Parallel Sepctral Clustering

新納浩幸

概要

基本は以下の論文のさわり部分だけの紹介

Song, Yangqiu, et al. "Parallel spectral clustering.“
Machine Learning and Knowledge Discovery in Databases.
Springer Berlin Heidelberg, 2008. 374-389.

＋

クラスタリングについての話題

クラスタリング

Rで学ぶクラスタ解析 [単行本]

新納 浩幸 (著)

単行本: 208ページ

出版社: オーム社 (2007/11)

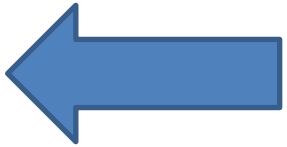
ISBN-10: 4274067033

ISBN-13: 978-4274067037

発売日： 2007/11

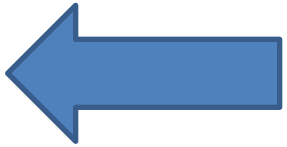
教師なし学習

教師付き学習



少数の訓練事例を利用

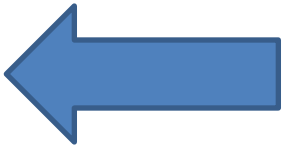
教師なし学習



クラスタリングと同義

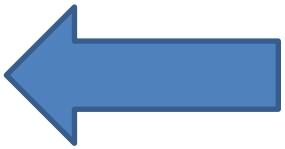
NLP のクラスタリング

単語のクラスタリング



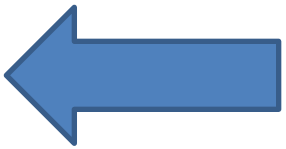
シソーラスの作成、辞書の作成

文書のクラスタリング



検索と関連あり

データのクラスタリング

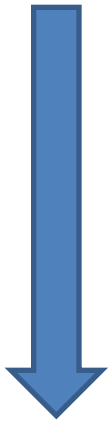


あるシステム内のモジュール

クラスタリング手法

K-means

精度はそこそこだけど頑健、実用的



改善手法としては
様々な手法が存在
有力なものも多数存在

スペクトラルクラスタリング

クラスタリング手法のブレークスルー
State-of-the-art の1つ

スペクトラルクラスタリング

色々なバリエーションがある

Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.

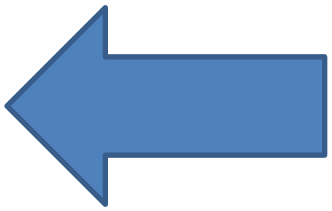
Normalized Sepctral Clustering が標準？

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems* 2 (2002): 849-856.

スペクトラルクラスタリングの問題

どのタイプのスペクトラルクラスタリングでも
Laplacian 行列 L の固有ベクトルを求める

データ数 n 、クラスタ数 k のとき Laplacian 行列の次元は $n \times n$
固有値の小さい順に k 個の固有ベクトル求める



データ数が 1000 位でも計算困難

Sparse Similarity Matrix

Laplacian 行列 L はデータの類似度行列 S から作成される

S をスパース行列に近似してから、
L を作って、L の固有ベクトルを求める

なかなか good な手法、本来、
別クラスタに属するデータ間の類似度は 0

Nyström 近似法

S（n×n）の固有値、固有ベクトルを
A（l×l）の固有値、固有ベクトルから近似的に
求める方法・・・（l << n）

$$S = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \qquad W = \begin{bmatrix} A \\ B^T \end{bmatrix}$$

C がない

$$S \approx \tilde{S} = WA^{-1}W^T = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix}$$

$\tilde{S} = \tilde{V}\tilde{\Sigma}\tilde{V}^T$　が求めるもの

$A = V_A \Sigma_A V_A^T$　を解く、これは可能、で・・・

$$\tilde{\Sigma} = \left(\frac{n}{l}\right)\Sigma_A$$

$$\tilde{V} = \sqrt{\frac{l}{n}} W V_A \Sigma_A^{-1}$$

終わり

どっちの対策が better？

Nyström 近似法

論文ではわずかだけどクラスタリング結果がよい・・・

並列の効果かどうかわからないが、速度も better

メモリの効率についても better

実は結構問題だけど私は気にしない
結局、速度の話になるので・・・

更に現実の問題

Sparse Similarty Matrix による対策

スパースな行列の固有値問題を解くプログラムが必要

Nyström 近似法による対策

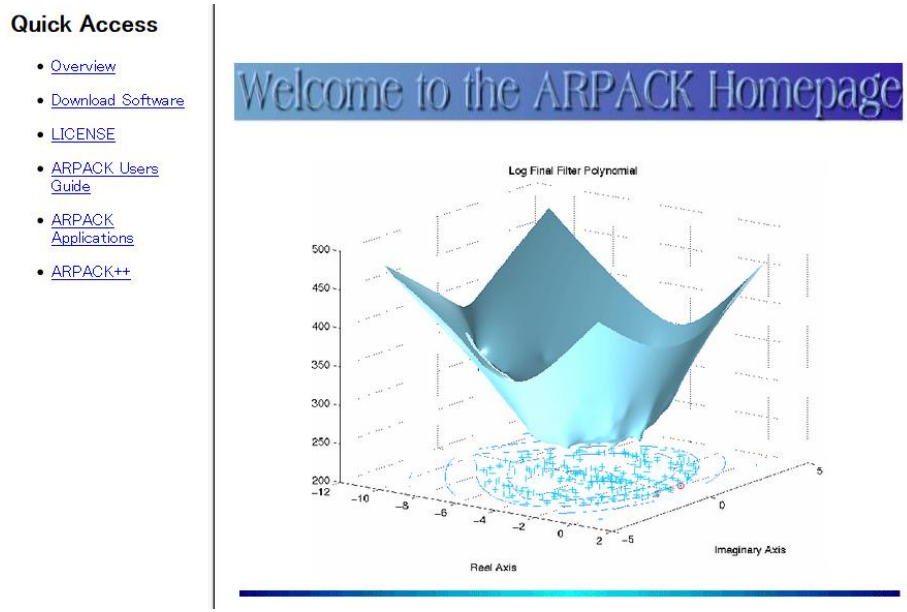
プログラムが面倒？？？

解決（1）

スパースな行列の固有値問題を解くプログラム

→ ARPPACK　**ARPACK**(**AR**noldi **PACK**age)

