ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Learning &
Adaptive Systems

Master Thesis

Spatial Summarization of Image Collections

Spring Term 2016

This dissertation is submitted for the degree of *Master of Science ETH in Computer Science*

Supervised by:
Prof. Dr. Andreas Krause
Dr. Sebastian Tschiatschek
Alkis Gotovos, M.Sc.

Author:
Diego Alfonso Ballesteros Villamizar

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

List of Figures

List of Tables

Preface

Acknowledgments and such …

Abstract

The Abstract …

Symbols

Symbols

η	Noise to data ratio
κ	Largest subset size
θ	Model parameters
σ	Gaussian distribution's standard deviation
$\mathcal{D}$	Set of data samples
K	Latent coherence dimensions
L	Latent diversity dimensions
M	Number of features
$\mathcal{N}$	Set of noise samples
$\mathbb{R}$	Real numbers
S, T	Subset
V	Ground set
Z	Normalization constant

Matrices

$\mathbf{C}$	Feature coherence weights
$\mathbf{D}$	Feature diversity weights
$\mathcal{I}$	Identity
$\mathbf{W}^C$	Item coherence weights
$\mathbf{W}^D$	Item diversity weights
$\mathbf{X}$	Item features

Vectors

$\mathbf{a}$	Utility weights for features
$\mathbf{u}$	Item utilities

Indices

i	Item
j	Feature

c	Coherence dimension
d	Diversity dimension

Acronyms and Abbreviations

ETH	Eidgenössische Technische Hochschule
FLID	Facility Location Diversity
FLDC	Facility Location Diversity and Coherence
FFLDC	Featurized Facility Location Diversity and Coherence
MLE	Maximum Likelihood Estimation
NCE	Noise Contrastive Estimation
LAS	Learning & Adaptive Systems
SGD	Stochastic Gradient Descent

Chapter 1

Introduction

Motivation (tourist routes and summarization), and quick idea of the methodology.

1.1 Contributions

Summarization of contributions.

Chapter 2

Related Work

2.1 Probabilistic Sub/super-modular models

Talk about the importance of these models and the recent interest in them, e.g. DPPs, Sampling.

2.2 Image collection summarization

Generally about the problem and the different approaches.

2.3 Tourist routes recommendation

The initial papers I found on this, basically Markov chains and clustering.

Chapter 3

Models

This chapter explores different probabilistic models for the problem, explains how they can be efficiently learned from data and presents the learning algorithm performance on synthetic data.

3.1 Submodular model: FLID

The Facility Location Diversity (FLID) model was first proposed in [1] and belongs to the class of log-submodular probability distributions over sets.

Definition 3.1.1. A distribution over sets $S \subseteq V$, where w.l.o.g $V = \{1, \ldots, |V|\}$, of the form $P(S) = \frac{1}{Z}\exp(F(S))$ is called a log-submodular probability distribution, if $F(S)$ is a submodular function [2].

Definition 3.1.2. A function $F : 2^V \rightarrow \mathbb{R}$ is submodular if it exhibits a "diminishing returns" property [3], namely:

$$\forall S, T \subseteq V : S \subseteq T, i \notin T \mid F(S \cup i) - F(S) \geq F(T \cup i) - F(T)$$

Submodular functions intuitively indicate that adding an item to a smaller set results in a larger gain than adding it to a larger one. This is a natural property in the context of summarization where adding more information to a large summary is less effective than adding it to a smaller one.
The FLID model consists of two terms, first a modular one which considers the utility or relevance of the items in the set, namely:

$$P(S) \propto \exp\left(\sum_{i \in S} u_i\right) \tag{3.1}$$

Where $u_i \in \mathbb{R}$ quantifies the utility of item i. In the context of location summarization, this utility could be proportional to the popularity of a place and how many times it has been photographed.
The diversity term is based on the idea of a latent concept space of dimension L where each item can be represented with a vector $\mathbf{w}_i^D \in \mathbb{R}^L_{\geq 0}$. The representation in this space allows the model to identify items that are similar and penalize sets that include them together. Formally, the diversity term for a set $S \subseteq V$ is:

$$\text{div}(S) = \sum_{d=1}^{L} \left(\max_{i \in S} w_{i,d}^D - \sum_{i \in S} w_{i,d}^D \right) \tag{3.2}$$

Putting this two terms together results in the FLID probability model proposed in [1].

