Graph-Based Term Weighting for Text Categorization

Fragkiskos D. Malliaros[1] Konstantinos Skianis[1,2]

[1]École Polytechnique, France

[2]ENS Cachan, France

SoMeRis workshop, ASONAM 2015

Paris, August 25, 2015

Outline

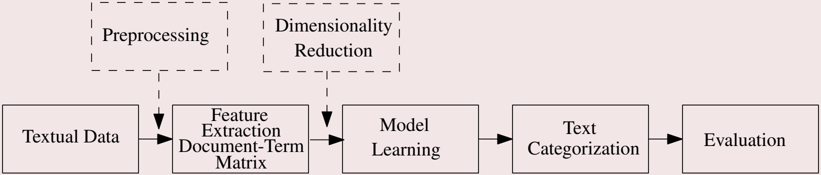
Outline

Introduction

- Online social media and networking platforms produce a vast amount of textual data
- Analyze and extract useful information from textual data is a crucial task
- **Text categorization (TC)** refers to the supervised learning task of assigning a document to a set of two or more pre-defined categories, based on learning models that have been trained using labeled data
- Plethora of applications
  - Opinion mining for risk assessment and management
  - Email filtering
  - Spam detection
  - News classification
  - ...

Text categorization: the pipeline

Basic pipeline of the text categorization task



Term weighting in the Bag-of-words model

Vector Space Model

- $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ denotes a collection of m documents
- $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ be the dictionary

Feature extraction

Every document is represented by a feature vector that contains boolean or weighted representation of unigrams or n-grams

- TF (Term Frequency), TF-IDF (Term Frequency - Inverse Document Frequency)

$$tf\text{-}idf(t, d) = tf(t, d) \times idf(t, \mathcal{D}),$$

$$\text{where } idf(t, \mathcal{D}) = \log \frac{m + 1}{|\{d \in \mathcal{D} : t \in d|\}}$$

Contributions of this work

- **Graph-based term weighting schemes for TC**
  - Propose a simple graph-based representation of documents for text categorization
  - Derive novel term weighting schemes, that go beyond single term frequency
- **Exploration of model's parameter space and experimental evaluation**
  - We discuss how to construct the graph
  - We examine the performance of the different proposed weighting criteria using standard document collections

Outline

Graph-of-words: overview

Why Graph-of-words?

- Capture relationships between terms
- Questioning the term independence assumption
- Already applied in other data analytics tasks (e.g., IR [Blanco and Lioma, '12], [Rousseau and Vazirgiannis, '13])

Representation of a document

Each document $d \in \mathcal{D}$ is represented by a graph $G_d = (V, E)$

- Nodes correspond to the **terms** t of the document
- Edges capture **co-occurence relations** between terms within a fixed-size sliding window of size w

Proposed graph-based term weighting method for TC

Input: Collection of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ and set (dictionary) of terms $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$

Output: Term weights $tw(t, d)$ for each term $t \in \mathcal{T}$ to each document $d \in \mathcal{D}$

1: **for** $d \in \mathcal{D}$ **do**

2: **(Graph Construction)** Construct a graph $G_d = (V, E)$. Each node $v \in V$ corresponds to a term $t \in \mathcal{T}$ of document d. Add edge $e = (u, v)$ between terms u and v if they co-occur within the same window of size w

3: **(Term Weighting)** Consider a node centrality criterion. For each term $t \in \mathcal{T}$, compute the weight $tw(t, d)$ based on the centrality score of node t in graph G_d and fill in the Document-Term matrix

4: **end for**

Graph construction: parameters of the model

- **Directed vs. undirected graph**
  - Directed graphs are able to preserve actual flow of a text
  - In undirected ones, an edge captures co-occurrence of two terms whatever the respective order between them is ✓
- **Weighted vs. unweighted graph**
  - Weighted: the higher the number of co-occurences of two terms in the document, the higher the weight of the corresponding edge
  - Unweighted (our choice due to the simplicity of the model) ✓
- **Size *w* of the sliding window**
  - We add edges between the terms of the document that co-occur within a sliding window of size w
  - $w = 3$ performed well in TC ✓
  - Larger window sizes produce graphs that are relatively dense

