KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Martin Váňa

Master's Thesis

Computer Science and Engineering
Software Engineering
2017/2018

Supervisor:
Doc. Ing. Josef Steinberger, PhD.

FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Incremental News Clustering

Abstract

The goal was to research model-based clustering methods, notably the Distance Dependent Chinese Restaurant Process (ddCRP), and propose an incremental clustering system which would be capable of maintaining growing number of topic clusters of news articles coming online from a crawler. LDA, LSA, and doc2vec methods were used to represent a document as a fixed-length numeric vector. Cluster assignments given by a proof-of-concept implementation of such system were evaluated using various metrics, notably purity, F-measure and V-measure. A modification of V-measure – NV-measure – was introduced in order to penalize an excessive or insufficient number of clusters. The best results were achieved with doc2vec and ddCRP.

Vector Representation for News Articles

Text, unlike image and audio, does not have a natural way of vector representation. Three methods (LSA, LDA, and doc2vec) were used to encode text as a fixed-length numeric vector in such a way that related documents are close in the result vector space.

Model–based Clustering

Model-based clustering methods assume that the observed data $\mathbf{X}$ are a result of a generative process. These models which have hidden (latent) variables are called Latent Variable Models.

The most widely used model, despite the fact that it might be an oversimplification of reality, is the Gaussian mixture model. In this model, each base distribution belonging to a certain cluster is the multivariate Gaussian with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Nonparametric Bayesian models whose parameter space has infinite dimension were used. Namely, it was the Chinese restaurant process (CRP) and the Distance Dependent Chinese Restaurant Process (ddCRP). The timespan of articles was used as a distance between two articles.

Evaluation of Clustering

The evaluation of clustering is the most difficult part of cluster analysis. Clustering is an unsupervised learning technique, therefore it is hard to evaluate the quality of the output of given methods.

Many criteria, such as Rand index or V-measure, are sensitive to the number of clusters. For that reason, Normalized V-measure (NV-measure) is introduced. It penalizes difference between the true class labels, C, and the number of clusters found, K, and is constructed in such a way which gives the V-measure values exactly when $K = C$. NV-measure is defined as follows:

$$\mathrm{NV}_{p,\beta}(\boldsymbol{\Omega}, \mathbf{C}) = \left(1 - \left(1 - \left(\frac{\min(K, C)}{\max(K, C)}\right)^{p}\right)^{\frac{1}{p}}\right) \mathrm{V}_{\beta}(\boldsymbol{\Omega}, \mathbf{C})$$

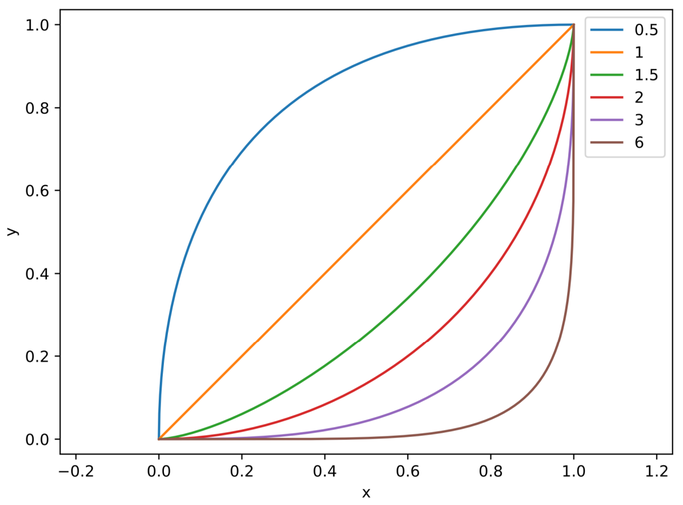

Figure 1: The influence of the parameter p

Incremental News Clustering System

As figure 2 depicts, first, the raw data are preprocessed and stored in a temporary corpus file. The documents grouped by day are then transformed to its vector representation by a pretrained model and passed to a selected clustering algorithm. Resulting cluster assignments are evaluated and stored in files.

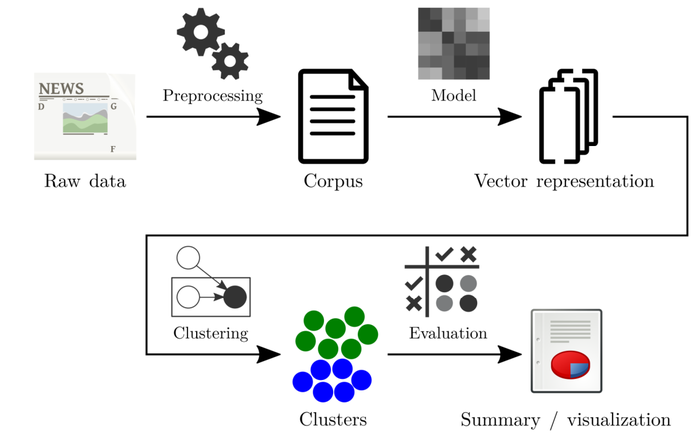

Figure 2: An illustration of the clustering system

Results and Discussion

The ddCRP performed much better than the CRP. It converged in fewer iterations and it was easier to find applicable hyperparameters. On the other hand, the ddCRP is computationally more costly.

As tables 1 and 2 show, a choice of the vector representation of news articles had a huge impact. Experiments show that doc2vec outperformed LSA and LDA in most of the evaluation metrics.

Table 1: A brief result table for D = 50

	CRP			ddCRP		
	LSA	LDA	doc2vec	LSA	LDA	doc2vec
purity	0.3878	0.4864	**0.5816**	0.5340	0.6803	**0.8061**
F_1	0.0695	**0.2535**	0.2327	0.2479	0.4937	**0.5002**
V_1	0.5812	0.6484	**0.6964**	0.6757	0.7860	**0.8506**
$NV_{1,1}$	0.4403	0.5187	**0.5394**	0.6740	0.7081	**0.7445**

Table 2: A brief result table for D = 100

	CRP			ddCRP		
	LSA	LDA	doc2vec	LSA	LDA	doc2vec
purity	0.3673	**0.5408**	0.5000	0.5272	0.6565	**0.7653**
F_1	0.0752	0.2506	**0.2677**	0.2315	0.4549	**0.5316**
V_1	0.5540	**0.6734**	0.6509	0.6764	0.7938	**0.8512**
$NV_{1,1}$	0.4858	0.5576	**0.5678**	0.6579	0.6870	**0.8072**