中山大學
SUN YAT-SEN UNIVERSITY

Integrating Semantics and Neighborhood Information with Graph-Driven Generative Models for Document Retrieval

Zijing Ou, Qinliang Su, Jianxin Yu, Bang Liu, Jingwen Wang, Ruihui Zhao, Changyou Chen, Yefeng Zheng

June 16, 2021

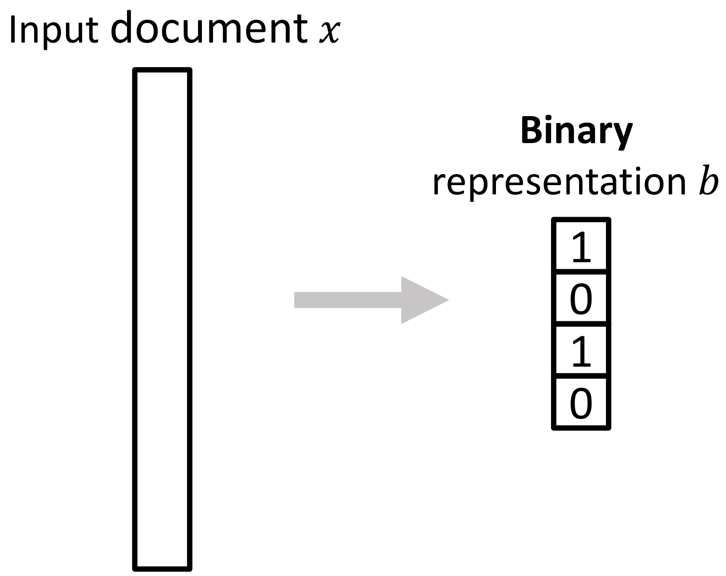
1 Learning Binarized Document Representations

Motivation:

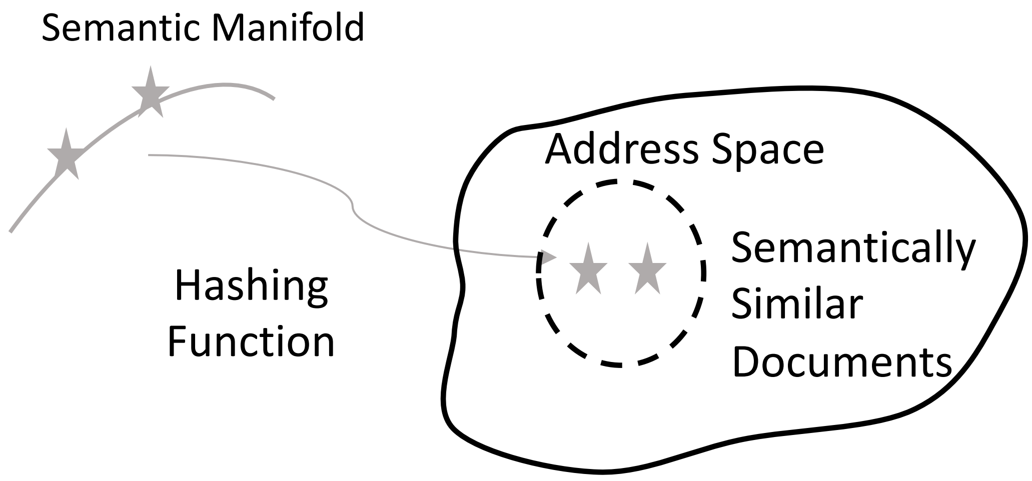
- **Reducing** **similarity-computation** and **storage-cost** in large scale information retrieval.

Goal:

- Learning a mapping from an input document x to a **Binary** representation b that capture its **semantic meaning**.

