Document Classification Based on Word Vectors

Liu Rong
CSLT RIIT TSINGHUA
2014-09-13
Outline

• Document Classification
  • Introduction
  • Approach
• Document Vector (Text Representation)
  • LDA
  • Word2Vec
• Experiment
• Conclusion
• References
Document Classification - Introduction

• Introduction
  • Task
to classify documents into predefined classes
• Relevant Technologies
  Text Clustering, Information retrieval,
  Information filtering, Information Extraction.
• Application
  QA, Categorize newspaper articles and newswires into topics.
  Organize Web pages into hierarchical categories.
  Sort journals and abstracts by subject categories
Document Classification - Introduction

- **Approaches**
  - **Rule-based**
    - Rule 1: "ball" ∈ d → t(d) = sports
    - Rule 2: "ball" ∈ d & "dance" ∉ d & game ∈ d & "play" ∈ d → t(d) = sports
  - **Machine learning-based**
    - **Text preprocessing**
      - removing stop word and predefined words
    - **Feature Extraction**
      - TF-IDF(Bag-of-word), LDA, LSI, **word2vec**
    - **Classifier Construction**
      - Native Bayes, KNN, SVM
    - **Classifier Evaluation**
• Introduction

• Each topic is a distribution over words
• Each document is a mixture of corpus-wide topics
• Each word is drawn from one of those topics
The goal of training is to get the $\alpha$ and $\beta$ when the corpus function get the maximum value.
Document Classification-LDA

• Document Vector
  • The topic distribution in a document
    
    \[
    \text{document vector} = \theta = [T_1, T_2 \cdots T_K]
    \]

  where \(T_k\) is the probability of \(k_{th}\) topic in a document

• Problem
  • Learning structure is uncertain
  • LDA is sensitive with initial value
  • High computational complexity
  • Loss semantic information that LDA don’t consider the word sequence
• One-hot Representation
  
  dog => [0 0 0 0 1 0 0 0 0 0 0]
  cat => [1 0 0 0 0 0 0 0 0 0 0]

• Distributed Representation
  
  dog => [0.792 -0.177 0.98 -0.9 .....]
  cat => [0.76 0.12 -0.54 0.9 0.65 ....]

• Method

  NNLM:
  C&W:
  M&H: Log-Bilinear /Hierarchical Log-Bilinear Model
  RNNLM:
  Huang: add document information
  Glove:
Document Classification-w2v

- Google Word2vec
  - Skip-gram

\[
\frac{1}{T} \sum_{i=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)
\]

where \( w_1, w_2 \cdots w_T \) is sequence words, \( c \) is size of context.

\[
p(w_O | w_I) = \frac{\exp \left( v_{w_O}^T v_{w_I} \right)}{\sum_{w=1}^{W} \exp \left( v_{w}^T v_{w_I} \right)}
\]

where \( v_w \) and \( v'_w \) are the input and output vector representation.

\( w \) is the number of words in the vocabulary.
Document Classification-w2v

- Document Vector

\[ \text{document vector} = \frac{1}{|d|} \sum_{w \in d} c_w \]

where \(|d|\) is the number of words in the document. \(c_w\) is the word vector of \(w\).

互联网的发展带来电子文本数据的快速增长。为提高信息检索和管理的效率，文本自动分类技术成为研究的热点。文本自动分类的基本方法是从一个训练文本集合学习一定分类规则或分类模型，使得依据该规则或模型对未知新文本进行分类时具有较好的分类精度[1]。一个典型的文本分类系统主要包括文本预处理、特征提取、分类模型训练、文本分类等四个模块，其中特征提取模块的任务是将以不等长散字符串形式存在的文本表示成可用以建立分类模型的文本特征向量，其性能的好坏对文本分类系统的效果以及效率有直接影响。
Data

SogouLab:
1. car, economics, IT, health, sports, travel, education, Recruitment, culture and military
2. train: 14301(65M), test: 1809

w2v: train word vector on People's Daily(5G)
w2v-ex: train word vector on train data of SogouLab(65M)

Tool

segment word: http://www.xunsearch.com/scws/index.php
word2vec: https://code.google.com/p/word2vec
classifier/weka: http://www.cs.waikato.ac.nz/ml/weka
LDA: http://www.cs.princeton.edu/blei/lda-c
Different dimensions of LDA and w2v

- The w2v get higher accuracy than LDA
- The w2v is more stable than LDA
- The w2v need more data with higher dimension

**w2v**: train word vector on People's Daily (5G)

**w2v-ex**: train word vector on train data of SougouLab (65M)

**lda**: train lda on train data of SougouLab
Different classification task

- w2v is equal with LDA from 2-classes to 4-classes
- w2v get higher accuracy from 5-classes to 9-classes
- w2v is more general

Document Classification-Experiment

- w2v: train word vector on People's Daily (5G)
- w2v-ex: train word vector on train data of SougouLab
- lda: train lda on train data of SougouLab

%Accuracy

<table>
<thead>
<tr>
<th>Class Number</th>
<th>w2v</th>
<th>w2v-ex</th>
<th>lda</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Document Classification- Experiment

- Different classes Accuracy
  
  - w2v is more stable

*Graph showing accuracy for different classes with labels and explanations for w2v, w2v-ex, and lda.*

- w2v: train word vector on People's Daily(5G)
- w2v-ex: train word vector on train data of SougouLab
- lda: train lda on train data of SougouLab
Document Classification-Experiment

- Efficiency

![Training Time Diagram](image1)

![Inference Time Diagram](image2)
Conclusion

Introduce the word vector to document classification and analysis the different of semantic generation between word vector and LDA.

Experiment show that document classification based on word vector superior to LDA in classifier accuracy, computational complexity, scalability field, processing capacity in complex classification task and representation of content.
Document Classification-Reference

Document Classification-QA

Question and answer
Document Classification-VSM

- Introduction
- Document Vector

\[ \text{document vector} = [T_1, T_2 \ldots T_K] \]

where \( T_j = n \times \log\left(\frac{M}{m}\right) \), \( n \) is TF, \( M/m \) is IDF.
Document Classification-Introduction

• Approaches
  • Rule-based
    Rule 1: "ball" ∈ d → t(d) = sports
    Rule 2: "ball" ∈ d & "dance" ∉ d & game ∈ d & "play" ∈ d → t(d) = sports
  • Machine learning-based
    • Text preprocessing
      removing stop word and predefined words
    • Feature Extraction
      TF-IDF(Bag-of-word), LDA, LSI, \texttt{word2vec}
    • Classifier Construction
      Native Bayes, KNN, SVM
    • Classifier Evaluation

diagram:

- preprocess
- Feature extraction
  - word2vec
  - LDA
  - ...
- Classifier
- Evaluation