Document Modeling

Power-law Distribution of Topic Popularity

Long-tail topics are important
- Amount is large
- Arguably more interesting
  - Example: in advertisement, a “lose weight” is more important than a “time” topic

Diversity Regularized RBM

Diversity Metric
- Measure the dissimilarity between two vectors
- Measure the diversity of a vector set

Diversity between Two Vectors
- Non-obtuse angle
  \[ \theta = \arccos \left( \frac{y \cdot x}{\|y\| \cdot \|x\|} \right) \]
- Invariant to scale, translation, rotation and orientation of the two vectors

Diversity of A Vector Set
- The diversity of a set of vectors \( A = \{a_i\}_{i=1}^K \) is defined as
  \[ \Omega(A) = \text{mean}(\Theta) - \text{var}(\Theta) \]
- Mean: summarize how these vectors are different from each other on the whole
- Variance: encourage the vectors to evenly spread out

Diversity Regularized RBM
\[
\max_{\bar{\Lambda}} \quad L(D; \bar{\Lambda}) + \lambda \Omega(\bar{\Lambda}) \\
\text{Optimization} \\
\max_{g, \bar{\Lambda}} \quad L(D; g, \bar{\Lambda}) + \lambda \Omega(\bar{\Lambda}) \\
\text{s.t.} \quad \forall i, \|g_i\| = 1, g_i \geq 0
\]

Evaluation

Experimental Setup
- Dataset: TDT, 20-Newsgroups, Reuters
- Evaluation: retrieval (precision), clustering (accuracy), perplexity, qualitative evaluation

**Retrieval Precision**

**Clustering Accuracy**

**Perplexity**