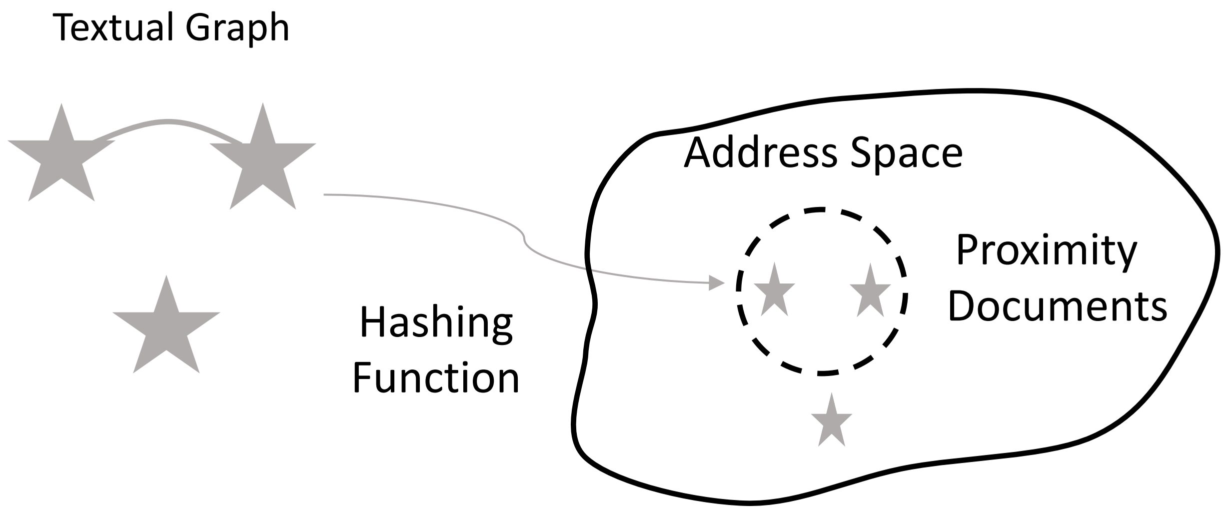


1 Semantics-Preserving Hashing



- Lean a binary code that preserves **semantic meaning**;
- The **more semantic information** is preserved, the **more similarity of codes** derived from semantic-similar documents appears.

1 Neighborhood-Preserving Hashing



- ➢ Lean a binary code that preserves **proximity**;
- ➢ Aim to retain the neighborhood information, such that **similar codes** can be produced for **neighboring documents**.

1 Learning Binarized Document Presentations

Semantics-Preserving Hashing (SPH)

- The document can be modeled by a **generative** model
$$p(x,z) = p_\theta(x|z)p(z),$$
- Semantics can be captured by **likelihood function** $p_\theta(x|z)$;
- Due to the $i.i.d.$ assumption, the model can be trained by maximized joint distribution of N documents:
$$P(X,Z) = \prod_{k=1}^{N} p_\theta(x_k|z_k)\, p(z_k).$$

Neighborhood-Preserving Hashing (NPH)

- Assume the **affinity matrix** $A = [a_{ij}]$ of the documents is available;
- Proximity can be preserved in the binary code b by optimizing the following minimization problem
$$\sum_{ij} a_{ij} ||b_i - b_j||^2 ;$$
- Therefore similar codes can be retained between neighboring documents.

How to simultaneously preserve the two types of information?

1 Heuristic Strategies

Combining two objectives together:

$$\underbrace{E_{q_\varphi(z,x)}[\log p_\theta(x|z)] - KL[q_\varphi(z|x)||p(x)]}_{\text{Semantics Preserving}} + \underbrace{\sum_{ij} a_{ij}||z_i - z_j||^2}_{\text{Neighborhood Preserving}}$$

Imposing codes to generate neighbors:

$$E_{q_\varphi(z,x)}[\underbrace{\boxed{\log p_\theta(N(x)|z)}}_{\text{Generate neighbors as well}}] - KL[q_\varphi(z|x)||p(x)]$$

- **Above methods lack basic principles to guide the integrated process of semantics and neighborhood information.**
- **How to simultaneously preserve the two types of information in a unified framework?**

2 Preserving Semantic and Neighborhood Information

Rewrite SPH:

- To simultaneously preserve semantic and neighborhood information, we first **rewrite** SHP method in compact form
$$p_\theta(X,Z) = p_\theta(X|\mathrm{Z})p_I(Z),$$
- For efficient training, the prior $p(Z)$ is generally selected as **diagonal Gaussian**
$$p_I(Z) = N(Z; 0, I_N \otimes I_d),$$
- Under the **variational inference** framework, by introducing approximate posterior $q_\varphi(Z|X)$, we can maximize the lower bound of $\log p(X)$ to train SHP model
$$L = E_{q_\varphi(Z|X)}[\log p_\theta(X|Z)] - KL[q_\varphi(Z|X)||p_I(Z)].$$

