情報検索とベーズ決定

Raphael Shu

Weblio Inc.

October 30, 2014

分類子と関連性

分類子

- 文献の集合：　S
- 文献の中にm個の分類子$\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_m$が存在する
  例：　$\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3 =$ 政治, 経済, 教育

Z_j関数

文献にある分類しが与えられたかどうかを判断する関数を定義する。

$$Z_j(s) = \begin{cases} 1 & \mathbf{l}_j \text{ は} s \text{ に与えられている} \\ 0 & \mathbf{l}_j \text{ は} s \text{ に与えられていない} \end{cases}$$

すると、文献の分類はm桁の2進数$(Z_1(s), Z_2(s), ..., Z_m(s)$で表示できる。

- このような2進数の数は全部$2^m - 1$個ある

例

分類子の集合を政治、経済、教育にする時、政治と経済に関連する記事は$(1, 1, 0)$の分類になる

関連性の尺度

関連性の尺度をq段階に分ける。

- 例えば、0は関連なし、1はやや関連ありとか

関連性の尺度も分類の表示に入れると、m+1桁になる。

例

政治と経済に強く関連する記事は(1, 1, 0, 5)の表示になる

ここで、分類子の組合せの数をiにして、関連性をk段階にすると、文献の集合Sは:

$$S = \bigcup_i \bigcup_k S_{ik}$$

文献からサンプリング

カテゴリーの確率

文献の集合Sからランダムサンプリングすることを考える。

- カテゴリーS_{ik}の確率: $\theta_{ik} = P(S_{ik})$
- カテゴリーiの確率: $P(S_i) = \sum_{k=0}^{q} \theta_{ik}$

サンプリング

- 文献Sの中に、文書の数をnにする
- カテゴリーs_{ik}に入る文書の数をr_{ik}にする
  - 当然、足し合わせるとnになる

$$\sum_{i=0}^{q}\sum_{k=0}^{q} r_{ik} = n$$

文献カテゴリーに応じる確率の例

表 5.2　ある図書館の文献構成

文献カテゴリー		閲覧	OR	図書館	閲覧・OR	関連性	確率	標本データ
S_0	S_{00}	×	×	×	×	なし	θ_{00}	r_{00}
	S_{01}	×	×	×	×	ややあり	θ_{01}	r_{01}
	S_{02}	×	×	×	×	あり	θ_{02}	r_{02}
S_1	S_{10}	×	×	×	○	なし	θ_{10}	r_{10}
	S_{11}	×	×	×	○	ややあり	θ_{12}	r_{11}
	S_{12}	×	×	×	○	あり	θ_{12}	r_{12}
	S_{20}	×	×	○	×	なし	θ_{20}	r_{20}
	…	…	…	…	…	…	…	…
	$S_{15,2}$	○	○	○	○	あり	$\theta_{15,2}$	$r_{15,2}$
	S						1	n（点）

各カテゴリーの確率とランダム・サンプルの結果が与えられている．各カテゴリーの定義は該当（○），非該当（×）で定められている．

カテゴリーの組合せのサンプリング確率

背景説明

分類子が下記の 3 つの場合

- 政治(確率: θ_0, 標本データ数: r_0)
- 経済(確率: θ_1, 標本データ数: r_1)
- 教育(確率: θ_2, 標本データ数: r_2)

政治或いは経済に関わる文書数の確率分布を知りたい。

- この例では、求めたいカテゴリーの組合せの確率のベクトルは$\theta = (\theta_0, \theta_1)$である。$r$をこの組合せのデータ数にする。

カテゴリーの組合せのサンプリング確率

- 各カテゴリーが独立し、関連性は 2 段階に分ける場合、$f(r|\theta)$は多項分布に従う。

$$f(r|\theta) = \prod_{i=0}^{p} \frac{n!}{r_{i0}!r_{i1}!}((\theta_{i0})^{r_{i0}})((\theta_{i1})^{r_{i1}})$$

