Latent Contextual Indexing of Annotated Documents

Christian Sengstock
Institute of Computer Science
Heidelberg University, Germany
sengstock@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University, Germany
gertz@informatik.uni-heidelberg.de

ABSTRACT

In this paper we propose a simple and flexible framework to index context-annotated documents, e.g., documents with timestamps or georeferences, by *contextual topics*. A contextual topic is a distribution over document features with a particular meaning in the context domain, such as a repetitive event or a geographic phenomenon. Such a framework supports document clustering, labeling, and search, with respect to contextual knowledge contained in the document collection. To realize the framework, we introduce an approach to project documents into a *context-feature space*. Then, dimensionality reduction is used to extract contextual topics in this context-feature space. The topics can then be projected back onto the documents. We demonstrate the utility of our approach with a case study on georeferenced Wikipedia articles.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Database applications - Data Mining; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Document Context, Georeferenced Data, Topic Models, Latent Semantic Analysis, Exploratory Data Analysis

1. INTRODUCTION

The Web provides a huge source of context-annotated documents, like georeferenced and/or timestamped Tweets, Flickr photos, or Wikipedia pages. Recent work used probabilistic topic models to extract geographical topics from Flickr photos [1] and to explore the geographic distribution of Blog topics [2]. By that, context-annotated document collections have been used to discover meaningful knowledge in the geographic context domain. The motivation of these approaches has mainly been driven by data exploration tasks. However, we believe that contextual topics have a much wider range of applications in contextual search, document labeling, clustering, and query result (re)ranking.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

In this paper, we outline a simple framework to mine contextual topics by projecting the context-annotated documents into a context-feature space, which is then subject to feature extraction. The extracted topics can then be projected back onto the documents, which we call *latent contextual indexing*.

Idea: Documents with context annotations can be embedded in the context domain. Then, document features (e.g., terms) tend to accumulate in subsets of the context-domain if there is a particular contextual meaning. For example, georeferenced documents containing the term *mountain* tend to accumulate in subsets of geographic space in which mountains occur.

We show that this simple scheme allows to discover meaningful contextual topics from highly noisy data sources as typical on the Web.

2. LATENT CONTEXTUAL INDEXING

We assume a context-annotated document collection, such as a set of georeferenced text documents. From now on, we assume the context-annotations to be points (locations) in the geographic domain. We are given the document-term matrix $A = \mathbb{N}^{m \times p}$ and the document-context matrix $B = \{0,1\}^{m \times l}$, with $B_{ij} = 1$ if document i occurs at location j.

Context-Feature Space: The locations are embedded in a geographic grid, such that each document falls into the cell(s) in which it occurs. Hence, the m columns of B can be seen as the set of grid cells. The context-feature count matrix C then is defined as

$$C := B^T A = \mathbb{N}^{l \times p} \tag{1}$$

A cell (row) of C contains all the features of the documents occurring in this cell. We say a cell (location) is represented in *context-feature space*.

Feature Extraction: The *contextual topics* are extracted using dimensionality reduction in the context-feature space, for instance using PCA, ICA, or NMF. We generally denote a feature extraction task into k topics as:

$$\tilde{C} = WH, W = \mathbb{R}^{l \times k}, H = \mathbb{R}^{k \times p} \tag{2}$$

Then, a contextual topic consists of two distributions: The context-topic distribution W, representing the topic weights over cells, and the topic-feature distribution H, representing the feature weights over topics.

Indexing: Note that the feature extraction happens in context-feature space. To project the topics back onto the documents, two projections are possible:

$$\tilde{A}^F = AH^T \tag{3}$$

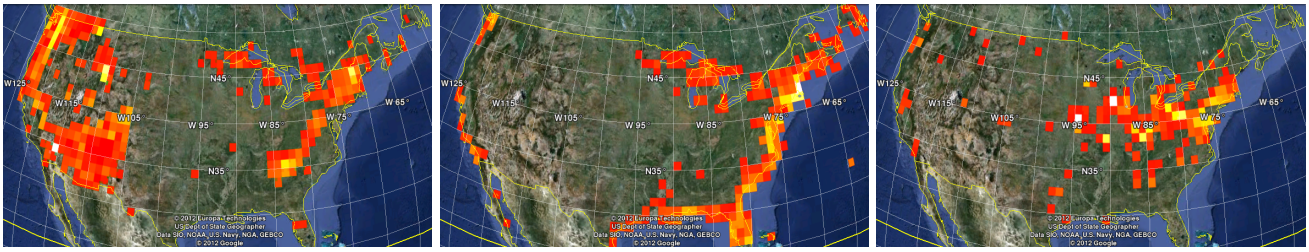

Figure 1: Context-topic distributions (W_1, W_2, W_3), correspond to topics in Figure 2 (please view in color).

$$\tilde{A}^G = BW \quad (4)$$

(3) is the feature-based projection: A document will have a high topic weight if its features correspond to the topic. (4) is the context-based projection: A document will have a high topic weight if the topic has high weight at the document context annotations.

3. EXPERIMENTS

We conducted experiments on 111,166 Wikipedia abstracts occurring within the US ($[-124,-54]\times[26,50]$), downloaded from *DBpedia*[1]. Only terms occurring at least 20 times in the collection have been kept and no stopwords have been removed, resulting in 17,266 document features (terms). The context domain is represented by a regular grid with step-width 1.0 degree over the US bounds, resulting in a context-feature count matrix C with 1728 rows (72×24 grid). We normalized C to use only the binary occurrence information of a feature.