Table 3.1: FLID probability distribution for the scenario in example 3.1.2

S	$P(S)$
$\{h,s\},\{h,f\}$	≈ 0.41
$\{h\},\{s\},\{f\}$	≈ 0.06
$\{\},\{s,f\},\{h,s,f\}$	≈ 0.00

$$P(S) = \frac{1}{Z}\exp\left(\sum_{i\in S} u_i + \sum_{d=1}^{L}\left(\max_{i\in S} w_{i,d}^D - \sum_{i\in S} w_{i,d}^D\right)\right) \tag{FLID}$$

In this model, $\mathbf{u} \in \mathbb{R}^{|V|}$ is the vector of utilities and $\mathbf{W}^D \in \mathbb{R}_{\geq 0}^{|V|\times L}$ is the diversity weight matrix where each row is the aforementioned $\mathbf{w}_i^D$ vector.

3.1.1 Partition function

In log-submodular probabilistic models, the normalization constant Z is known as the *partition function* [2] and its exact computation is known to be #P-complete [4]. However, it has been proven [1] that for FLID the partition function can be computed exactly in $O(|V|^{L+1})$ time, which can be efficient for $L \ll |V|$. This is an important property because the partition function is necessary to compute marginal probabilities and other quantities.

3.1.2 Example: Two landmarks

In order to illustrate the model, consider a town with 3 popular locations: A town hall (h), a statue (s) and a fountain (f). Data shows that visitors only take photos at the town hall and the statue, or at the town hall and the fountain. This can be modeled with FLID, introducing a latent concept that discourages taking photos of both the fountain and statue. Concretely, let $V = \{h,s,f\}$ and $\mathbf{u} = (2,2,2)$, indicating that all locations are equally popular. A suitable diversity weight vector would then be $\mathbf{W}^D = (0,20,20)^\intercal$. Table 3.1 shows the resulting probabilities of the subsets, accurately representing the aforementioned description of the problem.

3.2 Mixed model: FLDC

Diversity is an important property in the context of summarization [1], however coherence is also a desired property of summaries, especially in the context of structured timeline summarization [5]. Balancing coherence and diversity is considered a challenge because maxiziming only one of these properties may lead to poor results on the other one [6].
In order to model coherence in this model, the addition of a log-supermodular term analogous to the diversity term 3.2 is proposed.

Definition 3.2.1. A function $F : 2^V \to \mathbb{R}$ is supermodular iff $-F(S)$ is submodular.

The supermodular term encodes the items into another latent concept space, of dimension K, where sets containing items with high values in some latent dimension are rewarded. Hence modeling complementarity between items.
Concretely, consider a spatial summary of a city where people tend to stay close to the city center. A possible latent dimension could encode the distance to the center, and rewarding coherence on this dimension would create summarizes where all locations are close together which is the modeled behavior.

Table 3.2: Probability distribution for example 3.2.1

S	$P(S)$
$\{1,2\},\{3,4\}$	0.5
$2^V \setminus \{\{1,2\},\{3,4\}\}$	0.0

The extended model will be reffered to as the Facility Location Diversity and Coherence (FLDC) model and its probability distribution is:

$$P(S) = \frac{1}{Z}\exp\left(\sum_{i\in S} u_i + \text{div}(S) + \sum_{c=1}^{K}\left(\sum_{i\in S} w_{i,c}^C - \max_{i\in S} w_{i,c}^C\right)\right) \tag{FLDC}$$

Where $\mathbf{w}_i^C$ is the i-th row of the $\mathbf{W}^C \in \mathbb{R}_{\geq 0}^{|V|\times K}$ matrix and corresponds to the representation of item i in the concept space of dimension K.
It should be noted that the FLDC model is neither log-submodular or log-supermodular unless $K = 0$ or $L = 0$, respectively. However, it can be used and learned in a similar fashion as the FLID model.