- ここで、実際に関連性ありの確率θ_{i1}が関連性なしの確率より上まわるカテゴリーを検索すればよい。

ベイズ検索

ベイズ検索

問題設定

カテゴリー$S_{ik}$$n$確率$\theta_{ik}$をベイズ推定する。

- 各カテゴリー$S_{ik}$$n$の母数を$\alpha_{ik}$にすると、$\theta$の事前確率は下記の式になる。

$$w(\theta) \propto \prod_{i=0}^{p} (\theta_{i0})^{a_{i0}-1} (\theta_{i1})^{a_{i1}-1}$$

- rの確率分布

$$f(r|\theta) = \prod_{i=0}^{p} \frac{n!}{r_{i0}! r_{i1}!} ((\theta_{i0})^{r_{i0}})((\theta_{i1})^{r_{i1}})$$

- よって、観測された標本数rが得たとき、θの事後確率は

$$w(\theta|r) \propto \prod_{i=0}^{p} (\theta_{i0})^{a_{i0}+r_{i0}-1} (\theta_{i1})^{a_{i1}+r_{i1}-1}$$

ベイズ検索

- ここで、実際に母数がαから$\alpha + r$に変化した。
- 同じように、あるカテゴリを検索するかどうかを判断する時、確率の比例を使う。

$$E(\frac{\theta_{i1}}{\theta_{i0}}) = \frac{\alpha_{i1} + r_{i1}}{\alpha_{i0} + r_{i0} - 1} \geq c$$

説明

$E(\frac{\theta_{i1}}{\theta_{i0}}) = \frac{1}{5}$の意味は、カテゴリー$S_i$の文献を検索すると、関連文書 1 件を得るために 6 件の文書を検索する必要がある。

