The Evolution of Bayesian-related Research over Time: A Temporal Text Mining Task

Alta de Waal

Defence, Peace, Safety & Security

Council for Scientific and Industrial Research

Pretoria

South Africa

Abstract

Temporal Text Mining (TTM) is the task of discovering temporal patterns in text collected over time. This is useful in application domains where each entity of text in a text stream (usually a document or publication) has a meaningful timestamp [2].

For example, research papers in the field of Bayesian analysis have publication dates that can be viewed as time stamps. In this text stream, interesting temporal patterns could exist.

A research field such as Bayesian analysis could inherit patterns of change over time. These patterns could include evolution in research topics and sphere of interest of researchers.

Abstract (cont.)

The aim of this study is to evaluate the discovery of temporal themes in a text stream.

We evaluate a probabilistic model for unsupervised learning to solve the problem and a scheme for theme evolution visualisation is proposed.

The proposed methods will be evaluated on a collection of Bayesian Analysis abstracts (www.bayesian.org). The output will be a temporal summary of Bayesian related research themes and how they evolve over time, captured in a graph

Definitions

Data: Collection of time-indexed documents:

$$C = \{d_1, d_2, ..., d_T\}$$

Each document is a sequence of words from a vocabulary set:

$$V = \{w_1, w_2, ..., w_N\}$$

Collection *C* is partitioned into subcollections defined by time intervals. A subcollection of *C* consists of documents in the time span that defines the subcollection.

Themes

A theme in a text collection *C* is a probabilistic distribution of words that characterises a semantically coherent topic [2].

Theme *i* is represented by θ

Each theme in the text collection is represented by a **multinomial distribution**. This is also known as a **unigram language model**:

$$\{p(w \mid \theta)\}_{w \in V} \quad \text{where} \quad \sum_{w=1}^{N} p(w \mid \theta) = 1$$

Multinomial Mixture Model

$\theta_1,...,\theta_k$ are *k* multinomial models that represent *k* themes in document collection C

The multinomial mixture model is [3]:

$$P(C_d;\alpha,\theta)=\sum_{j=1}^{k}\alpha_k\frac{l_d!}{\prod_{w=1}^{n_W}C_d(w)}\prod_{w=1}^{n_W}\theta_{wj}^{C_d(w)}$$

l_d Length of document *d*

$C_d(w)$ Count of word *w* in document *d*

Multinomial Mixture Model (cont.)

$\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ Probabilities for *k* themes

$\theta = (\theta_1, \theta_2, ..., \theta_{n_{wt}})$ Probabilities for theme-specific words

The **log-likelihood** of collection C

$$\sum_{d \in C} \sum_{w \in V} \left[C_d(w) \log \left(\sum_{j=1}^{k} \alpha_k \prod_{w=1}^{n_W} \theta_{wt}^{C_d(w)} \right) \right]$$

Estimate model parameters:
EM (expectation-maximisation) Algorithm [4]

E-step: Compute the expected values p_{ij} of the hidden indicator variables z_{ij}.

z_{ij} = 1 if data was generated by the ith component

= 0 otherwise

M-step: Find new values of the parameters that maximises the likelihood of the data

Updating Formulas

$p(z_{d,w} = j)$ indicates word ***w*** document ***d*** is generated using theme ***j*** [2]

$$p(z_{d,w} = j) = \frac{\alpha_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)}{\sum_{j'=1}^{k} \alpha_{d,j'}^{(n)} p^{(n)}(w \mid \theta_{j'})}$$

$$\alpha_{d,j}^{(n+1)} = \frac{\sum_{d \in C} C_d(w) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} C_d(w') p(z_{d,w'} = j)}$$

Non-informative Priors:

Dirichlet : The conjugate prior distribution for the parameters of the multinomial distribution [6]:

$$p(\theta) = Diriclet(\theta \mid \beta_1, ..., \beta_k)$$

$$\beta_j > 0; \beta_0 \equiv \sum_{j=1}^{k} \beta_j$$

Data

Abstracts downloaded from abstract archive located at the Duke University Institute of Statistics and Decision Sciences (ISDS).