The contextual features have been extracted using Independent Component Analysis (ICA), reducing the 17,266 dimensional space to 10 dimensions. Figure 1 shows the geographic distribution of three selected topics, and Figure 2 shows those topics by their 10 highest weighted terms. Topic 1 clearly represents the mountains in the US, and Topic 2 the coastal regions. Topic 3 can be seen as a topic related to historic places, with high intensity in the east. We note that other meaningful topics are the major cities, the Canadian border, and California. Also a single non-geographic topic exists whose geographic distribution just follows the distribution of the document locations (not presented here).

Those extracted topics can be projected back onto the documents using equations (3) and (4). Figure 3 shows the top-8 Wikipedia articles regarding the feature-based indexing weights (left three value-columns). The context-based indexing weights are also shown (right three value-columns). The top-ranked Wikipedia articles for the three topics show meaningful results. Interestingly, not all articles have high feature-based and context-based topic weights (bold names). Articles like *Schofield_Pass_Nevada* have a high weight for the mountain topic regarding its terms. However, the article is not located in a cell with a high weight for that topic. Such differences give interesting opportunities to explore, search, and cluster the articles according to term- or context-based preferences. In general, the results show that the context feature space of annotated document collections (like Wikipedia abstracts) contains meaningful knowledge about the context domain.

[1]http://dbpedia.org

Topic 1: $\max_{10}(H_{1.})$	Topic 2: $\max_{10}(H_{2.})$	Topic 3: $\max_{10}(H_{3.})$
mountains mountain summit peak wilderness hiking range forest flows highest	bay coast islands peninsula beach island port coastal boat ocean	steel cementery trains 1900 joseph mills society pennsylvania tracks cost

Figure 2: Selected topic-feature distributions from Wikipedia using ICA.

Wiki Article	$\tilde{A}^F_{i1}$	$\tilde{A}^F_{i2}$	$\tilde{A}^F_{i3}$	$\tilde{A}^G_{i1}$	$\tilde{A}^G_{i2}$	$\tilde{A}^G_{i3}$
Humback_Mountain_Cs.	**0.16**	0.03	0.07	**5.78**	-0.16	-0.85
Schofield_Pass_Nevada	**0.15**	0.03	0.05	-0.06	-0.04	-0.26
Schoflied_Pass_Wyoming	**0.15**	0.02	0.06	0.15	-0.42	0.21
Red_Mountain_Cascades	**0.14**	0.04	0.07	**5.78**	-0.16	-0.85
Conjeos_Peak	**0.13**	0.03	0.05	0.58	-0.71	-0.08
Red_Mountain_Rossland	**0.12**	0.05	0.06	**1.42**	-0.30	0.00
Willow_Creek_Pass_Col.	**0.12**	0.04	0.05	**2.01**	-0.32	0.06
Stampeda_Pass	**0.12**	0.05	0.05	**5.78**	-0.16	-0.85
Bay_Island,_Bermuda	0.03	**0.12**	0.04	0.03	**3.17**	-0.97
North_Dumpling_Light	0.05	**0.11**	0.06	1.47	**6.31**	4.85
Long_Beack_Light	0.06	**0.11**	0.05	-1.71	**3.72**	0.24
Mapeque_Bay,_Prince_Edw.	0.03	**0.11**	0.06	-1.47	0.89	-0.53
Cornelius_Island	0.03	**0.11**	0.04	-1.08	**7.83**	4.15
Monomoy_National_W._R.	0.05	**0.11**	0.04	-0.46	**8.34**	0.29
Nosuch_Bay,_Bermuda	0.03	**0.11**	0.03	0.03	**3.17**	-0.97
Bedwell_Bay,_British_Colum	0.05	**0.11**	0.07	1.98	-1.52	-0.35
Mount_Vernon_Cementry	0.04	0.04	**0.09**	-0.53	-2.02	**8.65**
Boulevard_Heights,_St._L.	0.00	0.05	**0.09**	-2.17	-2.24	**8.32**
Acheson_Tunnel	0.06	0.05	**0.09**	-0.09	-0.56	**7.67**
Washington_Trust_Build.	0.04	0.07	**0.09**	-0.09	-0.56	**7.67**
St.Thomas_Syro-Malabar_Church	0.00	0.07	**0.09**	-0.12	1.40	**10.17**
Theatre_Passe_Muraille	0.03	0.05	**0.09**	-0.01	-8.97	-3.00
Crystal_Mall_British_Colum.	0.03	0.06	**0.09**	1.77	-0.61	-0.49
Reynolda_Gardens	0.02	0.07	**0.09**	1.12	-0.14	1.07

Figure 3: Top-8 weighted Wikipedia pages by indexed contextual topic. Feature-based indexing (left three values) and context-based indexing (right three values)

4. CONCLUSIONS

We proposed a simple approach to index documents by contextual topics and conducted experiments using a collection of georeferenced Wikipedia articles. The projection of documents into a context-feature space provides a rich framework. It allows for normalization, smoothing, and resolution settings. Here, we showed initial results using a simple geographic grid and ICA, which gives promising results. Currently we evaluate parameterizations and feature extraction approaches on various data sets, and work on applications of latent contextual indexing in the domain of search and recommendation.

5. REFERENCES

[1] Y. Zhijun, L. Cao, J. Han, et. al.: Geographical Topic Discovery and Comparison. In: Proceedings of WWW 2011, p. 247-256

[2] Q. Mei, C. Liu, H. Su, C. Zhai: A Probablistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In: Proceedings of WWW 2006, p. 533-542