3.2.1 Example: Disjoint pairs

As an example of the extended model, consider the distribution presented in table 3.2 for $V = \{1,2,3,4\}$. It represents a set of two disjoint pairs, which indicates there exists a diversity component between the two pairs whilist having a coherence component between the items contained in each pair.
Concretely, the weight matrices $\mathbf{W}^D, \mathbf{W}^C$ in figure 3.1 illustrate one possible instance of the model. The corresponding utility vector is $\mathbf{u} = \overrightarrow{0}$, because there is no indication that individual items are favored over the pairs. Note that this model is easily interpretable and accurately realizes the distribution.

3.3 Featurized model: FFLDC

An important characteristic of the FLDC and FLID models is that they are agnostic to the type of items in the ground set, this allows its application to a wide range of problems without prior knowledge. However the downside is that the model has no capability to make use of information about the items, if available, to improve the modeling of the data. Moreover, if a new item is added to the set there is no way to generalize the existing knowledge about similar items to it.
In order to solve these problems, a further extension to the model is proposed. Firstly, the information about each item ($i \in V$) is represented as a vector $\mathbf{x}_i \in \mathbb{R}^M$ where each component is a feature, e.g. for venues one feature could be its aggregated rating while another indicates whether it is indoors or outdoors.
Then, the utility vector $\mathbf{u}$ and weight matrices $\mathbf{W}^D, \mathbf{W}^C$ are factorized to include the feature matrix $X \in \mathbb{R}^{|V|\times M}$ as follows:

$$\mathbf{u} = \mathbf{Xa} \tag{3.3}$$

$$\mathbf{W}^D = \mathbf{XD} \tag{3.4}$$

$$\mathbf{W}^C = \mathbf{XC} \tag{3.5}$$

Where $a \in \mathbb{R}^M$ represents the contribution of each feature to the total utility of an item, whilist $\mathbf{D} \in \mathbb{R}^{M\times L}$ and $\mathbf{C} \in \mathbb{R}^{M\times K}$ encode the contribution of each feature to each latent

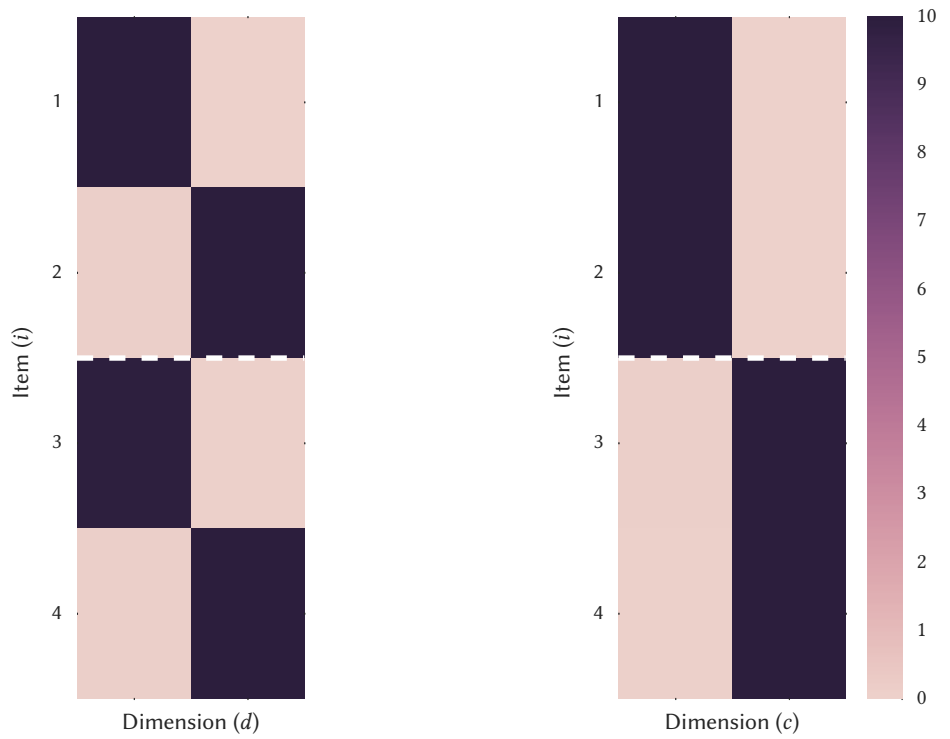

Figure 3.1: Diversity (left) and coherence weights (right) for FLDC model in example 3.2.1. The dotted line divides the paired items.

