Short-Text Representation using Diffusion Wavelets

Vidit Jain
Yahoo Labs Bangalore
viditj@yahoo-inc.com

Jay Mahadeokar
Yahoo Labs Bangalore
jaym@yahoo-inc.com

ABSTRACT

Usual text document representations such as `tf-idf` do not work well in classification tasks for short-text documents and across diverse data domains. Optimizing different representations for different data domains is infeasible in a practical setting on the Internet. Mining such representations from the data in an unsupervised manner is desirable. In this paper, we study a representation based on the multi-scale harmonic analysis of term-term co-occurrence graph. This representation is not only sparse, but also leads to the discovery of semantically coherent topics in data. In our experiments on user-generated short documents e.g., newsgroup messages, user comments, and meta-data, we found this representation to outperform other representations across different choice of classifiers. Similar improvements were also observed for data sets in Chinese and Portuguese languages.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms**]

Keywords

Representation discovery, text classification

1. INTRODUCTION

Text classification is important for efficient indexing for improving several tasks including retrieval, ranking, and recommendation. `tf-idf` representation works well for long documents and webpages, but performs poorly on short text documents and user generated content. Adapting these custom representations to a new data domain often requires significant effort for parameter selection or building new representations from scratch. The problem of performing text classification in diverse data domains is increasingly more important with the explosion of user generated content (e.g., tags, comments, tweets, and blogs) and non-English content on the Internet. It has become infeasible to customize representations for all of these diverse, large number of data domains; mining effective representations from the data in an unsupervised manner is therefore desirable. To this end, here we propose the use of multi-scale harmonic analysis of the data corpus to embed the terms in the corresponding vocabulary in a continuous Euclidean space. We show that the resulting representation is sparse and correlates well with the discovery of semantically coherent topics in data. Furthermore, this representation has little computational overhead over the standard `binary` and `tf-idf` based representations.

The main contribution of this paper is the study of harmonic analysis of co-occurrence matrix to compute effective representations for short, noisy text documents. Our experiments on five real-world data sets suggest that this data-driven representation learned in an unsupervised manner outperforms the standard alternative representations.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577345.

2. OUR APPROACH

Consider a graph $\mathcal{G}$ over the n terms in the vocabulary. The adjacency matrix W is given by the $n \times n$ term-term matrix computed using a similarity function (e.g., pair-wise co-occurrence statistics) defined for a pair of terms. Let D be the diagonal matrix whose entries are the row sums of W. Using the normalized term-term matrix as the operator $T = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, we compute the diffusion scaling functions Φ_l and wavelet functions ψ_l at different levels l using the diffusion wavelet tree construction algorithm [2]. At each level l, we also compute an *extended basis* Ω_l that facilitates the projection from the original subspace to the subspace spanned by the corresponding scaling functions Φ_l. Ω_l is computed in an inductive manner with

$$\Omega_0 = \Phi_0, \qquad \Omega_l = \Omega_{l-1} * \Phi_l. \tag{1}$$

Note that Ω_l is $n \times n_l$ matrix, where its i^{th} column gives the i^{th} extended basis function. The j^{th} row gives the representation of the j^{th} term in the vocabulary, which we refer to as $\mathbf{e_{l,j}}$. Each of these $\mathbf{e_{l,j}}$ vectors reside in $\mathbb{R}^{n_l}$, where n_l is the number of basis functions at level l. Also, the i^{th} value in each of these vectors indicates the weighted association between the term and the i^{th} topic.

Each of the train (or test) documents is represented as a weighted combination of the representations of the terms appearing in them:

$$\mathbf{v_d} = \frac{1}{|d|} \sum_{e \in d} term\_rep(e) * w(e), \tag{2}$$

where $w(e)$ refers the weight of the term e in the given document. In our experiments, we consider two standard choices for the computation of this weight: `binary`[1] and `tf-idf`. These two representations are referred to as **DW-b** and **DW-t**, respectively.

[1] 0 or 1 based on the presence of the term in the document

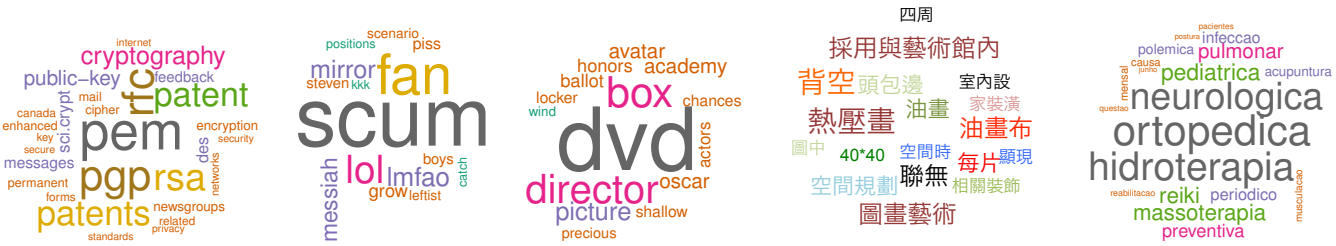

Figure 1: ***Example DW-bases for the different data sets interpreted as semantic topics.*** **Statistical topic model (e.g., LDA) fail to extract such coherent structure from short, noisy text documents.**