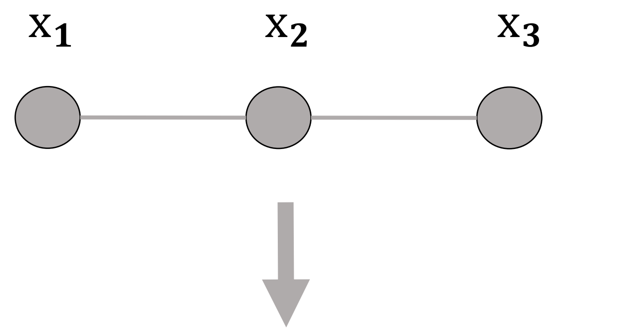
Note that:

- The **likelihood** $p_\theta(X|\mathrm{Z}) = \prod_{k=1}^{N} p_\theta(x_k|z_k)$ can effectively capture **semantic information**;
- Inspired by the property of covariance matrix in Gaussian, the **neighborhood information** can be introduced by using a ***non-diagonal*** **Gaussian prior**.

2 Preserving Semantic and Neighborhood Information

Introduce Neighborhood Information:

- Given an affinity matrix A, using $\lambda \in [0,1)$ to control correlation strength, the **neighborhood information** can be described as
$$I_N + \lambda A,$$
- To introduce neighborhood information, we can require that the representation Z are drawn from the following Gaussian
$$p_G(Z) = N(Z; 0, (I_N + \lambda A) \otimes I_d),$$
- However, compute the ELBO containing this **neighborhood aware prior** is **inefficient**, as the computation $KL[q_\varphi(Z|X)||p_G(Z)]$ involves the term of
$$((I_N + \lambda A) \otimes I_d)^{-1}.$$



$$\begin{array}{c} \\ z_1 \\ \\ z_2 \\ \\ z_3 \\ \end{array} \begin{array}{c} \begin{array}{ccc} \quad z_1 \quad & \quad z_2 \quad & \quad z_3 \quad \end{array} \\ \left(\begin{array}{cc|cc|cc} 1 & 0 & \lambda a_{12} & 0 & 0 & 0 \\ 0 & 1 & 0 & \lambda a_{12} & 0 & 0 \\ \hline \lambda a_{21} & 0 & 1 & 0 & \lambda a_{23} & 0 \\ 0 & \lambda a_{21} & 0 & 1 & 0 & \lambda a_{23} \\ \hline 0 & 0 & \lambda a_{32} & 0 & 1 & 0 \\ 0 & 0 & 0 & \lambda a_{32} & 0 & 1 \end{array}\right) \end{array}$$

2 Tree Approximations for Efficient Training

Approximation with One Spanning Tree :

- Let $G \triangleq (V, E)$ denote the corresponding graph of matrix A, where $V = \{1,2, \ldots, N\}$ is the set of documents indices; and $E = \{(i, j) | a_{ij} \neq 0\}$ is the set of connections between documents;
- From the graph G, a **spanning tree** $T \triangleq (V, E_T)$ can be obtained easily;
- We aim to propose a new prior that only captures **partial** neighborhood information, with the associated special structure being able to **facilitate** the training process.

2 Tree Approximations for Efficient Training

Tree-type Prior :

- To capture neighborhood information and facilitate training process, we construct a **tree-type prior** as

$$p_T(Z) = \prod_{i \in V} p_G(z_i) \prod_{(i,j) \in E} \frac{p_G(z_i, z_j)}{p_G(z_i) p_G(z_j)},$$

- $p_G(z_i, z_j)$ and $p_G(z_i)$ represent the one- and two- variable **marginal distributions** of $p_G(Z)$.