diversity and coherence dimension, respectively. The intuition behind this factorization is that the information about the items can enhance the latent representations, hence producing a richer model.
The extended model will be reffered to as the Featurized Facility Location Diversity and Coherence (FFLDC) model and its probability distribution is:

$$P(S) = \frac{1}{Z} \exp\left(\sum_{i \in S} \mathbf{X}_{i,*}\mathbf{a} + fdiv(S) + fcoh(S)\right) \tag{FFLDC}$$

$$fdiv(S) = \sum_{d=1}^{L}\left(\max_{i \in S} \mathbf{X}_{i,*}\mathbf{D}_{*,d} - \sum_{i \in S} \mathbf{X}_{i,*}\mathbf{D}_{*,d}\right) \tag{3.6}$$

$$fcoh(S) = \sum_{c=1}^{K}\left(\sum_{i \in S} \mathbf{X}_{i,*}\mathbf{C}_{*,c} - \max_{i \in S} \mathbf{X}_{i,*}\mathbf{C}_{*,c}\right) \tag{3.7}$$

Remark. If $X = \mathcal{I}$, then FFLDC is equivalent to FLDC.

The use of features also allows the extension of the model to previously unknown items, hence solving the aforementioned problem of generalization. This is because the parameters of the FFLDC model, i.e. $\mathbf{a}, \mathbf{D}, \mathbf{C}$, do not depend on the ground set V but rather on the space of features $\mathbb{R}^M$. If an item $j \notin V$ is considered, a model learned on only items in V can immediately be applied to the new set $V \cup \{j\}$, contrary to the case of FLID or FLDC where it would require adding a new row to the weight matrices and learning its components.

Table 3.3: Synthetic data for example 3.3.1

Locations visited (S)	$P(S)$
$\{0, 2\}$	0.29
$\{2, 3\}$	0.26
$\{2, 5\}$	0.14
$\{0\}, \{1\}$	0.05
$\{2\}, \{3\}, \{4\}, \{5\}$	0.04

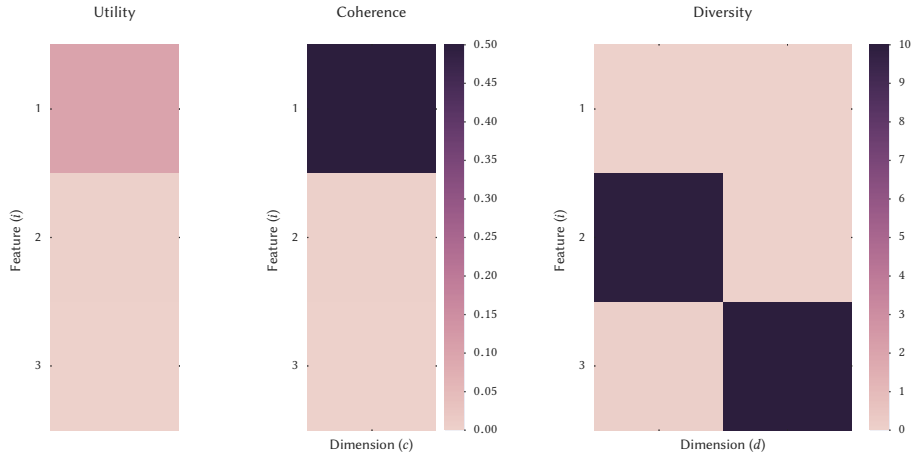

Figure 3.2: FFLDC sample model for example 3.3.1.

