Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Matthias Nitsche

Continuous Clustering for a Daily News Summarization System

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Matthias Nitsche

Continuous Clustering for a Daily News Summarization System

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Michael Neitzke
Zweitgutachter: Prof. Dr. Olaf Zukunft

Eingereicht am: 02. February 2016

Matthias Nitsche

Thema der Arbeit
Continuous Clustering for a Daily News Summarization System

Stichworte
Clustering, Cluster Analyse, Dokument Clustering, Vector Space Model Partitionelles Clustering, Hierarchisches Clustering, Probabilistic Topic Modelling, LSA, LSI, Latent Semantic Analysis, LDA, Latent Dirichlet Allocation, Text Zusammenfassung, Text Mining, Data Mining, Machinelles Lernen, Unüberwachtes Lernen, Information Retrieval, IR

Kurzzusammenfassung
Kontinuierliches Clustering für ein Newspaper Summarization System.
Matthias Nitsche

Title of the paper
Continuous Clustering for a Daily News Summarization System

Keywords
Clustering, Cluster Analysis, Document Clustering, Vector Space Model Partitional Clustering, Hierarchical Clustering, Probabilistic Topic Modelling, LSA, LSI, Latent Semantic Analysis, LDA, Latent Dirichlet Allocation, Summarization, Text Mining, Data Mining, Machine Learning, Unsupervised Learning, Information Retrieval, IR

Abstract
Continuous document clustering of newspaper articles.

Contents

List of Tables

List of Figures

1 Introduction

Why and relevance? What can be expected? What did I do? What did my work achieve? The Objective

1.1 Machine Learning

1.2 Structure of the thesis

What happens in chapter 2..n and how are they related?

2 Basics

"If I have seen further it is by standing on the shoulders of giants."

Isaac Newton

The goal of this section is to give some intuition and the necessary theoretical background for the following chapters. The areas where clustering problems arise are huge. It provides solutions to problems like market segmentation, classification, document organization or indexing.
Firstly we will have a look at the definition of clustering and summarization. How they are related and the variety of possibilities this imposes. Secondly the vector space model (VSM) is introduced. It contains all information about how to represent documents in a vectorized form. Of special interest are enhanced models which reduce the dimensionality of documents by singular value decomposition (SVD). Thirdly traditional clustering algorithms from the hierarchical (Ward, Birch) and partitional (K-Means, Mean-Shift) will be presented. Closely related are the generative models. These methods can be used as a kind of clustering algorithm and are highly useful in several steps of traditional clustering. They can be used as dimensionality reduction techniques as well. Lastly some quality measures of clusters based on internal measures (without ground truth labels) and external measures (with explicit labelling of the ground truth) are explained.

2.1 Clustering and Summarization

Clustering as defined by Aggarwal and Zhai (2012) is finding groups of similar objects in the data with a defined similarity function between objects. The granularity of the features can vary:

- *Sentence based* - A document d is split into sentences so clustering reveals the most coherent groups of sentences that are closely related.

- *Collection of documents* - A collection of documents d (corpora) is grouped to get groups of documents that are closely related
- *Stream of documents* - The same as clustering copora with the constraint that over time the size of documents grow.

Document clustering on large corpora can be seen as a summarization of the underlying concepts. The representation of documents as feature vectors is described with the vector space model in the next section.
Automatic text summarization on the other hand, is the process of reducing textual content to the most important concepts in a readable, formatted form to the user Mani (2001).
This results in a few possibilities where clustering works great as a preprocessing step for summarization.

First Clustering groups that have a *higher density of information* resulting in a grouped input for summarizers.

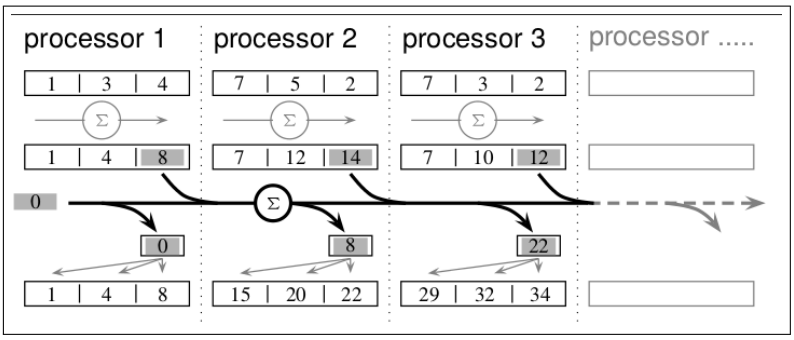
Second Grouping the *latent topics* accross and within documents to create a meta concept of closely related documents

Third Classify documents into *categories* in a semi supervised way to construct hierarchies of relationships

Fourth Finding *outliers* that will not highly contribute to the summarization

Aside clustering itself can be seen as a summarization as well. Clustering can lead to well formed topical browsers where users can interact with a graphical user interface to browse topics in a more coherent and semantic way see Stefanowski and Weiss (2003).

