4. Discuss the main ideas of two probability distribution models.

*** document representation**

*** query representation**

*** matching**

Firstly we consider two discrete probability distributions $P=(p_1,p_2,\dots p_n)$ and $Q=(q_1,q_2,\dots q_n)$. Suppose P is absolutely continuous with respect to Q, the divergence $I(P,Q)$ is defined by: $I(P,Q)=\sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}$ where $I(P,Q)\geq 0$ and $I(P,Q)=0$ if only if P and Q are identical. From the information theory point of view, the divergence $I(P,Q)$ can be interpreted as the difference of the information contained in P and that contained in Q about P. Based on Shannon's entropy function, we need a statistical measure of similarity between two arbitrary probability distributions. Given a discrete probability distribution: $p_i \geq 0 (i=1,2,\dots,n)$, $\sum_{i=1}^{n} p_i = 1$, then the entropy function is defined by $H(P)=H(p_1,p_2,\dots p_n)=-\sum_{i=1}^{n} p_i \log p_i$. We can verify that $H(P)$ has following properties: 1. $H(P)\geq 0$; 2. $H(P)=\log n$ if $p_1=p_2=,\dots,=p_n=\frac{1}{n}$; 3. $H(P)=0$ if $p_{i_0}=1$ and $p_i=0$ $(1\leq i\leq n; i\neq i_0)$. Given two probability distributions P and Q and $\lambda_1,\lambda_2\in[0,1]$, $\lambda_1+\lambda_2=1$, the increase of entropy for the composite distribution $\lambda_1 P+\lambda_2 Q$ is defined by $\beta(P,Q:\lambda_1,\lambda_2)=H(\lambda_1 P+\lambda_2 Q)-[\lambda_1 H(P)+\lambda_2 H(Q)]$ where the function $\beta(P,Q:\lambda_1,\lambda_2)\in[0,1]$. There exists a close relationship between the divergence $I(P,Q)$ and the entropy increase $\beta(P,Q:\lambda_1,\lambda_2)$: $\beta(P,Q:\lambda_1,\lambda_2)=\lambda_1 I(P,\lambda_1 P+\lambda_2 Q)+\lambda_2 I(Q,\lambda_1 P+\lambda_2 Q)$. Because $I(P,\lambda_1 P+\lambda_2 Q)$ is a measure of the difference between the composite distribution $\lambda_1 P+\lambda_2 Q$ and its component P, the function $\beta(P,Q:\lambda_1,\lambda_2)$ can be interpreted naturally as an indirect measure of difference between the two distributions P and Q. In

information retrieval we are searching for a symmetric dissimilarity measure between P_q and P_d and it can be defined: by choosing $\lambda_1 = \lambda_2 = \frac{1}{2}$, we can get $\beta(P,Q:\frac{1}{2},\frac{1}{2}) = H(\frac{1}{2}P+\frac{1}{2}Q) - \frac{1}{2}[H(P)+H(Q)]$ and the function $\beta(P,Q:\lambda_1,\lambda_2)$ satisfies following properties: 1. $\beta(P,Q:\frac{1}{2},\frac{1}{2}) \geq 0$; 2. $\beta(P,P:\frac{1}{2},\frac{1}{2}) = 0$; 3. $\beta(P,Q:\frac{1}{2},\frac{1}{2}) = \beta(Q,P:\frac{1}{2},\frac{1}{2})$. From above we know β is bounded, therefore a similarity measure between two probability distributions P and Q can be defined as: $SIM(P,Q) = 1 - \beta(P,Q:\frac{1}{2},\frac{1}{2})$.

5. Discuss the main ideas of relevance feedback with respect to the following two methods for representing your judgements on a sample set of documents.

a). binary relevance

b). user preference

The main task of relevance feedback is to learn a better query. From query vector to document, we have S and $\overline{S}$. For S we have S^+ and S^-. Then we can get a function $q' = f(\vec{q}, S^+, S^-)$ where S^+ means $\vec{q}$ did a good job and we keep it, S^- means $\vec{q}$ did a bad job and we improve it. We rank document by $f(\vec{d},\vec{q}) = \vec{d}\cdot\vec{q}$, for $\vec{d}_1 \equiv +$, $f(\vec{d}_1,\vec{q})$ is high enough so we can still increase it, for $\vec{d}_2 \equiv -$, $f(\vec{d}_2,\vec{q})$ is too high and we should decrease the value. We make q' to be more similar to document in S^+ and to be less similar to document in S^-. Then we have centre of the positive: $\frac{1}{|S^+|}\sum_{d\in S^+}\vec{d}$ and centre of the negative: $\frac{1}{|S^-|}\sum_{d\in S-}\vec{d}'$. Based on query $\vec{q}_0$, we have $\vec{q}_1 = \alpha\vec{q}_1 + \beta\frac{1}{|S^+|}\sum_{d\in S^+}\vec{d} - \gamma\frac{1}{|S^-|}\sum_{d\in S^-}\vec{d}'$ where special case is $\alpha = \beta = \gamma = 1$. For binary relevance we have R and $\overline{R}$ and $R \succ \overline{R}$. For user preference $\vec{d}_1 \succ \vec{d}_2$ and

acceptable ranking $f(\vec{d},\vec{q}) = \vec{d}\cdot\vec{q}$,we have $\vec{d}_1 \succ \vec{d}_2 \Rightarrow f(\vec{d}_1,\vec{q}) \succ f(\vec{d}_2,\vec{q})$. According to $f(\vec{d}_1,\vec{q}) > f(\vec{d}_2,\vec{q}) \equiv \vec{d}_1\cdot\vec{q} > \vec{d}_2\cdot\vec{q}$, we know q did a good job. With respect to $f(\vec{d}_1,\vec{q}) \le f(\vec{d}_2,\vec{q}) \equiv \vec{d}_1\cdot\vec{q} \le \vec{d}_2\cdot\vec{q}$, we get q did a bad job. For relevance feedback with user preference, we need to find a solution of a system of linear inequality which is $\vec{b}\cdot\vec{q} \succ 0$ for $\vec{b} \in B(\succ)$. If $\vec{b}\cdot\vec{q} = b_1q_1 + b_2q_2 + \ldots + b_mq_m = 0$, we call it hyper plane. Then the solution of $\vec{b}\cdot\vec{q} \succ 0$ for $\vec{b} \in B(\succ)$ can be the intersection of the positive sides of some hyper planes. If $\vec{q}_0\cdot\vec{b}_1 > 0$, $\vec{q}_0$ is correct which means nothing need to be done. If $\vec{q}_0\cdot\vec{b}_1 \le 0$, $\vec{q}_0$ made an error so it should be corrected by error correction strategy which is repeating checking correction of there is an error until no error for all $\vec{b} \in B(\succ)$.