3.3.1 Example: Rated locations

A simple town has 6 popular locations, each of them has been rated from 0 (terrible) to 5 (excellent). It is known that a typical visitor cares about these ratings but also is interested in visiting places that are outdoors and/or serve food. Given data from previous visitors, shown in table 3.3, the task is to model this behavior using the FFLDC model.
In this example, there is knowledge about the items and what features are relevant for the data. These features are summarized in the matrix $\mathbf{X}$ in equation 3.8. The first column corresponds to the aforementioned rating, the second and third are binary features indicating whether the location is an outdoor one and whether it serves food, respectively.

$$\mathbf{X} = \begin{pmatrix} 4 & 1 & 0 \\ 4 & 1 & 1 \\ 3 & 0 & 1 \\ 3 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & 0 \end{pmatrix} \tag{3.8}$$

Looking at the data and features, it is possible to draw a FFLDC model that encourages diversity on the second a third feature while assigning a positive utility and coherence values to the first feature. One such model is presented in figure 3.2 and accurately realizes the distribution form table 3.3.
With this model, if a new location j is considered then it is straightforward to estimate what its probability of being visited $P(j \in S)$ would be, only knowing its feature repre-

sentation.

3.4 Learning from data

Having described the models, the next step is to devise a way to learn them efficiently from data. As noted in [1], maximum likelihood estimation (MLE) is intractable in FLID due to the complexity of calculating the partition function for large L. This result can be extended to the FLDC model using the fact that in the algorithm presented in [1] the weights can be either positive or negative, therefore it can include the coherence weights as negative diversity weights.

Proposition 3.4.1. The time complexity of calculating the partition function for the FLDC model is $O(|V|^{L+M+1})$, using a modified version of the algorithm presented in [1].

Additionally, this result can be extended to the FFLDC case because an FFLDC model can always be converted to an equivalent FLDC model through the factorization in 3.3-3.5. This has a time complexity of $O(M|V|(L+K))$, corresponding to the matrix multiplications, which is considerably smaller than the complexity of the partition function computation for FLDC.
As an alternative to MLE, Noise Constrastive Estimation (NCE) has been proposed as a method for estimating unnormalized probablistic models from observed data [7]. The following section describes this method in more detail and its application to the models.

3.4.1 NCE Learning

The idea behind NCE, as presented in [7], is to transform the unsupervised learning task of estimating a probability density from data into a supervised classification task. In order to do this, the observed data $\mathcal{D}$, assumed to be drawn from an unknown distribution P_d, is compared to an artificially generated set of noise samples $\mathcal{N}$ drawn from a known distribution P_n. The classification task is setup to optimize the likelihood of correctly discriminating each sample as either data or noise.
Formally, denote $\mathcal{A}$ as the complete set of labeled samples, i.e. $\mathcal{A} = \{(S, Y_s) : S \in \mathcal{D} \cup \mathcal{N}\}$ where $Y_s = 1 \equiv S \in \mathcal{D}$ and $Y_s = 0 \equiv S \in \mathcal{N}$. Additionally, η is the ratio between noise and data samples, i.e. $\eta = {}^{|\mathcal{N}|}/_{|\mathcal{D}|}$.
The goal is to estimate the posterior probabilities $P(Y_s = 1 \mid S; \theta)$ and $P(Y_s = 0 \mid S; \theta)$, in order to discriminate noise from data samples. These are defined as:

$$P(Y_s = 1 \mid S; \theta) = \frac{\hat{P}_d(S; \theta)}{\hat{P}_d(S; \theta) + \eta P_n(S)} \tag{3.9}$$

$$P(Y_s = 0 \mid S; \theta) = \frac{\eta P_n(S)}{\hat{P}_d(S; \theta) + \eta P_n(S)} \tag{3.10}$$

It is worth noting that estimated distribution $\hat{P}_d$ is used instead of P_d, because the real density is unknown. As indicated in [7], $\hat{P}_d$ can be an unnormalized distribution for NCE where the partition function Z is included in the set of parameters θ as $\hat{Z}$.
Estimating the posterior probabilities is equivalent to maximizing the conditional log-likelihood objective in 3.11 [7].