Supervision As opposed to unsupervised learning strategies such as clustering, supervised learning classifies some input based on a provided ground truth. That is for an input x there are labels y that describe the class they are in. Supervision can be done by explicitly classifying the documents before the clustering. The input is then split into n classes. Then each class can be individually clustered. Often however this is no option. We need to manually label all documents. This can be time consuming and error prone. Often several labellers are needed to crossvalidate human bias. With this in mind there are two options on how to label unseen or new data:



- Use a supervised classification algorithm to automatically label unlabelled data. A prerequisite is to have a labelled training set and to have a lot of data. To name a few candidates: *Multinominal/Gaussian) Naive Bayes (NB)*, *Multivariate Logistic/Linear Regression*, *Neural Networks (ANN)*, *Support Vector Machines (SVM)* or *Random Forests (RF)*.

- Use an unsupervised clustering algorithm to automatically label unlabelled data. This can be done by first forming clusters and then merging the nearest clusters until k distinct categories remain. Usually the merging criterion can be controlled by some threshold and high variance documents are sorted out into an outlier cluster.

Usually by clustering we mean automatic detection of the grund truths. Often this is to shallow and does not lead to labels with a high confidence. In the domain of document clustering all information that provide some context are critical and should be used.

2.2 Vector Space Model (VSM)

2.2.1 Notation

2.2.2 Bag of words

2.2.3 Similarity and Distances

2.2.4 Dimensionality and Hashing

2.2.5 Enhancing the Vector Space Model

2.2.5.1 Singular Value Decomposition (SVD)

2.2.5.2 Latent Semantic Analysis (LSA)

2.2.5.3 Principal Component Analysis (PCA)

2.3 Clustering algorithms

2.3.1 Objective goal

EM, Cost functions, general clustering scheme

2.3.2 Hierarchical / Agglomerative clustering

Ward, Complete and Average Linkage

Birch

2.3.3 Partitional clustering

K-Means

Mean-Shift

2.3.4 Others

1. *Spectral* - x
2. *Density* - x
3. *Grid* - x

2.4 Generative Models

2.4.1 Topic modelling

Bayes Theorem

Multinominal Distributions

Dirichlet Distributions Chinese Restaurant Process

2.4.2 Methods

2.4.2.1 Latent Dirichlet Allocation (LDA)

2.4.2.2 Non Negative Matrix Factorization (NMF)

2.5 Clustering quality measures

2.5.1 Internal measures

Without labels of the ground truth

1. *Silhouette coefficient* - x
2. *Davies–Bouldin index* - x
3. *Dunn index* - x

2.5.2 External measures

With labels of the ground truth

1. *F-Measure* - x
2. *Jaccard index* - x

3 Data Pipeline

This section explains a general dataflow and the necessary steps to get data from an external source into a vectorized form. It will be held short giving intuition about a general setup, necessary preprocessing steps and the various problems that arise.

3.1 Python and Libraries

3.2 Pipeline

3.3 Problems

4 Feature Selection

"Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering."

Andrew Ng

This section is about engineering the right features for clustering that contain enough information to firstly represent the document itself and secondly to link documents to each other. To do so the document domain of newspapers will be discussed. The feature selection process of a document by taking several processing steps is examined. At the heart: How can we expand the knowledge of a feature set by adding semantics to it?

4.1 Document domain

4.1.1 Document structure

4.1.2 Categories

4.1.3 Metadata

4.2 Selection process

4.3 Adding Semantics

4.3.1 Syntactic parsing

4.3.2 Wikipedia and Wordnet

5 Clustering experiments

"An algorithm must be seen to be believed."

Donald Knuth

In this section two strategies for clustering are proposed. The Strategies were implemented with the help of Python and several open source libraries.

Due to time constraints there is only one model for covering a single day and one model covering multiple days of the scraped newspapers.

For more strategies see the github project.

5.1 Single day clustering

5.1.1 Implementation

5.1.2 Evaluation

5.1.3 Interpretation

5.2 Multiple days clustering

5.2.1 Implementation

5.2.2 Evaluation

5.2.3 Interpretation

5.3 Summarization

6 Results and Discussion

"Simple models and a lot of data trump more elaborate models based on less data."

Peter Norvig

In this section the results of this thesis are discussed. The implementations are compared and evaluated. What is problematic and what worked out well? What can we conclude by now?

6.1 Similarity Evaluation

6.2 Pros and Cons

6.3 Conclusion

7 Outlook

In this section we will summarize what has been done in the last chapters. Moreover further reading, related and future work is depicted.

7.1 Summary

7.2 Further Reading / Related Work

7.3 Future Work

7.4 Final Words

Bibliography

Aggarwal, C. and Zhai, C. (2012). A Survey of Text Clustering Algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, chapter 4, pages 77–128. Springer US, Boston, MA.

Mani, I. (2001). Summarization evaluation: An overview.

Stefanowski, J. and Weiss, D. (2003). Carrot and Language Properties in Web Search Results Clustering. In *Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003, Proceedings*, pages 240–249.

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 02. February 2016 Matthias Nitsche