External Topics

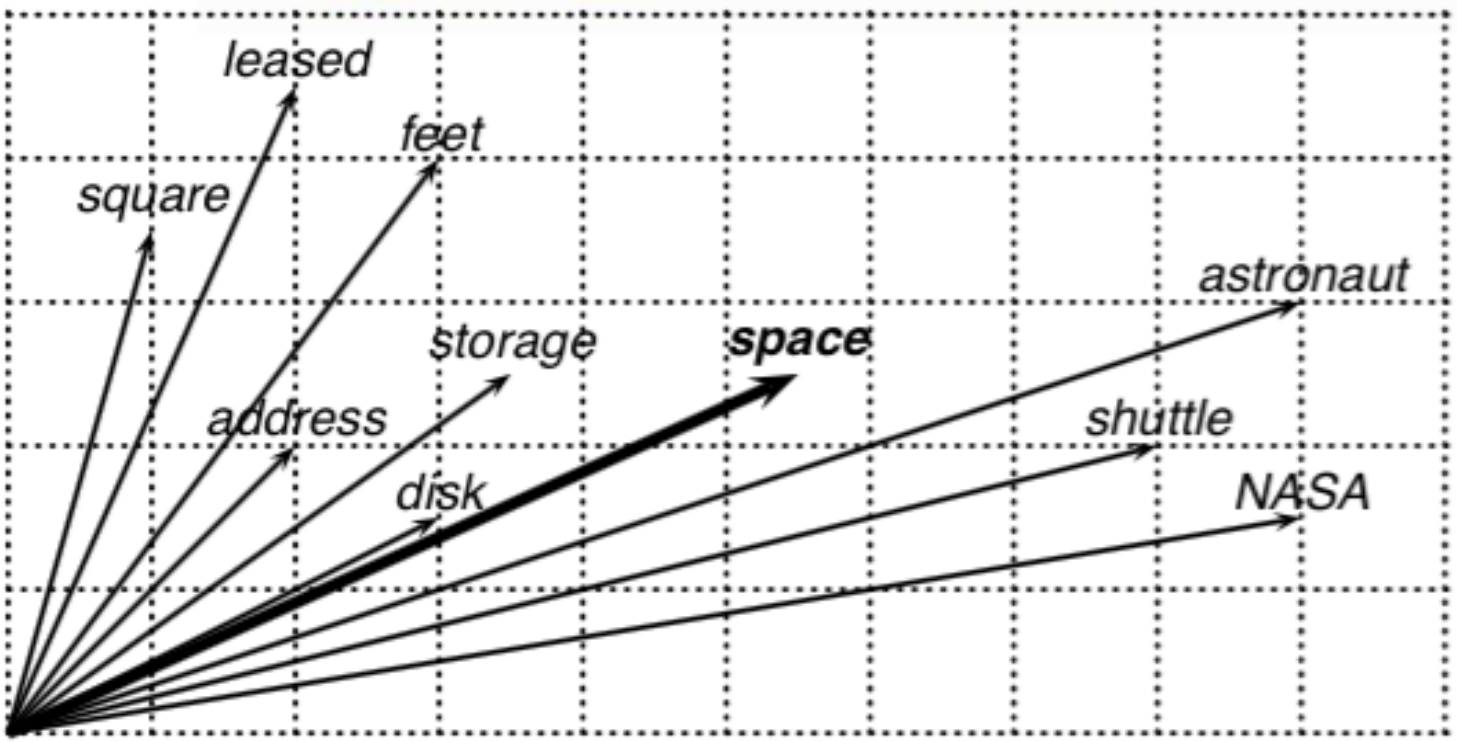
ベクトル空間モデルによる情報検索

潜在意味解析(LSA、情報検索分野ではLSI)

- ある単語の意味は文脈で決める
  - "You shall know a word by the company it keeps" -J. R. Firth (1890)
  - "The meaning of a word is defined by the circumstances of its use" -Wittgenstein (1889)
- 単語の共起行列

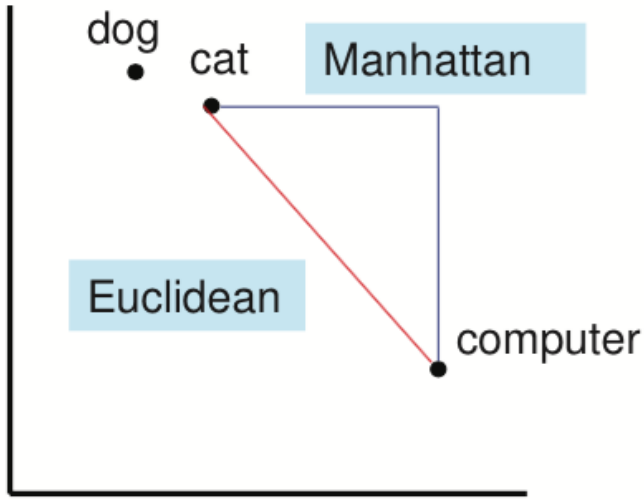
	car	Chomsky	corpus	emissions	engine	hood	make	model	noun	parsing	tagging	tires	truck	trunk	wonderful
car	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0
hood	1	0	0	1	1	1	1	1	0	0	0	1	1	1	0
Chomsky	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1