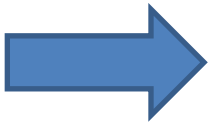
大規模固有値問題のために開発されたFORTRANサブルーチン群



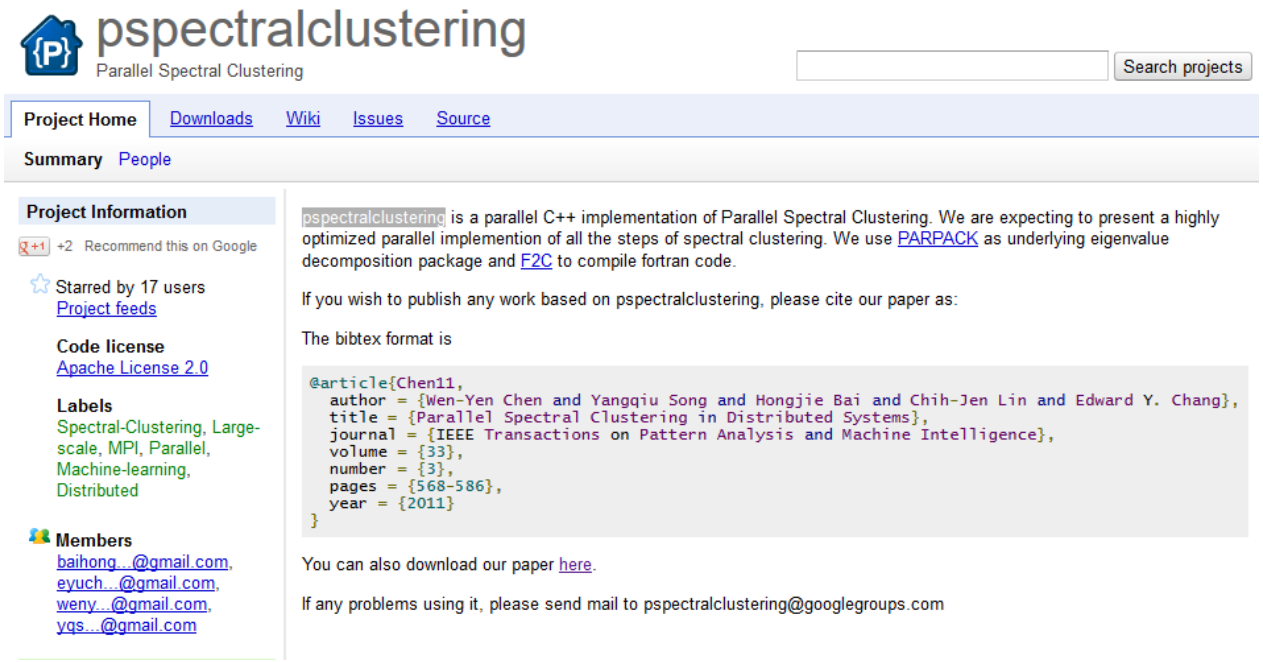
http://www.caam.rice.edu/software/ARPACK/

解決（2）

Nyström 近似法によるスペクトラルクラスタリング



pspectralclustering



pspectralclustering
Parallel Spectral Clustering

Search projects

Project Home | Downloads | Wiki | Issues | Source

Summary | People

Project Information

+2 Recommend this on Google

Starred by 17 users
Project feeds

Code license
Apache License 2.0

Labels
Spectral-Clustering, Large-scale, MPI, Parallel, Machine-learning, Distributed

Members
baihong...@gmail.com,
eyuch...@gmail.com,
weny...@gmail.com,
yqs...@gmail.com

pspectralclustering is a parallel C++ implementation of Parallel Spectral Clustering. We are expecting to present a highly optimized parallel implemention of all the steps of spectral clustering. We use PARPACK as underlying eigenvalue decomposition package and F2C to compile fortran code.

If you wish to publish any work based on pspectralclustering, please cite our paper as:

The bibtex format is

```
@article{Chen11,
  author = {Wen-Yen Chen and Yangqiu Song and Hongjie Bai and Chih-Jen Lin and Edward Y. Chang},
  title = {Parallel Spectral Clustering in Distributed Systems},
  journal = {IEEE Transactions on Pattern Analysis and Machine Intelligence},
  volume = {33},
  number = {3},
  pages = {568-586},
  year = {2011}
}
```

You can also download our paper here.

If any problems using it, please send mail to pspectralclustering@googlegroups.com

http://code.google.com/p/pspectralclustering/

pspectralclustering

実用的なスペクトラルクラスタリングの
コードはなかなかない。これは1つの可能性。

Linux 64 bit 環境の C++ でコンパイル
並列処理も可能だけど、必須ではない

現在、までコンパイル成功していません
誰かやって下さい

redsvd

このページは追加です

私の勉強不足でした。すみません。

http://code.google.com/p/redsvd/

By プリファードインフラストラクチャー
岡野原さん

これで全て解決します・・・

すごすぎです