2 Tree Approximations for Efficient Training

Tree-type Posterior :

- Following the tree-type priors, a **tree-type posterior** is also constructed as

$$q_T(Z|X) = \prod_{i \in V} q_\varphi(z_i|x_i) \prod_{(i,j)\in E} \frac{q_\varphi(z_i, z_j|x_i, x_j)}{q_\varphi(z_i|x_i\ )q_\varphi(z_i|x_i\ )},$$

- $q_\varphi(z_i, z_j|x_i, x_j)$ is defined to be Gaussian, with its mean defined as $[\mu_i; \mu_j]$ and the **covariance matrix** defined as

$$\begin{bmatrix} diag(\sigma_i^2) & diag(\gamma_{ij} \odot \sigma_i \odot \sigma_j) \\ diag(\gamma_{ij} \odot \sigma_i \odot \sigma_j) & diag(\sigma_j^2) \end{bmatrix},$$

- γ_{ij} **controls** the correlation strength between z_i and z_j, whose element are restricted in $(0,1]$ to ensure **positive correlation**.

Efficient Training:

- By using the tree-type prior and posterior, the ELBO can be expressed as

$$
\begin{aligned}
L_T = & \sum_{i \in V} E_{q_T(z_i|x_i)}[\log p_\theta(x_i|z_i)] - KL[q_\varphi(z_i|x_i)||p_G(z_i)] \\
& + \sum_{(i,j) \in E_T} KL\left[q_\varphi\left(z_i, z_j|x_i, x_j\right)||p_G\left(z_i, z_j\right)\right] \\
& - KL\left[q_\varphi(z_i|x_i)||p_G(z_i)\right] - KL\left[q_\varphi\left(z_j|x_j\right)||p_G\left(z_j\right)\right]
\end{aligned}
$$

- Therefore the ELBO is broken down into the terms involving **single** or **pairwise** variables.

2 Tree Approximations for Efficient Training

Extend to Multiples Spanning Trees :

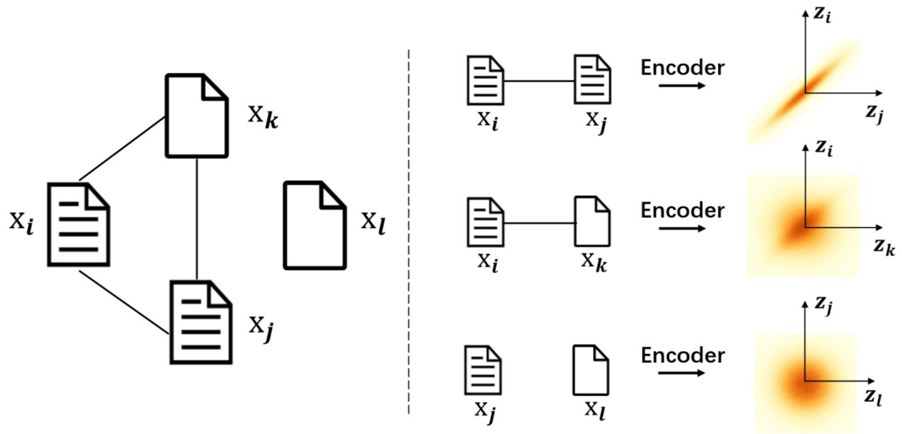
- From the graph G, we can construct a set of M spanning trees $T_{\mathrm{G}} = \{T_1, \ldots T_{\mathrm{M}}\}$.
- Based on the set of spanning trees, a **mixture-distribution** prior and posterior can be defined as

$$p_{MT}(Z) = \frac{1}{M}\sum_{T\in T_G} p_T(Z)\,, \qquad q_{MT}(Z|X) = \frac{1}{M}\sum_{T\in T_G} q_T(Z|X)\,;$$

- By applying log-sum inequality, the EBLO can be further lower bounded as

$$\begin{aligned} L_{MT} &= \frac{1}{M}\textstyle\sum_{T\in T_G} E_{q_T(Z|X)}[\log p_\theta(X|Z)] - KL[q_T(Z|X)||p_T(Z)] \\ &= \frac{1}{M}\textstyle\sum_{T\in T_G} L_T \end{aligned}$$

2 Details of Modeling



- **Variational Encoder** $q_\varphi(z_i|x_i)$
  - take single document as input, and outputs the mean and variance of Gaussian distribution $[\mu_i; \sigma_i^2] = f_\varphi(x_i)$.
- **Correlation Encoder**
  - take pairwise documents as input, and outputs the correlation coefficient $\gamma_{ij} = f_\varphi\left(x_i, x_j\right)$.
- **Generative Decoder** $p_\theta(x_i|z_i)$
  - take the latent variable z_i as input and output the document x_i.