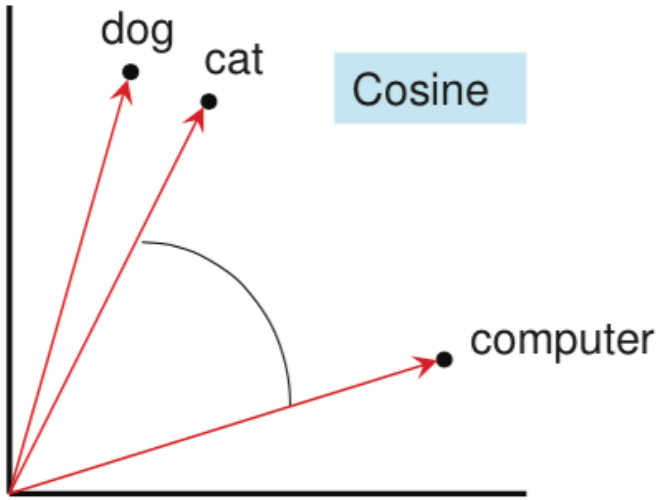
ベクトル空間モデルによる情報検索



単語の類似度を測る

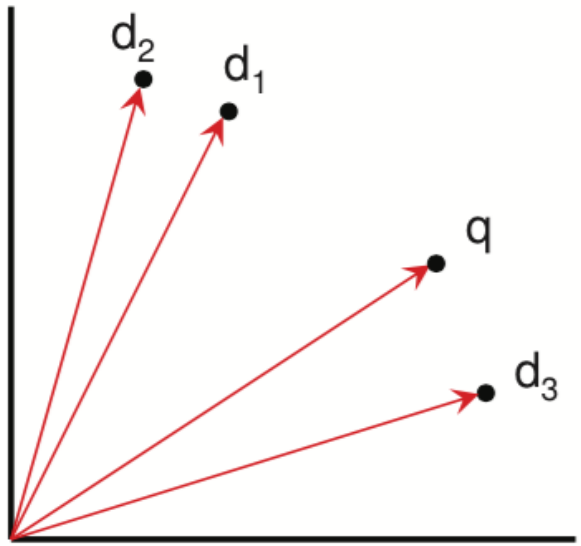
- 測りたい単語ペア(w_1, w_2)のベクトルをそれぞれ作って、多次元空間において、ベクトルの距離を測る
- 三種類の距離