$$g(\theta) = \sum_{S \in \mathcal{D}} \log P(Y_s = 1 \mid S; \theta) + \sum_{S \in \mathcal{N}} \log P(Y_s = 0 \mid S; \theta) \tag{3.11}$$

In the case of the models described before, the parameters to adjust in order to maximize this objective are:

- FLID [1]: $\theta = [\hat{Z}, \mathbf{u}, \mathbf{W}^D]$.
- FLDC: $\theta = [\hat{Z}, \mathbf{u}, \mathbf{W}^D, \mathbf{W}^C]$.
- FFLDC: $\theta = [\hat{Z}, \mathbf{a}, \mathbf{D}, \mathbf{C}]$.

Finally, a couple of important conditions for NCE to work [7], are:

1. The parameterized probability function $\hat{P}_d(S;\theta)$ must be of the same family as the real distribution P_d, i.e. $\exists\theta^* \mid \hat{P}_d(S;\theta^*) = P_d$.
2. The noise distribution P_n is nonzero whenever P_d is nonzero.

There are no constraints in the optimization method to use for this problem [7]. Stochastic Gradient Descent (SGD) will be used for the proposed models, this is the same method used in [1] for the FLID model. SGD is a gradient-based method that has proven effective in large-scale learning tasks due to its efficiency when the computation time is a limiting factor [8][9], hence making it appropriate for the scale of the problem considered in this work.
In each iteration, the sub-gradient of the objective function $g(\theta)$, this reduces to the computation of $\nabla P(Y_s = y \mid S;\theta)$. The following sections present the sub-gradient for each of the proposed models.

FLID

Show the gradient referencing from the AISTATS paper.

FLDC

Add the gradient term for the coherence term.

FFLDC

Derive the gradients with features, show that the are the same as FLDC with X = I.

3.4.2 Learning FLID

Show the results from learning the synthetic dataset from section 3.1.2.

3.4.3 Learning FLDC

Show the results from learning the synthetic dataset from section 3.2.1.

3.4.4 Learning FFLDC

Show the results from learning the synthetic dataset from section 3.3.1.

3.4.5 Adagrad

Quickly summarize the method, mention that it will be useful later.

3.4.6 Hyper-parameter sensitivity

Use the synthetic datasets to show the effect of the following parameters:

- Number of iterations
- Noise size
- Step size (with/without adagrad)

Chapter 4

Experimental Setup

4.1 Flickr Dataset

Dataset description and how it was collected.

4.2 Path finding

How the actual sets (paths) are built from the data, and the 2 sets of data to be explore 10 items (small) and 100 items (big).

4.3 NCE learning

Details on the implementation of NCE, noise generation.

4.4 Baselines

The baseline models (Markov, distance).

4.5 Evaluation

10 fold evaluation, accuracy and MRR.

Chapter 5

Results

5.1 Small dataset

Results on the small dataset. Difference between FLID, FLDC, FFLDC models. Different feature sets.

5.2 Large dataset

Results on the large dataset. Difference between FLID, FLDC, FFLDC models. Different feature sets.

Chapter 6

Conclusion

Bibliography

[1] S. Tschiatschek, J. Djolonga, and A. Krause, "Learning probabilistic submodular diversity models via noise contrastive estimation," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2016.

[2] J. Djolonga and A. Krause, "From MAP to marginals: Variational inference in bayesian submodular models," in *Neural Information Processing Systems (NIPS)*, December 2014.

[3] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, February 2014, ch. 3.

[4] M. Jerrum and A. Sinclair, "Polynomial-time approximation algorithms for the ising model," in *Automata, Languages and Programming*. Springer Berlin Heidelberg, July 1990, pp. 462–475.

[5] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2011, pp. 745–754.

[6] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 899–908.

[7] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 307–361, January 2012.

[8] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics*. Heidelberg: Physica-Verlag HD, September 2010, pp. 177–186.

[9] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the Twenty-first International Conference on Machine Learning*. New York, NY, USA: ACM, 2004, pp. 116–123.