Linear Optimal Topic Transport for Document Similarity

Columbia University Data Science Institute

Yilin Zhu, Parijat Dube

IBM

Contributions

Our research introduce linear optimal topic transport for measuring document similarities:

- **Efficiency**: improves computation while maintains accuracy
- **Scalability**: supports for large-scale applications
- **Adaptability**: enables topic weighting for personalized retrieval

Word Mover's Distance

Consider discrete OT, 1-Wasserstein distance between p and q is:

$$W_1(p,q) = \min_{\Gamma \in \mathbb{R}_+^{n\times m}} \sum_{ij} C_{ij}\Gamma_{ij}$$

$$\text{subject to } \sum_j \Gamma_{ij} = p_i \text{ and } \sum_i \Gamma_{ij} = q_j$$

Where C is the cost matrix with entries C_{ij} the distance between points x_i from distribution p and y_j from distribution q. Γ is the transport plan specifies how much mass is moved from each x_i to y_j to minimize the overall transport cost between p and q.

Let d^1 and d^2 represent normalized word counts across vocabulary, Word Mover's Distance is defined as $WMD(d^1, d^2) = W_1(d^1, d^2)$.

Hierarchical Optimal Topic Transport

If we represent document d^1, d^2 as distributions over topics, then **hierarchical optimal topic transport (HOTT)** is defined as

$$HOTT(d^1,d^2) = W_1\Big(\sum_{k=1}^{T} d_k^1 \delta_{t_k}, \sum_{k=1}^{T} d_k^2 \delta_{t_k}\Big)$$

where δ_{t_k} denotes a Dirac delta centered at topic t_k, weighted by d_k, the proportions of topic t_k contribute to the document.

Linear Optimal Transport

Let σ be a fixed reference measure on $\mathbb{R}^n$. The linear optimal transport embedding (LOT) is defined as a mapping $F_\sigma : P_2(\mathbb{R}^n) \mapsto L^2(\mathbb{R}^n, \sigma)$ where P_2 is probability space with second finite moment.

It takes a probability measure $\mu \in P_2(\mathbb{R}^n)$ to the optimal transport map T_σ^μ that pushes σ to μ, namely $F_\sigma(\mu) = T_\sigma^\mu$.

We approximate $HOTT$ in L^2-space and define **linear optimal topic transport distance (LOTT)** between document d^1, and d^2 as

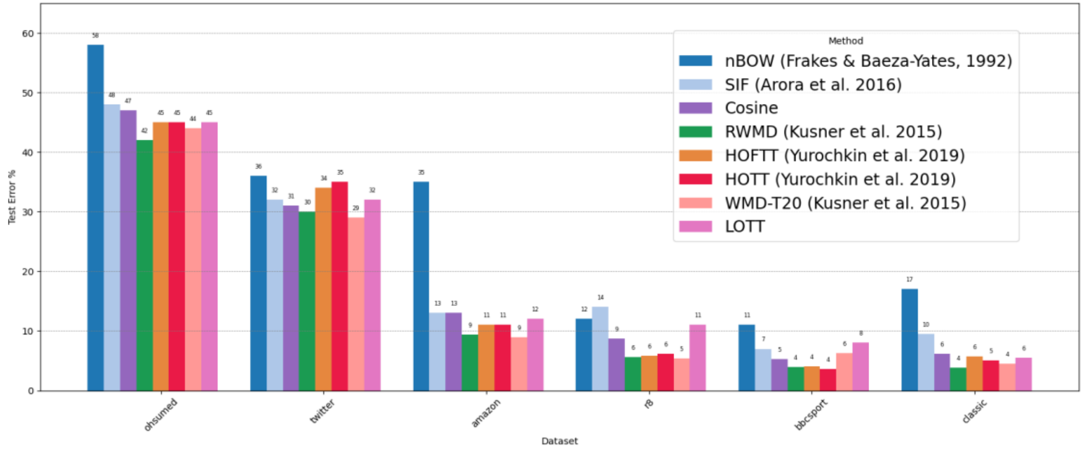
$$LOTT(d^1,d^2) = \|F_\sigma(d^1) - F_\sigma(d^2)\|_\sigma$$

where σ is gaussian and the embedding is in L^2-space.

Datasets

	Documents	Vocab	Words Overlaps	Uniquewords	Classes
BBCSPORT	737	3657	0.066	116.5	5
TWITTER	3108	1205	0.029	9.7	3
OHSUMED	9152	8261	0.046	59.4	10
CLASSIC	7093	5813	0.017	38.5	4
REUTERS8	7674	5495	0.06	35.7	8
AMAZON	8000	16753	0.019	44.3	4

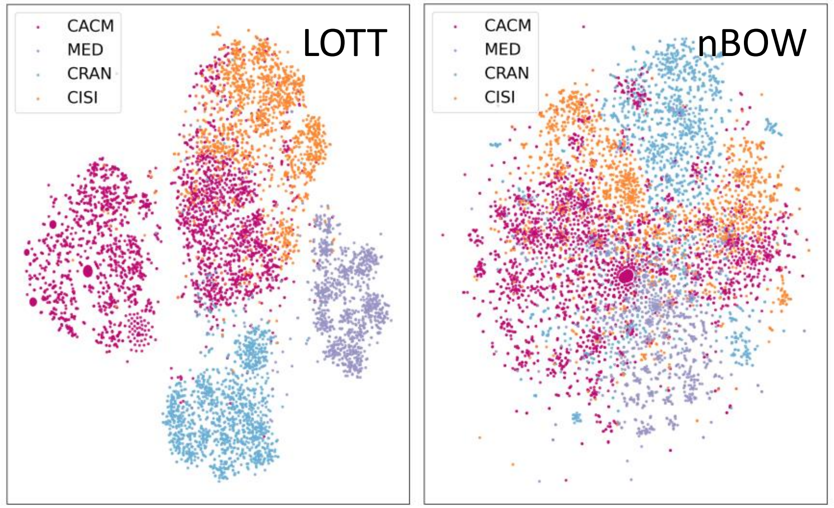
Classification Accuracy



Computational Efficiency

Dataset	RWMD	WMD	WMDT20	HOFTT	HOTT	LOTT
	Document pairs per second					
bbcsport	1494	526	1545	2016	2548	**129318**
twitter	2664	2536	2194	1384	1552	**201999**
ohsumed	454	377	473	829	908	**1402163**
classic	816	689	720	980	1053	**865099**
reuters8	834	685	672	918	989	**1312662**
amazon	289	259	253	927	966	**514521**

T-SNE Visualization



Sensitivity Analysis