Graph construction: parameters of the model

- **Directed vs. undirected graph**
  - Directed graphs are able to preserve actual flow of a text
  - In undirected ones, an edge captures co-occurrence of two terms whatever the respective order between them is ✓
- **Weighted vs. unweighted graph**
  - Weighted: the higher the number of co-occurences of two terms in the document, the higher the weight of the corresponding edge
  - Unweighted (our choice due to the simplicity of the model) ✓
- **Size w of the sliding window**
  - We add edges between the terms of the document that co-occur within a sliding window of size w
  - $w = 3$ performed well in TC ✓
  - Larger window sizes produce graphs that are relatively dense

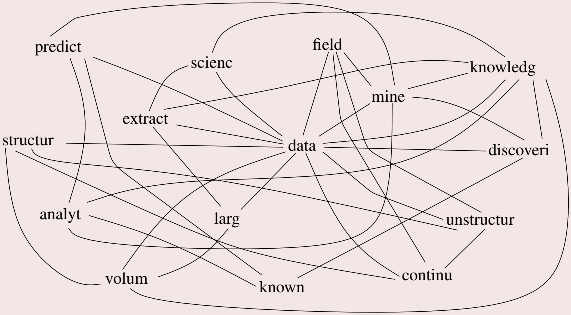
Graph construction: parameters of the model

- **Directed vs. undirected graph**
  - Directed graphs are able to preserve actual flow of a text
  - In undirected ones, an edge captures co-occurrence of two terms whatever the respective order between them is ✓
- **Weighted vs. unweighted graph**
  - Weighted: the higher the number of co-occurences of two terms in the document, the higher the weight of the corresponding edge
  - Unweighted (our choice due to the simplicity of the model) ✓
- **Size w of the sliding window**
  - We add edges between the terms of the document that co-occur within a sliding window of size w
  - $w = 3$ performed well in TC ✓
  - Larger window sizes produce graphs that are relatively dense

Example: text to graph representation

Graph representation of a document ($w = 3$; undirected graph)

Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured which is a continuation of the field of data mining and predictive analytics, also known as knowledge discovery and data mining.



Term weighting criteria

- Utilize **node centrality criteria** of the graph
  - The importance of a term in a document can be inferred by the importance of the corresponding node in the graph
- Consider information of the graph:
  - **Local:** degree centrality, in-degree/out-degree centrality in directed networks, weighted degree in weighted graphs, clustering coefficient
  - **Global:** PageRank centrality, eigenvector centrality, betweenness centrality, closeness centrality

$$\text{degree\_centrality}(i) = \frac{|\mathcal{N}(i)|}{|V| - 1}, \quad \text{closeness}(i) = \frac{|V| - 1}{\sum_{j \in V} dist(i, j)}$$

- Proposed weighting schemes for TC:
  - TW
  - TW-IDF

Experimental set-up

- **Datasets**
  1. *Reuters-21578 R8*: documents of Reuters newswire in 1987
     - # of **train** docs: **5, 485**; # of **test** docs: **2, 189**; **total**: **7, 674**
     - # of categories: **8**
  2. *WebKB*: academic webpages
     - # of **train** docs: **2, 803**; # of **test** docs: **1, 396**; **total**: **4, 199**
     - # of categories: **4**

- **Evaluation**
  - Linear SVM classifier
  - Train the model on the **train** documents
  - Report classification results from the **test** documents
  - Macro-averaged F1 score and classification accuracy

- **Baseline methods**
  - Traditional TF and TF-IDF weighting schemes vs. the proposed TW and TW-IDF (degree, in-degree, out-degree and closeness centrality; window-size=3)