3 Experiments: Setup

- **Datasets:** we evaluate the proposal method on three benchmarks: Reuters21579, 20Newsgroups, TMC;
- **TFIDF features** are utilized as input x for documents;
- The neighbors are selected as the **top-k** similar item for each document based on **cosine similarity of TFIDF**;
- We employ **precision** as the evaluation metric: the percentage of documents among the top 100 retrieved ones that belong to the same label (topic) with the query document.

3 Experiments: Semantic Hashing Evaluation

Method	Reuters				TMC				20Newsgroups				Avg
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	
SpH	0.6340	0.6513	0.6290	0.6045	0.6055	0.6281	0.6143	0.5891	0.3200	0.3709	0.3196	0.2716	0.5198
STH	0.7351	0.7554	0.7350	0.6986	0.3947	0.4105	0.4181	0.4123	0.5237	0.5860	0.5806	0.5443	0.5662
VDSH	0.7165	0.7753	0.7456	0.7318	0.6853	0.7108	0.4410	0.5847	0.3904	0.4327	0.1731	0.0522	0.5366
NbrReg	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.4120	0.4644	0.4768	0.4893	0.4249
NASH	0.7624	0.7993	0.7812	0.7559	0.6573	0.6921	0.6548	0.5998	0.5108	0.5671	0.5071	0.4664	0.6462
GMSH	0.7672	0.8183	0.8212	0.7846	0.6736	0.7024	0.7086	0.7237	0.4855	0.5381	0.5869	0.5583	0.6807
AMMI	0.8173	0.8446	0.8506	0.8602	0.7096	0.7416	0.7522	0.7627	0.5518	0.5956	0.6398	0.6618	0.7323
CorrSH	0.8212	0.8420	0.8465	0.8482	0.7243	0.7534	0.7606	0.7632	**0.5839**	0.6183	0.6279	0.6359	0.7355
SNUH	**0.8320**	**0.8466**	**0.8560**	**0.8624**	**0.7251**	**0.7543**	**0.7658**	**0.7726**	0.5775	**0.6387**	**0.6646**	**0.6731**	**0.7474**

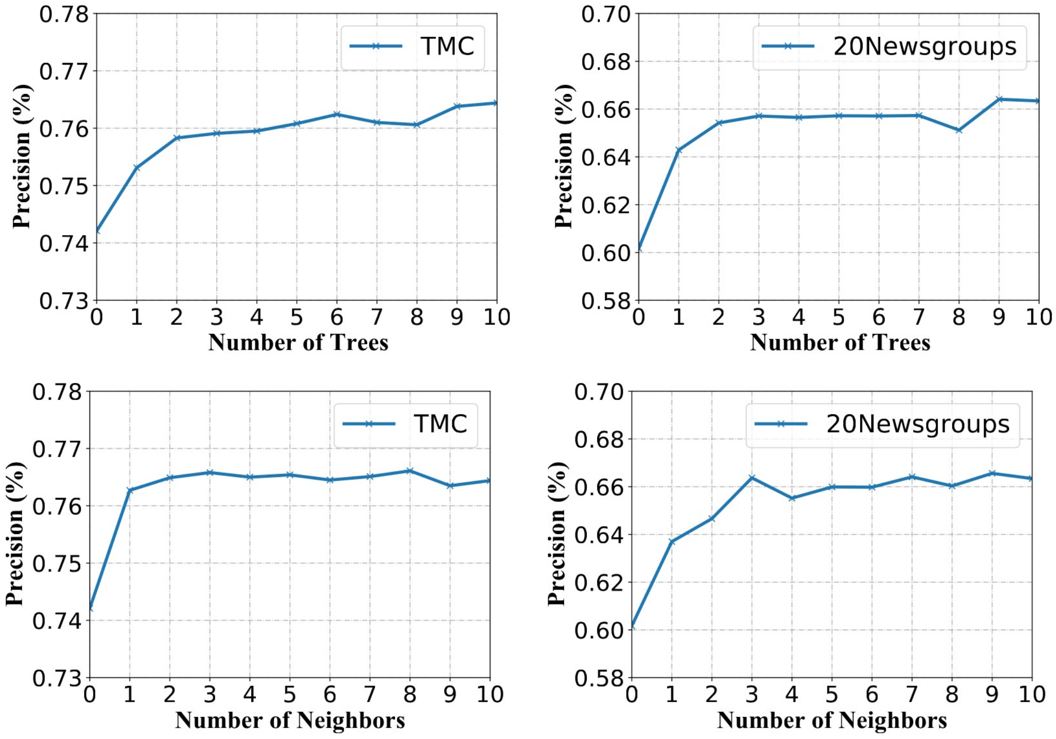
- **Perform on three datasets consistently better than baselines in most experimental settings.**