Website: www.bayesian.org

236 abstracts from 1995 to 2003

Experiment

Partition data into three time intervals

1995-1997	1998-2000	2001-2003
d = 93	d = 106	d = 37

Results: 1995 - 1997

Theme1		Theme2		Theme3		Theme4	
individual	0.0134	asymptotic	0.0149	Markov	0.0128	Clinical	0.0183
inherent	0.0132	Bernoulli	0.0146	Hierarchical	0.0128	Cancer	0.011
Carlo	0.0129	approximations	0.0128	Models	0.0128	Heterogeneity	0.0104
chain	0.0124	analytically	0.0094	resolutions	0.0128	heterogeneous	0.0103
Monte	0.012	orthogonal	0.0081	smoothly	0.0127	diseases	0.0103
estimated	0.0116	mean	0.0081	identification	0.0127	components	0.01
namely	0.0077	strategies	0.0079	build	0.0126	multivariate	0.01
passing	0.0076	averaging	0.0079	crucial	0.0126	time-varying	0.0099
investigation	0.0076	approximated	0.0079	Gibbs	0.0125	undergoing	0.0099
non-Bayesian	0.0076	evaluate	0.0079	flexible	0.0125	human	0.0099

Results: 1998-2000

Theme 1		Theme2		Theme 3		Theme 4	
EM	0.0065	Jeffreys	0.0087	focus	0.0118	prognosis	0.0116
Gaussian	0.0064	minimizing	0.007	learning	0.0084	relationship	0.0102
elucidation	0.0063	few	0.0069	described	0.0083	generated	0.0101
classification	0.005	Maximizing	0.0065	informative	0.0076	mass	0.01
BATS	0.005	desirable	0.0061	Though	0.0066	heterogeneity	0.0077
Metropolis-Hastings	0.0049	possible	0.0054	Southern	0.0063	group	0.0071
sampling	0.0048	predictive	0.0054	customers	0.0063	Dirichlet	0.0055
simulation	0.0041	approximations	0.0048	structuring	0.0062	Steroid	0.0052
non-Bayesian	0.0038	simply	0.0048	EXPLORATION	0.0061	population-based	0.0052
Unsupervised	0.0037	accept-reject	0.0047	scheme	0.006	Surveillance	0.0052

Results: 2001-2003

Theme 1		Theme 2		Theme 3		Theme 4	
appropriate	0.0076	goodness-of-fit	0.0076	empirical	0.0087	earlier	0.0077
common	0.006	investigated	0.0073	components	0.0085	Pareto	0.0074
observation	0.0058	involved	0.0064	Student-t	0.0071	quantities	0.0072
estimates	0.0058	Wishart	0.0062	economic	0.0066	sampled	0.0072
taken	0.0057	Significance	0.0058	computation	0.0064	Conjugate	0.006
Joint	0.0054	unknown	0.0055	Gibbs	0.0063	require	0.006
geographical	0.0053	Bayesian	0.0054	algorithms	0.0056	measured	0.0057
Markov	0.0048	Ilorin	0.0054	extensions	0.0056	Jeffreys'	0.0057
smoothing	0.0048	flexibly	0.0054	multivariate	0.0053	subjective	0.0057
Mapping	0.0045	fasting	0.0053	Developments	0.0049	powerful	0.0054

Theme Summary

	Theme 1	Theme 2	Theme 3	Theme 4
95-97	•Parameter estimation	•Approximations	•Sampling methods •Hierarchical models	•Clinical trials •Heterogeneous data
98-00	•Parameter estimation •Learning algorithms	•Approximations •Algorithms	• ??	•Clinical trials •Heterogeneous data •Dirichlet distribution
01-03	•Parameter estimation	•Distributions •Goodness-of-fit	•Sampling methods •Gibbs	•Prior distributions

References:

1. Airoldi E M, Anderson A G, Fienberg S E and Skinner K K, 2006. Who wrote Ronald Reagan's Radio Addresses? Bayesian Analysis 2006, Volume 1, Number 2, pp. 189-383.
2. Mei Q and Zhai C, 2005. Discovering Evolutionary Theme Patterns from Text – An Exploration of Temporal Text Mining. KDD'05, August 21-24, 2005. Chicago, Illinois, USA.
3. Rigouste, L, Cappe, O and Yvon, F, 2005. *Evaluation of a Probabilistic Method for Unsupervised Text Clustering.*
4. Stuart R. and Norvig P. 2003. Artificial Intelligence A Modern Approach. Pearson Education Inc., New Jersey, pp. 496-511.
5. Zhai C, Velivelli A and Bei Y, 2004. A Cross-Collection Mixture Model for Comparative Text Mining. KDD'04, August 22-25, 2004. Seattle, Washington, USA.
6. Gelman *et al. Bayesian Data Analysis.* Chapman & Hall, 1995.