Experimental results

Reuters-21578 R8 and WebKB datasets

Weighting	F1-score	Accuracy
TF	**0.9127**	**0.9634**
TW, degree	**0.8991**	**0.9611**
TW, in-degree	**0.8037**	**0.9438**
TW, out-degree	**0.8585**	**0.9546**
TW, closeness	**0.9125**	**0.9625**
TF-IDF	**0.8962**	**0.9616**
TW-IDF, degree	**0.9175**	**0.9661**
TW-IDF, in-degree	**0.8985**	**0.9629**
TW-IDF, out-degree	**0.8854**	**0.9625**
TW-IDF, closeness	**0.8846**	**0.9547**

Reuters-21578 R8

Weighting	F1-score	Accuracy
TF	**0.8741**	**0.8853**
TW, degree	**0.8962**	**0.9032**
TW, in-degree	**0.8286**	**0.8545**
TW, out-degree	**0.8365**	**0.8603**
TW, closeness	**0.8960**	**0.9004**
TF-IDF	**0.8331**	**0.8538**
TW-IDF, degree	**0.8800**	**0.8882**
TW-IDF, in-degree	**0.7890**	**0.8381**
TW-IDF, out-degree	**0.8049**	**0.8474**
TW-IDF, closeness	**0.8505**	**0.8674**

WebKB

Outline

Conclusions and future work

Contributions:

- Introduce a new paradigm for TC
- Potential of graph-based weighting mechanisms in TC

Future work:

- Exploration of parameter's space: many diverse centrality criteria can be applied in order to weight the terms
- Graph-based inverse collection weight: a more thorough theoretical analysis of its properties is also an interesting future direction
- Graph-based dimensionality reduction: extend the task of dimensionality reduction to the graph representation of the documents

References I

R. Blanco and C. Lioma
Graph-based term weighting for information retrieval.
Inf. Retr., 15(1), 2012.

C. M. Bishop
Pattern Recognition and Machine Learning (Information Science and Statistics).
Springer-Verlag New York, Inc., 2006.

D. Easley and J. Kleinberg
Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, 2010.

S. Hassan, R. Mihalcea, and C. Banea
Random walk term weighting for improved text classification.
Int. J. Semantic Computing, 1(4), 2007.

T. Joachims
Text categorization with suport vector machines: Learning with many relevant features.
In *ECML*, 1998.

M. Lan, C.-L. Tan, H.-B. Low, and S.-Y. Sung
A comprehensive comparative study on term weighting schemes for text categorization with support vector machines.
In *WWW*, 2005.

C. D. Manning, P. Raghavan, and H. Schuütze
Introduction to Information Retrieval.
Cambridge University Press, 2008.

R. Mihalcea and P. Tarau
Textrank: Bringing order into text.
In *EMNLP*, 2004.

References II

G. Paltoglou and M. Thelwall
A Study of Information Retrieval Weighting Schemes for Sentiment Analysis.
In *ACL*, 2010.

F. Rousseau and M. Vazirgiannis
Graph-of-word and TW-IDF: new approach to ad hoc IR.
In *CIKM*, 2013.

F. Rousseau, E. Kiagias, and M. Vazirgiannis
Text categorization as a graph classification problem.
In *ACL*, 2015.

G. Salton and C. Buckley
Term-weighting approaches in automatic text retrieval.
Inf. Process. Manage., 24(5), 1988.

A. Schenker, M. Last, H. Bunke, and A. Kandel
Classification of web documents using a graph model.
In *ICDAR*, 2003.

F. Sebastiani
Machine learning in automated text categorization.
ACM Comput. Surv., 34(1), 2002.

Thank You !!

Fragkiskos D. Malliaros
Data Science and Mining Group (DaSciM)
École Polytechnique, France
fmalliaros@lix.polytechnique.fr
www.lix.polytechnique.fr/∼fmalliaros

Konstantinos Skianis
Data Science and Mining Group (DaSciM)
École Polytechnique, France
kskianis@lix.polytechnique.fr
www.lix.polytechnique.fr/∼kskianis