Data set	#w/d	#d	#classes	source
Newsgroup	147	20K	20	20-Newsgroup
Comments	27	12K	5	Yahoo News
Meta-data	32	42K	171	Yahoo Videos
Products	145	40K	75	Yahoo Shopping
CADE-12	9	30K	12	CADE

Table 1: ***Data sets used in our experiments.***

3. OBSERVATIONS

We consider three collections of short, noisy documents – i.e., newsgroup messages, comments [3], and meta-data. The messages are categorized into the newsgroup they belong to; the comments are categorized based on their content and presentation quality; and the meta-data (title, description, and keywords) is categorized into a pre-defined taxonomy. To further evaluate the proposed unsupervised discovery of effective representations, we consider two non-English data sets: product descriptions[2] in Chinese language and short web documents[3] in Portuguese language. In each data set, we restrict our analysis to the $10K$ most frequent terms.

Table 2 shows that the DW based representations outperform other representations. For all data sets and representations, SVM outperformed naïve-Bayes classifier. These data sets have only about 25K training examples for 10K dimensional data. In the absence of densely sampled training data, discriminative models generalize better than the generative models. We also observed that the binary representation (and hence DW-b) performs better than `tf-idf` for short documents or documents that do not follow natural language statistics (e.g., documents with high frequency of repetitive words, misspellings, and colloquial usage).

Latent Dirichlet allocation (LDA) [1] with 100 topics has similar model complexity (i.e., non-zero values) as DW. With 400 topics, the LDA representation improved performance for SVM, but was still worse than DW. Also, the DW representation is compressible as the F1-score remains almost the same even when only 0.1% most significant values in representation are used. Similar observations were made on other data sets in this work and on the TDT2 data set [4].

Figure 1 shows some example basis that can be interpreted as semantic topics. We asked domain experts to assess the semantic coherence of the obtained basis functions for non-English data sets. They confirmed that most of the basis (78% and 58%, respectively) correlated well with semantically coherent topics while some of them corresponded to general text structure in the respective languages.

(a) 20 Newsgroups

	Naïve Bayes	SVM
Binary	0.668	0.736
`tf-idf`	0.490	0.785
LDA-100	0.608	0.672
LDA-400	0.536	0.697
DW-b	**0.732**	0.828
DW-t	0.646	**0.850**

(b) Naïve Bayes on other data sets

	Comments	Meta-data	Products	CADE
Binary	**0.497**	0.378	0.652	0.361
`tf-idf`	0.106	0.042	0.506	0.081
DW-b	0.367	**0.386**	0.687	**0.373**
DW-t	0.457	0.342	**0.722**	0.369

(c) SVM on other data sets

	Comments	Meta-data	Products	CADE
Binary	0.561	0.603	0.848	0.461
`tf-idf`	0.520	0.508	0.802	0.483
DW-b	**0.580**	**0.609**	**0.868**	0.503
DW-t	0.569	0.546	0.833	**0.524**

Table 2: ***Categorization results.*** **These weighted-F1 scores are computed for a 2:1 train-test split.**

4. CONCLUSION

We have shown that diffusion wavelets based representations outperform other standard representations for categorizing short and noisy, user-generated text documents. The proposed representation is data driven and learned in an unsupervised manner as has been illustrated through experiments on documents from non-English languages. These experiments suggest that the proposed representations are feasible to learn for a diverse, large number of data domains, and which are likely to outperform the standard representations for categorization tasks in these domains.

5. REFERENCES

[1] Blei et al. Latent Dirichlet allocation. JMLR 2003.
[2] Coifman & Maggioni. Diffusion wavelets. ACHA, 2006.
[3] Jain & Galbrun. Topical organization of user comments and application to content recommendation. WWW, 2013.
[4] Wang & Mahadevan. Multiscale Analysis of Document Corpora Based on Diffusion Models. IJCAI 2009.

[2]from Yahoo Taiwan Shopping inventory
[3]`http://web.ist.utl.pt/~acardoso/datasets/`