

Latent Semantic Mapping: Principles & Applications

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2007

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Latent Semantic Mapping: Principles & Applications
Jerome R. Bellegarda

ISBN: 978-3-031-01428-4 paperback
ISBN: 978-3-031-01428-4 paperback

ISBN: 978-3-031-02556-3 ebook
ISBN: 978-3-031-02556-3 ebook

DOI: 10.1007/978-3-031-02556-3

A Publication in the Springer series

SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING #3

Lecture #3

Series Editor: B.H. Juang, Georgia Tech

Library of Congress Cataloging-in-Publication Data

Series ISSN: 1932-121X print
Series ISSN: 1932-1678 electronic

First Edition

10 9 8 7 6 5 4 3 2 1

Latent Semantic Mapping: Principles & Applications

Jerome R. Bellegarda

Apple Inc.

SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING #3

ABSTRACT

Latent semantic mapping (LSM) is a generalization of latent semantic analysis (LSA), a paradigm originally developed to capture hidden word patterns in a text document corpus.

In information retrieval, LSA enables retrieval on the basis of conceptual content, instead of merely matching words between queries and documents. It operates under the assumption that there is some latent semantic structure in the data, which is partially obscured by the randomness of word choice with respect to retrieval. Algebraic and/or statistical techniques are brought to bear to estimate this structure and get rid of the obscuring “noise.” This results in a parsimonious continuous parameter description of words and documents, which then replaces the original parameterization in indexing and retrieval.

This approach exhibits three main characteristics:

- *discrete* entities (words and documents) are mapped onto a *continuous* vector space;
- this mapping is determined by *global correlation patterns*; and
- *dimensionality reduction* is an integral part of the process.

Such fairly generic properties are advantageous in a variety of different contexts, which motivates a broader interpretation of the underlying paradigm. The outcome (LSM) is a data-driven framework for modeling meaningful global relationships implicit in large volumes of (not necessarily textual) data.

This monograph gives a general overview of the framework, and underscores the multifaceted benefits it can bring to a number of problems in natural language understanding and spoken language processing. It concludes with a discussion of the inherent tradeoffs associated with the approach, and some perspectives on its general applicability to data-driven information extraction.

KEYWORDS

natural language processing, long-span dependencies, data-driven modeling, parsimonious representation, singular value decomposition.

Contents

I.	Principles	1
1.	Introduction	3
1.1	Motivation	3
1.2	From LSA to LSM	4
1.3	Organization	7
1.3.1	Part I	7
1.3.2	Part II	8
1.3.3	Part III	8
2.	Latent Semantic Mapping	9
2.1	Co-occurrence Matrix	9
2.2	Vector Representation	10
2.2.1	Singular Value Decomposition	10
2.2.2	SVD Properties	11
2.3	Interpretation	11
3.	LSM Feature Space	15
3.1	Closeness Measures	15
3.1.1	Unit–Unit Comparisons	15
3.1.2	Composition–Composition Comparisons	16
3.1.3	Unit–Composition Comparisons	16
3.2	LSM Framework Extension	17
3.3	Salient Characteristics	18
4.	Computational Effort	21
4.1	Off–Line Cost	21
4.2	Online Cost	22
4.3	Possible Shortcuts	22
4.3.1	Incremental SVD Implementations	23
4.3.2	Other Matrix Decompositions	23
4.3.3	Alternative Formulations	24

5.	Probabilistic Extensions	25
5.1	Dual Probability Model	25
5.1.1	Composition Model	25
5.1.2	Unit Model	26
5.1.3	Comments	27
5.2	Probabilistic Latent Semantic Analysis	27
5.3	Inherent Limitations	29
II.	Applications	31
6.	Junk E-Mail Filtering	33
6.1	Conventional Approaches	33
6.1.1	Header Analysis	33
6.1.2	Rule-Based Predicates	34
6.1.3	Machine Learning Approaches	34
6.2	LSM-Based Filtering	35
6.3	Performance	38
7.	Semantic Classification	41
7.1	Underlying Issues	41
7.1.1	Case Study: Desktop Interface Control	41
7.1.2	Language Modeling Constraints	42
7.2	Semantic Inference	43
7.2.1	Framework	43
7.2.2	Illustration	43
7.3	Caveats	44
8.	Language Modeling	49
8.1	N -Gram Limitations	49
8.2	MultiSpan Language Modeling	50
8.2.1	Hybrid Formulation	50
8.2.2	Context Scope Selection	51
8.2.3	LSM Probability	52
8.3	Smoothing	52
8.3.1	Word Smoothing	53
8.3.2	Document Smoothing	53
8.3.3	Joint Smoothing	54

9.	Pronunciation Modeling	55
9.1	Grapheme-to-Phoneme Conversion	55
9.1.1	Top-Down Approaches	55
9.1.2	Illustration	56
9.1.3	Bottom-Up Approaches	57
9.2	Pronunciation by Latent Analogy	58
9.2.1	Orthographic Neighborhoods	58
9.2.2	Sequence Alignment	59
10.	Speaker Verification	63
10.1	The Task	63
10.2	LSM-based speaker verification	64
10.2.1	Single-Utterance Representation	64
10.2.2	LSM-Tailored Metric	65
10.2.3	Integration with DTW	67
11.	TTS Unit Selection	71
11.1	Concatenative Synthesis	71
11.2	LSM-Based Unit Selection	72
11.2.1	Feature Extraction	72
11.2.2	Comparison to Fourier Analysis	73
11.2.3	Properties	74
11.3	LSM-Based Boundary Training	75
III.	Perspectives	77
12.	Discussion	79
12.1	Inherent Tradeoffs	79
12.1.1	Descriptive Power	79
12.1.2	Domain Sensitivity	80
12.1.3	Adaptation Capabilities	81
12.2	General Applicability	81
12.2.1	Natural Language Processing	81
12.2.2	Generic Pattern Recognition	82

13. Conclusion	85
13.1 Summary	85
13.2 Perspectives	86
Bibliography	89
Author Biography	101

List of Figures

1.1	Improved Topic Separability Under LSA	5
1.2	Typical Hand-Labeled Versus LSA Topics	7
2.1	Illustration of Equations (2.5) and (2.6)	12
2.2	Singular Value Decomposition (SVD)	12
3.1	LSM Framework Extension	18
6.1	LSM-based Junk E-mail Filtering	36
6.2	Finding the Semantic Anchors	37
6.3	Vector Representation for a New E-mail	37
7.1	Traditional Desktop Interface Control	42
7.2	LSM-based Desktop Interface Control	44
7.3	Semantic Inference in Latent Semantic Space $\mathcal{L}$ ($R = 2$)	45
9.1	Top-down Grapheme-to-phoneme Conversion	56
9.2	Bottom-up Grapheme-to-phoneme Conversion	58
9.3	Pronunciation by Latent Analogy Framework	59
9.4	Orthographic Neighbors in Orthographic Space $\mathcal{L}$ ($R = 2$)	60
10.1	Performance Space of LSM+DTW Approach	68
11.1	Conventional Feature Extraction Framework	72
11.2	LSM-based Feature Extraction Framework	73
11.3	Iterative Training of Unit Boundaries	76
12.1	Finding the Encoding Anchors	83

List of Tables

1.1	Information Retrieval Terminology	4
1.2	Two Typical LSA Word Clusters.....	6
6.1	Comparison with Naive Bayes Approaches	38
9.1	Example of Sequence Alignment for “ <i>Krishnamoorthy</i> ”	61