ベクトル空間モデルによる文書検索

- 文脈の代わりに、文書内の単語でベクトルを作って、単語の重要度(TF-IDF)で重みを付ける
  - 文書と検索クエリ



次元圧縮問題

問題

学習用テキストデータのサイズがでかいので、ベクトルが膨大しすぎて、計算が難しい

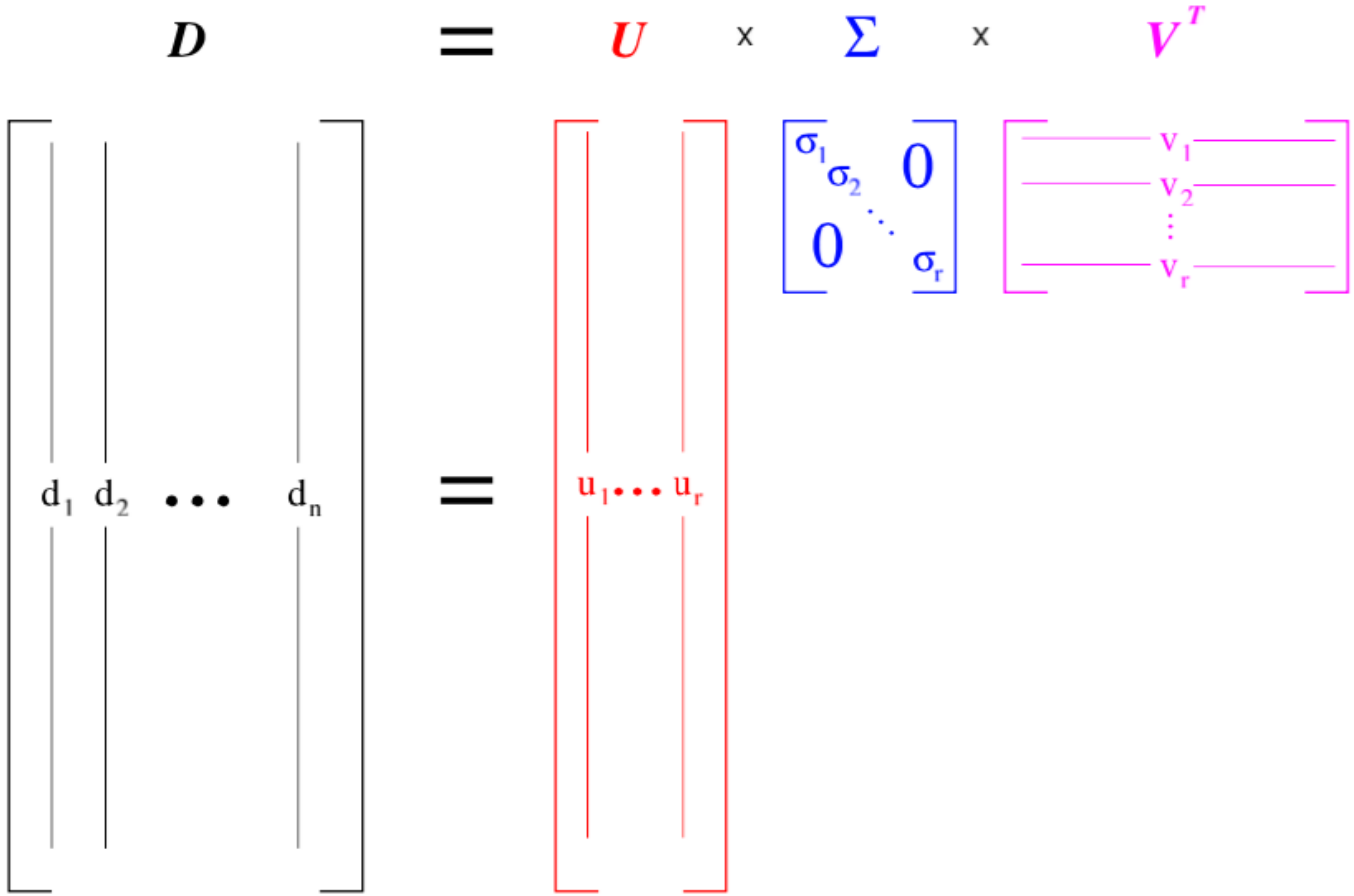
次元削減

- 特異値分解(SVD)

特異値分解(SVD)