3 Experiments: Impact of Introducing Correlations

Ablation Study		16bits	32bits	64bits	128bits
Reuters	$\mathrm{SNUH}_{\mathrm{ind}}$	0.7823	0.8094	0.8180	0.8385
	$\mathrm{SNUH}_{\mathrm{prior}}$	0.8043	0.8295	0.8431	0.8460
	SNUH	**0.8320**	**0.8466**	**0.8560**	**0.8624**
TMC	$\mathrm{SNUH}_{\mathrm{ind}}$	0.6978	0.7307	0.7421	0.7526
	$\mathrm{SNUH}_{\mathrm{prior}}$	0.7177	0.7408	0.7518	0.7528
	SNUH	**0.7251**	**0.7543**	**0.7658**	**0.7726**
NG20	$\mathrm{SNUH}_{\mathrm{ind}}$	0.4806	0.5503	0.6017	0.6060
	$\mathrm{SNUH}_{\mathrm{prior}}$	0.5443	0.6071	0.6212	0.6014
	SNUH	**0.5775**	**0.6387**	**0.6646**	**0.6731**

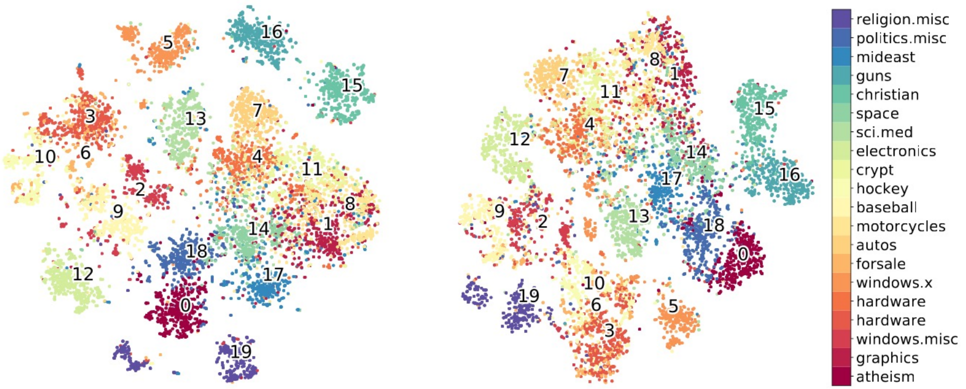
- By taking the **correlations** into account in the prior and posterior, significant **improvements** of SNUH can be observed.

3 Experiments: Effect of Spanning Trees



- Compared to not using any correlation, **one tree** alone can bring significant performance gains.

3 Experiments: Qualitative Analysis



(a) SNUH (b) AMMI

Distance	Category	Title/Subject
query	**hockey**	**NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-21-93**
1	hockey	NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-19-93
10	hockey	NHL Summary parse results for games played Thur, April 15, 1993
20	hockey	AHL playoff results (4/15)
50	forsale	RE: == MOVING SALE ===
70	hardware	Re: Quadra SCSI Problems?
90	politics.misc	Re: Employment (was Re: Why not concentrate on child molesters?

3 Conclusions

- We proposed an effective and efficient semantic hashing method to preserve both the semantics and neighborhood information of documents.
- we applied a graph-induced Gaussian prior to model the two types of information in a unified framework.
- To facilitate training, a tree structure approximation was further developed to decompose the ELBO into terms involving only singleton or pairwise variables.
- Extensive evaluations demonstrated that our model significantly outperforms baseline methods by incorporating both the semantics and neighborhood information.

Thanks for Listening !