情報を失わずに高次元ベクトルを低次元に投射する。

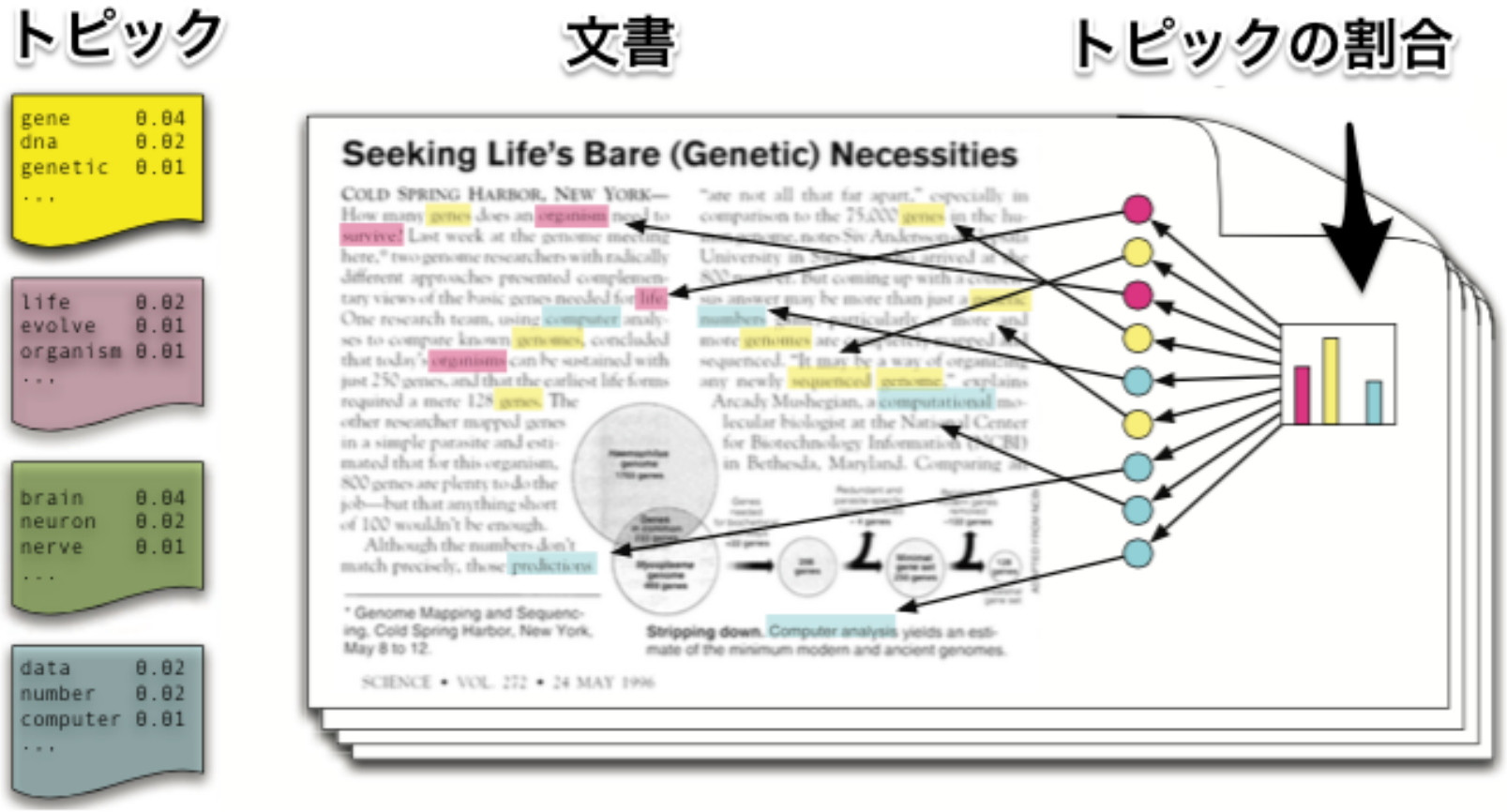
- $D = U\Sigma V^T$



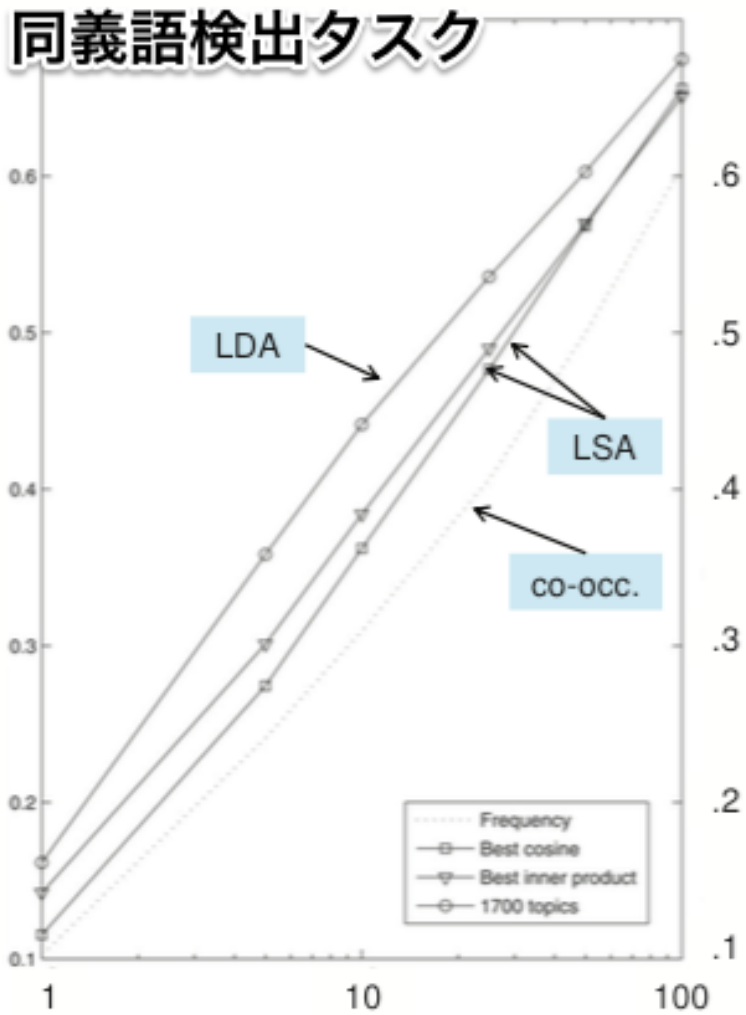
トピックモデル

潜在的ディリクレ配分法(LDA)

- 文書はいろんなトピックが確率的に組み合わせたもの



まとめ



- LSAは情報検索と自然言語処理の分野でよく使われる
- LSAはLDAより実用的
- 論文書くならLDAがもっとマシ

End

Merry Halloween